Estimating the Eligible to Naturalize Population: Developing a Robust Method

Katherine M. Condon, U.S. Citizenship & Immigration Services
Ronald E. Wilson, U.S. Citizenship & Immigration Services

This paper is being presented to inform interested parties of ongoing research on immigrant population estimation. The intent of the presentation is to encourage discussion on method improvements and the use of unique data sources. Any views expressed on technical or operational issues are those of the authors and not that of the U.S. Citizenship and Immigration Services.

INTRODUCTION

Estimating the eligible to naturalize (EtN) population accurately and properly requires a detailed and thorough process. Currently, multiple divisions within the U.S. Citizenship and Immigration Services (USCIS), as well as the Office of Immigration Statistics (OIS) of the Department of Homeland Security (DHS) produce EtN estimates from administrative records of the lawful permanent resident (LPR) living in the United States, with the purpose of assessing resource allocation in operational planning. While each division's method produces similar estimates, none completely capitalize on USCIS's data sources in refining the method to obtain the most accurate estimate possible. We present a method in this paper that maximizes USCIS own data to refine estimates of EtNs. The range of detailed immigrant beneficiary characteristics across multiple beneficiary types allow for using sound social science methods and research to bring a high level of accuracy to the EtN estimates. To do so, the estimation must be done in several stages.[1]

Refining the estimates facilitate the fine tuning of resource allocation within USCIS, but also in the estimation of LPRs who most likely will naturalize[2], as well as any other estimations based off of EtN population estimates.

BACKGROUND

Multiple agencies, inside and outside of USCIS, use Federal government administrative data sources to estimate the level of immigration to the United States. Such sources are the U.S. (i) Census Bureau's decennial censuses, (ii) Department of Housing & Urban Development's American Housing Survey, and (iii) Department of Labor, which include characteristics on place of birth, citizenship, and year of entry into the United States. These data provide essential information on the total foreign-born population, naturalized citizens, and non-U.S. citizens residing in the United States. While these data sources are used in support of their overall missions for operational planning and forecasting, the broad categories of non-U.S. citizen data do not help overcome the limits of error introduced into estimates from using non-immigration specific data.

Estimation by USCIS is guided by the Immigration and Nationality Act (INA), which produces detailed data sets of immigrant characteristics for foreign nationals the allow us to specifically identify the

[1] We envision the full paper will be divided into the following parts that describes each stage. These parts include: (i) concept behind the stage; (ii) any underlying empirical social science research that underpins the stage; (iii) the data used; and (iv) statistical mechanics in creating the estimates at that stage.
[2] Not every LPR will naturalize, to which there will be variation in the patterns of timing to naturalize.

number of LPRs in the U.S. that serves as solid foundation for improving the estimation of LPRs eligible to naturalize. The following is an overview of the data sources and method for producing a highly refined estimate of the EtN from the national level.

DATA AND METHOD

The main objective of this work is to refine the process of estimating with respect to both the data sources, as well as the methods. While there are a number of differences between other estimation methods of EtN population, it is important to state that there is a basic overlap between the current method to estimate and the one proposed in this paper.

Further, we propose that separate population estimates of eligible to naturalize LPRs be developed by year of admission and class of admission. With regard to year of admission: (a) LPRs who were admitted to the United States before 1973 and (b) LPRs who were admitted after 1973. This was done, because USCIS computerization of their administrative data did not start until 1973. With regard to the class of admission, there will be separate estimates for (a) LPRs admitted for humanitarian reasons, e.g., refugees and asylees; and (b) all other. These estimate sets are added together to obtain the overall EtN population estimates as of July 1, 2015. We will also take into consideration out-migration and mortality.

The following are the general stages used to produce our estimates: 1) Standardizing the data; 2) Assembling of base data set; 3) Establishing a Refined Clock Start Date; 4) Identifying Eligible LPRs; 5) Separating the Data Set by Class of Admission; and 6) Accounting for Out-migration and Mortality.

Standardizing the Data Sets

Since the data come from different data systems in both OIS and USCIS, the naming conventions of the variables used differ. We established a set of common variable names to combine the same information from different field names across the systems into one common field. Most of the required characteristics across the data sets were straight-forward when combining into a single standardized field, such as (i) Class of Admission, (ii) Sex, (iii) Date of Birth, (iv) Date of Naturalization, (v) Country of Birth, (vii) State, and (viii) ZIP Code. However, the standardization of the date on which eligibility begins is based on the path used in becoming an LPR.[3]

There are two main paths followed to become an LPR, which are an adjustment of status from already being in the U.S., and being a new arrival from a foreign country. An (ix) Approval Date was created to capture either date an immigrant becomes an LPR is based on which path is used. If the immigrant is an adjustment, then the date eligibility is based off is the approval date of their adjustment. If a new arrival, then the date is based off the entry date into the U.S. The OIS data contained Admission variable that was put into the Approval Date variable. But for both paths, there is a subcategory humanitarian immigrants (asylees and refugees) whose eligibility dates are not the adjustment approval or arrival dates. The approval dates for the immigrants need to be adjusted for determining when they are eligible to naturalize. The details about the adjustment of this subcategory of immigrants for establishing eligibility date for is discussed in more detail below in regards to how we established the "Clock Start Date" on which being eligible to naturalize begins.

We needed to create additional variables to not only ensure the back tracing of a record to the original data source, but also needed information toward estimating the EtN. Finally one data issue we

[3] Additional variables were standardized, such as Marital Status, Name, Received Date, Form Number, and Receipt Number for the purpose of managing the processing of the data. But the variables listed in the narrative are the primary variables used in the estimation.

encountered is with regard to a small number of records that either have the sex of the LPR as unknown or is missing.[4]

Data Set Assembly

With both the USCIS and OIS data sets standardized, each data set needed to be truncated and combined. The OIS data set contains all LPRs from 1973 to 2015. The USCIS data ranged from 1993 to the present.[5] Because USCIS data systems only came on line after its creation, it took time for the agency's data systems to mature and contain all the records necessary conducting any analysis. In our analysis, we identified the year 2010 to be the first year to contain complete data.

LPRs who arrived before 1973 and who have not naturalized are estimated using the aggregated data from the most recent 5-year estimate from the Census Bureau's American Community Survey (ACS) to supplement the estimates derived from individual records after the processing the OIS and USCIS data. These data, though, were not combined with the OIS and USCIS data because they are aggregated and could not be processed by individual records and only added in aggregate at the end.

As such, we combined the USCIS and OIS data sets at July 1, 2010 to make a comprehensive data set that contained all the information needed to filter the data, identify existing naturalizations, and identify eligibility start dates based on class of admission. The begin date of this data set was truncated at July 1, 1973 and truncated at June 30, 2015. This July 1 date aligns with the Census Bureau's population estimate dates. Using this date ensures consistency when combining USCIS/OIS EtN estimates with the EtN estimates from ACS data prior to 1973, as well as aligning with the yearly mortality rate dates.

Finally, because the USCIS and OIS data contained all LPRs across time, we matched and removed the records for LPRs who had naturalized. We did this in two stages. First, since the OIS data already identified those LPRs who had naturalized, we removed those records that had a naturalization date. Second, we then matched the naturalization records in USCIS databases to the entire data set for those who had naturalized with (1) the N-400 (Application for Naturalization) and (2) those who applied for citizenship certificate with N-600 (Application for Certificate of Citizenship).[6] We applied the N-400 and N-600 records to not only identify those who naturalized in the USCIS data, but also to pick up any additional naturalization in the OIS that might have not been recorded correctly.

Establishing an Eligibility Clock Start Date and Identifying Eligible LPRs

Depending on the pathway into LPR status, the start date towards eligibility to naturalize varies by the initial classification for LPR status. All LPRs have a waiting period before and LPR can become eligible for naturalization. In general, whether an adjustment of status or new arrival, the waiting period is five years from first date the immigrant officially becomes an LPR. The exceptions are those LPRs who married a U.S. Citizen (USC) or are protected under the Violence Against Women Act (VAWA), to which the waiting period is three years. Asylees, refugees, parolees, or who have a cancellation or removal are

---

[4] While these records make up less than a fraction of LPRs use in our estimation, we, nonetheless, impute the sex for those records to preserve as many records as possible. We impute sex for these records by proportionally assign the 1.1% sex ratios, which migration research shows that it is about is about 1.01 to 1.02 males to females. We use simple random sampling to draw a 51.5% of cases with either missing or unknown sex and assign those cases to be male, with remaining 48.5% to be female.

[5] USCIS and OIS were established in 2002 under the Homeland Security Act of 2002 out of what had previously been known as Immigration and Naturalization Service (INS). (https://www.dhs.gov/office-immigration-statistics) The Department of Homeland Security's Office of Immigration Statistics (OIS) leads the collection and dissemination to Congress and the public of statistical information and analysis useful in evaluating the social, economic, environmental, and demographic impact of immigration laws, migration flows, and immigration enforcement. While the Department of Homeland Security's U.S. Citizenship and Immigration Services (USCIS).

[6] The N-600 was used to capture any derivative LPRs who became eligible and officially obtained documentation of becoming a naturalized citizen. Not all derivatives will apply for a certificate, making for an incomplete filter on all naturalizations. Identifying this group remains difficult to estimate because there is no source to identify these LPRs.

all subject to having their approval dates "rolled back" to an earlier date. Asylees have their clock start date adjusted to one year prior to the approval of the application for asylum. Refugees, parolees, and those with removal cancellations all have their dates rolled back to either the date of entry into the U.S. or 10 years prior to the time of their approval, whichever date occurs earlier.. The approval dates for each LPR are transferred into a Clock Start Date field, adjusting for special immigrant classes, then projecting forward the waiting period to establish the date at which the LPR is eligible to naturalize.[7]

Accounting for Out-migration and Mortality

Neither USCIS nor OIS track deaths or outmigration as part of an LPR's administrative record. This leaves the need to apply probabilities of dying or leaving the U.S. to the LPR data set to reduce the number to a more realistic number of LPRs in the U.S. who might naturalize. For estimating out-migration[8] to reduce the data set to only those who remain in the U.S. in order to then calculate mortality. We use USCIS data from I-407 (Record of Abandonment of Lawful Permanent Resident Status[9]) records in aggregate to produce probabilities of out-migration by age to reduce the number LPRs who can apply for naturalization.[10]

For each year, starting in 1973, the out-migration probabilities are applied to each age group, and then those who "stayed" are survived forward one year using NCHS life tables to age them forward. This will be done as an iterative annual process.

CALCULATING THE FINAL NATIONAL-LEVEL ESTIMATE

After processing the data, we summarize the estimates in the third to last year (2012) of the final year (2015) of the outmigration and mortality data sets across the age groups, leaving out those under age 18. The year 2012 year was used because no LPR within three years of the last recorded year are eligible to naturalize. We then combined the three summarized estimates for 1973 to 2012 with the American Community Survey estimates from 1973 and prior to create the final EtN national-level estimate.

Local-level estimates can also be derived from the process, particularly for state and metropolitan areas. The process can be followed up to the stage applying the out migration and mortality probabilities. Estimates at that stage can be split out by the desired local-level geography, which the remainder of the process can be followed on the estimates for those geographies to produce more localized estimates for more detailed planning.

While outside agencies cannot make use of USCIS administrative data, our hope is that our method will prompt creative thinking from outside agencies toward finding data sources that allow them to follow this method and improve their estimation efforts.

---

[7] We keep the under age 18 population during the production because they are essential for using our cascading approach to estimating mortality and out migration.

[8] Prior to this step for accounting of out-migration and mortality, we separated LPRs by class of admission, i.e., asylees, refugees, parolees, and those with removal cancellations, to only apply the mortality probabilities. Given the nature of their immigrating to the U.S., it is unlikely that many will want to return to their country of birth/nationality. As such, very few of these immigrants are likely to out-migrate and we do not include them in our out-migration calculations. We, then, summarize the individual records from the base data set into a new data set of age groups and apply the out-migration probabilities derived from the I-407 population only to LPRs with class of admission as "Other".

[9] The purpose of the I-407 form is to record in USCIS data systems that an LPR has decided to voluntarily abandon their status as an LPR from having left the country and not returned after a given period of time.

[10] The distribution of ages in these records appears consistent with migration research in accordance with risk of out-migration based on (i) age to do so, (ii) financial means to do so by age, and (iii) intent/motivation to do so by age. This distribution can be used to create a series of emigration probabilities by age.