*Hundreds of people demonstrated on Thursday outside the Johannesburg High Court to protest the suspension of Zwelinzima Vavi, general secretary of the Congress of South African Trade Unions (COSATU). "We are here to give him our support. We shall protest until the court overturns his suspension," [protester Patrick Malume] said.[[1]](#footnote-1)*

*27 March 2014*

*Johannesburg, South Africa*

*A teenager was shot dead during a violent protest in a region of South Africa which has been rocked by days of rioting… "It was a protest action, there was a crowd or a mob, and somebody took out their firearm and shot at the young man and he died on the scene," North West police spokesperson Thulani Ngubane said.[[2]](#footnote-2)*

*10 April 2014*

*Christiana, South Africa*

Two protests in South Africa occurred within two weeks of each other in Spring 2014, both relating to trade worker factionalization spurred by dissatisfaction with the current ruling party. The protest in Johannesburg was a peaceful demonstration outside the High Court, complete with flags and signs. The protest in Christiana was a riot, complete with firebombs, looting…and a death. What happened in Christiana that did not happen in Johannesburg to result in a casualty? I attempt to quantitatively approach this question by analyzing social conflict events throughout the African continent.

1. Data and Research Questions

Social conflicts—defined to be conflict events including, but not limited to, protests, riots, strikes, inter-communal conflict, and government violence against civilians[[3]](#footnote-3)—are differentiated from full scale intra-state war and mobilized conflict, but this differentiation does not make social conflict events less ubiquitous, disruptive, or dangerous. First released in March 2011, the Social Conflict in Africa Database (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 SCAD provides data on 7957 African social conflict events occurring in 48 countries between 1990 and 2011. Figure 1 and Table 1, below, attempt to give an overarching overview of the conflicts included in the SCAD.

**FIGURE 1 & TABLE 1**

Secondary datasets are merged into the SCAD to add more information about country demographics; these datasets include the World Religions (WRP) and National Material Capabilities (NMC) datasets from the Correlates of War Project and the national freedom scores from the Polity IV Project. All datasets used in this project are available for direct download from the websites listed at the end of this paper and in local cached form from the project’s GitHub account.

In this course, we reversed the ‘traditional’ scientific process in that we started from data and then developed research questions. I started out the process with the general idea of investigating differences between social conflict events. After seeing my available data, I decided to focus on the following two main areas of inquiry and investigation:

1. What differentiates an episode of social conflict that results in deaths from an episode of social conflict that does not result in deaths?
2. Is there a way to predict the number of deaths that will result from an episode of social conflict?

2. Who Cares? Possible Impacts Of This Study

Burdened with colonization until the mid 20th Century followed by, in many cases, a transition to independence hijacked by dictatorial rule, African countries are still embroiled in a transition from colonial rule to true independent, democratic governance (**SOURCE**). As such, there is a relatively wide body of literature discussing African social conflict in the context of an overall process of political regime change; however, the body of literature on African conflict as a unique entity is almost nonexistent (Scherrer 4).

Nevertheless, there is a somewhat substantial theoretical and case study based literature about social conflict in general, not necessarily relating to the African continent. Scholarship in this category does not agree on a concise set of factors resulting in the escalation of social conflict events. Some researchers such as peace scholar Christian Scherrer and anthropologist Jay O’Brian use ethnic and/or religious identity to explain differences in conflict severity, citing the claim, “Few of the nation-states created by Europe in Africa bore any relationship to any [natural ethnic or religious divides]” (O’Brian 63). On the contrary, geographer Adrian P. Wood argues, “Shortages of natural resources lead to competition which may result in conflict” (83). In yet another contrary argument, sociologist Ralf Dahrendorf identifies economic disparities across a country’s social classes as a potentially major factor, and economist Massimo De Angelis expounds on this idea by giving an example from the United States: “The Great Depression, with its historically high levels of unemployment, did not make the American working class more docile. On the contrary, it sparked open insurrection: ‘Don’t starve—fight!’ was one slogan” (Dahrendorf 52; De Angelis 51). As sociologist Nigel Fielding writes, conflicts can be attributed to any, some, or all of “class, ethnicity, gender and sexual politics, region, nation, employment status, age, and ideology” (5). There is no consensus on how to explain conflict.

Since the release of the SCAD dataset in 2011, a handful of quantitative studies specifically focusing on African conflicts have been published using the SCAD as a primary source. These quantitative studies, like the earlier-referenced literature, identify a variety of factors influencing conflict escalation ranging from climate change to food price spikes to the occurrence of elections (Devlin, Franck, & Hendrix; Smith; Salehyan & Linebarger). However, data on external variables such as weather patterns, food prices, and election schedules are not part of the SCAD, and a unifying theme amongst these studies is a lack of reproducibility. Most of the papers are published without code and, in many cases, without links to specific datasets used for analysis, making it impossible to further investigate their results in a study such as this. Therefore, a reproducibly transparent quantitative exploration of the factors that precede the escalation of African social conflict events would most likely be a welcome addition to the academic debate.

3. Data Cleaning & Manipulation

All conflicts in the SCAD dataset are included in this study, with the exception of 119 conflicts that do not start and end in the same calendar year. Such conflicts were excluded from the analysis due to the complexities of merging secondary data sources delineated by calendar year onto multiyear conflicts; these multiyear conflicts represent under 1.5% of the entire dataset, so removing them from the study should have a relatively minor effect on the analyses. The NMC and Polity IV datasets provide information on a country per year level of specification, and variables such as democracy scores, iron production, and total country population were directly merged into the SCAD with little manipulation. The World Religion dataset measures the number of people who subscribe to a specific religion at five-year intervals for every country. Consequently, I computed the dominant religion in every county per five years, forward filled the data to create country-year measurements, and merged onto SCAD. The rest of the variables were used with little manipulation, except for the creation of a death/no death indicator.

4. Methods & Results

The work described here approaches the two research questions sequentially, first conducting a broad descriptive analysis of a variety of factors that could potentially influence a conflict’s number of deaths. Then I proceed to the modeling phase of the project, breaking the modeling task into two subtasks: first I will attempt to model the death/no death indicator variable, and then I will use the results of that previous modeling effort and attempt to model the actual number of deaths. Before investigating any of the research questions, let’s first attempt to get a sense for the distribution of the number of deaths and associated conflict characteristics. **TABLE 1 and FIGURE 2 go here with captions.**

*4.1 Differences between Death and No Death Conflicts*

Before we start exploring factors that might be different between conflicts resulting in deaths and those not, just simply look at a distribution of conflicts by death, both aggregated into no death/at least one death and split out by the number of deaths. There are many more no death conflicts than conflicts resulting in at least one death, but the difference is not so large as to create a sample size too small to adequately model in later phases of the project. The plot on the right, showing the actual number of deaths, reveals more potential issues: the distribution of deaths is extremely zero-inflated and overdispersed. The distribution of the data is not, for lack of a better term, very ‘nice,’ which could pose some problems later in the process.

**FIGURE 2 GOES HERE**

Now that we have a good idea of the overall distribution of the data, let’s do some exploratory analyses across some of the variables included in the literature to see if changes in a certain factor might show a difference in occurrence of a death. First analyze magnitude variables such as duration of the conflict and the number of participants; one might hypothesize that the larger a conflict in duration or in participants, the more possibilities for a death. Figures 3 and 4, below, show that this intuition is, in fact, not supported by the data.

**Figures 3 & 4**

Now consider the dominant religion in a particular country; as recounted in the literature review portion of this paper, scholars such as Christian Scherrer suggest that conflicts can escalate along ethnic divides, so one could hypothesize that countries that have a particularly dominant religion might sustain more conflicts resulting in death. Once again, Table 2 illustrates that this hypothesis is not supported by the dataset. The distributions across the religions look essentially the same across death and no death conflicts (**CHI SQ TEST?**).

**Table 2 goes here**

Potentially of more promise is geography within the African continent. Certain countries constantly appear in the news related to headlines about tragically blood-filled conflicts—genocides in Rwanda and Sudan come immediately to mind—while other countries do not make the news nearly as often. One could hypothesize, then, that geography or country lines could differentiate between a death and not. The SCAD provides latitude and longitude data for each conflict, so assessing this proposition is relatively straightforward. Figures 5, attempts to graphically show the conflicts superimposed on maps of the African continent. There seems to possibly be some differentiation between the patterns of death and no death conflicts in Figure 5.

**FIGURE 5**

However, superimposing the conflicts on top of each other and rounding coordinates to the nearest degree shows a different story where much of the continent contains both death and no death conflicts, shown by overlapping red and blue points in Figure 6. There are two notable exceptions: a pocket of red in central Africa corresponds to the Sudan, and a pocket of blue in southern Africa corresponds to the successful, peaceful democracy Namibia. Otherwise, however, geographic and country boundaries do not seem to provide much information as to what differentiates death from no death.

**FIGURE 6**

In a last effort to find a variable that shows some difference in the frequency of deaths, let’s analyze the scholarly idea that countries of a different regime type could allow conflicts to escalate to different extents. One might suppose that strong democracies would encourage free speech and therefore be more inclined to foster peaceful social conflict, while oppressive governments might violently stamp out conflicts when they first begin. Table 3 shows the frequency of death and no death conflicts split out by regime type, and a pattern finally emerges in the data, though not exactly the original hypothesis.

**TABLE 3**

In fact, as can be easily seen by the row percentages printed in the table cells, the data actually suggest that democracies are more likely to result in *death* conflicts and autocracies *no death*. Why could this occur? Perhaps, autocratic government are more likely to encourage low-level dissent because citizens of those countries know the government will retaliate if things get out of hand, and the opposite argument holds for democracies. It is also certainly possible that this counterintuitive pattern is somewhat meaningless, either a false pattern in otherwise random data or a pattern that is a symptom of another, unanalyzed variable. In any case, this is an interesting result that will be investigated further in the modeling efforts of this project. All in all, individual variables do not seem to offer a lot of insight into what results in a conflict’s death.

*4.2 Modeling*

With limited success in exploring individual variables, let’s move on to the modeling phase of this project and attempt to explore multiple factors in interaction with each other. Randomly split the data into training and testing sets (70-30 percent, respectively), and first try to model the death/no death indicator variable before moving on to modeling the more complicated absolute number of deaths in a given conflict. Both modeling efforts will employ similar modeling technique wherein we will first select a subset of variables through regression techniques to build a model and then we will use that selected subset of variables to train models using other methods in an effort to assess model robustness.

4.2.1 Death/No Death Indicator

Using the training set, start by running logistic regression on the indicator death/no death variable. Use a modified best subset selection technique to do model selection, selecting an initial set of variables based on the results from the previous analyses done in this paper as well as the literature. Make adjustments to the variables included in the model, and choose the model with the smallest prediction residual sum of squares. The author recognizes that this is a somewhat rough variable selection process, and potential limitations and improvements will be discussed later in the paper. All categorical variables were transformed into dummy variables before beginning this process. The following five variables were chosen to be included in the final model:

* Location (categorical)—where was the conflict located within a country’s borders?
* Event Type (categorical)—organized/spontaneous, riot/strike/protest, etc.
* Was the Central Government a Target? (0/1)
* Primary Issue Motivating the Conflict (categorical)
* Composite Index of National Capability score (numeric)—attempts to represent a country’s hard power

Train a k-nearest neighbor model (KNN) using the same set of variables, setting k=10 through a similar rough trial-and-error process (again, limitations and improvements to this process will be discussed).

**tables 4-5 go here**

Tables 4 and 5, above, are contingency tables showing the prediction accuracy of the two modeling methods attempted in this section of analysis. Overall, both methods to extremely well at predicting the outcomes of conflicts in the test set. The KNN model does slightly better at predicting the no death conflicts whereas the logistic regression does better at predicting the death conflicts; however, prediction is almost at or above 60% for all categories in all methods. Given that we are working with complex real-world data, such accuracy can be considered relatively successful, and the comparative success across two methods suggests that there might be some important information conveyed in the five selected variables.

4.2.2 Absolute Number of Deaths

With some success in modeling the death/no death indicator variable, move on to attempt to model the absolute number of deaths in a given conflict. Use the same training and testing sets as before, and apply a similar variable selection process using Poisson regression. Interestingly, the exact same set of five variables minimized prediction error for in this modeling effort as did for modeling the death/no death indicator, perhaps lending more support for the importance of those variables. Again, train other models with the same variables using different techniques to test robustness, this time a classification/regression tree (CART) in addition to a KNN (k=10).

**figure 7 goes here.**

Figure 7, above, shows the results of these three models by plotting the natural log of fitted values (with an unlogged reference axis on top) against the size of the residual. All three methods do a decent job at predicting. The Poisson Regression and KNN models appear to be similar, with a decent amount of spread in the residuals and a tendency to overpredict as the fitted values get large. The KNN has a large spike of residuals for conflicts that have a fitted value of 0 deaths; this could be an argument for the KNN model being the least successful of the three, as it isn’t very comforting to know that the model is inaccurate at predicting conflicts with no deaths. The decision tree, by contrast to the other two methods, seems to have residuals closely centered around zero with very little overarching pattern, suggesting that the tree could be a very successful model.

However, all three methods have some very large residuals in the thousands, and exploration of those residuals shows that they are, in fact, the same set of conflicts across all three models (denoted in color by the legend in Figure 7). Investigation into these particular conflicts does not yield any revelations into why they might be consistently poorly predicted. The data for the conflicts does not seem unreasonable, so I am unwilling to completely throw them out as outliers. While the decision tree seems to do an exceptional job at predicting on the test set, the presence of these outliers without a clear explanation as to why raises some concerns that something could be missing from the model.

5. Discussion

Overall, the analyses up until this point seem to be somewhat successful. We have built a set of models that rely on the same five variables—location, event type, central government target, primary issue, & CINC score—and seem relatively robust across methods at predicting on the test set. Notably, the variable that seemed to have an impact during the first portion of this study, the Polity IV score for regime type, was not included in the variable selection process. We had previously noted that the patterns highlighted in the analysis of regime type were counterintuitive to preconceived expectations and had wondered if the variable was truly a good predictor of deaths in a conflict. The fact that the variable was not selected in either modeling effort suggests that perhaps the skepticism in exploratory analyses was warranted.

Of more note than what was omitted is what variables were selected into the model. The five selected variables do not come from any single school of scholarship. The CINC score focuses on factors strictly related to the country’s government, both related to natural resources and government policy. The other four variables focus on the conflict itself, but they focus on different subareas of interest: geography, social concerns spurring the conflict, and conflict logistics. In short, our analysis here does not support any one particular scholar’s argument but rather supports the continued argument. Conflict escalation seems to be as complex as the scholarship might suggest, a product of multiple factors that measure different aspects of the sociopolitical environment surrounding a conflict.

6. Future Extensions and Final Conclusions

While the conclusions in this study show promise, more work needs to be done before I can confidently add these results to the academic literature surrounding social conflict in Africa. On one hand, future analyses should try to merge or generate more variables to better assess some of the existing hypotheses surrounding social conflict escalation: representations of ethnic groups within a country, weather patterns, election data, and historical food prices are some that come to mind. On the other hand, the actual statistical analyses conducted here—particularly concerning variable selection and model testing—are rough at best. I conducted analyses on this project before covering more advanced modeling techniques such as stepwise variable selection and an in-depth discussion of cross validation techniques. While the original project proposal was approved without the implementation of these techniques, it would be much better to re-run these analyses using a more organized variable selection process such as stepwise selection and a more accurate model assessment process such as cross-validation. Making such changes would allow us to have much more confidence in the usefulness and robustness of the results presented here.

Limitations withstanding, this analysis represents a useful reproducible quantitative analysis of African social conflict. In many ways, these conclusions support the variety of theoretical hypotheses surrounding conflict escalation: a combination of varied factors seems to impact the number of deaths that will result from an episode of social conflict. These specific conclusions can be verified and refined through the study modifications suggested earlier. Perhaps more important than the specific conclusions about the conflicts is the research process itself. In stark contrast to previous studies using the SCAD, this study has its entire research workflow—code, plots, datasets, documentation, etc.—available on GitHub for other students and researchers to explore and build upon. Such a transparent and reproducible workflow lays the groundwork for future work and is an important first step towards regularly using quantitative analyses to analyze sociopolitical issues such as social conflict in Africa.

**UN idea – stop future conflict before it escalates**

Works Cited

*Datasets:*

Grieg, J. Michael, et. al. *National Material Capabilities Dataset*. Version 4.0. 30 June 2010. Web. Accessed 25 February 2014. <http://www.correlatesofwar.org/COW2%20Data/ Capabilities/nmc4.htm>.

Hendrix, Cullen S., and Idean Saehyan. *Social Conflict in Africa Database (SCAD)*. Version 3.0. 29 March 2012. Web. Accessed 5 February 2014. <www.scaddata.org>.

Maoz, Zeev, and Errol A. Henderson. *World Religion Project*. Version 1.1. 17 August 2013. Web. Accessed 25 February 2014. <http://www.correlatesofwar.org/COW2%20Data/ Religion/Religion.htm>.

Marshall, Monty G., Ted Robert Gurr, and Keith Jaggers. *Polity IV Project: Political Regime Characteristics and Transitions, 1800–2012*. Version 2012. 21 April 2013. Web. Accessed 25 February 2014. <www.systemicpeace.org>.

*Citations:* ***need to add epigraphs and social conflict definition from scad website***

Dahrendorf, Ralf. *The Modern Social Conflict: The Politics of Liberty*. 2nd ed. New Brunswick, NJ, USA: Transaction, 2008. Print.

De Angelis, Massimo. *Keynesianism, Social Conflict and Political Economy.* London: Macmillan, 2000. Print.

Devlin, Colleen, Brittany Franck, and Cullen S. Hendrix. “Trends and Triggers: Climate Change and Interstate Conflict.” *Climate Change and African Political Stability*. December 2013: Research Brief No. 21. Web.

Fielding, Nigel G. *The Police and Social Conflict*. 2nd ed. London: Glasshouse, 2005. Print.

Maoz, Zeev, and Errol A. Henderson. “The World Religion Dataset, 1945-2010: Logic, Estimates, and Trends.” *International Interactions*, 39.3 (2013): 265-291.

O’Brian, Jay. “Ethnicity, National Identity and Social Conflict.” *Social Science and Conflict Analysis: Proceedings of the International Conference* The Role of Social Science in Conflict Resolution *Held in Helsinki 9*–*13.1.1993*. Ed. A. Hurskainen and M. Salih. Helsinki: Helsinki UP, 1993. 60–82. Print.

Salehyan, Idean, and Christopher Linebarger. “Elections and Social Conflict in Africa.” *Climate Change and African Political Stability*. February 2013: Research Brief No. 6. Web.

Salehyan, Idean, Cullen S. Hendrix, Jesse Hamner, Christina Case, Christopher Linebarger, Emily Stull, and Jennifer Williams. "Social Conflict in Africa: A New Database.” *International Interactions*. 38.4 (2012): 503-511.

Scherrer, Christian P. *Genocide and Crisis in Central Africa: Conflict Roots, Mass Violence, and Regional War*. Westport, CT, USA: Praeger, 2002. Print.

Singer, J. David, Stuart Bremer, and John Stuckey. "Capability Distribution, Uncertainty, and Major Power War, 1820-1965." *Peace, War, and Numbers.* Ed. Bruce Russett. Beverly Hills: Sage, 1972. 19-48.

Smith, Todd G. “Food Price Spikes and Social Unrest in Africa.” *Climate Change and African Political Stability*. April 2013: Research Brief No. 11. Web.

Wood, Adrian P. “Natural Resource Conflicts in South-West Ethiopia: State, Communities, and the Role of the National Conservation Strategy in the search for Sustainable Development.” *Social Science and Conflict Analysis: Proceedings of the International Conference* The Role of Social Science in Conflict Resolution *Held in Helsinki 9*–*13.1.1993*. Ed. A. Hurskainen and M. Salih. Helsinki: Helsinki UP, 1993. 83–102. Print.

1. http://www.worldbulletin.net/news/132179/hundreds-of-south-africans-protest-outside-johannesburg-court [↑](#footnote-ref-1)
2. http://www.globalpost.com/dispatch/news/afp/140410/teenager-shot-dead-during-safrica-riots [↑](#footnote-ref-2)
3. https://www.strausscenter.org/scad.html [↑](#footnote-ref-3)