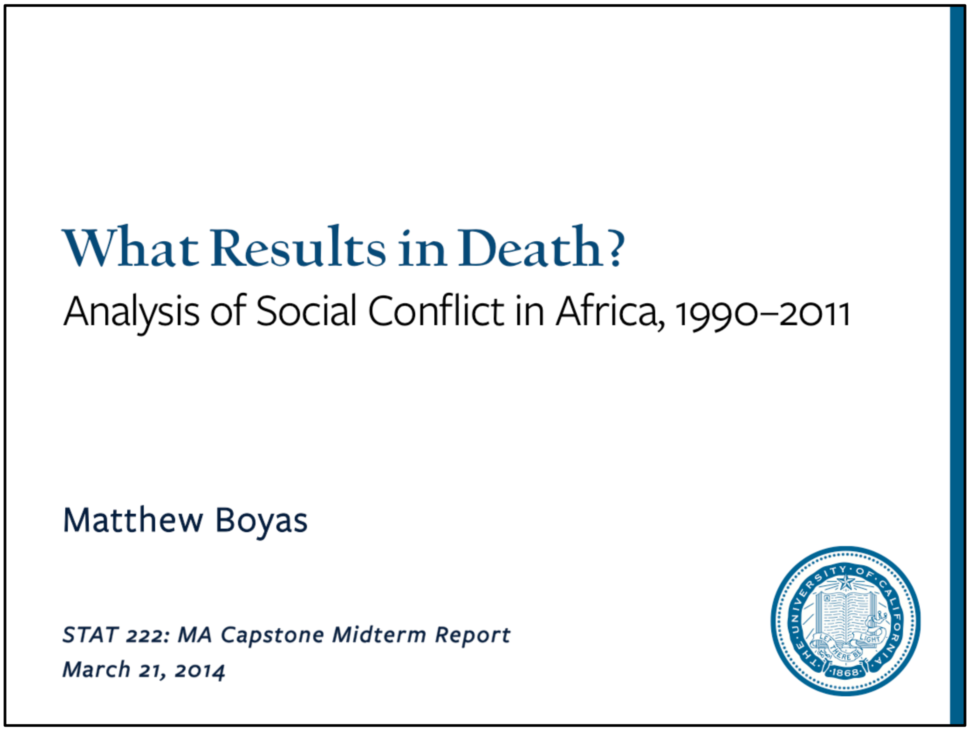
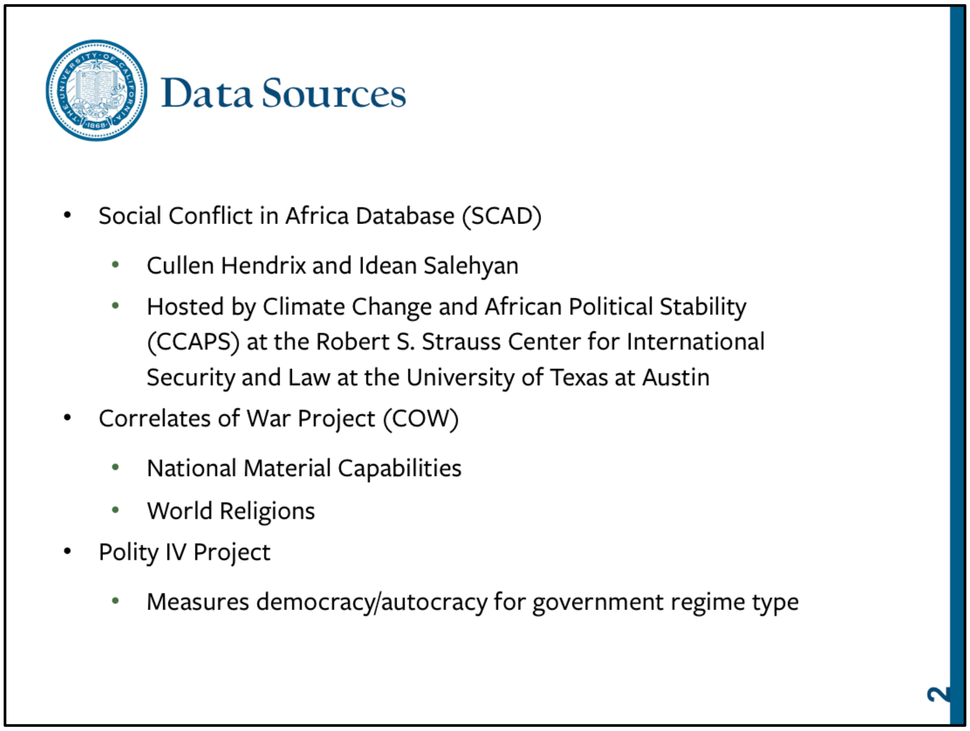


What Results in Death?

Analysis of Social Conflict in Africa, 1990–2011

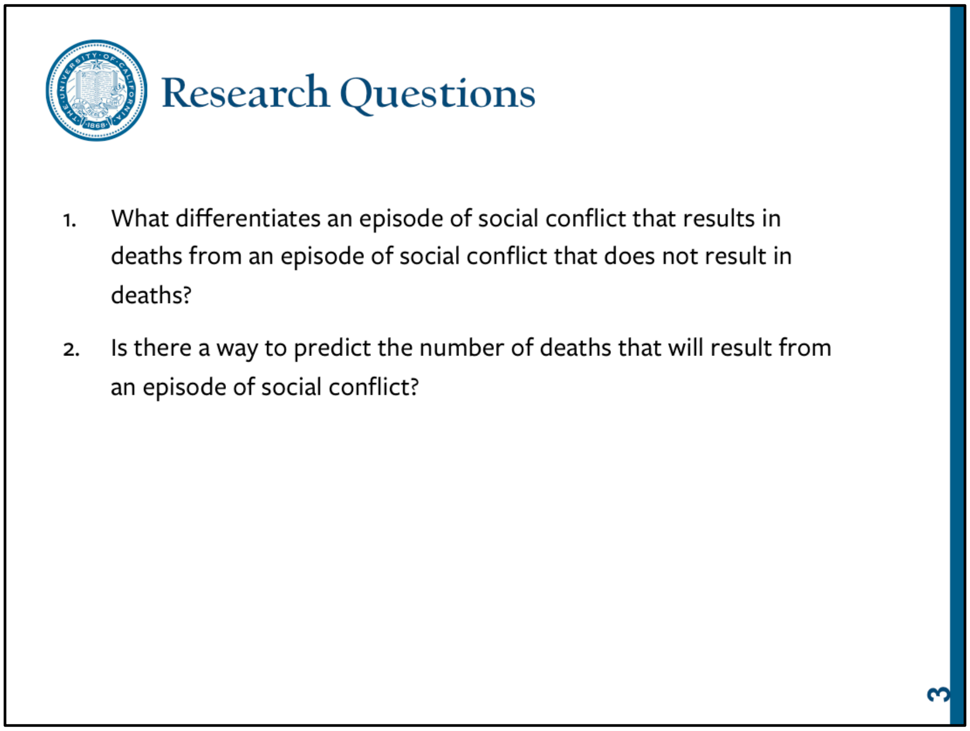
Matthew Boyas

STAT 222: MA Capstone Midterm Report
March 21, 2014



Data Sources

- Social Conflict in Africa Database (SCAD)
  - Cullen Hendrix and Idean Salehyan
  - Hosted by 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
- Correlates of War Project (COW)
  - National Material Capabilities
  - World Religions
- Polity IV Project
  - Measures democracy/autocracy for government regime type



Research Questions

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?

3

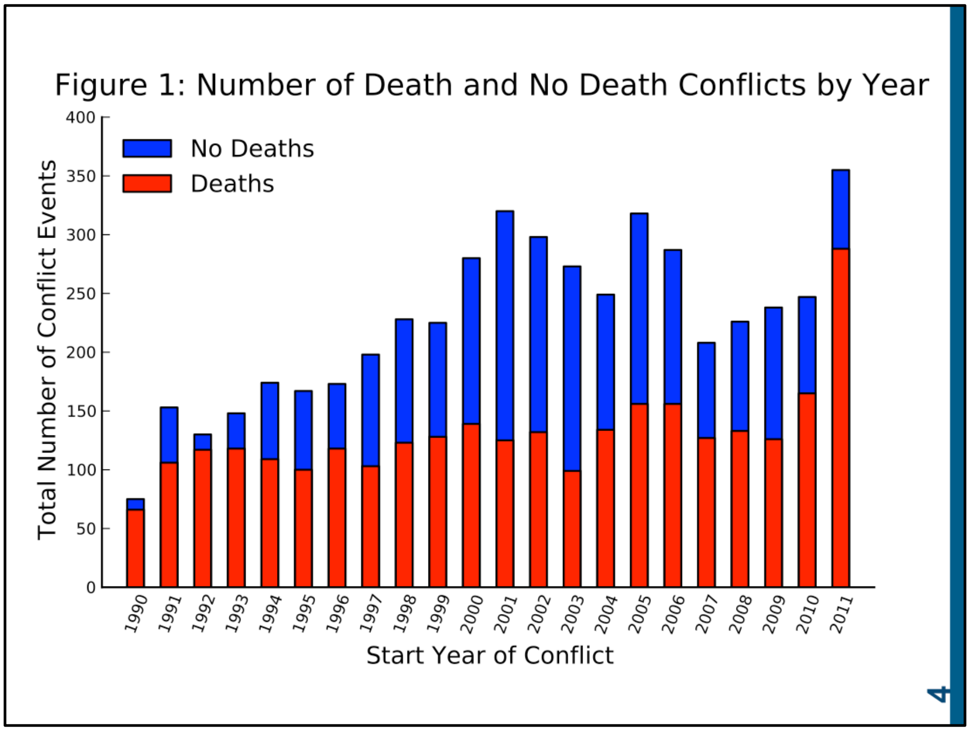


Figure 1, above, is a split bar-chart, with one bar per year, visually showing the number of conflicts resulting in deaths relative to the total number of conflicts per year. Each bar represents all conflicts in the specified year and is split into two pieces, one piece for conflicts resulting in deaths and one piece for conflicts resulting in no deaths. Note that with the exception of the spike in death conflicts in 2011 (most likely attributable to the Arab Spring), the number of conflicts result in deaths remain somewhat constant across the years of analysis. Additionally, the number of conflicts resulting in no-deaths show a somewhat increasing trend as more governments democratize following the political revolutions of the 20th Century until the late 2000s, when dissatisfaction with incomplete democratization begins to set in—which, in turn, led to the Arab Spring.

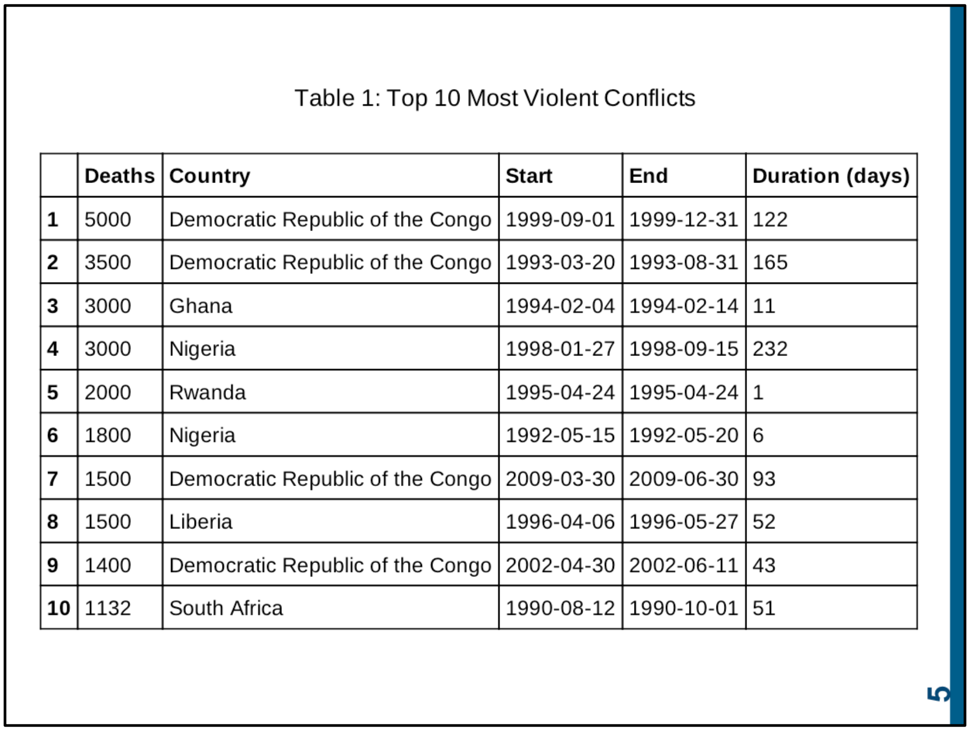


Table 1: Top 10 Most Violent Conflicts

	Deaths	Country	Start	End	Duration (days)
1	5000	Democratic Republic of the Congo	1999-09-01	1999-12-31	122
2	3500	Democratic Republic of the Congo	1993-03-20	1993-08-31	165
3	3000	Ghana	1994-02-04	1994-02-14	11
4	3000	Nigeria	1998-01-27	1998-09-15	232
5	2000	Rwanda	1995-04-24	1995-04-24	1
6	1800	Nigeria	1992-05-15	1992-05-20	6
7	1500	Democratic Republic of the Congo	2009-03-30	2009-06-30	93
8	1500	Liberia	1996-04-06	1996-05-27	52
9	1400	Democratic Republic of the Congo	2002-04-30	2002-06-11	43
10	1132	South Africa	1990-08-12	1990-10-01	51

Table 1 lists the top 10 most violent conflicts in terms of absolute deaths. Note the variation even amongst these 10 records: 6 different countries, 10 different years, 7 different months, and durations ranging from one day to almost 8 months. There does not seem to be a unique profile that identifies a high-mortality conflict, which could create issues during the modeling phase of this project.

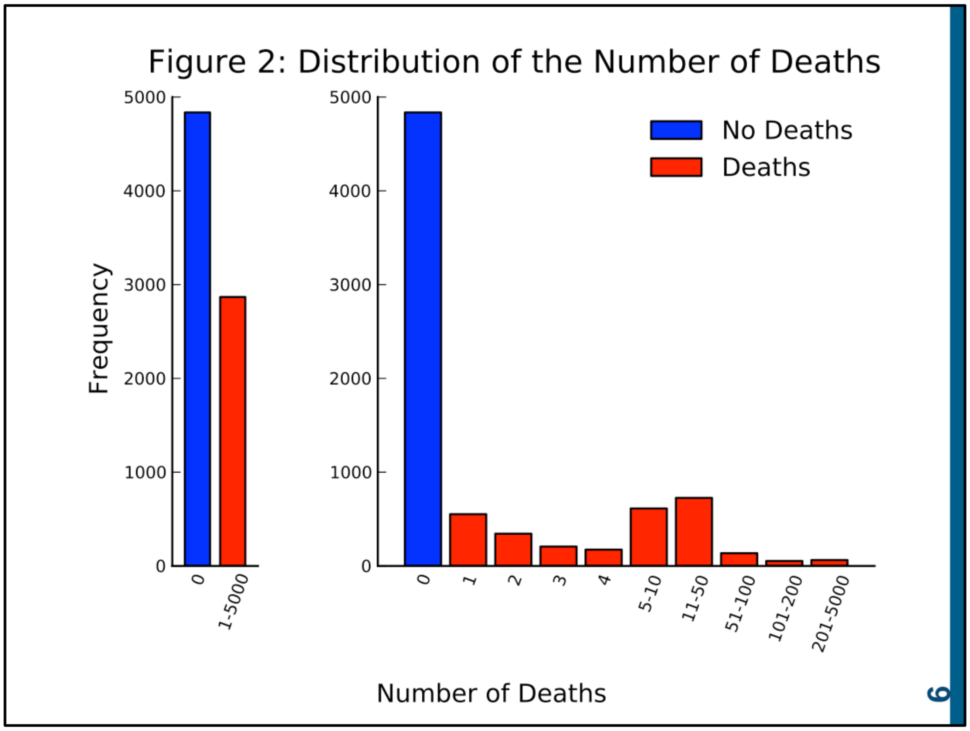
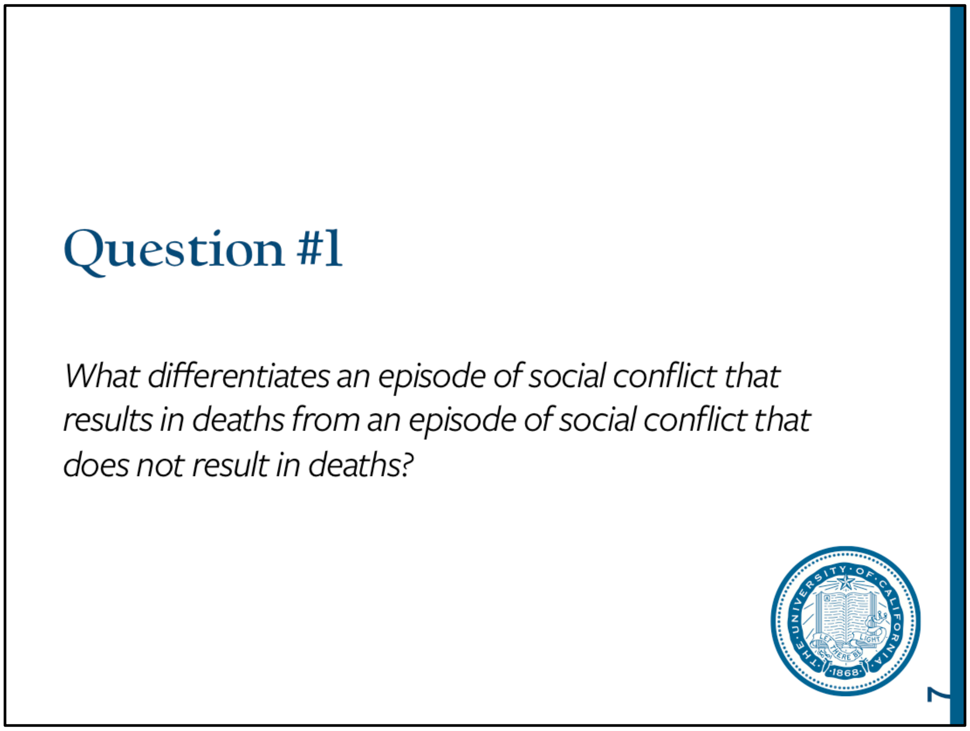


Figure 2: Distribution of the Number of Deaths

Figure 2 is a set of two bar charts intended to show the frequencies of the number of deaths in all of the African social conflicts from 1990–2011. The left plot shows the number of conflicts that resulted in no deaths and at least one death. The right plot shows the same frequency bar for no death conflicts but bins the conflicts resulting in at least one death. Note the extreme overdispersion and zero-inflation on this more detailed plot; while there are over 2000 conflicts resulting in at least one death, the actual number of deaths ranges from one death to 5000 deaths. This zero-inflated, overdispersed distribution of the actual number of deaths could be a problem for the modeling effort.



Question #1

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

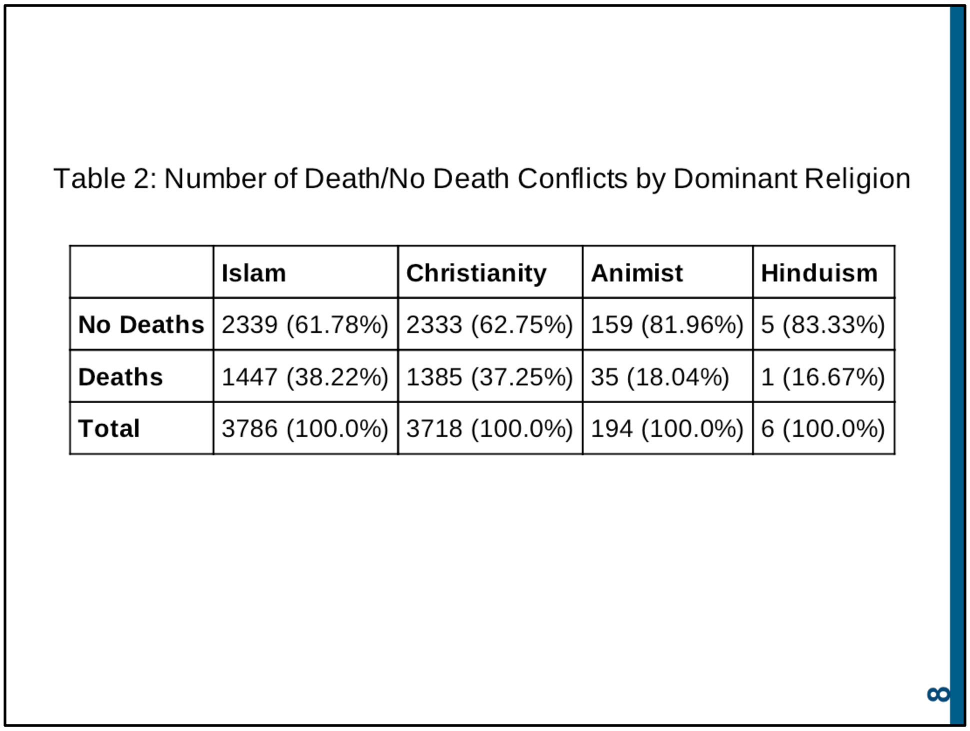


Table 2: Number of Death/No Death Conflicts by Dominant Religion

	Islam	Christianity	Animist	Hinduism
No Deaths	2339 (61.78%)	2333 (62.75%)	159 (81.96%)	5 (83.33%)
Deaths	1447 (38.22%)	1385 (37.25%)	35 (18.04%)	1 (16.67%)
Total	3786 (100.0%)	3718 (100.0%)	194 (100.0%)	6 (100.0%)

Table 2 shows the number of conflicts resulting in deaths / no deaths broken out by the dominant religion of the country (with column percentages in parentheses). There does not seem to be any pattern here. Christianity and Islam are the two major religions in Africa, and they show very similar numbers in the table output. The high percentages of No Death conflicts in both Animist and Hindu dominated countries can probably be explained by the relatively small number of observations in both columns.

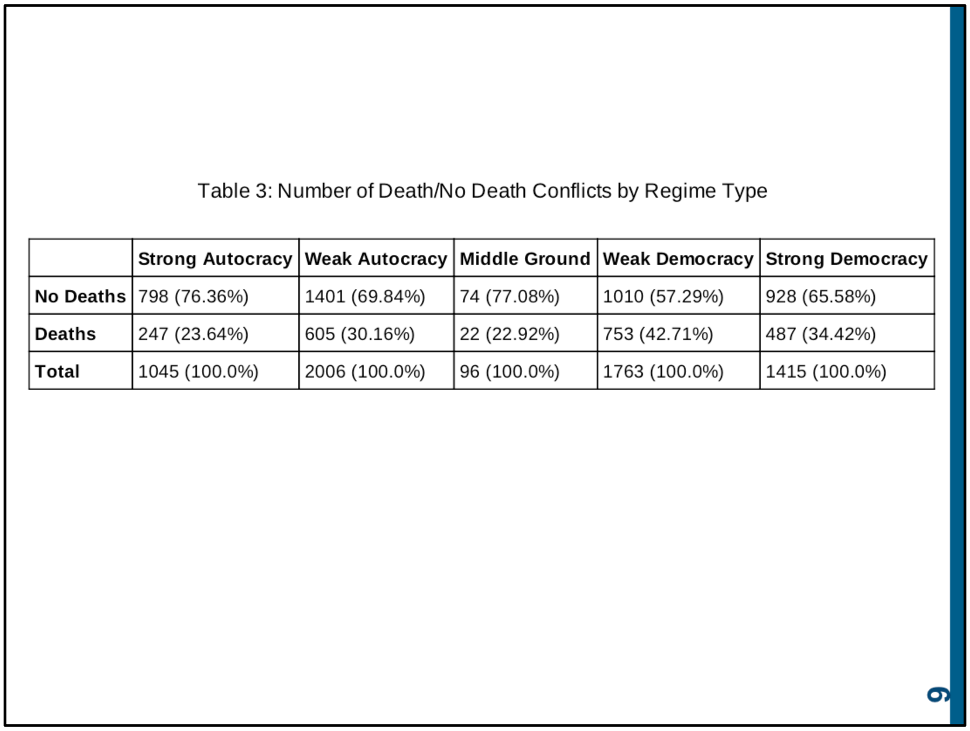


Table 3: Number of Death/No Death Conflicts by Regime Type

	Strong Autocracy	Weak Autocracy	Middle Ground	Weak Democracy	Strong Democracy
No Deaths	798 (76.36%)	1401 (69.84%)	74 (77.08%)	1010 (57.29%)	928 (65.58%)
Deaths	247 (23.64%)	605 (30.16%)	22 (22.92%)	753 (42.71%)	487 (34.42%)
Total	1045 (100.0%)	2006 (100.0%)	96 (100.0%)	1763 (100.0%)	1415 (100.0%)

Table 3 shows the number of conflicts resulting in deaths / no deaths broken out by the regime type of the country (with column percentages in parentheses). Regime type is measured on a continuum binned from the Freedom House Polity score which ranges as integers from -10 (full autocracy) to +10 (full democracy). Countries in the 'Middle Ground' group have a Polity score of 0, dead center between autocracy and democracy. The above table seems to show that democracies, particularly weak democracies, have more death-resulting conflicts than do autocracies. This counterintuitive observation could potentially be explained by the idea that democracy fosters a political environment where citizens can speak their minds, which can lead to more situations of death. This table shows that perhaps there is a relationship between regime type as measured by Polity score and the occurrence of deaths in a conflict, something that will be investigated further in the modeling phase of this project.

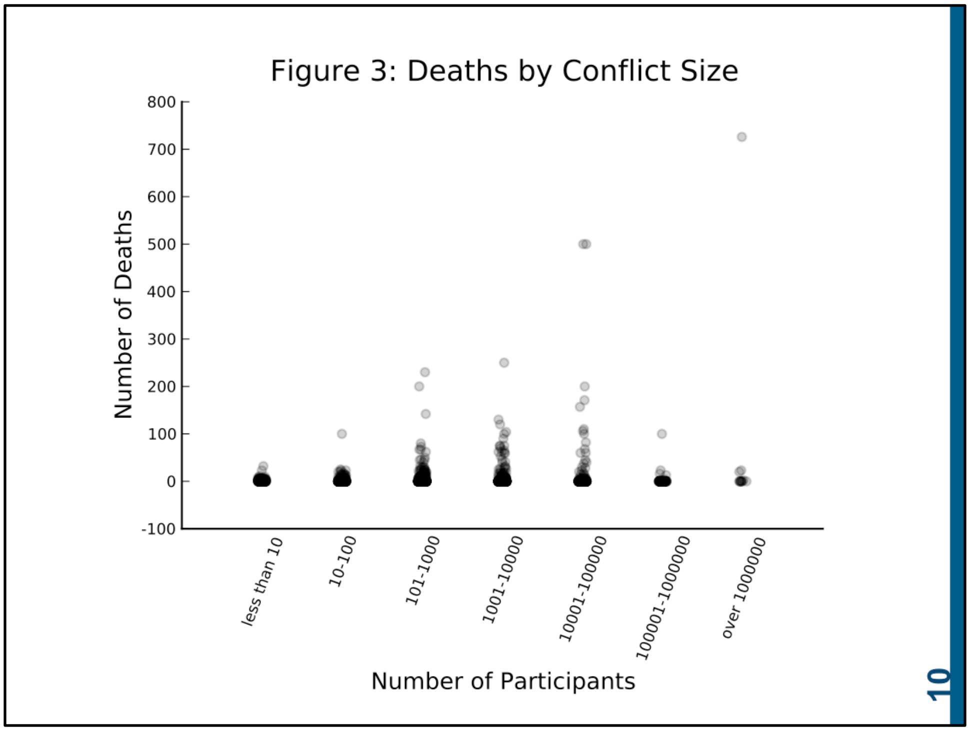
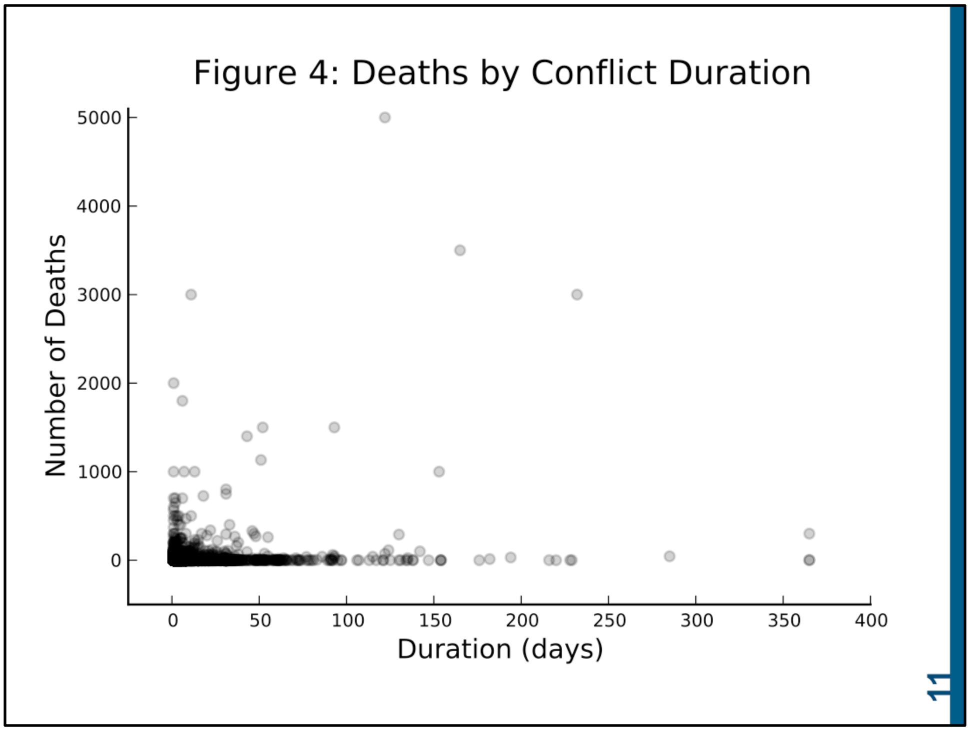


Figure 3: Deaths by Conflict Size

Figure 3, above, is a scatterplot showing the number of participants (binned by the original coders of the SCAD dataset) vs. the continuous variable number of deaths. The points are translucent and jittered slightly to try to show the density. It was expected that the plot should show an increasing relationship between participants and number of deaths, simply based on the idea that a conflict involving more people has higher odds to result in a death. The plot does not necessarily show this relationship, though the normal-esque relationship is probably an artifact of the dataset having significantly more observations in the center than at the tails of the number of participants variable.



Motivated similarly as Figure 3, Figure 4 is a scatterplot of the duration of each conflict (in days) against the number of deaths. It intuitively makes sense that there should be an increasing relationship between these variables, as the longer a conflict, the more chances for a death to result. However, the plot does not support such a conclusion; taking into account the fact that we have many more shorter conflicts in the dataset, there does not seem to be that much of a relationship between duration and the number of deaths.

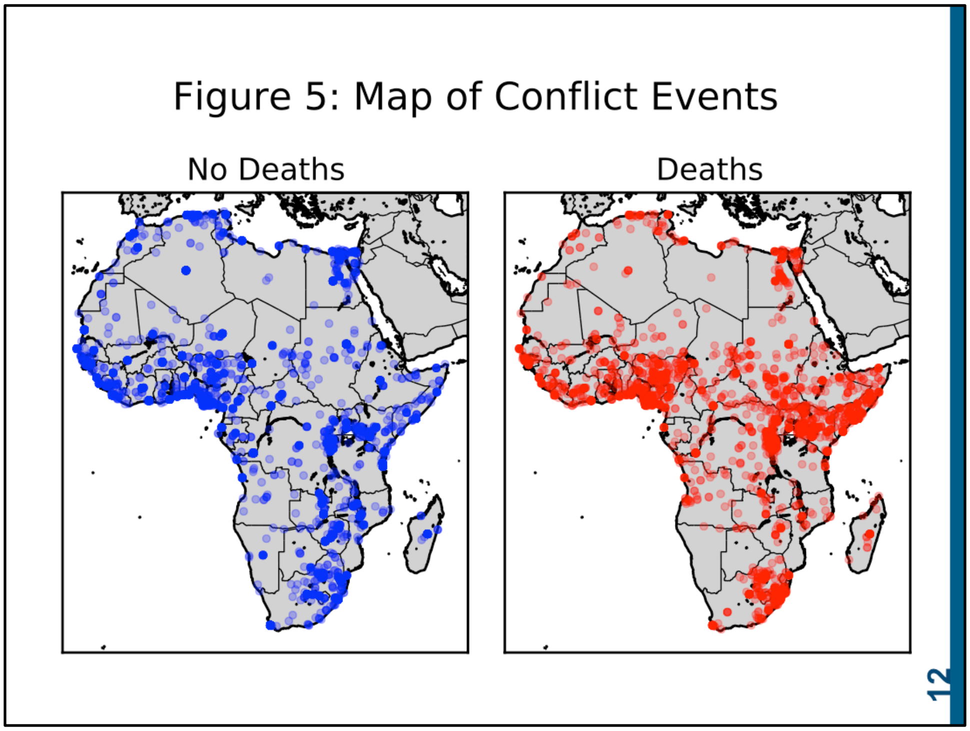


Figure 5: Map of Conflict Events

Figure 5 above attempts to visualize an answer to the question, "Are deaths more likely to happen in certain countries?" In short, the answer seems to be, “Not really.” Each dot represents an individual unique conflict ID, and the dots are translucent to represent the density of conflicts. At first glance, the two maps look effectively the same, with the bulk of all conflicts occurring in West and East Africa. While certain countries are certainly more prone towards *conflicts* than other countries, deaths do not seem to follow a clear geographic pattern.

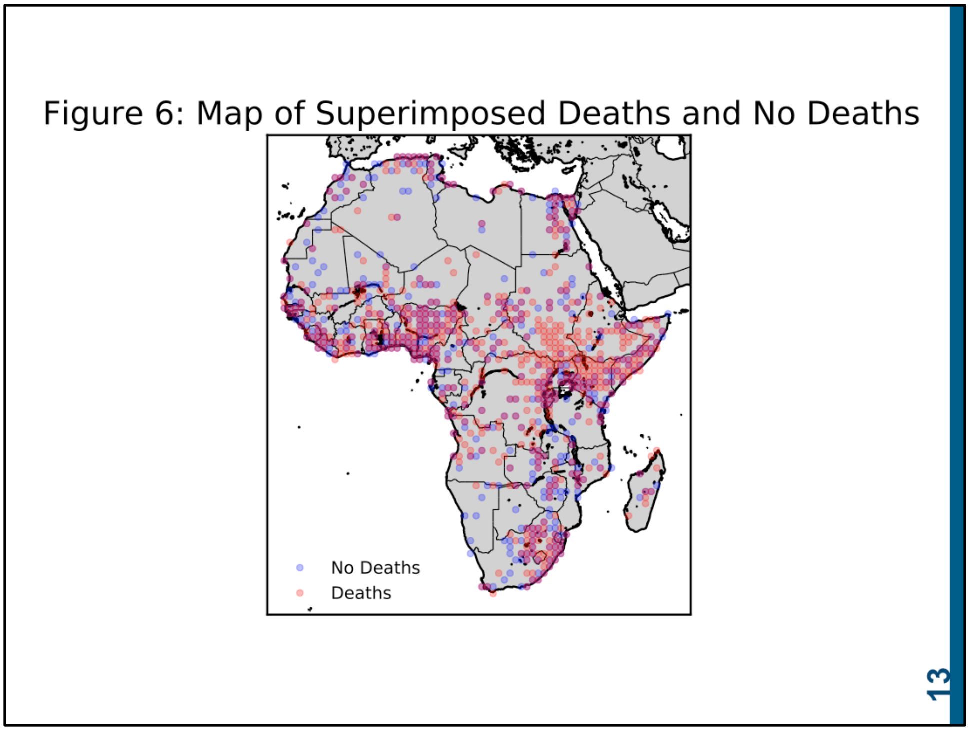
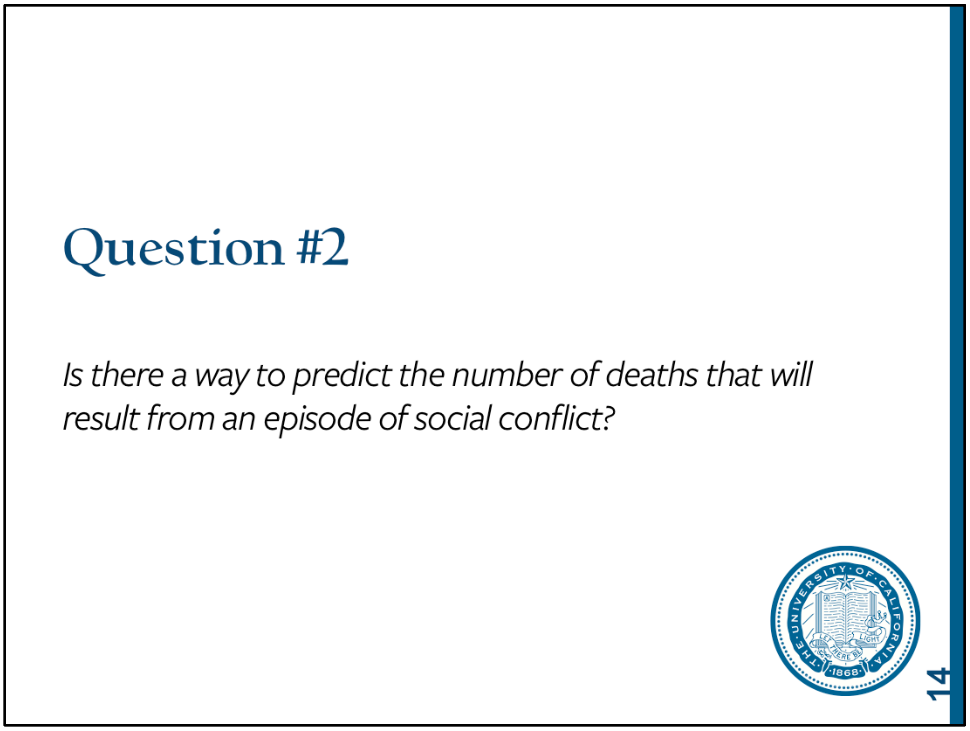


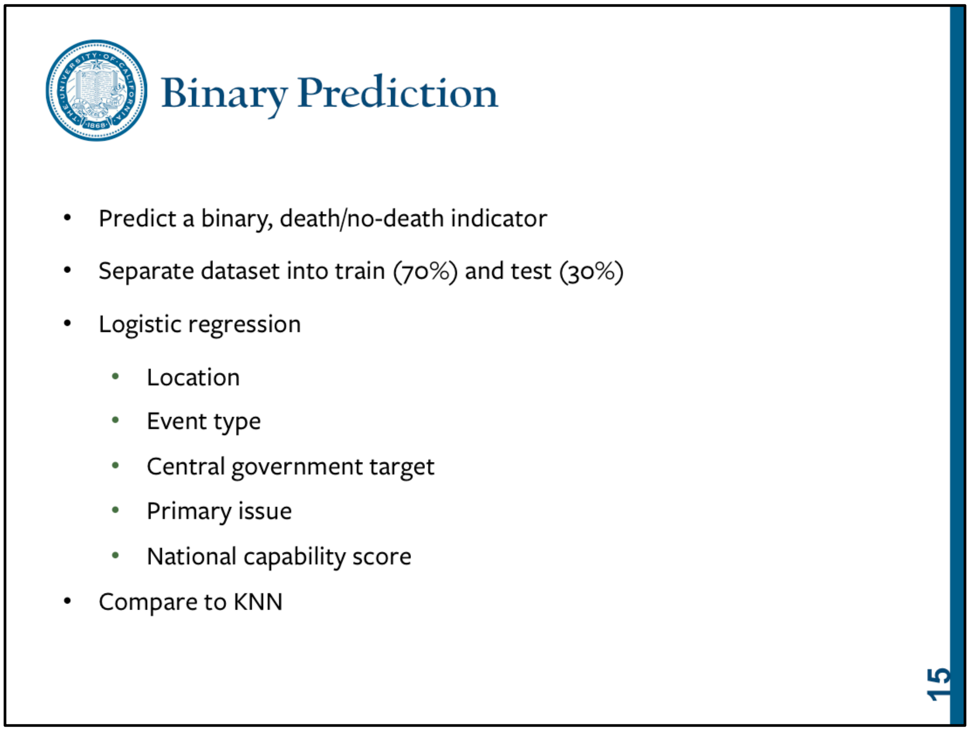
Figure 6: Map of Superimposed Deaths and No Deaths

Figure 6 above attempts to show the difference between the two maps included in Figure 5. Because of the high density of points map obscuring what type of conflict occurs where, I rounded each set of latitude/longitude coordinates to the nearest degree and then removed duplicate sets of coordinates *after* subsetting by the death/no-death indicator variable. Then plotting the points in red and blue with translucency reveals a very interesting picture. Most of the African conflicts with deaths occur on top or nearby conflicts with no deaths with two exceptions: there is a large pocket of death-resulting conflicts in Eastern Africa, which roughly corresponds to the conflict-ridden South Sudan. There is also a large pocket of no death conflicts in Southern Africa, corresponding to Namibia, a very successful democracy. While this map illustrates more of a geographic pattern than does the previous figure, the patterns still do not seem to be wholly significant.



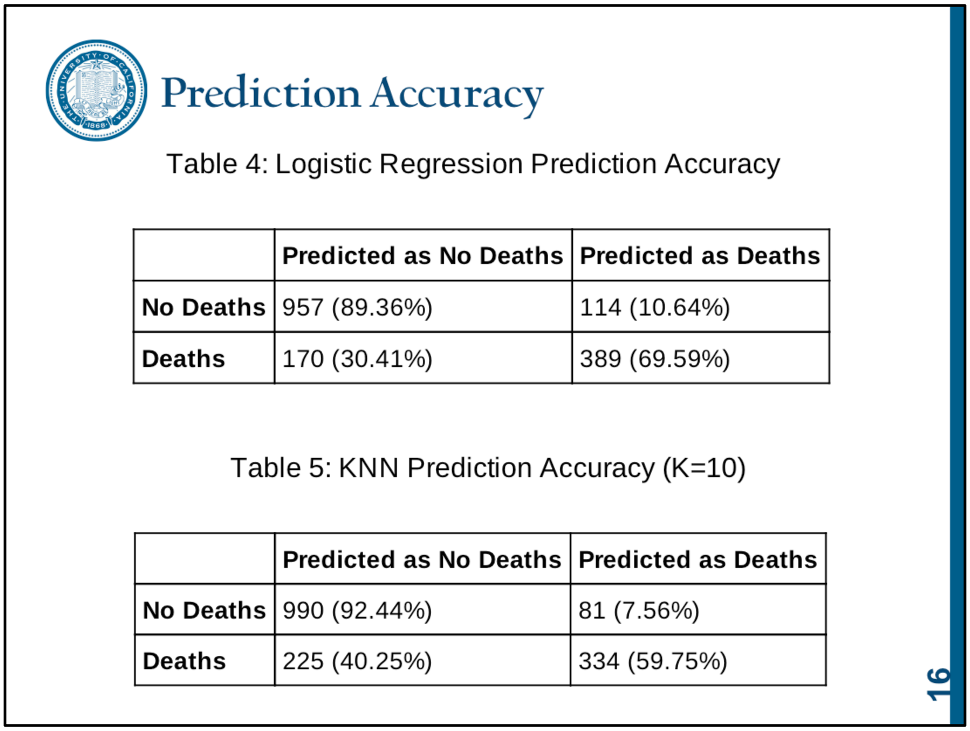
Question #2

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



Binary Prediction

- Predict a binary, death/no-death indicator
- Separate dataset into train (70%) and test (30%)
- Logistic regression
  - Location
  - Event type
  - Central government target
  - Primary issue
  - National capability score
- Compare to KNN



Prediction Accuracy

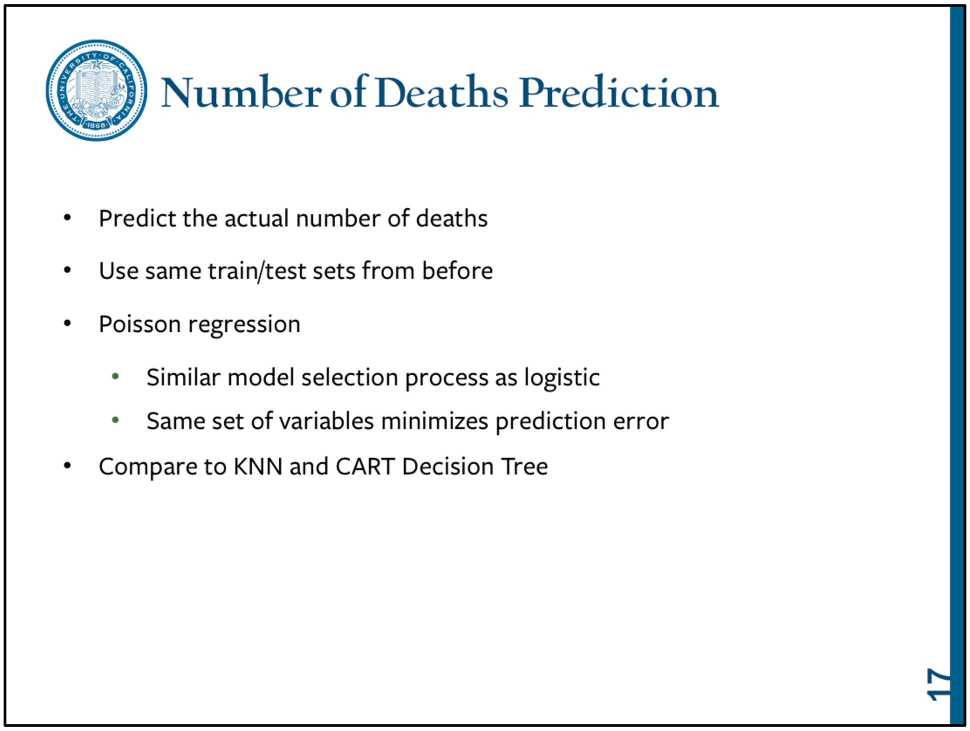
Table 4: Logistic Regression Prediction Accuracy

	Predicted as No Deaths	Predicted as Deaths
No Deaths	957 (89.36%)	114 (10.64%)
Deaths	170 (30.41%)	389 (69.59%)

Table 5: KNN Prediction Accuracy (K=10)

	Predicted as No Deaths	Predicted as Deaths
No Deaths	990 (92.44%)	81 (7.56%)
Deaths	225 (40.25%)	334 (59.75%)

Tables 4 and 5 show prediction rates on the test set for the logistic and KNN models, in absolute numbers with row percentages in parentheses. Both models have very similar accuracy rates at predicting the test set. For this specific test set, the KNN model more accurately predicts the No Deaths (with a 92% success rate) and the Logit model more accurately predicts the Deaths (with a 69% success rate). However, the differences in accuracy do not seem to be significant enough to prefer one model over another purely based on prediction power. It is interesting that we get similar prediction results using different techniques on the same data, which leads me to believe that perhaps there is something particularly interesting about this combination of variables for modeling the death/no death indicator variable.



Number of Deaths Prediction

- Predict the actual number of deaths
- Use same train/test sets from before
- Poisson regression
  - Similar model selection process as logistic
  - Same set of variables minimizes prediction error
- Compare to KNN and CART Decision Tree

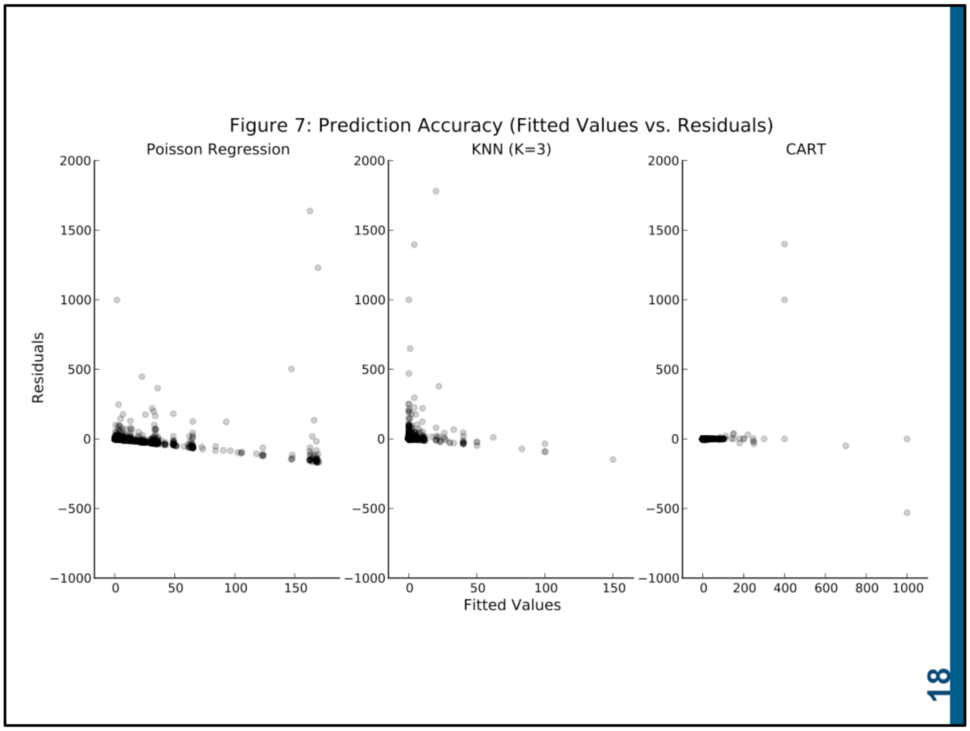
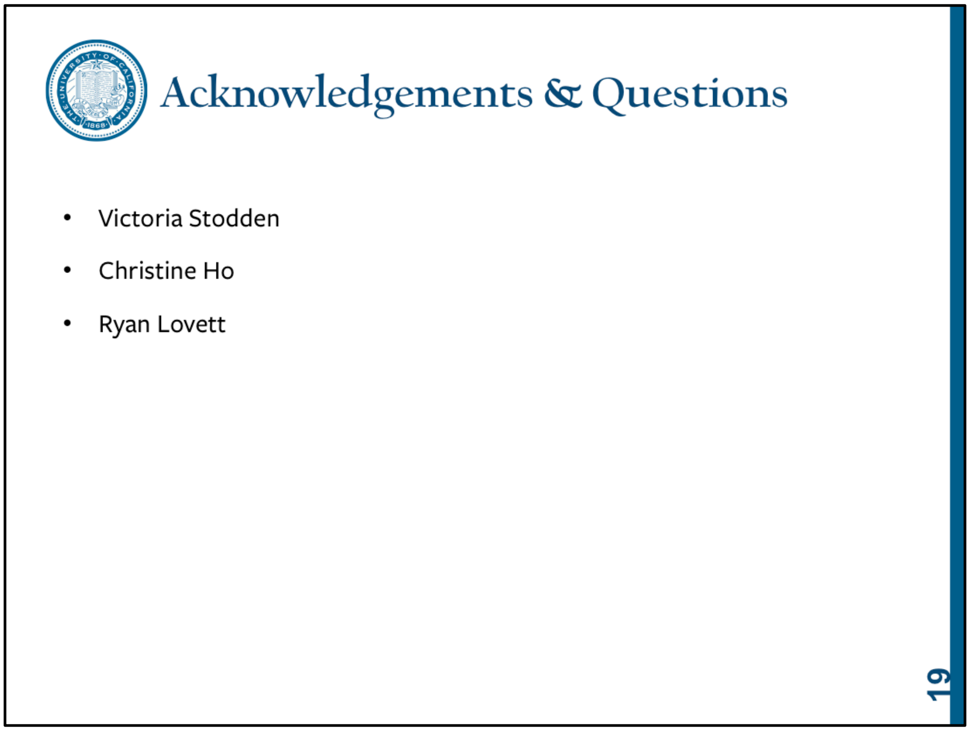


Figure 7: Prediction Accuracy (Fitted Values vs. Residuals)

Figure 7 shows the Fitted Value vs. Residual plots for predicting the number of deaths using the same set of variables from the same test dataset and three different modeling techniques: Poisson Regression, KNN (K=3), and CART Decision Tree. It is interesting that the three models have very different prediction accuracies, generally unlike what we saw earlier with the binary models. The CART tree encompasses about half the mean squared prediction error as the KNN model, which, in turn, has less than half of the prediction error of the Poisson Regression. This analysis lends more support for the information encoded in these variables but also suggests that more work should be done with both parametric and nonparametric methods to fit the optimal model—all of these models seem to have problems predicting in some way or another. The decision tree, the best model here, seems to have very good prediction accuracy for values up to around 500; however, values higher than that can result in large residuals, something that could be a problem depending on the model's intended application.



Acknowledgements & Questions

- Victoria Stodden
- Christine Ho
- Ryan Lovett

Matt Boyas
Stat 222 Midterm Report—Deviations from Proposal

This document is intended to accompany the included work and serves to explain some of the deviations from my original research proposal.

(1)
It made sense to me to lump together my first and third research questions, for as I began investigating question #1, I was also testing many of the variables that I wanted to look at in question #3.

(2)
Some of the titles of plots and tables have been altered or completely changed because I discovered that the titles no longer really worked once I had an actual figure/table at which to look.

(3)
The originally-proposed Figures 3 & 4 were changed to scatterplots. I was not sure how to calculate the original proportions for conflicts with zero deaths, and the scatterplots seemed to convey a very similar idea.

(4)
I chose not to include the regression summary due to the fact that three of the five variables are categorical, split into 7-10 binaries. As a result, there are a lot of variables listed with names that do not mean all that much, so I plan on just summarizing the regression summary in the text of my paper and perhaps including the entire summary output as an appendix.

(5)
Many of the proposed figures/tables relating to the regression were somewhat hypothetical, and as my modeling methods changed, so did some of the plots. For example, I chose not to include normal Q-Q plots due to not running linear regression. I also did not include fitted value vs. residual plots for the training dataset in favor of focusing on clear plots/tables regarding prediction accuracy (I did not want to have too many similar looking figures/tables in my paper for fear that overall meaning and flow of the paper would be affected.)

(6)
The figure to show prediction accuracy from the number of deaths model was modified to be a fitted value vs. residual plot for the three modeling methods that I attempted.