BRIEFING NOTE:PREDICTIVE ANALYTICS PILOT
ANTICIPATING FINANCIAL REQUIREMENTS IN SOMALIA TO COMBAT FOOD INSECURITY

BY MANU SINGH
Predictive Analytics Fellow
Centre for Humanitarian Data

Summary

There is increasing interest in the humanitarian sector to improve preparedness of future famine risks. A key element of this preparation is setting rules for triggering financing which should be modeled with rigorous data and methodology. In this study we explore the feasibility of predicting humanitarian financing requirement in locations of severe food insecurity. An early estimate of the financial needs will be critical in preparation and on-ground interventions. Additionally, an effective implementation of the same dollar amount in a famine like crisis can significantly alleviate human suffering and result higher efficiency of humanitarian financing. The pilot study in Somalia outlined below shows very promising results.

Introduction

There is an increasing interest in predictive analytics and broadly the application of statistical analysis in the humanitarian sector. Despite the interest there are certain roadblocks such as a lack of clear understanding of statistical procedures, unavailability of clean data, data privacy issues, inertia to change and a generally high cost of initial investment. As a predictive analytics fellow stationed in The Hague for June-July of 2018, I was asked to demonstrate the value addition of predictive analytics to the humanitarian space. My personal objective in addition to this was to also understand the state of data as it exists and dispel some myths commonly associated with the discipline.

During my many fruitful interactions with various UNOCHA (United Nations Office for the Coordination of Humanitarian Affairs) colleagues, UNHCR (United Nations High Commissioner for Refugees) personnel, and local country teams I uncovered many opportunities where predictive analytics could add immense value. This value additional spectrum extends all the way from decision makers in head offices to field officers dealing with small daily data inconsistency issues. One such relevant venue for exploration was understanding the expenditure of OCHA controlled financial resources; could this procedure be made more scientific and less emotional? This question formed the basis of my fellowship and I dedicated the rest of my time in understanding how we can study past allocation patterns under various circumstances of food insecurity to predict future needs. Although this paper outlines only the pilot in Somalia, the theoretical framework can be extended to other country contexts just as easily.

The two OCHA-controlled funds studied in this analysis are:
Country-based Pooled Funds - (CBPF) - allow donors to pool their contributions to deliver coordinated and timely assistance.

Central Emergency Response Funds (CERF) has been around for 12 years and is responsible for time- critical emergency assistance for underfunded emergencies or rapid response.

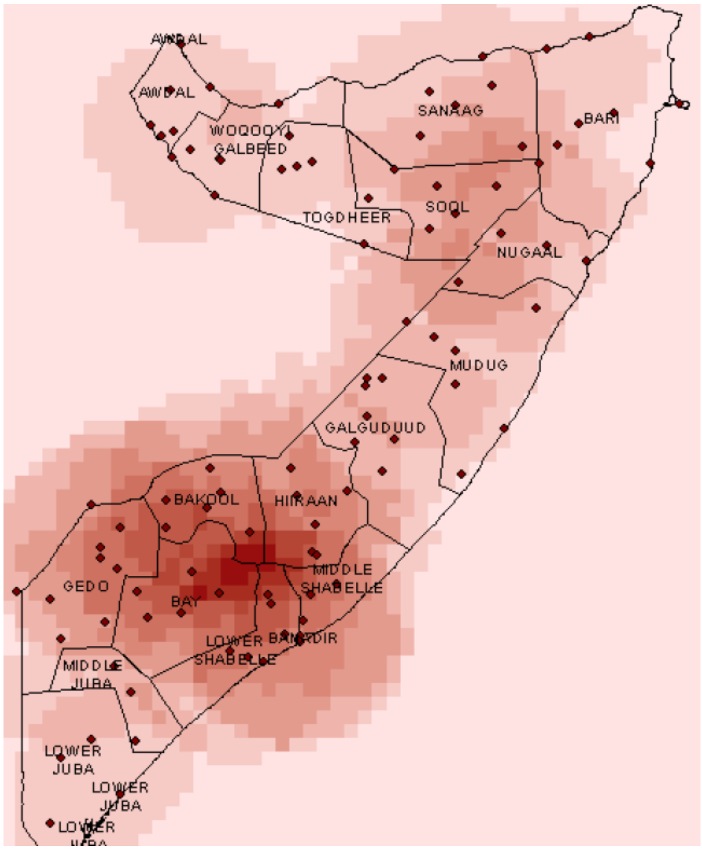

Figure 1: Distribution of CERF funds by location in Somalia in 2017. Here a darker colour represents a concentration of funds.

Modeling Strategy

The objective was to analyze spending per location per unit time so in the future if large food insecurity is projected in any area[1] we can estimate based on historical spending patterns what future needs would look like. As the foundation of this analysis is food insecurity, all available data on IPC (Integrated Phase Classification) levels was merged with the spending information. Here the former is the most important independent variable and the latter is the dependent variable. A large number of control variables are also accounted for in the analysis to improve model performance.

Data Requirements and Challenges

A large number of datasets were pulled in from various resources to make the model richer. Control variables are important because even though we are not directly interested in these numbers they are bound to have an effect on the dependent variable. Its imperative to net out the effects of what may be potential confounder variables in our study. With this in mind the following datasets and control variables were added to the model. Unless stated explicitly most of these datasets are sourced from the HDX platform.

[1] There is a parallel effort by theWorld Bank to predict phases of food insecurity and the underlying population numbers that are affected. An ongoing collaboration effort focuses on integrating their outputs of food insecurity using the FAM (famine action mechanism) model into our analysis of required financing. The study is still in its early stages.

centre for humdata
OCHA

Population Numbers - To estimate population densities and number of beneficiaries in each district or unit of analysis.

Road Density - Its important to understand that availbility of relief aid is highly dependent on the state and availbility of existing infrastructure. Thus a measure of primary and secondary road network density per district was added to the model.

Airport Density - Total number of airports (including paved and unpaved runways) per district.

Settlement Villages - A finer measure of population distribution is village settlements. For this analysis as the level of aggregation was district this measure was secondary but it will be very useful as the model gets more fine grained.

Violence - Fatalities and number of violent events in all the districts are tracked in a time-series from the Armed Conflict Location and Event Database. Violence restricts the ability of aid to reach the affected communities.

Health Indicators - Some health measures that reflect persistent food insecurity such as stunting in children under 5, maternal and child mortality rates are included from the WHO (World Health Organization) database.

Global Food Prices - A time-line tracking the prices of basic food products (grains, sugar, meat etc.) and some essential non- food items (such as cooking fuel, soap, water etc.) are tracked. A sharp uptick in prices of these commodities is reflective of impending shortages.

3W Data Matrix - Data to understand who is doing, what and where. Although this data is crucial to understand field conditions, beneficiary demographics and spread, in its current state the data needs a little more restructuring to be used in a machine learning model.

Vulnerability and Risk - A number of hazard exposure, risk and vulnerability measures from the INFORM dataset were incorporated in the model.

The next couple items are on the wish list but are not added to the model due to availability issues and time constraints :

IDP Numbers - To account for the constantly shifting population densities and a large number of internally displaced persons. One constant census number doesn't always reflect the truth. For example in case of a crisis the actual population densities in city centers is significantly higher than that calculated by the census.

High Frequency Survey Data - To closely understand situations at a more granual level than just the district.

Normalized Difference Vegetation Index (NDVI) - Incorporating some agricultural and vegetation index will make the prediction task easier.

Integrating data from so many sources comes at its own cost. IPC level information which is crucial to our analysis is published only every 6 months. Thus, richer granular data from multiple sources needed to be aggregated up reducing the number of data points. A large part of the puzzle also is getting an accurate measure of the number of individuals in need. In this analysis I use census numbers to get an estimate of population distribution. But in later stages of the project I learn that in countries like Somalia a large part of the population is constantly on the move thus population densities are always shifting.

An additional caveat worth mentioning is that needs may not always be captured by studying spending. For example, a village in distress could be inaccessible due to conflict or lack of infrastructure. Even though such hypothetical village has great need, but because funding hasn't been allocated in the past for whatever reason, a statistical model will continue to underplay its needs in the future. In the present iteration of the model I try and control for factors such as accessibility and conflict, there are other factors at play such as shifts in political control, amongst others, which aren't captured. Thus, if in the past money was not spent effectively the model will continue to learn from those bad decisions.

Lastly, data quality which is not an unknown issue, proved to be a bigger hurdle than anticipated. Lack of built in data integrity checks, non- uniform formats, and low priority of information management add considerable person hours to the task of data modeling. The financial databases which are the backbone of the analysis have sources of errors that are easy to fix. Some such errors are multiple records of the same project with different locations, the entire population of a district being counted as beneficiaries for certain project or locations and dates incorrectly recorded. On the other hand there are deeper data issues which can be addressed in the future iterations of the project. Examples of these data issues are lack of subnational data for multiple projects and only one or two rounds of high frequency surveys covering the whole country.

Once the data was cleaned and merged, I trained two models to study spending in every district of Somalia during each of the harvest seasons (Deyr - Winter rainfall season; Gu - Summer rainfall season) from 2011-2017 to make predictions for 2017-2018. This step was designed to validate model performance against actual spending information which is available for 2017 and part of 2018. Two models are used, one which is more interpretable and one which is less so. In a study like this absolute accuracy is less essential than model interpretability.

Random Forest

Random Forest is a versatile supervised learning algorithm. It is easy to use, usually fast and not prone to over-fitting

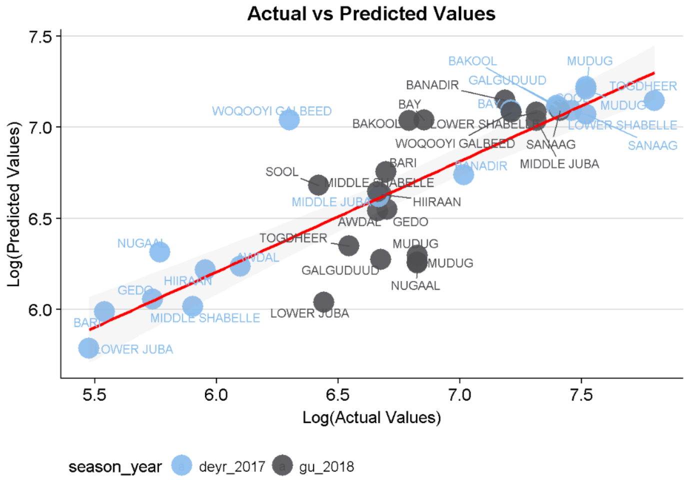

Figure 2: Plot of actual values vs predicted values on a log10 scale. Most of the predictions are close to their actual values. The results for 2018 (black dots) are worse than the results for 2017 (blue dots)

Elastic Nets

Elastic Net is a technique that falls under the group of Generalized Linear Models. This technique is very useful for modeling datasets that have highly correlated variables. The results from both the models above were comparable in error and gave consistent outputs.

Results

The preliminary results from both of the models look promising. The errors are low, and the models have captured a good amount of variance.

In Figure 2 for example, I show the results of Random Forest Model making predictions for the winter rainfall season (also known as Deyr) in 2017 and spring rainfall season (also known as Gu) in 2018 in blue and black respectively. With our current model setting we were able to assess variable importance.

It is interesting to note that the food prices are one of the most important variables in the prediction task next to only the number of beneficiaries. Measures of violence were also very predictive. Thus in the long run if we wish to improve our model it would be very beneficial to invest time in cleaning and getting good quality data on these measures. It is interesting to note that the predictions for 2017 are much more accurate than the predictions for 2018. This is evidenced by the fact that the black 2018 points are clustered around the center instead of being along the red line which is ideal. This finding suggests that with the current models we can accurately predict up to 6 months in the future. Predictions made further out start getting noisy in the current setting. This would be a venue for future improvement.

Conclusion and Recommendations

This pilot project brought forth a number of insightful learning opportunities and obstacles alike. On the one side, the humanitarian community is primed to receive and act on the fruits of quantitative and statistical analysis, but on the other, data quality continues to be a major drawback. But even with these drawbacks there is a lot of value which can be added intermediately. My study shows that even though we are not at a point where we can let an algorithm automate humanitarian funding, we can can assist decision makers with such models. There is immense scope of learning from past mistakes, make data-driven, consistent, and evidence-based decisions which are unbiased by emotions. Modeling financial spending can highlight previously unexplored issues such as gaps between actual needs and allotted funding. Moreover, this is just one step forward in the world of predictive analytics. With advancements in computing technology, availability of innovative datasets such as social media web-scrapes, sensor data, mobile phone data to name a few, models will continue to get better and more accurate. Thus, even though we are in the nascent stages of predictive analytics and application of machine learning in the humanitarian workspace, investments in this area are bound to have far reaching benefits.

The next steps in my research are to expand this model to other countries that are facing similar food insecurity. Although each country context and every crisis is different, it is worthwhile investing time in developing a strong theoretical foundation through collaboration and peer review.

Acknowledgments

I would like to express my gratitude to Dr. Dirk-Jan Omtzigt the Head of Information Management and Analysis Unit with OCHA Nairobi for his guidance and field expertise. I would also like to thank Juan Chaves-Gonzalez with Humanitarian Financing & Resource Mobilization Division in OCHA NY, for all his help in accessing financial data, and Dr. Eleonore Fournier-Tombs Data Scientist with the Centre for her technical expertise. Last but far from least I would like to thank Sarah Telford and Stuart Campo, from the Centre in The Hague for their unending support, words of encouragement and pragmatic guidance throughout the fellowship.