DATA 1030 Project 2 Final Report: Movement Matters

Jason Terry

1 Introduction

For the past several years, the world has been plagued with a series of refugee crises from various countries. The severity and impact of these cannot be understated. Not only are they humanitarian and human rights disasters, but they have a tendency to significantly influence the politics of the countries that receive them. We need look no further than United States to see how divisive the prospect of taking in refugees can be. Such refugees have been described as invaders and criminals, views that have alarmlingy-large support. It can drive hardline, isolationist, far-right, xenophobic, anti-immigration, and often racist policies and ideologies. This can also be seen in the resurgence of an organized and empowered neo-Nazi party in Germany that is largely due to the massive influx of immigrants from nearby countries. With millions of people?s lives on the line combined with the high-stakes political consequences, it is important to understand refugee movement and use that understanding to make predictions of the future.

This project was an attempt to do just that. By using monthly data about refugee movement from specific origin countries to specific destination countries combined with yearly data on acceptance rates for destinations and battle-related deaths in origins, predictions of the number of accepted refugees for the current month and following two months were made. These predictions were made for every pairwise combination of origin and destination countries. The time period considered was 2000 to 2018. The origin countries considered were Afghanistan, Syria, Sudan, Myanmar, and the Democratic Republic of the Congo. These were chosen due to their high levels of violence and fleeing refugees at some point during the time period, and the fact that they span two continents (Africa and Asia) and three distinct regions (Central Africa, the Middle East, and Southeast Asia). The destination countries included in the study were the United States, Canada, the United Kingdom, Germany, France, and Italy. Similar reasons to the origin choices were used to choose these, i.e. geographic spread and a large number of refugees at some point.

The results of this study were fairly decent. Some of the models performed quite well for most combinations, however, there were instances of extremely poor performance for countries with extremely large and varied refugee numbers. Regardless, most results were close enough that informed policies and decisions could be made off of the predictions.

2 Data

Four different datasets were used. The first two came from the UN refugee agency, and arm of the United Nations High Commissioner for Refugees (UNHCR). One dataset was the number of refugees arriving each month in a given country, from a given country. The UNHCR dataset from this source was yearly data on refugee application statuses from a given country, into a given country. Features included numbers such as the number of applications, the number of rejected applications, the number of accepted applications, and the number of applications with no decision. This purpose of this dataset was to serve as a proxy for refugee policies in the destination countries. A high rate of rejection implies that there may be some negative feelings towards refugees at that time and vice versa.

The third dataset used was from the World Bank. It is the annual GDP for each origin and destination country over the entire time period. This dataset only has 12 rows and 19 features, so it was extremely small. It was easily cleaned by hand - this mostly involved making the names consistent. However, the GDP data for Syria is missing for years following 2010 as a result of the extreme violence there. Those years were imputed by assuming 3.7% growth, which was the growth for the final year available and a seemingly-reasonable assumption for a country in such a situation.

The final dataset that was used came from the Uppsala Conflict Data Program (UCDP), which is a project run by the Uppsala Universitet in Sweden. The data in question keeps track of yearly fatalities due to armed conflict for all ongoing conflicts. This was chosen because the assumption that violence and violent deaths corresponds to refugee movement seems valid. EDA showed this was the case, which will be discussed later.

None of these datasets were very large - on the order of hundreds of kilobytes. The yearly data was especially small since only 19 years were considered. The battle death dataset had many features, but the number of armed conflicts going on a given time is typically fairly low. The refugee status data had fewer features, but this was compensated by the fact that it included all pairwise combinations of the six destinations and five origins. Data on the number of refugees was the largest as it was monthly and also included all pairwise combinations. All of this data was merged on year and destination/origin countries. When doing this, special care had to be taken with the countries that have changed their names in recent memory (Democratic Republic of the Congo, Myanmar, and Sudan) since the battle deaths data often included the old names. Furthermore, that data often used names for conflict zones that are region-specific. A particularly bad example is ”Eastern Zaire,” which corresponds to the Democratic Republic of the Congo. The names were not consistent across datasets either (e.g. ”Syrian Arab Republic” and ”Syria,” which also had to be taken into account. Some features were removed during the merging process, which resulted in data slightly larger than the monthly refugee movement data. This data was then grouped into pairwise combination datasets, which were quite a bit smaller (30x smaller as that is the number of combinations). All cleaned datasets were stored in .csv files and manipulated using Pandas.

Even though the datasets were relatively small, they still contained a lot of relevant data. However, there were some serious limitations. The most limiting was the fact that the UNHCR did not contain any data on Italy prior to 2003. Accordingly, all values were set to zero during that time, even though that was clearly not the case. Furthermore, there were months when the UNHCR deemed it prudent to not publish the data because of efforts to protect the anonymity of the refugees. These cases were imputed with the previous month's values. The battle death main limitation is that it only publishes data where a government is involved in the conflict. While this is typically the case, it is not always so. Examples include tribal, guerrilla, gang, and ethnicity-based conflicts, which were present in the Sudan and Democratic Republic of the Congo in the early to mid-2000s. For this reason, the Latin American countries that I wanted to look at (El Salvador, Honduras, Venezuala, and Columbia) were not viable as most of the violent deaths fall into one of the non-government conflict categories, which meant that the data on those deaths was not available from this source. These were of particular interest because they send large numbers of refugees to the US and Canada when compared to the other origin countries considered, and therefore can be assumed to have a large impact on refugee-related policies.

3 EDA

The EDA process was quite helpful in discovering which features were important enough to be used. As expected, a fairly strong correlation between battle deaths and refugees was found. This is perhaps most clearly seen in the UK and Germany data. Evidence of this correlation is shown in Figure 1. All of these figures are snapshots of interactive figures. Syria has by far the largest fluctuations in both refugees sent and conflict deaths, the correlation of which is clearly seen in the figures.

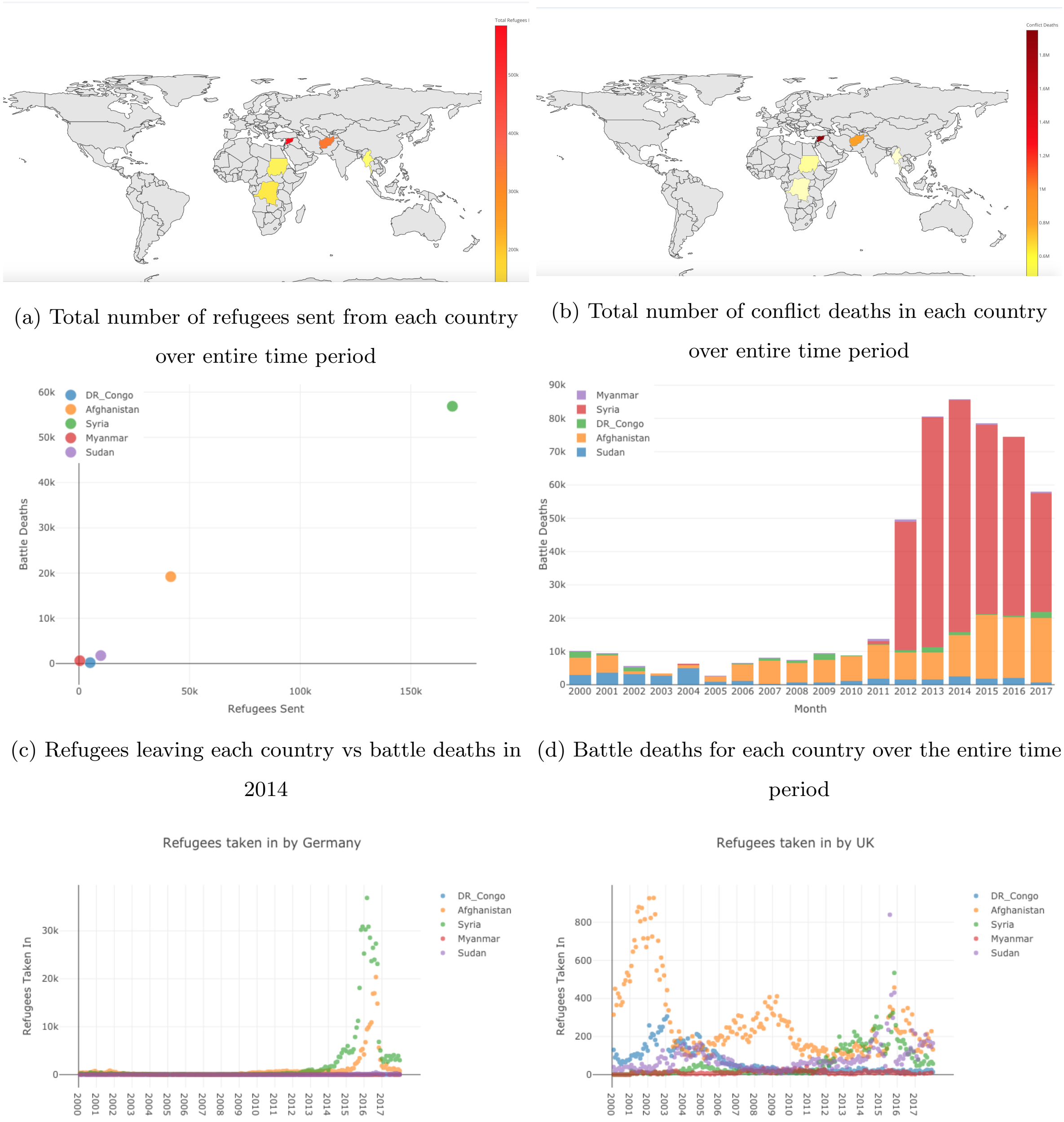
Another, and fairly unexpected though entirely reasonable, feature that was found to have a strong correlation with the number of refugees was the distance between the countries. This feature was somewhat stumbled upon by accident as I made a map of the refugee routes. Figure 2 shows the routes of refugees from and to each country. The opacity of the line is proportional to the relative number of total refugees sent to that country from a given country, i.e. the country that receives the most refugees from an origin has the highest opacity. A clear correlation can be seen. Other than Myanmar to the United States, the North American countries have a significantly lower opacity than the European destinations. This intuitively makes sense, but it was not something that I considered until I happened upon it. The strong correlation made it a very useful feature.

In addition, I made the number of refugees accepted for each of the last two months as features. This was particularly important for the baseline model, which only considered those features, but was useful for all other models. It was expected - and confirmed - that these correlate strongly with the follow three month's values.

The GDP data also proved to have somewhat of a correlation. This data was used to make features corresponding to the annual GDP for both the origin and the destination. Plots of the relationships can be found in Figure 3.

4 Previous Work

The closest work that I have been able to find to this project comes from Suleimenova, Bell Groen (2017). Unbeknownst to me when I started this project, they attempted to predict the where refugees from a given conflict will go, though their project was more focused on specific types of conflicts, e.g. violence against



(a) Total number of refugees sent from each country over entire time period

(b) Total number of conflict deaths in each country over entire time period

(c) Refugees leaving each country vs battle deaths in 2014

(d) Battle deaths for each country over the entire time period

(e) Refugees taken in by Germany from each country over the entire time period

(f) Refugees taken in by UK from each country over the entire time period

Figure 1: Evidence of correlation between violence and refugees

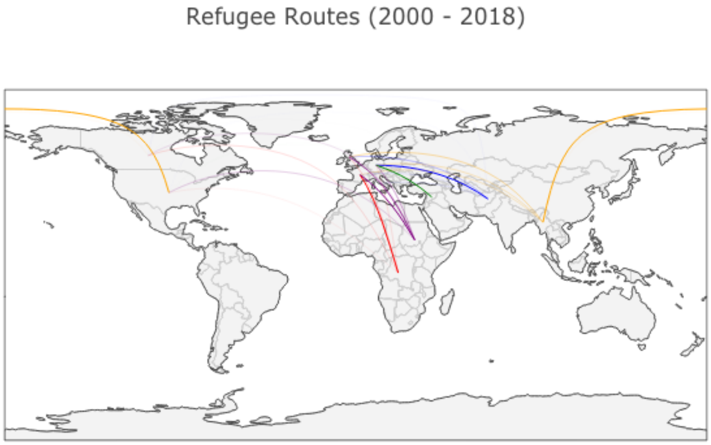
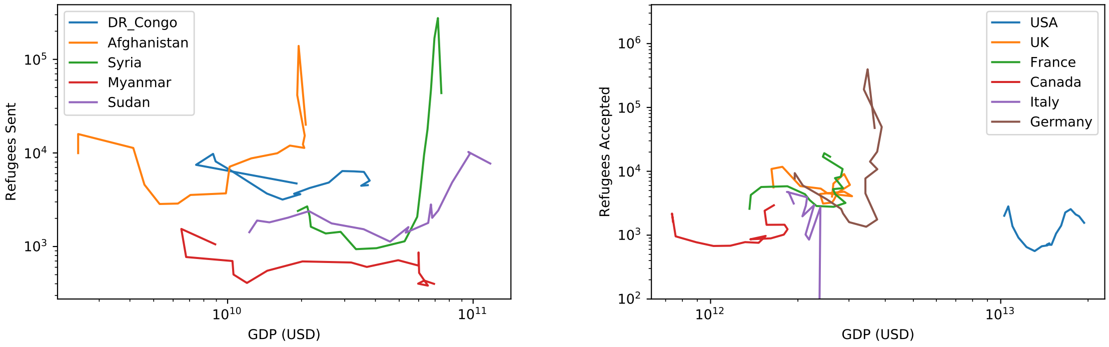


Figure 2: Refugee Routes. More opaque lines indicate that a higher number of refugees from a given origin arrived in that destination.



(a) Annual GDP vs total sent refugees

(b) Annual GDP vs total accepted refugees

Figure 3: GDP and refugee movement trends

protestors instead of all violence, and specific refugee camps instead of entire countries. They used some different data, and did not use the UCDP data. Finally, they only predicted 12 days into the future, whereas this projects attempts to forecast 3 months into the future. While their limited range of projection is probably more accurate than my long-term projections, I think that predicting 12 days into the future is not very helpful as far as policy or future expectations are concerned whereas 3 months into the future will give a better, albeit probably less accurate, idea of what to expect. This seems like information that will be more actionable for policy.

Their models worked fairly well. It was able to accurate predict the destinations of refugees after 12 days with 75% accuracy. This will be used as the benchmark against which I will assess the performance of my model.

5 Models

A number of models were considered. The simple baseline model was simply fitting a line from the previous two months and extrapolating. The advanced models included linear regression, Bayesian ridge regression, k-nearest neighbors regression, random forest regression, and boosted tree regression using XGBoost.

The aims of this project necessitated running 90 instances of each model. This is because there are three response variables of interest (the number for this month and the next two) and 30 pairwise combinations of destinations and origins. When running each instance of each model, hyperparameters were found using ten-fold grid search cross validation. Even though each dataset is fairly small, the combination of the cross validation and number of runs required led to the entire process to take several hours.

Initial EDA and model running led to the conclusion that using the log (base 10) of the number of refugees led to better results than using the actual value. This performance increase was measured by RMSE, which is the metric by which all models were judged. The numbers were converted from their logs before the metrics were calculated since this gave a clearer indication of the actual error.

Before any models or cross validation took place, 20% of the data was removed to be kept as out-of-sample data. When doing cross validation, the training data was again split 80/20 into training and testing data. This split was unique for each of the ten cross validation runs. Once cross validation was over and the best parameters found, the model was finally run on the data that it had never seen before: the remaining 20% of the data.

6 Results

The results of these models were fairly decent. For the most part, the values for the next to months were particularly better than I expected. There were some instances with terrible performance though. This only happened for months when there were an extremely large number of incoming refugees following a large fluctuation in battle deaths. It is most severe with Germany and Syria. When linear regression was used, there was on month with an error on the order of 10^{26}, a truly absurd number that indicated abysmal performance. However, for most months in the out-of-sample time range, the difference was much smaller, though some months had errors on the order of tens of thousands. This isn't as bad as it sounds because Germany took in as many as 36,860

Response Variable	Baseline	Linear	Bayesian Ridge	K-nearest Neighbors	Random Forest	Boosted Tree (XGBoost)
This month	140.7	419.1	100.6	26.0	31.4	32.3
Next month	410.3	453.1	78.3	32.3	28.9	37.6
Two months later	1665.9	152.3	83.2	26.6	29.8	30.9

Table 1: Median RMSE for each model and response variable

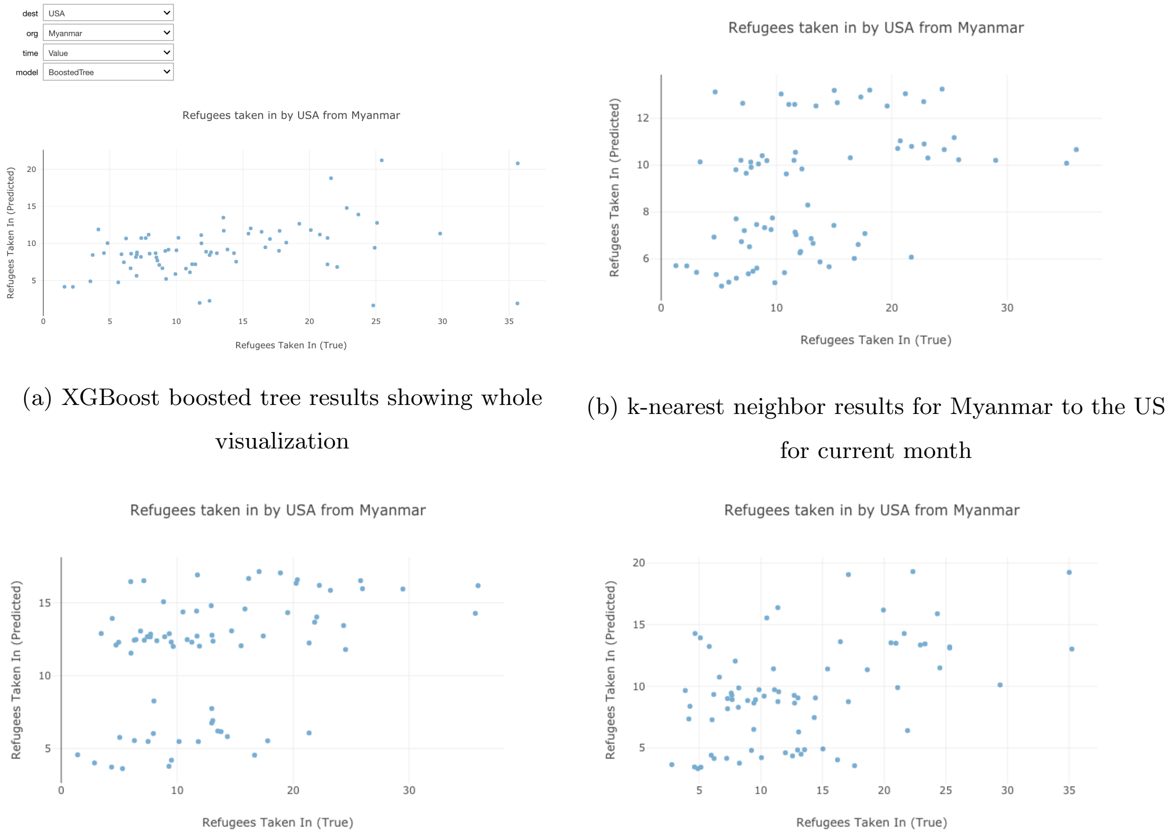
refugees from Syria in a month, but it is still over 30% error, which is not acceptable for making informed policy decisions. Because a handful of months were enormously wrong, but most weren't, it was decided that the metric to judge a model would be the median RMSE for a response variable rather than the average. These results are shown in Table 1.

As Table 1 shows, the baseline model, linear regression, and Bayesian ridge regression performed very poorly for all response variables. The baseline model performed particularly poorly, and its performance significantly increased as it had to predict further into the future. This makes sense because all it did was fit a line, which would be expected to generally be a worse prediction for points further away from the data used to make the line.

However, k-nearest neighbors, random forests, and boosted trees performed quite well on the whole. Their median RMSEs were below 40 for every response variable. This is deemed acceptable for making informed policy decisions as a difference of ± 40 refugees isn't so large that the decisions would be qualitatively different. Using the median RMSE values, k-nearest neighbors has been deemed the best overall model because it has the lowest median RMSE for two of the response variables and is fairly close to the best for the other. However, it is entirely possible to use random forests to predict that value for the following month, as that is the best predictor for that response, and k-nearest neighbors for the other two targets.

In order to visualize the results, I made in interactive plotly figure that allows the user to choose the destination, origin, response variable, and model of interest using dropdown menus. Using these values, a plot of the true number of refugees vs the predicted number of refugees was shown. This allows easy visual comparisons by simply analyzing the slope. Ideally, the data would fall on a straight line with a slope of one, so looking at where the data falls relative to that line is helpful. Examples of this visualization are provided in Figure 4. It is simply impractical to show all results, so only refugees coming from Myanmar and entering the US will be shown. These are fairly typical for situations with a somewhat low number of incoming refugees.

As Figure 2 shows, the results are fairly decent, though consistently underestimated. This is troublesome for policy because it is probably better to expect more refugees than you receive rather than the other way around, but since the number of refugees each month typically isn't very large and underestimation can now be assumed, the errors shouldn't result in qualitatively wrong policies in most situations. However, this comes with the important caveat that performance becomes significantly worse when the true number of refugees taken in



dest: USA
org: Myanmar
time: Value
model: BoostedTree

(a) XGBoost boosted tree results showing whole visualization

(b) k-nearest neighbor results for Myanmar to the US for current month

(c) k-nearest neighbor results for Myanmar to the US for next month

(d) k-nearest neighbor results for Myanmar to the US for two months later

Figure 4: Result visualization

is very high. In those situations, the results should not be trusted. This is quite unfortunate as far as policy is concerned because those are the situations in which effective policy is most critical.

Comparing these results to those found in Suleimenova, Bell Groen (2017) shows that these models do not outperform theirs. They are able to achieve 75% accuracy when projecting 12 days into the future. This is significantly better than a typical case in this study, which will generally sit between 40% - 60%. However, the gross difference isn't normally extraordinarily high and these models predict two months into the future rather than less than two weeks, so they are still considered to be useful.

7 Conclusion

This project resulted in models that can give fairly accurate predictions of the number of refugees taken in by a country for the current month and following two months when that number is low. The overall most effective model was k-nearest neighbors, though random forest regression slightly outperformed it when trying to predict the number for the next month. Given the success for months with a small number of accepted refugees, this model is considered to be accurate enough to make informed policy decisions in those situations. However, on the whole, the models performed incredibly poorly for months with an extremely large number of refugees. Given that result, the results for months with that quality should not be trusted or used in any sort of policy decision. Unfortunately, these are the times when effective policy is typically at its most important. However, these situations only occurred in European countries, and were most drastic for Syria and Germany as the number of refugees reaches over 36,000 for some months.

Because of that limitation, it does not seem like my work provides useful enough predictions to help the political landscape in Germany and, to a smaller extent, France and Italy as these are the countries that generally have the most refugees. However, the predictions for the UK, Canada, and the US were good enough in most cases to make informed policy decisions. Whether or not this is enough to have any significant sort of influence is outside of the scope of this investigation, though the results show that the number of refugees - which should correspond to their effect on a country to some extent - is typically fairly low for these combinations of countries. Given this fact, it seems that some of the anti-refugee rhetoric in the destination countries, especially for origins with a majority Muslim population, may be overblown. The predictive and descriptive power of the models in these situations may therefore serve as a solid basis for policy discussions. However, it is important to note that much of the current anti-refugee rhetoric in the United States - which has undoubtedly led to increased political polarization - revolves around refugees from Latin America, which were not included in this study.

Please consider my machine learning and visualizations for grading.

8 Sources

Diana Suleimenova, David Bell & Derek Groen, "A generalized simulation development approach for predicting refugee destinations." Scientific Reports, 7:13377, 2017

UNHCR Month Asylum Seekers Data. Accessed at http://popstats.unhcr.org/en/asylum_seekers_monthly in November 2018.

UNHCR Yearly Asylum Seekers Decision Data. Accessed at http://popstats.unhcr.org/en/asylum_seekers in November 2018.

UCDP Yearly Battle Death Data. Accessed at http://ucdp.uu.se/downloads/d8 in November 2018. World Bank Data. Accessed at https://data.worldbank.org/indicator/NY.GDP.MKTP.CD in December 2018.