Producing Refugee and Asylee Estimates from Survey Data through the Incorporation of Administrative Data

Michael Bowerman, US Census Bureau
UNECE Work Session on Migration Statistics
24-26 October 2018

Introduction

- Nearly three million refugees have entered the United States since 1980
- Over 600,000 individuals have been granted asylum to the United States since 1990
- The United States Census Bureau collects little information on this large subcomponent of international migration
- This presentation focuses on identifying refugees and asylees on Census Bureau datasets by training a logistic regression model on administrative data

Data Sources

- American Community Survey (ACS)
  - Annual household survey
  - Large sample (3.5 million households)
  - Rich data source (demographic, geographic, economic info)
  - No question to identify refugees and asylees

Data Sources

- Survey of Income and Program Participation (SIPP)
  - Longitudinal survey
  - Small sample size (53,000 households)
  - Very detailed information
  - Question included that identifies refugees and asylees

Data Sources

- Legal Permanent Resident file (LPR)
  - Very large data set
    - Includes every person granted legal permanent resident status to the United States each fiscal year
    - Accessed yearly sets between 2012 and 2015, though they exist to 1973
  - Can directly identify refugees and asylees
  - Little socioeconomic detail included

Data Sources

- We desire detailed characteristics from the ACS and SIPP about the refugee/asylee population in the US
  - LPR has the ideal sample size, but lacks detail
  - SIPP has the detail we'd like, but small sample
  - ACS has the detail we'd like and a large sample, but no questions on refugee/asylee status

Data Sources

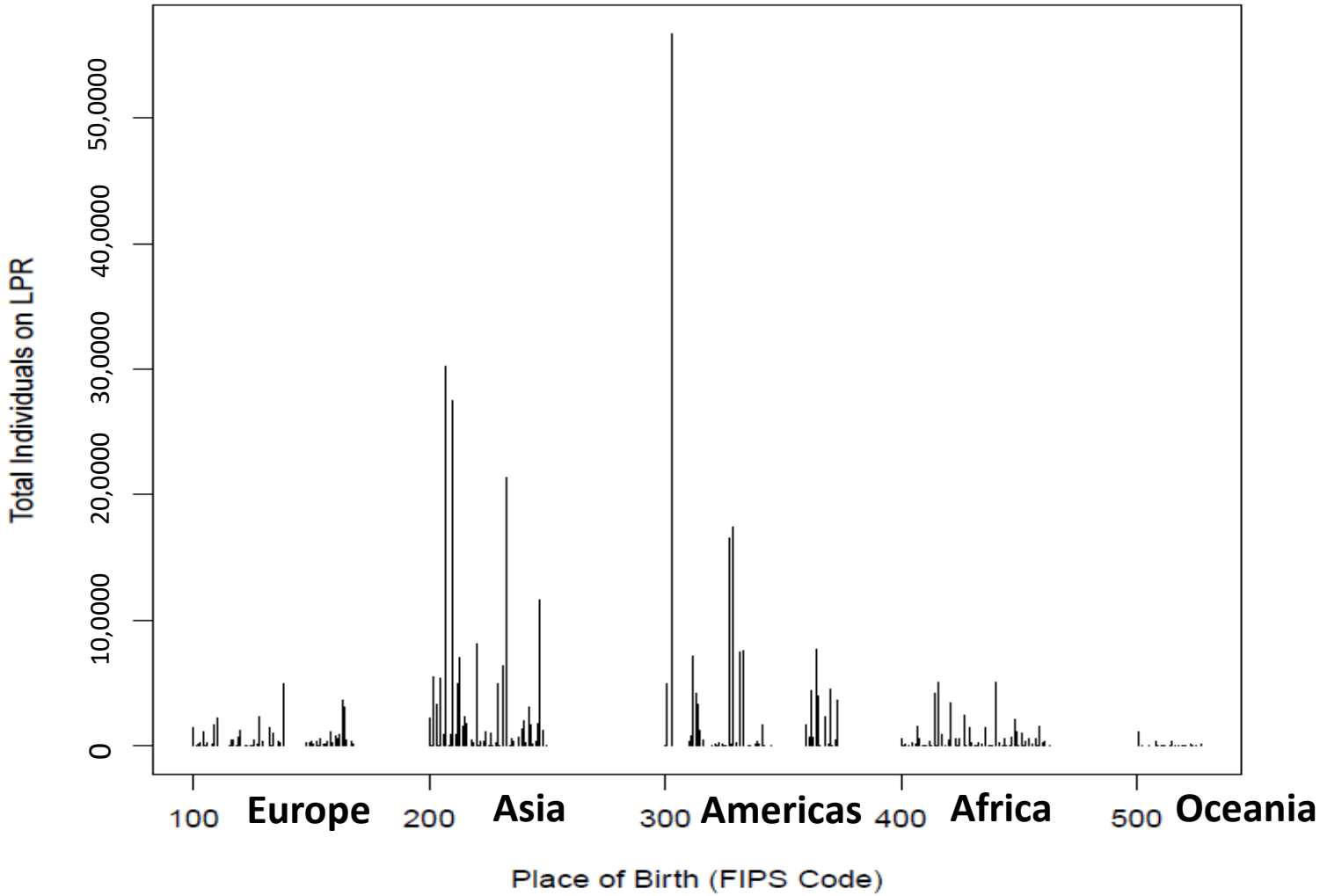
- Solution:
  - Identify refugees/asylees on the 2015 1-year ACS by training a model on the LPR files from 2012 to 2015

Variables on LPR

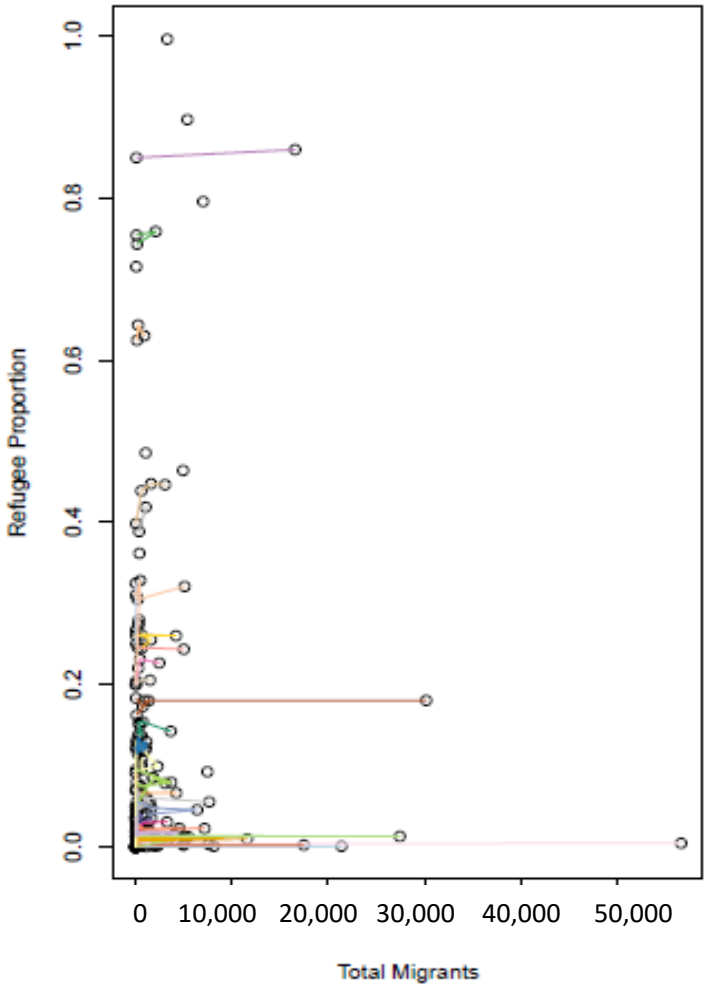
Variable Name	Details
Class of Admission	Type of visa the individual was admitted on – used to separate individuals into refugees/asylees and others
Country of Birth	Place of birth, with outlying US territories included
Country of Citizenship	Codes in this variable are nearly identical to Country of 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

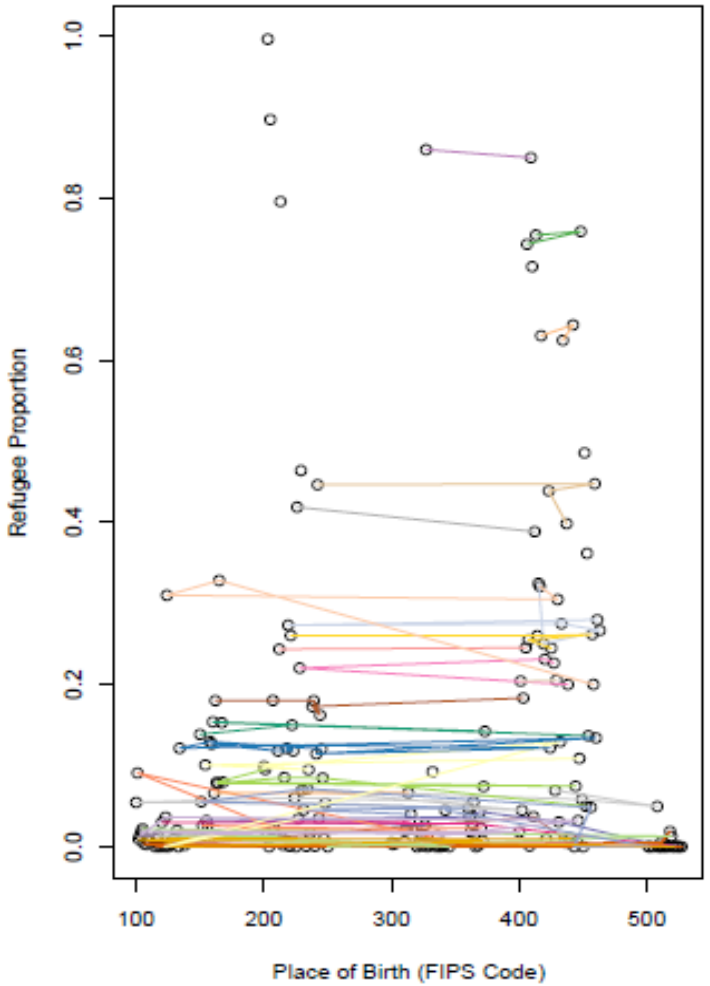
LPR: Country of Birth Variable

- 217 countries represented
- Some sample sizes for countries under 10
- Cannot use variable in this form in the model



Country of Birth Clustering

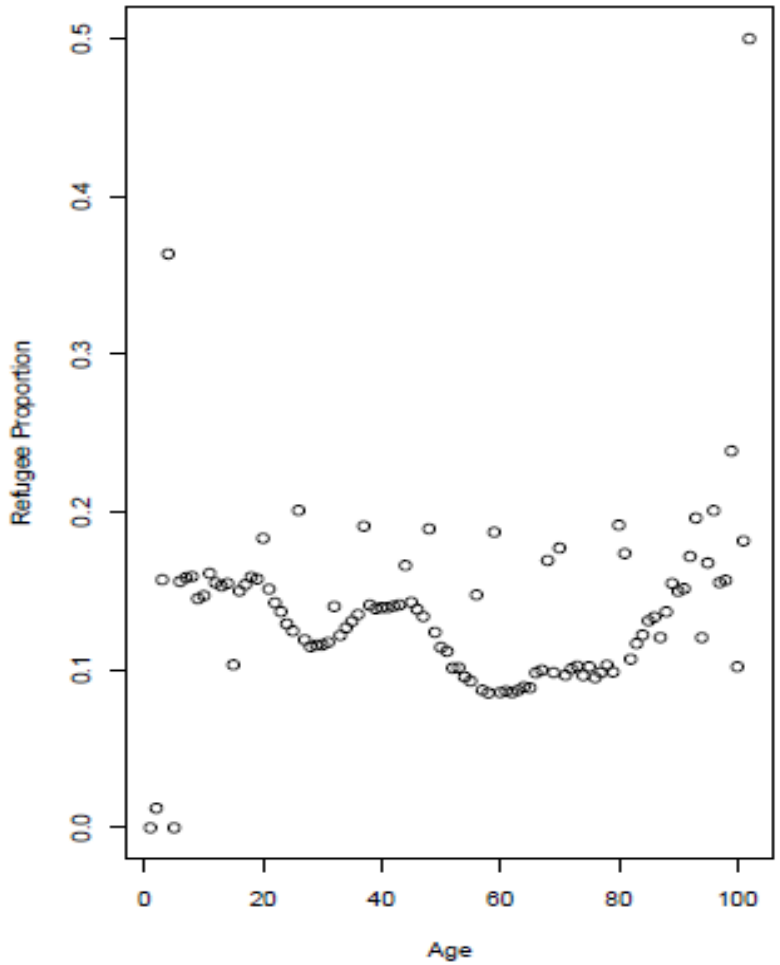


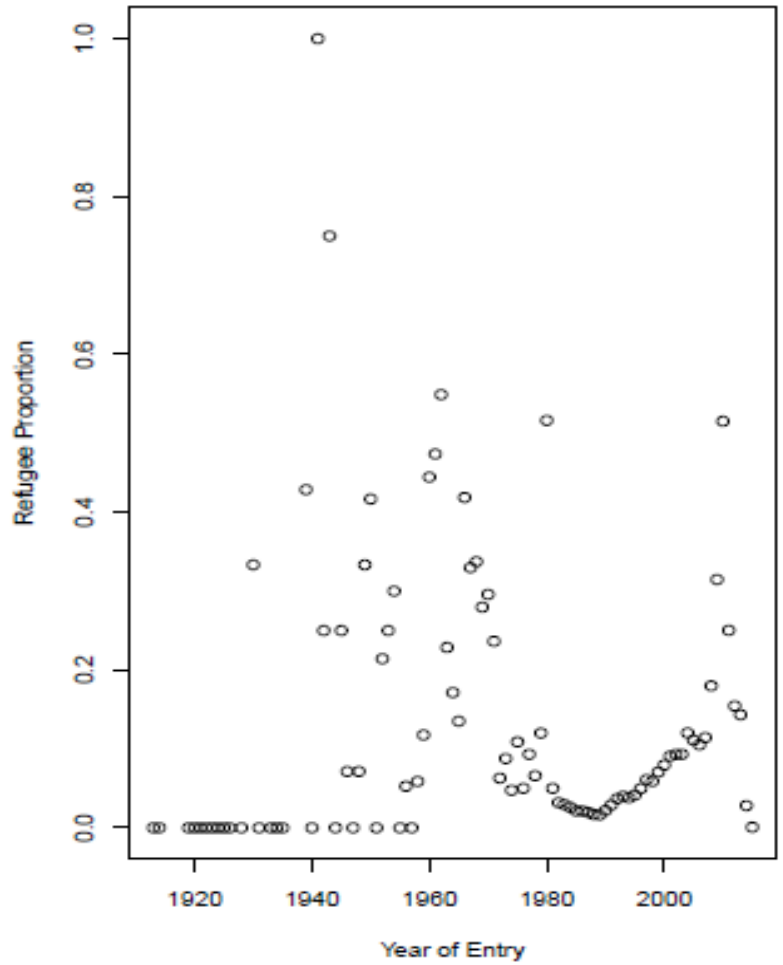


- Assign countries of birth on LPR to a cluster based on each country's propensity to send refugees/asylees to the United States
- Assignments made using semi-parametric beta-binomial mixture model with Dirichlet process prior
- Model identified 37 clusters

LPR: Age and Year of Entry Variables

- No clear relationship, linear or nonlinear, between age and refugee/asylee probability or year of entry and refugee/asylee probability
- These variables will be included in model using spline functions





Logistic Regression Model

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

- p_i is the probability that individual i is a refugee or asylee
- $S_{age}(age_i), S_{year}(year_i)$ are the values of the spline functions at the age and year of entry of individual i
- sex_i, mar_i are the sex and marital status of individual i
- $cluster_i$ is the cluster assignment of the country of birth of individual i
- ϵ_i is $N(0, \sigma^2)$ random error

Rejection Resampling

- We want to assign refugee/asylee status to individuals on the ACS
- Simple up-or-down decision at .5 would be influenced by model error
- Instead, simulate the assignment using a rejection resampling algorithm

Rejection Resampling Algorithm

For individual i:

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

Results

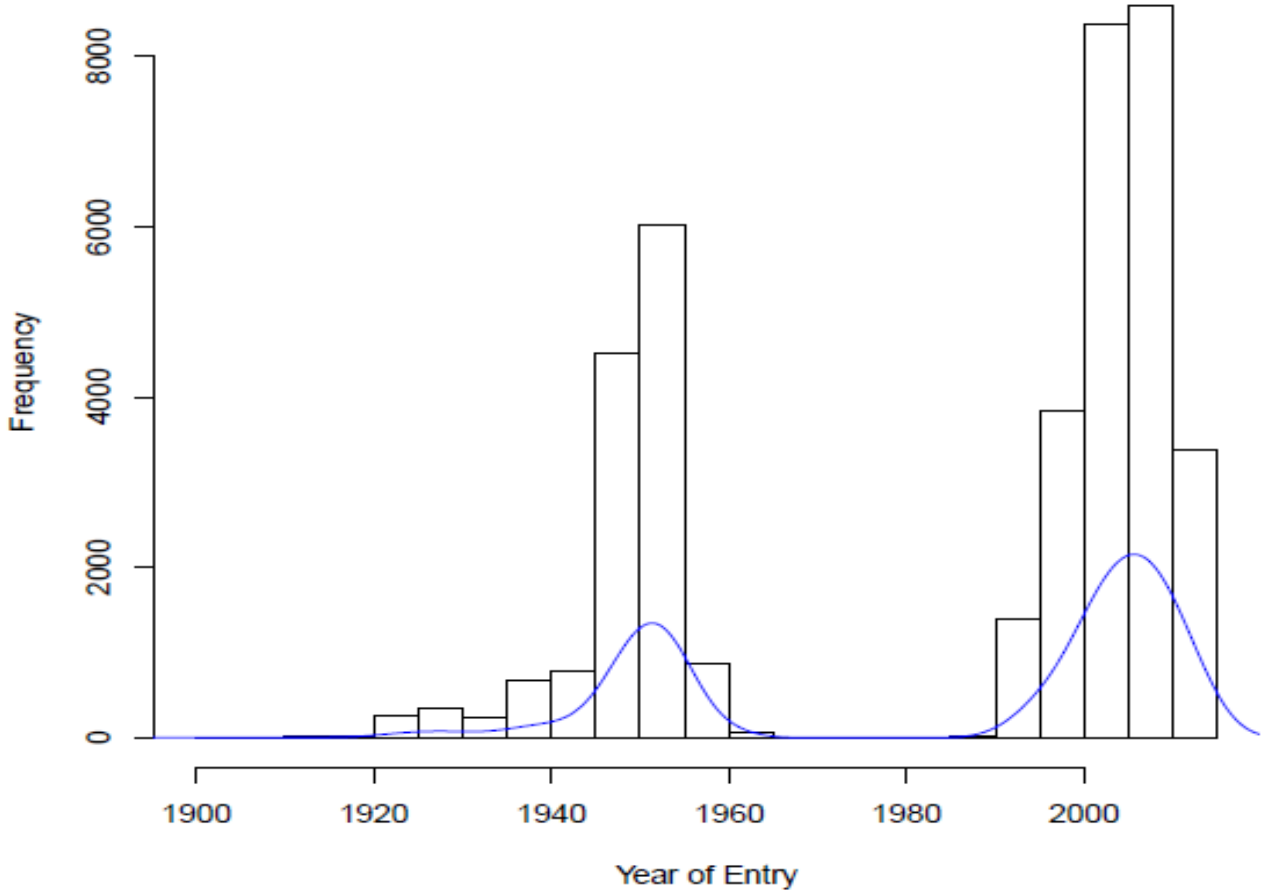
- Model was trained on LPR data
- Refugee/asylee probabilities for individuals on the 2015 1-year ACS were predicted
- Rejection resampling algorithm was used to create flags
- According to 2015 DHS Yearbook of Immigration Statistics, 2.5 million refugees and asylees entered the US between 1990 and 2015
- This method assigned 1.7 million individuals on the ACS refugee/asylee status who entered the country after 1990

Results

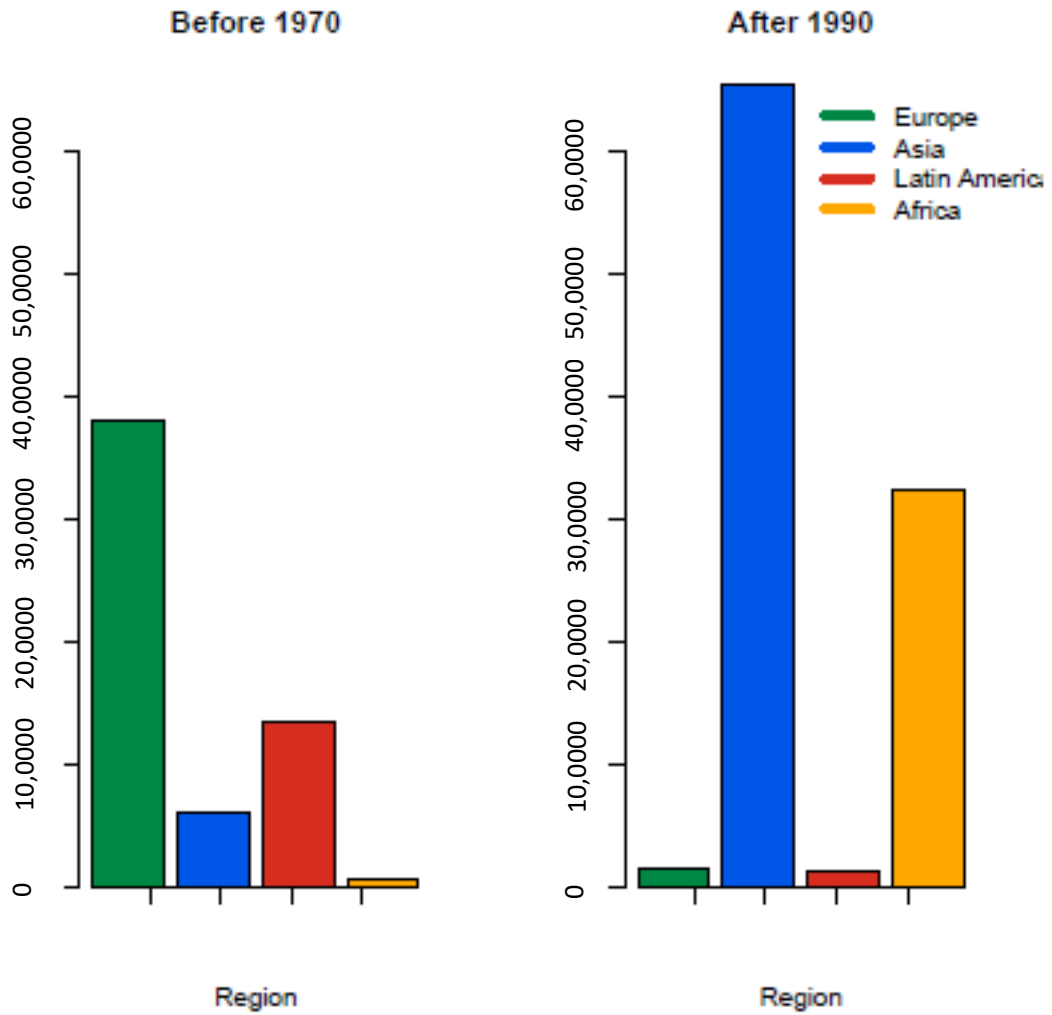
Percent in Category										
	Sex		Age			Marital Status				
	Male	Female	<25	25 to 65	>65	Married	Widowed	Divorced	Separated	Never Married
Refugee/Asylee	39%	61%	20%	38%	42%	39%	16%	9%	3%	33%
Other Foreign Born	48%	52%	14%	70%	16%	60%	5%	7%	2%	25%

Results

- Clear peaks following 1940 and after 2000
- Missing refugees from 1960-1990



Results


Before 1970
After 1990
0
10,0000
20,0000
30,0000
40,0000
50,0000
60,0000
Region
Region
Europe
Asia
Latin Americ
Africa

Results

Top Refugee/Asylee Countries of Birth	Number of Refugees/Asylees	Top non-Refugee/non-Asylee Countries of Birth	Number of non-Refugees/non-Asylees
Cuba	699,700	Mexico	11,560,000
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	925,000
Italy	53,930	Canada	764,000

Results

Top Refugee/Asylee Countries of Birth	Number of Refugees/Asylees	Top non-Refugee/non-Asylee Countries of Birth	Number of non-Refugees/non-Asylees
Cuba	699,700	Mexico	11,560,000
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	925,000
Italy	53,930	Canada	764,000

Conclusion

- The logistic regression model with rejection resampling algorithm reasonably assigns refugee/asylee status to individuals on the ACS
- Some inconsistencies:
  - Possibly improperly assigning status to individuals from Europe
  - Missing refugees/asylees between 1970 and 1990
  - Missing recent wave of refugees/asylees from Central America

Future Study

- Separate models for refugees and asylees
  - Profiles may be different, especially based on country of birth
- Incorporate more data
  - Earlier years of LPR data, back to 1973
  - Find separate dataset with refugee/asylee status
    - More variables would make assignments more reliable

Questions?

Michael Bowerman

United States Census Bureau

Net International Migration Branch

michael.d.bowerman@census.gov