August 24th, 2018

Refuge
Predicting Refugee Migration

Prepared by: Jose Pauig

Abstract

By the end of 2014, the estimate for total internally displaced persons and refugees was sixty million - a staggering amount. If machine learning has a role in aiding the resettlement process it might lie in predicting the future movements of refugees. If we knew in advance where refugees will originate, could the functions governing application, training, and support - the systems which drive successful resettlement - operate more effectively?

IPYNB files accompany this report and include the code which is the basis of this analysis.

Objective:

1. To build a model which takes as input a country represented by its demographic, economic, health, and trade measures and outputs a decision as to whether the country will generate refugees in a particular year, and;
2. To build a model which takes as input a country represented by its demographic, economic, health and trade measures and outputs a prediction on the number of refugees generated in a particular year and the destinations (at a state level in the US) of those refugees.
3. To see if different clusters of countries generate different volumes.

Sourcing and Cleaning our Data

Information representing the demographic, economic, health, work and trade measures of all countries recognized by the UN was sourced from: http://hdr.undp.org/en/data. The dependent variables for our models will comprise of this information.

Information about the volume, origins and destinations of refugees arriving in the US for each of the five years ending 2015 were taken from Buzfeed News' Github repository: https://github.com/BuzzFeedNews. It was originally collected from The Department of State's Refugee Processing Center.

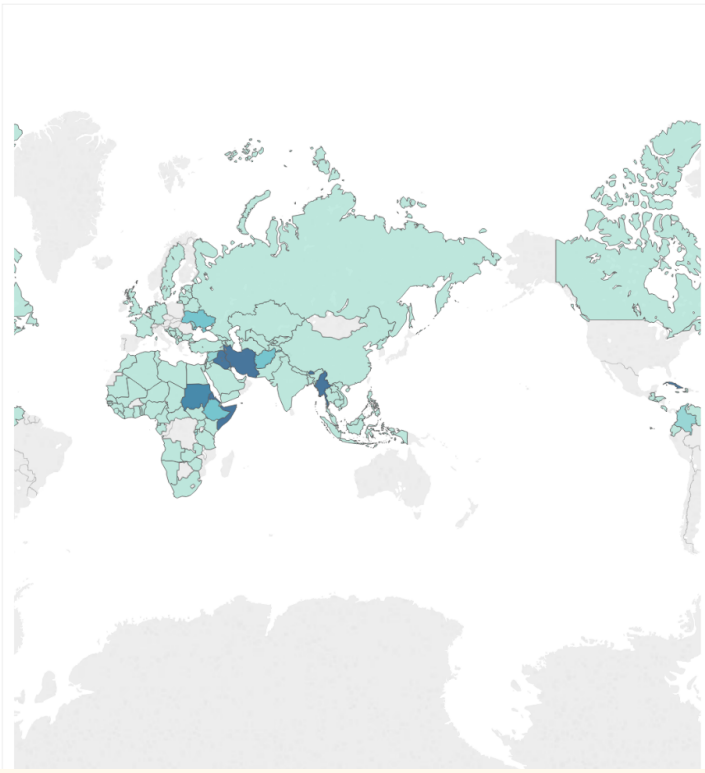
As the UN dataset contained missing information - missing countries, fields with missing values - certain countries and fields were discarded. For example any measure for which more than 20 countries were lacking information was removed from the dataset. Where a country or field was missing only a few values, the respective mean replaced the missing data.

Gaining Familiarity over our Data

Plotting the locations of origin countries that generate refugees, it becomes apparent that the highest volume origins generally cluster in East Africa and in two areas in Asia. Certainly some of the countries which output the greatest volumes of refugees come as no surprise. Burma which struggles with an ethnic conflict involving its Rohingya population, tops Iraq (from which most refugees claim to be escaping ISIS held territory). That Bhutan displaced 51 thousand refugees over our time period may come as a surprise. These countries raise two considerations: (1) There are a small handful of countries which generate the bulk of refugees, and (2) conflicts that are specific to these countries may explain the resulting displacement.

The Origins of Refugees in the USA

Origins of 2011-2015 Refugee Arrivals (USA)

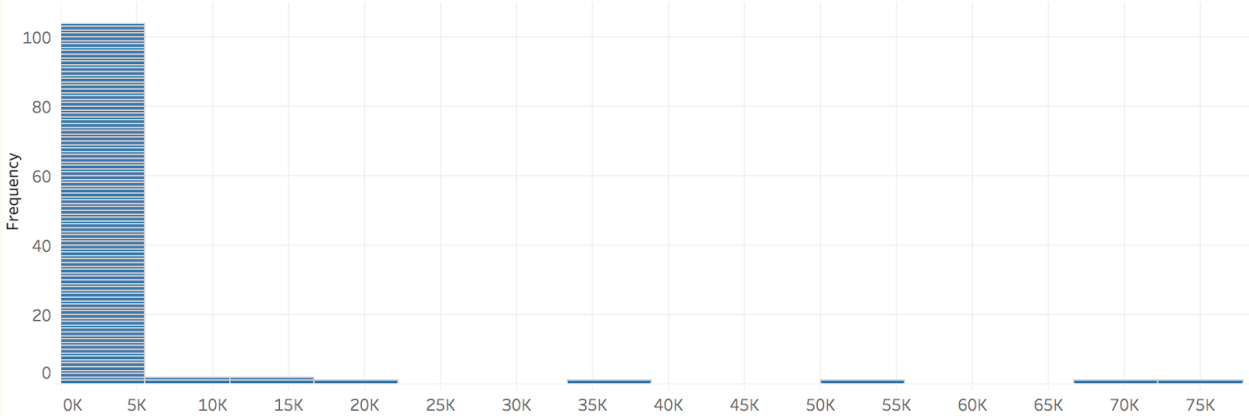


List of Arrivals by Country (US: 2011-15)

Origin	
Burma	77,790
Iraq	70,784
Bhutan	50,963
Somalia	33,837
Dem. Rep. Congo	18,606
Cuba	13,688
Iran	12,232
Eritrea	8,042
Sudan	6,634
Afghanistan	3,353
Ethiopia	3,338
Ukraine	3,202
Syria	2,227
Burundi	1,790
Colombia	1,253
Moldova	1,153
Pakistan	921
Russia	898
Central African Re..	896
Palestine	624
Rwanda	592
Vietnam	386
Congo	369
Uzbekistan	358
Sri Lanka (Ceylon)	354
Liberia	308

Given that our target variables are strictly positive we might expect a right skewed distribution for arrivals. As pictured below, significantly all countries output a low volume if any at all.

Histogram of Arrivals (2011-2015)



Using a Gapminder plot, we can visualize how arrivals change over two dimensions and over time. Included in the accompanying ipynb files is the gapminder file used in this analysis.

Building a Classifier to Identify Refugee Generating Countries

Using a Random Forest Classifier, we get R2 = 0.9 on a validation set (which arbitrarily was set to the 2015 data). We can therefore provide a reasonable prediction on whether a country will generate refugees.

Building a Regressor to Predict Volumes and Destination States of Refugees from an Origin Country

Using a Random Forest Regressor - which outperforms a gradient boosting regressor and linear regressor with regularization - we only get R2 = 0.2, a relatively weak result. A log transformation on the dependent variable improves R2 dramatically but abstracts the problem such that it loses its value as measured against our original objective.

Mapping the residuals, we know that we underperform disproportionately on those countries which generate the highest refugees. As we noted (above) the countries generating the greatest volumes are involved in specific conflicts which may not necessarily be represented by measures in our independent variable. I believe that improving the performance of the model would require additional fields which would do so.

Unsupervised Clustering to Identify Groups of Countries

Using k_means where k=5, identifies one small group of 10 countries including China, Armenia, Denmark, Albania, Lebanon, Argentina, El Salvador, Antigua and Barbuda, Burundi, and Albania. These countries do not share a common geography, level of development or any other immediately apparent attribute. Why they do cluster is a similarity in (1) low frequency of Adolescent Birth, and (2) low frequency of maternal mortality.

Though the within group distributions of arrivals for each of the 5 country clusters are not normal (as we expect the distribution is very much right skewed as a result of those countries which do not generate any refugees at all). But excusing the very far left of the distributions, they are somewhat smooth. If we apply an Anova test, we get a significant p-value which indicates that these groups do in fact generate different volumes.