Revised Research Proposal
STAT 222: MA Capstone

Matt Boyas

February 19, 2014

Note: This document constitutes a revised version of the original research proposal submitted in class on February 3. I tried to address the comments relating to the regression analysis and the corresponding plots at the end of this revised draft; everything else remains unchanged. The original graded proposal is attached.

1 Dataset

I plan on using the Social Conflict in Africa Database (SCAD) as my primary data source. The SCAD was prepared by Cullen Hendrix and Idean Salehyan for the program on Climate Change and African Political Stability (CCAPS) at the Robert S. Strauss Center for International Security and Law at the University of Texas at Austin. The data and codebook are made available online for research and other purposes.[1] The SCAD project authors describe the dataset:

> The Social Conflict in Africa Database (SCAD) includes protests, riots, strikes, inter-communal conflict, government violence against civilians, and other forms of social conflict not systematically tracked in other conflict datasets. SCAD currently includes information on over 7,900 social conflict events from 1990 to 2011.[2]

Depending on the progression of the research, it could be advantageous to use secondary data sources to add more information into the SCAD. Possible additional information and sources could include:

- information on a country's democracy/freedom level taken from either the Polity IV Project[3] or the Freedom House Freedom in the World Scores[4]
- information including data on major armed conflicts, territorial disputes, alliances, and world religions taken from from the Correlates of War Project[5]

Like the primary SCAD data, these supplementary datasets are made available online for research purposes. Merging information from these – or other – sources into the SCAD will be undertaken as required by the project after consultation of Victoria and/or Christine.

2 Research Question

Using SCAD as a primary souce of data, I intend to address the following primary research question:

> *What differentiates an episode of social conflict that results in deaths from an episode of social conflict that does not result in deaths?*

[1] `https://www.strausscenter.org/scad.html`
[2] Ibid.
[3] Prepared by The Center for Systemic Peace, `http://www.systemicpeace.org/polity/polity4.htm`
[4] Prepared by Freedom House, `http://www.freedomhouse.org/report/freedom-world-aggregate-and-subcategory-scores`
[5] `http://www.correlatesofwar.org`

Decisions about secondary research avenues will be made in consultation with Victoria and/or Christine after significant progress has been made on the primary research question. Secondary research questions, directly related to the conclusions of the primary research question, could include the following.

- *Is there a way to predict the number of deaths that will result from an episode of social conflict?*

  Note that this model could either forecast the number of deaths in absolute terms, or it could predict the number of deaths relative to the size of the conflict, which could be measured by the number of participants or the length of the conflict.

- *Are certain countries in Africa more susceptible to social conflict resulting in deaths? If yes, what might account for the differences between countries?*

  If analysis shows that certain countries are more susceptible to social conflict resulting in deaths, possible factors to investigate to explain the differences between countries could include, but are not limited to, the dominant country religion, central government type, and/or country freedom level.

3 Figure/Table Titles & Captions

Possible figures include:

- Title: Deaths and No Deaths in African Social Conflict Events, 1990–2011

  Caption/Description: This figure is a split bar-chart, with one bar per year, visually showing the number of conflicts resulting in deaths relative to the total number of conflicts per year. Each bar represents all conflicts in the specified year and is split into two pieces, one piece for conflicts resulting in deaths and one piece for conflicts resulting in no deaths.

- Title: Distribution of the Number of Deaths in African Conflicts, 1990–2011

  Caption/Description: This figure is a histogram showing the distribution of the number of deaths in all of the African social conflicts from 1990–2011.

- Title: Distribution of the Number of Deaths Per Conflict Participant, 1990–2011

  Caption/Description: This figure is a histogram showing the proportion of the number of deaths relative to the number of participants (i.e., deaths per participant) in the conflict in all of the African social conflicts from 1990–2011.

- Title: Distribution of the Number of Deaths Relative to Conflict Length, 1990–2011

  Caption/Description: This figure is a histogram showing the proportion of the number of deaths relative to the length of the conflict in days (i.e., deaths per day) in all of the African social conflicts from 1990–2011.

Possible tables include:

- Title: Top 10 Most Violent Conflicts, 1990–2011

  Caption/Description: This table shows the top 10 most violent conflicts (ranked by number of deaths) along with important accompanying information such as country, dates, and the major conflict issue.

- Title: Deaths and No Deaths by Political Regime/Dominant Religion/Freedom Level, 1990–2011

  Caption/Description: This table shows the number of death-resulting conflicts and the number of zero-death resulting conflicts split by political regime type, dominant country religion, and/or country freedom level.[6]

[6]The specifics of this particular table depend on what additional data source(s), if any, I decide to merge into the SCAD. If I end up adding multiple variables that could be appropriate for such a table, then I will include multiple tables in the final paper.

Additional figures and tables will appear depending on the success of the modeling effort and the type(s) of model(s) created. My current plan is to investigate the development of two models: (1) a logistic regression model to predict the indicator variable of deaths/no deaths in a conflict and (2) a linear regression to predict the number of deaths in a conflict. I could include the following tables/plots for both models:

- Table Title: Regression Information

  Caption/Description: This table will summarize the regression output, including information such as the coefficients, significance levels, estimate of MSE, and R^2.

- Figure Title: Residuals vs. Fitted Values

  Caption/Description: This scatterplot compares the residuals from the regression to the fitted values. I can use this plot to assess potential issues of nonlinearity, unequal error term variances, and/or outliers.

- Figure Title: Normal Q-Q Plot

  Caption/Description: This plot will be a quantile-quantile plot, comparing the theoretical normal quantiles to the sample residual quantiles. This plot will help me assess the assumption of normality.

I also plan on dividing my data into training and validation sets to assess model prediction accuracy. Possible tables and figures for model validation could include:

- Table Title: Model Selection Process, Logistic/Linear Regression

  This will probably be two tables – one for the logistic regression and the linear regression – showing the mean squared prediction error estimates for versions of the models including different covariates. These tables will only be included in the paper if there are highly interesting results; otherwise, I will probably say that I chose the model that minimized the estimate of mean squared prediction error.

- Table Title: Predicted Values from Validation Dataset, Logistic Regression

  This item will be a table summarizing the prediction power of the logistic regression model. I will have one row for the No-Death conflicts and one row for the Death conflicts, and I will print the percentage of conflicts in that specific category that the model correctly classified.

- Figure Title: Predicted Values from Validation Dataset, Linear Regression

  This item will be a scatterplot showing the observed number of deaths on the x-axis, sorted in increasing order, and the corresponding predicted values on the y-axis. A 45-degree line will be superimposed over the points to show what the plot should look like if the model were 100% accurate.