Raman Kadariya

Prof. Martin Barron

FA22-MSBR-70260-02

12/12/2022

**Missing Migrant Project Modeling and Predictions**

The goal of this project is to understand the circumstances behind deaths and missing migrants’ cases and find ways to prevent them in the future. It could be achieved by implementing the insights gained by analyzing the dataset. The data is extracted from the Missing Migrant Project, from the United Nations International Organization for Migration (IOM) website which tracks migrants who have died or gone missing on their journey to their destination. Through the process of cleaning, analyzing, and modeling, I hope to understand overall patterns in the data, and use that knowledge to stop the recurrence of similar cases.

I would be more confident in the performance of my modeling and analysis if there were other similar r projects available for comparison. But, except for the United Nations report, there have not been significant third-party r projects based on the missing migrant dataset. There are around four projects that I found on Kaggle, and most were not in-depth, or used minimal analysis. One of the projects with some work done, used a different method of writing the code than how we were taught throughout the Machine Learning course. However, their data visualizations were insightful, such as the map of the North American region, the European region that had the highest number of incidents reported. But I took a different direction as I did not compare regions, but the highest reported cases, highest reported causes, event timeline, gender differences, and any existing similarities or anomalies in the data that would help explain why during a certain period, at certain region had the highest number of similar cases versus other places.

The project can be divided into two parts: data visualization, and statistical analysis. I use the function ggplot2 to plot visuals such as boxplot, density plot, world map, etc. For statistical analysis, I perform summary statistics, linear regression, and k-mean clustering. Going into the project, my expectation was to achieve a clean outcome that could be easily interpreted, but the data had to be vigorously cleaned, the null values had to be omitted, certain columns (cause of death) had to be exported into new columns, dummy variables had to be added and additional data frames had to be created. Still, the results from the project are not very accurate to the point that they could be implemented in the real world. However, it does give a good insight into refugees’ journeys and hardships around the world since year 2014.

The first part of the project is the visualization of the data. Visualization is especially useful in revealing data features that statistical analysis may miss, such as unusual distributions of the data, patterns, gaps, missing values, evidence of rounding, implicit boundaries, outliers, etc. I use “region of the incident” on a world map ggplot to see how the deaths and missing cases are distributed throughout different regions around the globe. The Central Mediterranean region seems to have most of the cases as it is having the highest spread of incidents represented by colored dots. I use scatterplots and line charts to understand the number of cases at different periods each year, in different regions. It seems, 2014 had the highest number of known cases, whereas 2019 and 2020 had smaller cases, which might be due to the Covid-19 pandemic.

The Mediterranean has the highest number of known cases, followed by the Northern African region. It also seems that most cases have originated from the African continents, which may be due to the ongoing regional conflicts. I turned my focus on the Mediterranean region as it seems to be at the crossroads of migratory patterns, and the month of September seems to have the highest number of cases, which could be due to the hotter climate, making traveling on boats safer than in winter. The Mediterranean and Aegean Seas seems to be the most popular route for refugee using boats as they are located between Western European countries such as Greece, Italy, etc; and Turkey, the Middle East, and Northern Africa.

I also use a boxplot to visualize the difference between the number of known survivors and the number of known dead cases and turns out most do survive their journey, but still, many migrants lose their lives in the process. So, I use a density plot to understand who is dying and missing more frequently and it turns out to be women and children, and they are also in higher numbers overall compared to men. It could especially be harder for children to safely journey to their destination as they have weaker bodies compared to adult men.

Next, I use summary statistics to get quick snapshots of the data to understand facts on the ground. For example, there is a total 22956 number of known dead or missing migrant cases, whereas there is 77641 number of survivors, which means more migrants survive their perilous journey than those who die or go missing. Overall, since 2014 most known migrants’ origin is from Latin America/Caribbean, followed by Sub-Saharan Africa, and Central America. Over the period, the most known leading cause of death is drowning, followed by Vehicle accident / death linked to hazardous transport, and Harsh environmental conditions / lack of adequate shelter, food, and water. However, the most interesting insight is that overall, Arizona, USA has the most reported know death cases, followed by the Milak border at Iran (but unknown location and cases are highest, but have no official reporting).

Then, I perform two different types of statistical analysis methods to understand any patterns in the data. The first is linear regression, which is a statistical model that analyzes the relationship between a response variable (e.g., y) and one or more variables and their interactions (e.g., x or explanatory variables). First, I create a new dataframe c\_dat, consisting only of important predictors that I want to use for statistical analysis. I also create a new column called “next year death and missing” as the response variable. I plot a histogram of the data using the response variable and since it is very right skewed, I convert the response variable to a logarithmic variable. There are still lots of values at 0, which is the number of survivors in the data. I carry out four different types of linear regression models, first is lm\_full (Linear Model), second is lm\_bwd (Backward Selection/Elimination), third is gam (Generalized Addictive Model), and fourth is nn (Neural Net), and check summary for each model, compare the best model to predict response variable.

However, the results are very mixed, where the lm\_full showed that “Northern Africa” had the smallest p-value, which means the highest co-relation, lm\_bwd showed “Region of Incident” had the highest AIC, meaning the highest co-relation, and gam too had “Region of Incident” with the smallest p-value. So, “Region of Incident” seems to have the highest co-relation to the response variable (log\_next\_year\_deaths\_or\_missing). In terms of error values, nn seems to have the least RMSE (36.01834), whereas both lm\_full and gam have the same MAE (12.64979). I also produce gam and nn plots, which seems to be crowded, so limiting the explanatory variables could give a better understanding of the correlation. So, the regions where the incidents of deaths and missing cases are happening could be a better predictor of future reoccurrence. Comparing this outcome to our previous visual results, Central Mediterranean should be the area of focus for preventing future incidents, migration routes consisting of the Northern African region.

To further understand links and relationships among the predictors and the response variable, I use the second statical analysis method, an unsupervised machine learning known as K-mean clustering. Clustering groups together a set of objects in a way that objects in the same cluster are more like each other than objects in other clusters. In K-mean clustering, clusters are represented by a central vector or a centroid. This is an iterative clustering algorithm in which the notion of similarity is derived by how close a data point is to the centroid of the cluster. I create a new data frame called off\_dat consisting of important predictors (Cause of death and missing is expanded into separate columns using the unique values as separate columns and separate rows as their count). Altogether, I create three different clustering, the first includes 4 clusters, the second includes 8 clusters, and the last includes 6 clusters. The optimal number of clusters is 2, which is derived by using “Average Silhouette Width”, and “The Gap Statistic”, but if we carefully follow the rising points on the graphs, it is evident that the optimal number of clusters is 6, not 2.

When running clustering with k=6, the second cluster shows more accidental deaths in the Southern Asia region, and most took place in 2020. The third cluster shows that the Mediterranean had the most drowning cases, followed by children with the highest death or missing cases between 2014-16. Also, when running clustering with k=4, the fourth cluster also shows that the Mediterranean had the most drowning cases, followed by children with the highest death or missing cases between 2014-16.

The culmination of visualization and statistical analysis, I can conclude that the Mediterranean region is the deadliest region for migrants as it has the highest number of known deaths and missing cases, which includes a higher number of cases involving children, and drowning being the leading cause of death. It could also be the case that the Central Mediterranean region is the most crowded route for migrants/refugees, followed by the Southern US border. So, targeted migrant-friendly policies and funding should use towards preventing migration crises in Sub-Saharan Africa (resolving regional conflicts), and the Mediterranean (food, shelter, health support, etc). Finally, the modeling is limited as it needs more cleaning and conversion to other forms to carry out other machine learning models such as random forest, xgboost.

**Works Cited**

https://data.unhcr.org/es/situations. *OPERATIONAL DATA PORTAL (REFUGEE SITUATIONS)*. 2022.

https://missingmigrants.iom.int/. *MIssing Migrants Project*. 2022.