Question

As measured by the number of global conflicts and fatalities from war, has the world truly become a safer place following the end of the Cold War (from 1991 to present)? We sought to answer this question by looking at various secondary research questions covering dimensions of global conflict and arms proliferation.

1. What is the predicted trend in the number of global conflicts?
2. What is the predicted trend in the number of total fatalities from global conflict?
3. What is the predicted trend in the number of civilian fatalities from global conflict?
4. What factors (e.g. parties involved, region, type of violence, etc.) predict whether or not there are civilian fatalities in conflict events? (Old question)
    a. What are the predicted probabilities civilians died in observed conflict events, as predicted by the author's definition of type of violence? (New question)
5. What factors (number of conflicts, nation states involved, etc.) predict the total amount of arms exported globally?
6. What factors (number of conflicts, etc.) predict the total amount of arms imported globally?

Data

The Swedish University of Uppsala's UCDP Georeferenced Event Dataset (GED) Global version 17.1. According to researchers:

> This dataset is UCDP's most disaggregated dataset, covering individual events of organized violence (phenomena of lethal violence occurring at a given time and place). These events are sufficiently fine-grained to be geo-coded down to the level of individual villages, with temporal durations disaggregated to single, individual days (Sundberg et al., 2017).

The UCDP Georeferenced Event (GED) Dataset is housed in a 65.3mb CSV file, with 42 columns and 135,181 rows of data.

The Stockholm International Peace Research Institute (SIPRI) Arms Transfers Database and Military Expenditure Database "contains information on all transfers of major conventional weapons from 1950 to the most recent full calendar year" (SPIRI, 2017). Military expenditure data contains information about per capita military expenditure and military expenditure as a share of GDP for every country. The arms export data file contains 28 columns and 119 rows, and the arms import file contains 28 columns and 224 rows.

Results

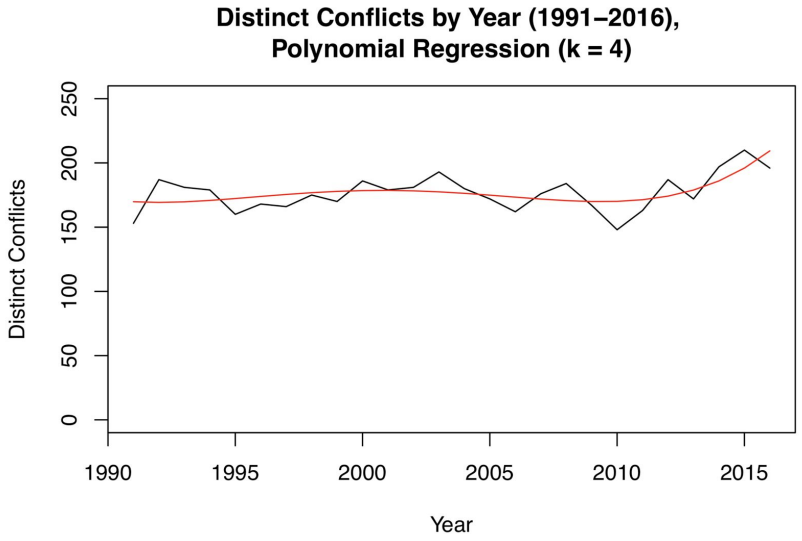
As measured by the number of global conflicts and fatalities from war, has the world truly become a safer place following the end of the Cold War (from 1991 to present)?

As measured by the number of global conflicts and fatalities from war, it is inconclusive that the world truly is a "safer" place following the end of the Cold War (from 1991 to present). Evidence that we have examined using datasets from UDP and SIPRI suggest that notions that the world is becoming "safer" may be overstated or outright incorrect in the context of distinct wars, and civilian and total deaths. The evidence scrutinized suggests it is equally plausible we may be headed towards a period of greater geopolitical conflict, violence, and arms proliferation.

Our results are inconclusive given the complexity of geopolitics and the ambitious nature of the topics we have sought to address, and we stress that any causal or conclusive findings cannot be drawn from the current project. With that said, we believe our findings may begin to help counter notions that the world is becoming "safer".

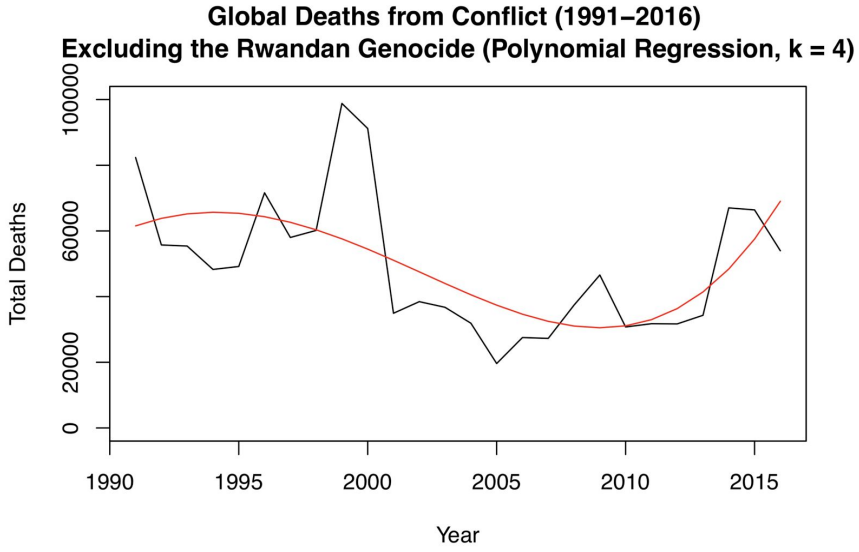
1. What is the predicted trend in the number of global conflicts?

As evidenced using polynomial regression, the trend is that the number of conflicts are on the rise from the end of the Cold War (using the GED dataset). Using a hands-on approach to examine the dataset of interest and the claims of one of our motivating sources for this project, we have examined the claim that conflicts are on the rise using polynomial regression to fit a trend line. Our model was statistically significant ($p < .05$), with an adjusted R-squared of 0.27. Given the pervasiveness of violent conflict in international news headlines, our model may also be practically significant, as this model suggests more conflict rather than less.



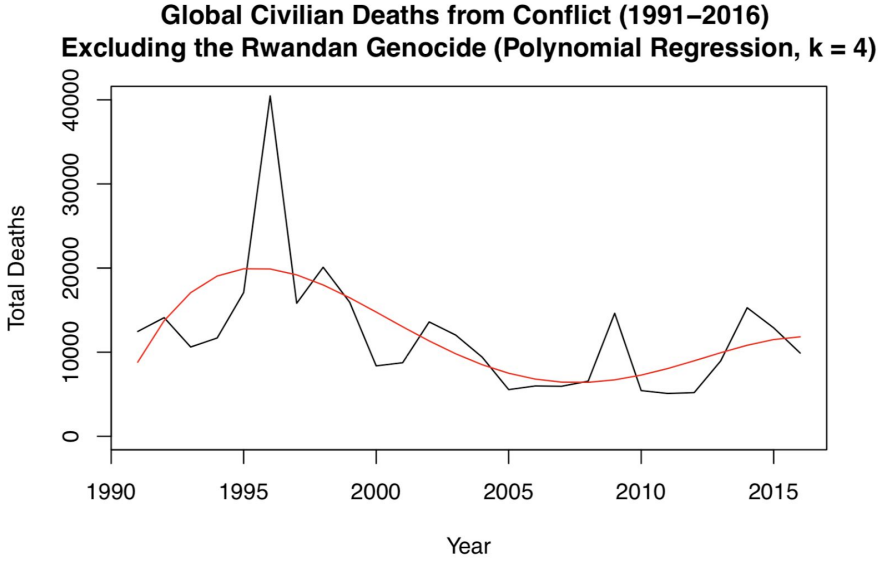
2. What is the predicted trend in the number of total fatalities from global conflict?

Using polynomial regression on the GED, and by removing the outlier of the Rwandan genocide, total deaths appear to also be on the rise from the mid to late 2000's onwards. Our model was statistically significant ($p < .05$), with an adjusted R-squared of 0.31.



3. What is the predicted trend in the number of civilian fatalities from global conflict?

Using polynomial regression on the GED, and by removing the outlier of the Rwandan genocide, civilian deaths also appear to be on the rise from the mid to late 2000's onwards. Our model was statistically significant ($p < .05$), again with an adjusted R-squared of 0.31.



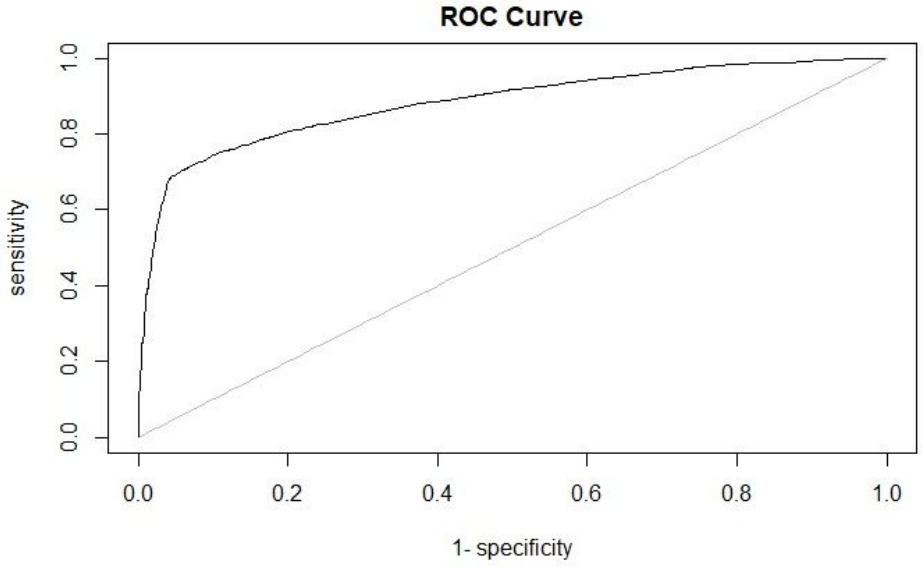
4. What factors (e.g. parties involved, region, type of violence, etc.) predict whether or not there are civilian fatalities in conflict events? (Old Question)

What are the predicted probabilities civilians died in observed conflict events, as predicted by the author's definition of type of violence? (New Question)

We initially sought to look at factors that may predict civilian death in conflict events. Given our lack of subject domain expertise, and to try to address raised uncertainties regarding dataset quality, we narrowed the scope of this question.

As a sanity check on an aspect of the dataset's internal validity, we modeled the predicted probability of civilian death in observed conflicts events using stepwise logistic regression. Our predictor variable was type of violence (defined by the dataset authors), as well as supposedly non-civilian variables (again, as defined by dataset authors). Non-civilian variables included the government or aggressor ("side a"), fatalities sustained by "side a," and region. What we found was that the category of one-sided violence against civilians was the most statistically significant coefficient ($p < 0.001$), when logistically modeling civilian death.

Running a plausible event of violence on the model, by plugging in Islamic State, zero fighters having been killed, the region being the Middle East, and, most importantly, the type of violence being one-sided, the model computed an 85% predicted probability of civilian death. We computed an ROC-AUC of 0.88 ("good") on the model (see below).



Upon cross fold validation (k = 10), the model yielded an ROC-AUC of 0.88 ("good"; sensitivity = 0.68, specificity = 0.96). This was an attempt at internally validating the type of violence variable of the dataset. Given its importance to our research questions and the news source nature of this dataset, we believe the classification accuracy of this model may lend some legitimacy to the dataset and, by extension, our previously modeled trend of civilian deaths.

5 & 6. What factors (number of conflicts, nation states involved, etc.) predict the total amount of arms exported globally? What factors (number of conflicts, etc.) predict the total amount of arms imported globally?

Upon combining the features of country, year and civilian deaths from the GED data set with the arms export, arms import, per capita military expenditure, and military expenditure as a share of GDP from the SIPRI dataset, we have explored and discovered some interesting relationships. Here are some statistically significant findings, though correlational in nature. To differentiate these findings from data dredging, we have chosen to report findings where we hypothesized there may be a plausible relationship.

1. Number of arms imported by a country is positively correlated with the number of civilian deaths due to conflict in that country (with statistically significant p-values). This finding may be important in understanding the role of arms imports in civilian deaths due to conflict.
2. Number of arms exported is negatively correlated with the number of civilian deaths due to conflict in a country, with statistically significant p-values. This finding may be important in exploring how lesser arms procurement may relate to lesser civilian deaths.
3. Civilian deaths due to conflict is not correlated per capita with military expenditure (and is without statistical significance). Inferring that military expenditure per civilian is somehow related to a state's duty or obligation to protect its citizens from physical harm, then a lack of negative correlation between civilian deaths and military expenditure could, in some contexts, be interpreted as alarming.

As we will discuss in our future directions section, we also attempted topic modeling text analysis to gain further insight into state intentions for arms procurement.

Methods and Approach

Questions 1-3

For questions 1-3, Cold War years were removed from the GED, then the dependent variables of interest were aggregated. For the predicted trend in the number of global conflicts, conflicts were aggregated by distinct year and conflict name. For the predicted trend in total and civilian fatalities by year, total deaths and total civilian deaths were aggregated by year.

We then produced and plotted linear, polynomial, and LOESS regression models for each question. Loess regression was used for exploratory data analysis only. We opted for polynomial regression as the data was non-linear and these models yielded the highest adjusted R-squared and lowest p-values. We then interpreted the polynomial regression models in the context of the questions asked. A challenge faced was choosing the right regression model, and improving iteratively on the LOESS and polynomial regressions.

Question 4

A major challenge faced was the sheer number of variables in the dataset and the task of predicting civilian casualties. Given a lack of graduate level, domain expertise in the subject of international conflict, we decided to change our approach after consultation. We narrowed the scope of the question and tried to quantify the construct validity of a key aspect of the database. The revised question was: What are the predicted probabilities civilians died in observed conflict events, as predicted by the author's definition of type of violence?

We began by removing Cold War years from the dataset. We then repurposed the dataset by defining a dependent variable by "dummy coding" number of civilian deaths in 129 thousand events as a binary variable (0 or 1). We then performed stepwise logistic regression on the model described above, and no variables were taken out. We then ran a test case described in the results. Finally, we sorted model coefficients by p-value ascending and cautiously interpreted the findings. To evaluate model performance, we computed the ROC-AUC, plotted the ROC Curve for the model, and performed cross-validation (k = 10).

Questions 5-6

The extract, transform, load process combining the GED and SIPRI datasets was scripted in Python using pandas and numpy libraries. To understand correlation between the various features, normality tests and Spearman correlation tests were performed.

For the development of a topic modelling prototype, as we will discuss more in our future directions, documents were obtained in PDF format and as html files. PyPDF and textract libraries were employed to process, tokenize and extract keywords from these PDF files. Text processing to extract keywords and summary from newspaper articles, and html files were processed through the newspaper library in Python. TFIDF method was also employed to verify keywords from PDF files extracted previously through textract.

Limitations

Questions 1-4

A chief limitation of our answers is that they depend on the comprehensiveness and quality of the GED dataset. This is open to debate. Moreover, it appears the dataset was generated from largely unclassified resources. For example, there is the limitation that militaries and intelligence agencies may conceal the true number of deaths in an event as a state secret. We have sought to explore a small aspect of the dataset's quality by looking at the construct validity of a single variable in the dataset.

A major limitation of our answers in examining trends in the number of global conflicts, as well as total and civilian casualties from 1991 to 2016 to present, is that we cannot make any causal inferences. Our biases are that we are using a dataset whose authors have proclaimed in the news that the number of conflicts are on the rise (though again, we sought to explore and model this claim in this project).

For question 4, the modeling of civilian deaths as a function of type of violence and other non-civilian variables, it should be cautioned that the model should not be used beyond the scope of this project. It is only an attempt at assessing the construct validity of the "type of violence" variable.

Questions 5-6

The SIPRI database had transactions recorded between 'Unknown rebel groups' and 'Unknown suppliers'. This transaction raises possible concerns about the data collection method and about the accuracy of the data. Given these and other concerns highlighted, we caution against making strong inferences using the results produced.

While gathering documents to be added into the corpus for our topic modelling prototype, it was discovered that many countries did not have an authoritative source discussing their military goals and approaches. This led to a weak corpus and results that did not have central topics.

Related Work

The current project differs from existing approaches in that it is combining two major datasets related to global conflict and the international arms trade. Furthermore, we found conflicting answers to our main research questions, with the data source that we used stating that the number of global conflicts and combat fatalities are on the rise (Koffman, 2015), and others indicating that the number war deaths and global conflicts is decreasing (Raphael, 2014; Ashford, 2016). Therefore, we sought to clarify the global trend in the number of wars and fatalities from war. Finally, our approach differs from many approaches in that we tried to use empirical evidence help answer if the world is becoming a safer place, whereas several publications we found on this topic did not utilize statistical modeling or predictive analytics to draw conclusions (e.g., Pinker, 2016). Upon closer scrutiny of the datasets used for this project, it became evident that there are a wealth of peer-reviewed publications using both the SIPRI and GED datasets used in this project.

Future Work

To better understand the motive behind military expenditure of various countries, an attempt to study their policy documents, military strategies, and newspaper reports through topic modelling was developed in the current project, which included the development of a working prototype. However, due to the lack of a strong corpus containing documents commenting on the military strategies of various countries, this attempt did not yield promising results. This is a plausible direction for future work if we were given the opportunity.

References

1. Ashford, E. (2016). We're Seeing a Trend Toward Less Violence in the World. Retrieved from https://www.nytimes.com/roomfordebate/2016/09/06/is-the-world-becoming-safer/were-seeing-a-trend-toward-less-violence-in-the-world
2. Koffman, L. (2015). Global conflicts on the rise. Retrieved from http://www.uu.se/en/media/news/article/?id=4906&typ=artikel
3. Pinker, S. (2016). Despite The Headlines, Steven Pinker Says The World Is Becoming Less Violent. Retreived from http://www.npr.org/2016/07/16/486311030/despite-the-headlines-steven-pinker-says-the-world-is-becoming-less-violent
4. Raphael, T.J. (2014). The world is actually safer than ever. And here's the data to prove that. Retreived from https://www.pri.org/stories/2014-10-23/world-actually-safer-ever-and-heres-data-prove
5. Sundberg, Ralph, and Erik Melander, 2013, "Introducing the UCDP Georeferenced Event Dataset", Journal of Peace Research, vol.50, no.4, 523-532 Croicu, Mihai and Ralph Sundberg, 2017, "UCDP GED Codebook version 17.1", Department of Peace and Conflict Research, Uppsala University