Analyzing and Predicting Political Instability and Conflict

Noah Dasanaike, Luann Jung, John Posada

Massachusetts Institute of Technology

Introduction

Political instability is an incredibly complex issue that is difficult to measure in a way that is comparable across countries or regions of the world. In this project, we wanted to explore the causes of political instability, and decided to approach the problem by analyzing violent conflict events reported by various sources across the world, and attempting to develop a statistical model to predict events of conflict using large amounts of past indicator data about a region. To do this, we utilize different indicator data to test new models for their predictive ability.

The previous literature has found significant difficulty in accurately predicting changes in conflict over time. From a comparative perspective, this arises from variance not only between countries, but between years. Models that were highly accurate in the years preceding 9/11 failed to predict the Global War on Terror; very few academics foresaw the collapse of the Soviet Union; and the Arab Spring has produced a wealth of literature in attempting to explain our inability to predict the inconsistent outcomes ranging from democracy to persistent authoritarianism (Bowlsby 2020). Additionally, because of the paradox that countries with the most conflict are those with the least data availability, we decided to be modest in our expectations.

Sources of Data

Conflict Data
We source our conflict event data from three projects. Each records information about the type of event that occurred, the actors involved, their geolocation, and the dates of occurrence.

- GDELT: event encoding from translated news reports using natural language processing
- ACLED: manual data entry but heavily skewed towards the reports used for data disaggregation
- UCDP: data acquired from search strings run through the Dow Jones Factiva aggregator

Indicator Data
Our indicator data captures a large amount of demographic, social, economic, and political information about countries. As such, we draw this data from several organizations which track these variables over time, but we can not list all the variables we used here.

- World Bank: approximately 562 development indicator variables of which any with entirely missing data were dropped
- V-Dem Project: several variables identified as being relevant to classifying national regime types, which we use to categorize country-years as either democratic or authoritarian

Methodology

Clustering
We believe that assuming there is a uniform relationship between the indicators and the conflict experienced in a country across all countries is faulty. To alleviate this, we perform k-means clustering on the countries to produce 5 clusters based on the *amount and type* of conflict they experience on a given year. This is utilized in performing data imputation. On a theoretical basis, this allows us to group countries based on the similarity of the conflicts they experience (i.e. countries at war are less similar to countries with ongoing protests).

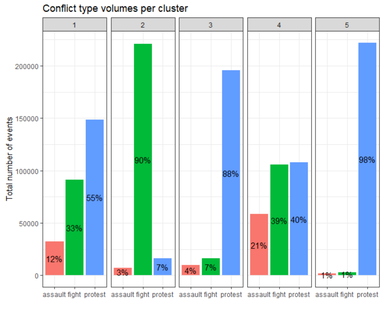

Fig. 1: Distribution of conflict types across the 5 country clusters

Preparation of Data
Multiple sources of our data were incomplete. Adapting the method used by the ViEWS project, we alleviate this problem using multiple imputation in order to ensure time analysis and prediction is as accurate as possible using the data. We impute via category means based on the clusters produced in the aforementioned k-means clustering.
In order to establish uniformity across our conflict data, we extended CAMEO codes (categorizations of conflict) to our three sources of conflict data. This unified our conflict data into one set which was categorized based on three CAMEO codes:

- *PROTEST*: Acts of protests, riots, etc.
- *ASSAULT*: Single-actor violence, terrorist acts, etc.
- *FIGHT*: Acts of war, international conflict, etc.

Models

ARIMA

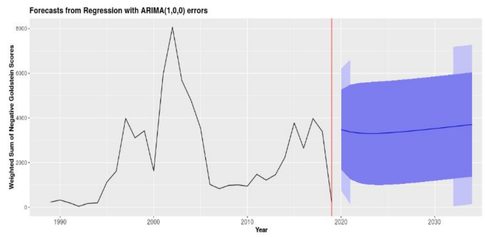

Fig. 2: ARIMA time series forecasting (2020 to 2035) for Uganda

Our first attempt at modeling involved ARIMA time series forecasting of conflict intensity based on a weighted sum based on Goldstein scores using the indicator variables as external regressors. We imputed missing data by means but ran into issues with very large confidence intervals for the predictions due to there not being enough data to distinguish any significant periodicity or trends. From this step, we decided to adjust the imputations in the data to be based on within-cluster means from our k-means clustering.

Stepwise Regression
For our first attempt at a non-ARIMA method, we fit a linear regression on the outcome variable using all of our indicators, then used bidirectional elimination to select the variables that produce the strongest statistical significance. However, this resulted in extreme over-fitting and was thus discarded.

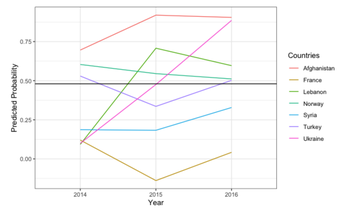

Fig. 3: Stepwise Linear Regression for 2014 to 2016

Random Forest
In an attempt to deal with our non-linear, large-variable data, we then split our data into sequential time folds, determined the best `mtry` parameter (number of variables available for splitting at each tree node), then fit using the `ranger()` function. We used `ranger()` rather than another implementation of random forest because of its speed and high memory efficiency compared to other methods. However, these results were highly skewed towards peace.

Elastic Net
Finally, we tried an Elastic Net model, wherein our training data was all countries from 1989 to 2013 and the testing data was countries from 2014 to 2016. This allowed us to incorporate a measure of correlation across time in our analysis. First, we tuned the alpha hyper-parameter using a predetermined fold-id, which was created by sampling random numbers across the response variable. Then, we used iterated cross-validation to determine the smallest lambda value; this method produced an alpha value of 0.9. From there, we determined the ideal cut-off value for the probabilities using accuracy and AUC.

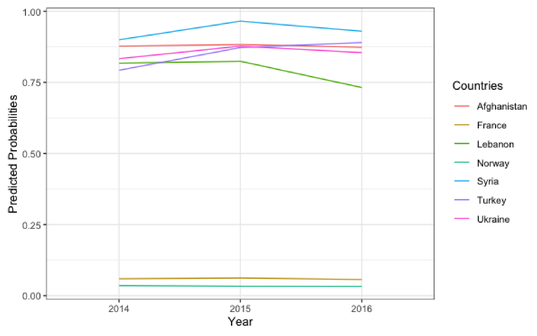

Fig. 4: Elastic Net for 2014 to 2016

Prediction

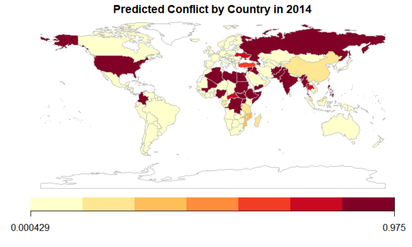

Fig. 5: 2014 predictive probabilities for conflict

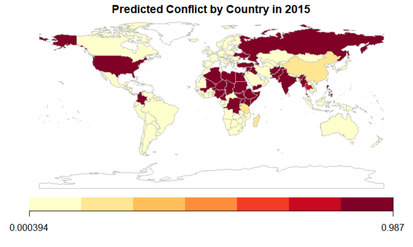

Fig. 6: 2015 predictive probabilities for conflict

Evaluation
In **Figures 5 and 6**, we can see heatmaps for 2014 and 2015 showing the intensity of the predictive probability of conflict for each country in that year. When trained on all countries' data from 1989 to 2013 and tested on all countries' data from 2014 to 2016, we obtained the following accuracy values and contingency matrix:

Accuracy	Precision	Recall	AUC
0.84	0.80	0.53	0.74

Fig. 7: Accuracy metrics

	Actual Peace	Actual Violence
Predicted Peace	401	21
Predicted Violence	72	82

Fig. 8: Contingency matrix for 2014-2016 predictions

One thing to note is that we seemed to have a high number of false positives. However, we will address this in our Remarks section regarding the reliability of the data.

Most Predictive Variables
Some of the variables that were most predictive in our model were the following:

Variable	Coefficient
Price level ratio of PPP to exchange rate	-0.248
Per capita GDP growth	-0.213
Initial household funding of tertiary education	0.128
Current year's violence value	4.536

Fig. 9: Coefficients for some variables in the Elastic Net Model

It is very reasonable and expected that the most informative variable in our model for predicting the next year's probability of violence was the current year's violence classification.

Remarks

Varying Accuracy Across Years
An interesting issue we discovered in the development of each of our models is the apparent extreme variation in accuracy across years. Despite this, previous studies have found that this sort of behavior is to be expected (Bowlsby 2020), and is inherently an issue with the fact that new conflicts are incredibly difficult to predict in this manner. We point out that all major conflict predicting models were unable to predict events such as the collapse of the USSR, and the events of Arab Spring.

Unreliable Data
Our models relied heavily on the data we acquired on whether or not a country was in conflict in a given year. We believe the over-reliance on this binary categorization hindered the true interpretability of the outputs of each model. When we attempted to instead output a continuous predicted probability, we found a similar polarization of the measure. There were numerous occurrences of countries such as India and Yemen being classified as peaceful in years that they were historically not – contributing to our supposed false positive rate.

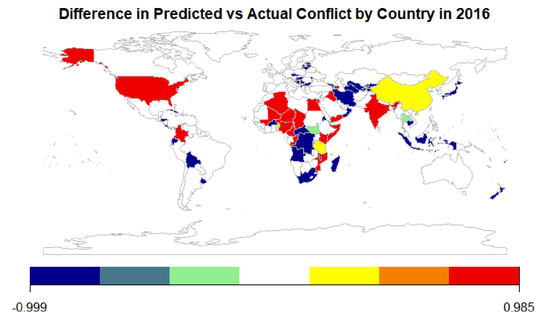

Fig. 10: Difference between model predictions and actual data in 2016

Summary and Next Steps

Overall, we were able to apply numerous models towards predicting conflict worldwide at a country-level granularity. We ran into a variety of issues with missing data, incorrect data, and also the unexpected nature of conflicts. However, our work provides a good heuristic for estimating predicted conflict in future years.
In developing this project further, we would work towards developing models for individual countries to better account for the variation in trends by country. Additionally, we would attempt to collect and analyze data with a higher resolution in order to predict regional conflict within countries. Finally, we would spend more time interpreting the predictions of our model in the context of historical events.