16 October 2018

United Nations Economic Commission for Europe

Conference of European Statisticians

Work Session on Migration Statistics
Geneva, Switzerland
24-26 October 2018
Item 3 of the provisional agenda
Integration of data from censuses, administrative sources and surveys for measuring migration

Estimates of Refugees from Survey Data by Incorporating Administrative Data

Note by the United States Census Bureau*

Abstract

The flow of refugees and asylees is increasingly a major force in international migration. Despite its prominence, little data is collected in the United States regarding these migrants after they enter the country. The US Census Bureau does not collect data in its largest survey, the American Community Survey (ACS), which could directly identify refugees and asylees. In order to produce estimates about these populations, a data set that contains limited demographic information on individuals granted legal permanent resident status in the United States is used to train a logistic regression model. This model is then used to predict the probability that a foreign-born individual on the ACS entered as a refugee or asylee. To make binary assignments indicating refugee/asylee status, a rejection resampling algorithm is employed. Using these assignments, the ACS can be used to examine the demographic characteristics of the refugee/asylee population in greater detail. This paper serves as a demonstration of how large administrative data sets can be used to expand the information one can glean from a smaller but more detailed survey set.

*Prepared by Michael Bowerman, Statistician, Net International Migration Branch, U.S. Census Bureau

I. Introduction

1. Following the passage of the Refugee Act of 1980, the U.S. Department of State began publishing data on the arrivals of refugees resettled to the United States. Since that time, the Department of State recorded nearly three million refugee arrivals. In the period between 1990 and 2016, over 600,000 individuals were granted asylum to the United States [1]. According to the Census Bureau's Population Estimates Program, the United States experienced a net international migration of 1.1 million individuals between July 1, 2016 and July 1, 2017 [2]. In FY 2016, approximately 105,000 refugees and asylees were granted legal permanent resident status in the United States.

2. Although this large number of refugees and asylees have entered the United States, little is known about this population after they enter. The U.S. Census Bureau does not include information on refugee or asylee status in the majority of its surveys and the census, with the exception being the Survey of Income and Program Participation (SIPP). Unfortunately, the SIPP does not have a large enough sample size for reliable estimates of a group whose size is small relative to the total population of the United States. Because of these limitations, the outcomes of these individuals after they enter the United States is not well known. The motivation behind this research is to be able to use the American Community Survey (ACS), which has a much larger sample size than the SIPP, to make estimates about the stock of refugees and asylees in the United States. To accomplish this, a large administrative dataset containing all individuals granted legal permanent resident status in the United States between 2010 and 2015 was used to train a logistic regression model in order to predict the probability that an individual on the ACS gained entry as a refugee or asylee. After this predicted probability was obtained from the logistic regression model, a rejection resampling algorithm was used to assign refugee status as a binary variable containing refugee/asylee assignment. Based on these assignments, details about the refugee and asylee population in the United States could be estimated.

II. American Community Survey (ACS)

3. The primary data source used to make annual net international migration estimates by the U.S. Census Bureau is the ACS, which is a yearly household survey that had a sample size of five million individuals in 2016 [3]. To identify immigrants on the ACS, four variables are typically used: place of birth, residence one year ago, year of entry to the United States, and citizenship status. For instance, individuals whose place of birth was abroad and have a citizenship status of naturalized citizen or noncitizen are considered part of the foreign-born stock. The flow of foreign born is estimated by selecting individuals identified as foreign born on the ACS and whose residence one year ago was abroad or who listed their year of entry to the United States as within the previous year. There are a variety of variables on the ACS that may be of interest in studying the refugee population, including detailed demographic, economic, and geographic information. However, there is no question on the ACS that directly identifies a refugee or asylee. To access this rich data source in making estimates of the refugee population, an outside data source must be used to assign refugee status to the ACS sample.

III. Legal Permanent Resident File (LPR)

4. When individuals receive legal permanent resident status in the United States, either by entering directly as legal permanent residents via application to the Department of State, or by adjusting from temporary to permanent status via application to the Department of Homeland

Security (DHS), they are recorded in DHS's LPR file, which the Census Bureau receives from the Office of Immigration Statistics. Details on individuals in the dataset are sparse, however; the LPR contains relatively few variables, and there are issues with missing data. Variables included in the LPR are detailed in Table 1.

Table 1. Variables on the LPR

Variable Name	Details
Class of Admission	Type of visa the individual was admitted on – used to separate individuals into refugees/asylees and others
Country of Birth	Place of birth, with outlying US territories included, translated to FIPS codes*
Country of Citizenship	Codes in this variable are nearly identical to Country of Birth variable
Country of Last Residence	Codes in this variable are nearly identical to Country of Birth variable
Occupation	Variable sparsely populated
Sex	
Marital Status	Recorded as single, married, separated, divorced, or widowed
In Care of Address	State and ZIP code where the migrant is resettled
Birthdate	Year of birth
Date of Entry	Year of entry to the United States

*Federal Information Processing Standard (FIPS), a unique country code. Details can be found here: https://www.census.gov/geographies/reference-files/2016/demo/popest/2016-fips.html

5. LPR files are tabulated every fiscal year from data collected by the U.S. State Department on permanent visas issued and from DHS on adjustments from temporary status. In this analysis, the LPR files from fiscal years 2012 to 2015 were used, which include a combined 4.1 million individuals. Using the Class of Admission variable on the LPR, refugees and asylees in the dataset can be identified easily.

6. As previously stated, the LPR contains little detail on individuals in the file, and, further, it is missing a significant amount of data for some variables. First, a large portion of the occupation variable is missing. Because the majority of the occupation variable is missing, it cannot be used to model the probability that an individual is a refugee – which is unfortunate, as this variable could be matched to the ACS. About 0.2% of the sex variable is missing as well; because it seems to be at random, these entries will be removed from the analysis. We will also remove individuals on the file with missing age or marital status, as it makes up such a small proportion of the file. No entries were missing for other variables of interest.

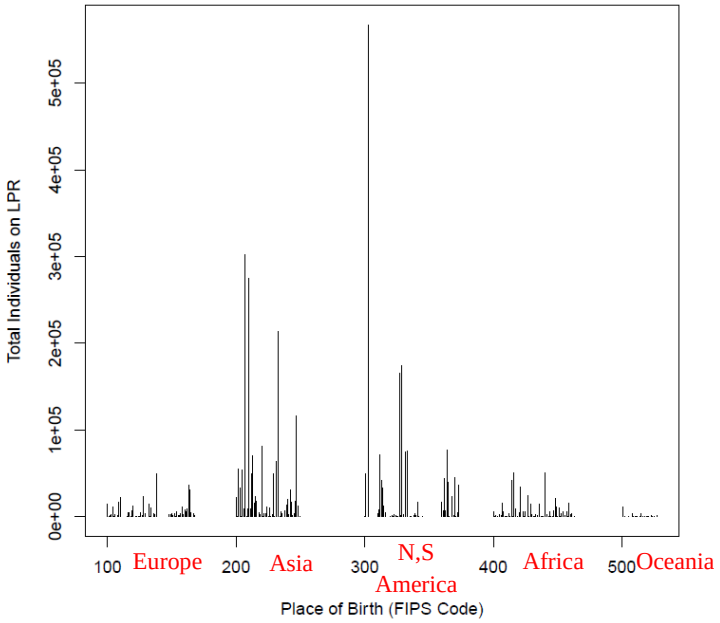
IV. Identification of Refugees/Asylees on the ACS using the LPR

7. On the LPR, there are five variables that might be informative in predicting refugee or asylee status: age, sex, marital status, place of birth, and year of entry. The other variables proved to be too sparse, in the case of occupation, or indiscernible from place of birth on the LPR, in the cases of citizenship and country of previous residence. Using these variables, a logistic regression model was fit to the LPR data to predict the probability that an individual is a refugee or asylee. Next, the binary outcome of an individual being a refugee or asylee is simulated using a rejection resampling algorithm. Before attempting to use the LPR data, though, some data variable manipulation is necessary.

A. Place of Birth

8. There are 216 distinct countries of birth represented on the LPR. A histogram showing the number of cases on the LPR by place of birth is shown in Figure 1. Many of the countries of birth on the LPR had sample sizes that were small enough to elicit concern regarding the representativeness of the sample. For instance, the number of refugee and asylee entries was less than thirty for 75 of the countries on the LPR. Because of this, using a country of birth variable in modeling the probability that a foreign-born individual entered the United States as a refugee or asylee would be misleading.

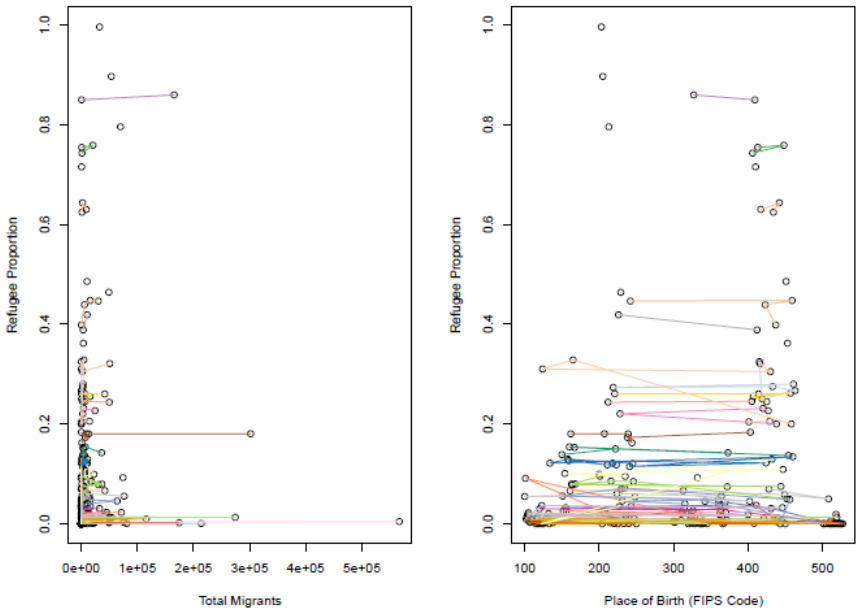
Figure 1. Histogram of number of cases by place of birth on LPR, by FIPS code



9. Instead of using a raw place of birth variable, we clustered countries on the LPR based on the total number of LPR entries from each country in tandem with the number of refugee entries from that country using a semiparametric beta-binomial model with a Dirichlet process prior on the mixing distribution. The generation of the posterior sample was done using the R package *DPpackage* [4].

10. This model clusters countries of birth based on the total number of individuals from that country on the LPR and the number of refugees and asylees from that country, in essence separating countries by their propensity to send refugees and asylees to the United States. The model generated a posterior sample separated into 37 clusters. So, instead of modeling based on place of birth, which would require 215 additional variables, we model based on cluster assignment, requiring only 36 additional variables. The cluster assignments are visualized in Figure 2; the points on the graph represent each country represented on the LPR, and the lines show countries connected in each cluster.

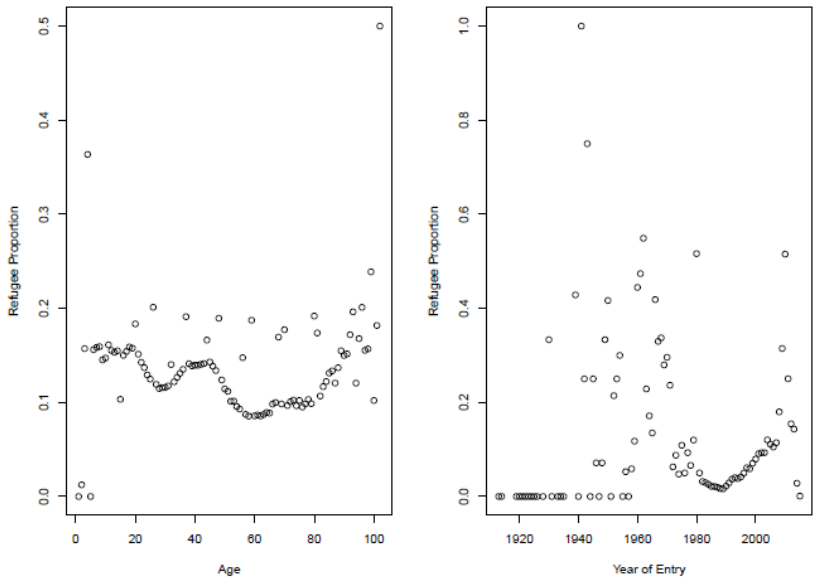
Figure 2. Country of birth cluster assignments on the LPR



B. Age and Year of Entry

11. Figure 3 shows plots of relationships between age and year of entry and the proportion of the LPR that are refugees or asylees. The patterns over time for these relationships does not show a clear pattern that might be resolved by transforming the data. Simply modeling year and age as class variables risks overfitting, and modeling them as raw numeric variables is clearly improper – the relationship between age or year of entry and refugee/asylee probability cannot be modeled by a straight line. Instead, the respective relationships between age and year of entry and the probability that an individual is a refugee or asylee will be included in the model using natural cubic splines with 20 knots, or a piecewise polynomial on the same range as the original variables, where a separate cubic polynomial is fit between each of the 20 knot points. Instead of requiring nearly 100 classes for year and 100 classes for age, refugee/asylee probability will be modeled based on a polynomial fit to a short subinterval of year and age.

Figure 3. Age vs. refugee proportion on the left; Year of entry vs. refugee proportion on the right



C. Logistic Regression Model

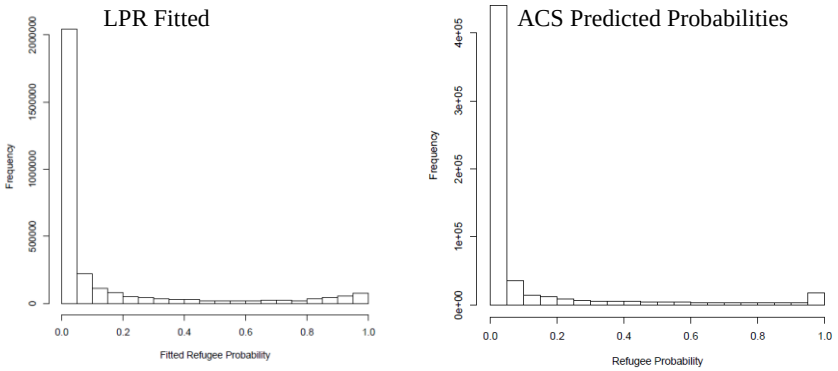
12. After the data pre-processing was complete, a logistic regression model was fit to the LPR data, modeling refugee/asylee probability based on sex, marital status, cluster assignment based on country of birth, age, and year of entry to the United States:

$$logit(p_i) = sex_i + mar_i + cluster_i + S_{age}(age_i) + S_{year}(year_i) + \epsilon_i,$$

where p_i is the probability that individual i is a refugee or asylee, $S_{age}(age_i)$ is the value of the natural cubic spline at value age_i, the age of individual i, and $S_{year}(year_i)$ is the value of the spline function at $year_i$, and ϵ_i is $N(0, \sigma^2)$ random error. This model was a better fit than the nested models. A histogram of the fitted probabilities that each individual on the LPR is a refugee/asylee can be found in Figure 4 (left panel).

13. This logistic regression model was then applied to the 2015 1-year ACS. Foreign born were selected from the ACS sample using the citizenship and country of birth variables. Individuals were then assigned to the same country of birth clusters that were used to fit the logistic model to the LPR data. Using the model, the probability that each individual on the ACS is a refugee was calculated. A histogram for the fitted probabilities for the ACS can be found in Figure 4 (right panel).

Figure 4. Fitted probabilities that an individual is a refugee or asylee on LPR (left) and ACS (right)



D. Rejection Resampling

14. Predicted probabilities that an individual on the ACS is a refugee or asylee was obtained from the logistic regression model, but a decision still must be made whether or not an individual is a refugee. To make this decision, a rejection resampling algorithm was used. This avoids a simple cut-off at some value, which would be unduly influenced by model error. This algorithm proceeds as follows:

For individual i:

1. A draw a is made from $Uniform(0,1)$
2. If $\hat{p}_i > a$, where $\hat{p}_i$ is the fitted probability for individual i, accept the sample
3. Repeat steps 1 and 2 for 1000 iterations
4. If $\frac{number\ of\ accepted\ samples}{number\ of\ iterations} > .5$, the individual is flagged as a refugee/asylee

15. To test this approach, a sample is taken from the LPR used to fit the original logistic regression model. In this case, one million of the 4.1 million individuals were retained as a test set. The logistic regression model was fit again using the remaining 3.1 million individual training set. Following this fit, predicted probabilities were calculated using the test set, and the rejection resampling algorithm was used to assign refugee flags to the test set. Using this method, it was found that only 92.9% of the test set was assigned refugee or non-refugee status correctly. This algorithm was then carried out for every foreign-born, non-native individual on the ACS to assign refugee status.

V. Results

16. Foreign-born, non-native individuals on the 2015 1-year ACS were allocated to either the refugee or non-refugee group using the rejection resampling method and predicted refugee probability from the logistic regression model. The weights from the ACS were then employed to make national estimates.

17. Some interesting differences were noted in the demographic characteristics of the group assigned as refugees or asylees on the ACS and all other foreign born. Table 2 shows some of these differences.

Table 2. Differences between foreign-born groups in the 2015 ACS assigned as refugees/asylees and other foreign born

Percent in Category										
	Sex		Age			Marital Status				
	Male	Female	<25	25 to 65	>65	Married	Widowed	Divorced	Separated	Never Married
Refugee/Asylee	39%	61%	20%	38%	42%	39%	16%	9%	3%	33%
Other Foreign Born	48%	52%	14%	70%	16%	60%	5%	7%	2%	25%

18. The refugee group is more female than others in the ACS. Their age distribution is also more widely distributed, whereas the large majority in the non-refugee/asylee group is between 25 and 65. Refugees and asylees are more likely to be widowed, as well, and non-refugees/asylees are more likely to be married.

19. From the 2015 DHS Yearbook of Immigration Statistics, 590,000 individuals were granted asylum in the United States between 1990 and 2015. From 1990 to 2015, 1,907,000 refugees entered the United States, for a total of 2,517,000 refugees and asylees entering the United States over those two periods. Using the method described above and the weights in the ACS, an estimate of 1,667,000 refugees and asylees were living in the United States in 2015 whose year of entry was after 1990. This is a may be reasonable when mortality and low levels of emigration are taken into account. In total, this method estimated that 2,869,000 refugees and asylees were living in the United States in 2015. According to our estimate, 6.6% of the stock of foreign born in the United States in 2015, as estimated from the ACS, were refugees and asylees.

20. A histogram of the years of entries for individuals on the ACS who were predicted to be refugees or asylees is shown in Figure 5. The trend is particularly interesting, showing peaks in refugee and asylee entry after World War II and again after following the wars in Iraq and Afghanistan.

21. The distributions of countries of birth differ significantly for the two peaks. Before 1970, the majority of refugees and asylees were coming from Europe, particularly from Germany, Italy, and Austria, and Mexico. After 1980, the trend shifted to the majority of refugees and asylees arriving from the Middle East, following the wars in Afghanistan and Iraq, and Cuba. Notably missing are refugees and asylees who arrived from the 1960s to the mid-1990s, especially refugees and asylees from Southeast Asia and former members of the Soviet Union. This is likely because they are not represented on the versions of the LPR available for this research. The spike around 1950 is rather strange, as the majority of individuals on the LPR entered the US after 2000. There are two likely reasons for this hump: the first, that individuals on the LPR who entered before 1960 were often refugees and asylees, which resulted in individuals whose year of entry on the ACS was before 1960 being assigned as refugees or asylees; the second, that these individuals who emigrated from Europe had demographic profiles that matched those refugees and asylees who entered more recently.

Figure 5. Year of entry to the United States for individuals on 2015 ACS predicted to be refugees or asylees; density[1] in blue

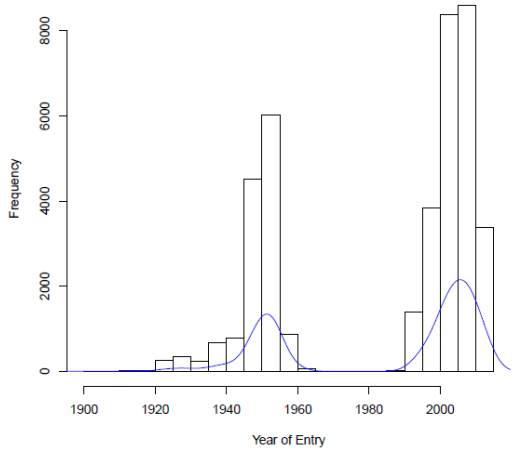
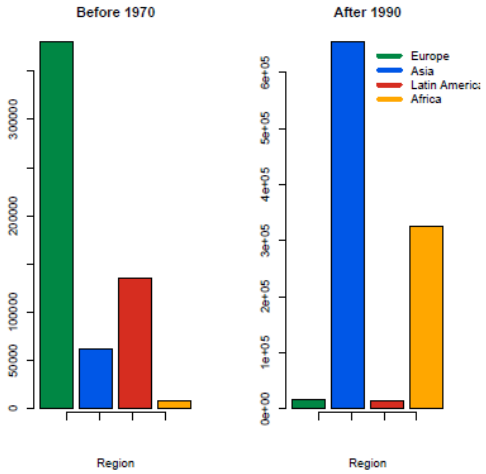


Figure 6. Region of birth of individuals on 2015 ACS assigned refugee or asylee status by period of entry



22. Further, Figure 7 shows a comparison of two plots. The first, in black, is the density of the frequency of regions of birth for individuals on the ACS assigned refugee status by this method whose year of entry was between 2010 and 2013 – these years being chosen as they have the best coverage of the foreign-born population for the years of the LPR used to fit the logistic regression model. The blue curve is the density of the frequency of region of nationality for individuals on the Worldwide Refugee Admissions Processing System (WRAPS)[2] who were resettled in the United States between 2010 and 2013 [6]. The distributions, again, agree, which suggests that the method produces valid estimates for the refugee stock in the United States.

[1] Empirical estimate of the probability distribution.

[2] WRAPS is a computer system used to process and track refugees who are resettled to the United States under the U.S. Refugee Admissions Program.

Figure 7. Densities of frequencies of regions of origin from WRAPS and ACS estimate

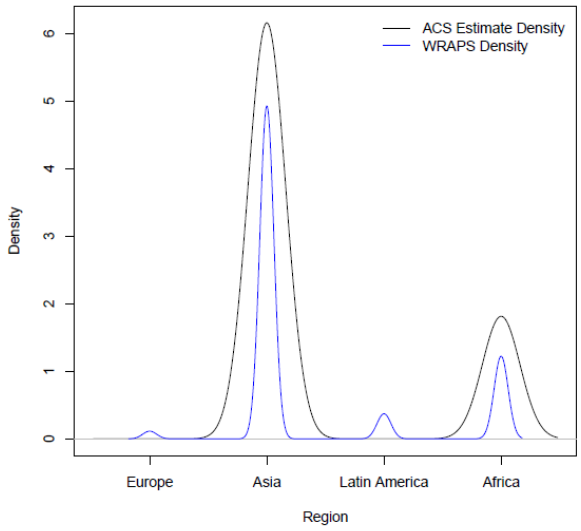


Table 3. Ordered table of most common places of birth for estimates of assigned refugees/asylees and non-refugees/non-asylees on ACS

Top Refugee/Asylee Places of Birth	Number of Refugees/Asylees	Top non-Refugee/non-Asylee Places of Birth	Number of non-Refugees/non-Asylees
Cuba	699,700	Mexico	11,561,000
China	178,300	India	2,387,000
Iraq	143,100	Philippines	1,975,000
Myanmar	112,900	China	1,928,000
Germany	111,600	El Salvador	1,349,000
Mexico	91,140	Vietnam	1,299,000
Ethiopia	74,040	Dominican Republic	1,059,000
Somalia	70,330	South Korea	1,057,000
Thailand	70,120	Guatemala	925,500
Italy	53,930	Canada	764,000

23. The countries of birth with the largest estimated numbers of refugees and asylees is compared to those for non-refugees and non-asylees, estimated from the ACS by this method, are shown in Table 3. The two lists are quite dissimilar, with only Mexico and China appearing on both lists – though this is likely due to the large number of Mexican-born and Chinese-born individuals on the ACS. Some notable countries appear on the non-refugee/non-asylee list that one might not expect, those being El Salvador, the Dominican Republic, and Guatemala. The number of individuals from these countries seeking asylum in the United States have seen surges in recent years [5], so we expected to see them among the top refugee/asylee places of birth. It is possible that these individuals, as asylum seekers, have not yet appeared on the LPR, as only individuals granted asylum are present on the dataset. They will appear on the set as they are granted asylee status in the United States, so the estimation method does not predict individuals whose place of birth is one of these countries are refugees or asylees. Again, Italy, Germany, and Mexico appear among the top refugee/asylee countries of birth, which may not be correct. More investigation is necessary to find the cause of this, whether the method is correctly assigning these individuals refugee/asylee status. The incorporation of more data sets may be necessary to resolve the misclassification, if they are indeed misclassified.

VI. Conclusion

24. The method described here, combining a logistic regression model with a decision made using a rejection resampling algorithm, provides an estimate of the refugee and asylee population from the American Community Survey that is consistent with data on the admittance of refugees and asylees and data on the resettlement of refugees in the United States. This is particularly valuable in that a sparse data source like the Legal Permanent Resident file can be employed to make estimates from the much more detailed ACS, which provides an avenue to make detailed demographic and economic estimates of the refugee and asylee populations currently living in the United States. Further, this method provides a stock estimate of the refugee and asylee population, whereas only flows of refugees and asylees are provided by the LPR.

25. Further research may explore getting access to other administrative data sets directly from the State Department to make estimates of other groups that cannot be directly surveyed in the ACS. More pertinent to this paper in particular would be the inclusion of earlier years of the LPR to train the logistic regression model – this model was trained on LPR sets from 2012 to 2015, though data sets exist for earlier years. This may improve the precision of the estimates, especially for the population of refugees in the wave following World War II and in the 1970s, where it was observed that many of the foreign born from Italy, Germany, and Mexico were assigned refugee/asylee status, possibly incorrectly, and refugees from Southeast Asia in the 1970s were missed.

VII. References

[1] Office of Immigration Statistics. 2016. *2015 Yearbook of Immigration Statistics*. Department of Homeland Security.

[2] Population and Housing Unit Estimates. 2018 https://www.census.gov/programs-surveys/popest/data/tables.html. US Census Bureau, Department of Commerce.

[3] United States Census Bureau. 2018. https://www.census.gov/programs-surveys/acs/about.html. Department of Commerce.

[4] Alejandro Jara, Timothy Hanson, Fernando Quintana, Peter Mueller, Gary Rosner. 2011. *DPpackage: Bayesian Semi- and Nonparametric Modeling in R*. Journal of Statistical Software, 40(5), 1-30.

[5] UNHCR Population Statistics. 2018. http://popstats.unhcr.org/en/asylum_seekers_monthly. The UN Refugee Agency.

[6] Refugee Processing Center. 2018. http://ireports.wrapsnet.org/. Department of State.