Machine Learning and Conflict Prediction: A Use Case

Chris Perry
International Peace Institute, New York

ABSTRACT
For at least the last two decades, the international community in general and the United Nations specifically have attempted to develop robust, accurate and effective conflict early warning system for conflict prevention. One potential and promising component of integrated early warning systems lies in the field of machine learning. This paper aims at giving conflict analysis a basic understanding of machine learning methodology as well as to test the feasibility and added value of such an approach. The paper finds that the selection of appropriate machine learning methodologies can offer substantial improvements in accuracy and performance. It also finds that even at this early stage in testing machine learning on conflict prediction, full models offer more predictive power than simply using a prior outbreak of violence as the leading indicator of current violence. This suggests that a refined data selection methodology combined with strategic use of machine learning algorithms could indeed offer a significant addition to the early warning toolkit. Finally, the paper suggests a number of steps moving forward to improve upon this initial test methodology.

Introduction

For at least the last two decades, the international community in general and the United Nations specifically have attempted to develop robust, accurate and effective conflict early warning system for conflict prevention. From Boutros Boutros-Ghali's An Agenda for Peace (UNSC 1992) on through to the present day, Secretaries General, practitioners and academics have continually documented and called to attention the need for comprehensive early warning systems to collate, analyze and disseminate information and data on sociopolitical and armed conflict dynamics. Indeed, as recently as September 2011, Secretary General Ban Ki-Moon and the UN Security Council reiterated this need once again through the Preventative Diplomacy: Delivering Results report (UNSC 2011a). The president of the Security Council at the time stated that a 'key component…of a comprehensive conflict prevention strategy include[s] early warning [mechanisms]' (UNSC 2011b). The need for an early warning system within the UN system is therefore well established. However, early warning systems face a number of practical hurdles to implementation. Two of these are access to open data and technical limits on making sense of that data once obtained.

One potential and promising component of integrated early warning systems lies in the field of machine learning. Machine learning is a branch of computer science that leverages algorithms, or a set of step-by-step computational procedures, to perform actions without explicitly being programmed to do so. Many of the methodologies that undergird the field of machine learning are not new, with some of the newest methodologies like support vector machines and random forests developed in the mid to

late 1990's. Indeed, some of these methodologies have been utilized in limited cases for early warning over the past decade.[1] However, advances in data management, predictive analytics, and parallelized data processing[2] have made these methods more widely accessible for use in both big and small data analysis. Additionally, all provided enormous benefits to a wide variety of private (as well as some public) sector applications. For example, the 2012 election campaign of US President Barack Obama successfully utilized a range of data science[3] techniques to analyze voter behavior, manage get out the vote campaigns, and better target potential swing voters. Companies like Netflix, Amazon, and Google have utilized predictive analytics for a number of years to predict consumer behavior and better target recommendations or advertisements. In both the macro- and microcosms, parallelization has had huge impacts for everything from human genome sequencing to analysis of the massive amounts of data returned from deep field telescopy. And of course the recent revelations regarding the NSA surveillance, including the now infamous PRISM program, indicate that all of these techniques are being utilized in intelligence gathering by some nation states.

Data science is increasingly being applied to the domain of international development and international relations, where the field holds 'new opportunities for humanitarian and development assistance in the most complex and dangerous environments' (Kilcullen and Courtney). For example in poverty mapping, researchers in Spain developed a predictive model using anonomyzed call detail records (CDRs) from an unnamed South American city to map poverty. The model was applied using a variety of predictors derived from aggregating call locations and call characteristics and was tested against existing survey derived municipal poverty maps. The result showed a high degree of accuracy (Soto et al. 2011). Similarly, one proposal to the NetMob 2013 special session on the D4D challenge proposed using CDRs for subnational poverty mapping in developing countries. The paper tested this hypothesis on historical data in Cote d'Ivoire, again with a high degree of accuracy (Smith et al. 2013). More recently, a number of researchers have attempted to apply machine learning towards predicting outbreaks of violence.[4]

Indeed, it appears that the application of machine learning techniques could provide significant contributions to tactical early warning systems and conflict prevention strategies in particular, if leveraged intelligently as part of a larger system of 'intelligence.' However, questions remain: is it possible to use a combination of open source geospatial and national statistics to develop subnational indicators of vulnerability to violent events? Is there a way to scale and incorporate additional data as it becomes available? Can the aggregation and modeling process be parallelized to handle larger datasets in real-time? Finally can these modern computational techniques even be used to predict conflict and feed into 3rd and 4th generation early warning systems?

This paper outlines an initial approach to test exactly that. First the paper offers a definition of machine learning terminology as an entry point as well as an outline of the analytical methods used. Then, in order to address the lack of highly granular and timely global statistics, this paper proposes a mixed method approach of using GIS data processing techniques to aggregate national, subnational and satellite data to the district level. The resulting data is then used as a test case for the application of two predictive machine-learning algorithms, which is reported in the final section.

Machine Learning Methodology

For the intents of this paper, the problem can be defined as a supervised classification machine-learning problem. As the previous sentence is probably not terribly meaningful to conflict analysis practitioners, a few definitions are in order at the outset.

Machine learning can be defined as 'systems [or algorithms] to automate decision making and classification of data' (Warden 2011:31). These algorithms use existing or incoming data to extract structure, generalize results and make predictions about future data. Machine learning tasks are typically divided into two categories: *supervised* and *unsupervised* learning. Unsupervised learning is an attempt to provide structure to relatively unstructured data. In unsupervised learning, the dependent variable is not known at the outset. For example, education researchers have used cluster analysis, a common unsupervised learner, to identify groups of students with similar demographic and socioeconomic properties. Likewise, police can use geo-referenced data on crime to identify 'clusters' of types of crime over time in order to better formulate policing strategies.

Supervised learners on the other hand, of which the current paper is an example, are typically used for making predictions from existing data.[5] The algorithm uses sample data in which the dependent variable is already known to extrapolate onto data in which the dependent variable is not know. A *classification* problem revolves around identifying which category or set of categories a given observation belongs to based on a set of predictive variables or features, often called the *feature space*. An algorithm used to identify is called a *classifier* and the category scheme is referred to as the *target class*.

One example of supervised classification learning is optical character recognition. In this use, the incoming feature space is a set of the pixel positions of known letters. The target class is simply what letter a group of pixels represents. Another example is spam filters in which the incoming data consists of emails pre-classified as spam and not-spam. In both these cases, incoming data is divided randomly into a *training set* and a *test set*. The training set is used to train the model for prediction and the test set is used to test the final model for predictive accuracy and generalizability.

There are a wide variety of learning algorithms to choose from for supervised classification applications. This paper represents the first step in an iterative process of developing a workable model of conflict prediction. As such, many of the methodological decisions that underpin this stage were made in the interest of expediency. Data was selected for completeness, algorithms were selected for speed, etc. Future research will refine this methodology to create better and more accurate models. However, this stage is rather more about testing the feasibility of even designing such a machine learning system. Due to these limiting considerations two initial options were selected: naïve Bayes and Random Forests.

The naïve Bayes algorithm provides a relatively simple test application that in many cases gives tolerable performance in terms of both speed and accuracy. The naïve Bayes algorithm is a simple probabilistic classifier. That is the algorithm assumes that the presence of a particular feature is unrelated to the presence of any other feature, given the target class variable. For example, an animal may be a classified as a cat if it is a

mammal, domesticated and 10 lbs. A naïve Bayes algorithm would consider each features contribution to the probability that this animal is a cat independently regardless of the presence or absence of other features.

The second method, the random forest algorithm, provides a more sophisticated learner and offers the possibility of model improvement. Random forests have gained popularity for speed and accuracy performance as well as ability to handle data with either a large number of observations or variables. Random forests are a form of ensemble learner that extends the decision tree learner algorithm. Ensemble methods refer to the use of multiple models that are then recombined to increase the predictive performance over any of the single models. Decision trees create a model of that tries to predict the value of a target variable based on a range of input variables. The trees are composed of a root, children and leaf elements. Each split corresponds to a given input variable and each child element corresponds to that variables possible values. Each leaf element corresponds to the value of a target variable given the all values of the elements as you traverse the path from the root to the leaf. The tree is 'learned' by deriving splits based on some test that determines which input variable best splits the data at that level. Random forests expand on this methodology by iterating over the data to create multiple trees. Each tree is derived from a sub-sample of the training data and a sub-selection of the predictor variables. The final classification is determined by an aggregation of 'votes' from each of the component trees. Random forests are highly popular due to a number of factors. In a wide variety of use cases, they have proven highly accurate. They are also efficient on datasets with a large number of observations and variables. They are also useful in determining the importance of component variables, which can be useful in designing future research.

Data

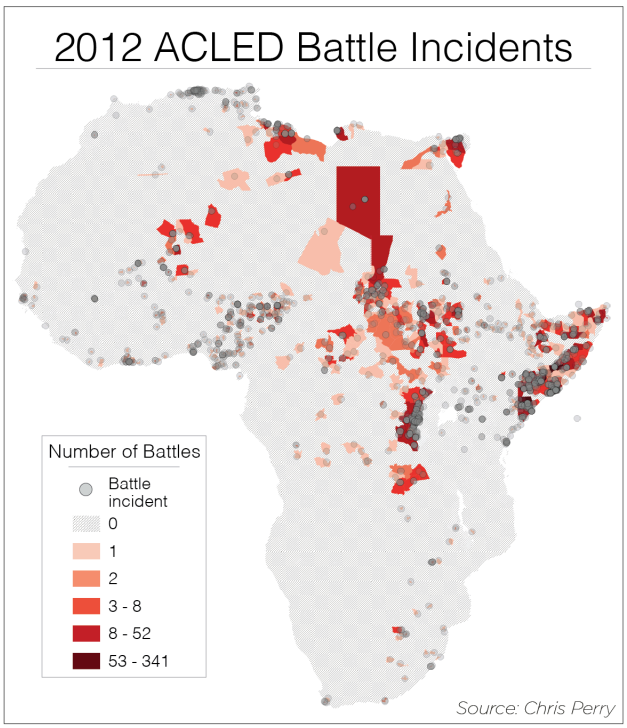
Much of the previous research involving international comparisons of vulnerability and conflict have taken the nation state as the unit of observation.[6] However, national aggregation of data risks loosing site of the nuances and context inherent in conflict analysis. For instance, a national aggregation of risk factors and incidence of violence in Nigeria over the last decade would risk collating the outbreaks of violence in the north and the south, both of which have very different historical factors and characteristics.

This project tries to square the circle by taking a more granular view, while maintaining some of the advantages of political boundaries. The unit of analysis is the third and second level subnational administrative boundaries derived from the National Administrative Boundaries GIS dataset provided by Columbia's Global Rural-Urban Mapping Project (GRUMP) (CIESIN et al. 2011). This data includes 399,747 non-overlapping polygons covering the globe. Each polygon corresponds to the geographic area of a specific subnational district or county. For instance the second level administrative boundaries in the United States includes all counties in every state in the country. The third level administrative boundaries of Canada would include all rural and urban jurisdictions in each district or county of each province in the country.

Research at this stage is limited for the reasons discussed above regarding experimentation. In the interest of maintaining a scope conducive to experimentation, the

area of study was limited to continental Africa. This results in 33,752 subnational units covering fifty-five countries. Future research will experiment with hierarchical modeling of subnational and national units as well as a variety of levels of aggregation on their own and an expanded scope.

Figure 1: 2012 ACLED battle incidents

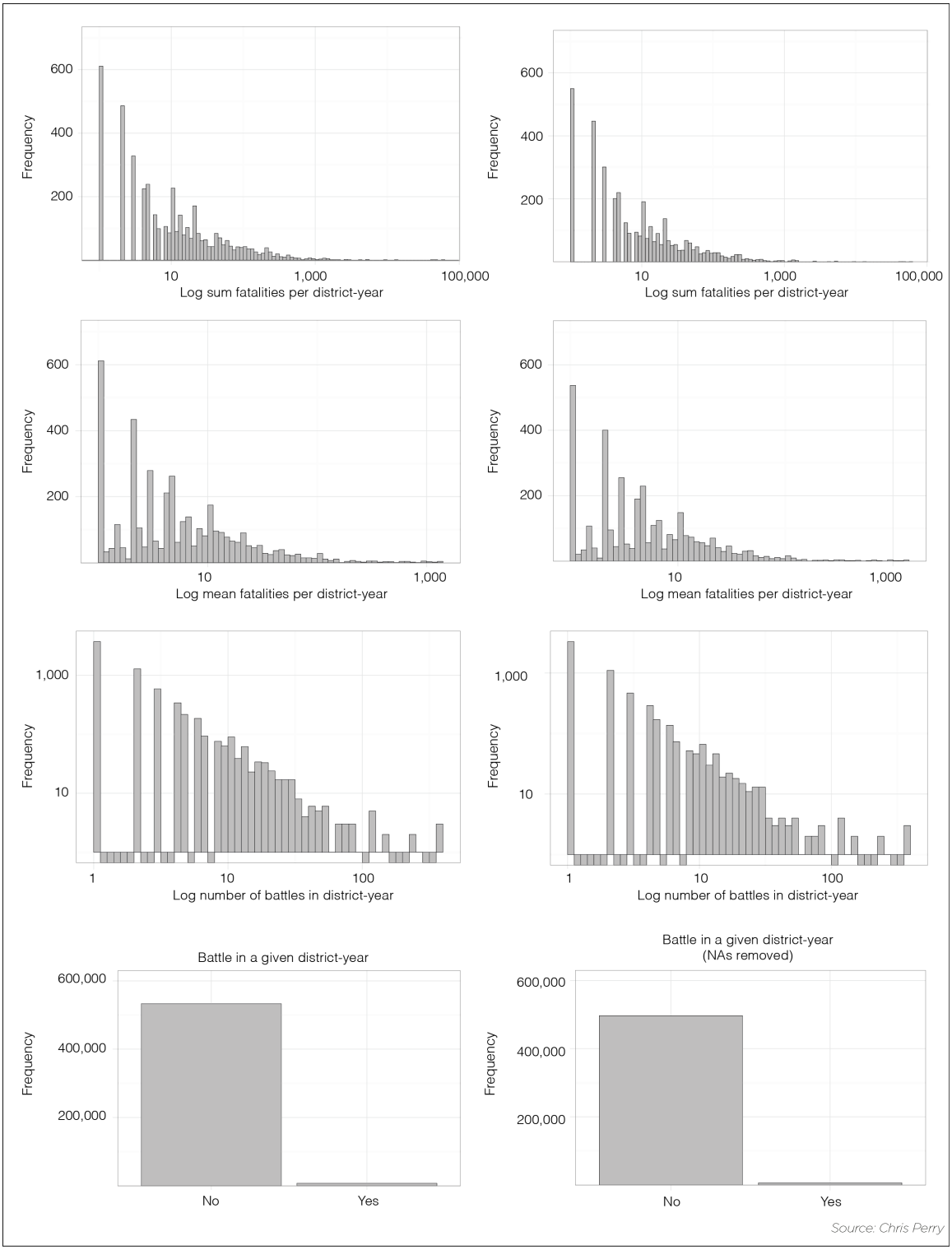


A necessary component of a classification machine-learning problem is the target class, in this case the instance or intensity of conflict in a given district. Over the last few years, a number of geospatially tagged datasets of violence have been developed. For the purposes of testing, we chose the Armed Conflict Location and Event Dataset (ACLED) (Raleigh et al. 2010). Designed for disaggregated conflict analysis and crisis mapping, the dataset includes reported political and armed conflict in over 50 developing countries, though we only look at battle incidents and fatalities. Temporal coverage spans from 1997 to near real time, but in order to match additional data, the scope of this paper is cut at 2012. Data was aggregated using a vector join function that yielded the number of battles, as well as minimum, maximum, mean, median, and sum of fatalities. This was done for each year from 1997 to 2012. This creates a dataset of 33,752 districts over sixteen years or 540,032 district-year observations.[7]

While there are a number of ways to derive a target class from this data, this paper uses two variations. The first is a simple binary class that indicates whether or not any battles occurred in a given district in a given year. The second is a numeric count of the number of battles that occurred in a given district in a given year. During initial testing, two other target variables were attempted: the sum of fatalities in a given year in a given district and the mean fatality per battle in a given year in a given district. Neither of these classes performed well either as continuous variables or as binned variables. However, they were a useful proxy for previous levels of violence that was in turn used as predictor variables. These variables also provide a useful baseline for predicting conflict based solely on previous conflict. The distributions of all four conflict variables are provided below.

The next step in the machine learning process is the selection of predictor variables. Predictors at this stage of experimentation were chosen based both on the literature surrounding conflict analysis and based on data availability. It is important to note at this stage that machine learning is an iterative process. While this paper establishes the foundation, future research will focus on refining the input data and improving the speed and accuracy of a variety of learners.

Figure 2: Frequencies of target classes before and after the removal of missing variables



Certain levels of ethnic fractionalization have been found to be a predictor of conflict in a number of econometric studies.[8] For this paper, supranational ethnic composition is taken from the Geo-referencing of Ethnic Groups (Weidman et al. 2010) dataset. The data

consists of 8,969 GIS shapefile polygons and includes features referencing majority ethnic composition drawn from the classical Soviet Atlas Narodov Mira. An intersection function was used to create a count of the number of different ethnic groups that overlap a given district. The values gained ranged from 0 to 18 with a mean of 1.286.[9]

Population density can be an important predictor of the level of violence if only because it is statistically more unlikely to find human on human violence in and area sparsely populated. Satellite derived population data was taken from Columbia's Gridded Population of the World (GPW) project and uses both the past and future population grid counts (CIESIN et al. 2005). The data consists of 29,652,480 raster grids, each associated with a value that indicates that grid cell's population level. The data comes in five-year increments starting in 1990 and projects to 2015. Past data is correlated to match UN population data revisions. This was processed to the district level using a zonal statistics function, which derives continuous variables on the sum and mean of values contained in a given district. The data was further processed to give the population value of the year of observation as well as the value lagged by one and two years. Mean values, which denote the average value of a raster cell in a given district, ranged from 0 to 751,416 with a mean of 6,550. "Sum" values, which denotes an estimate of the total population of a given district, ranged from 0 to 10,462,972 with a mean of 24,306.

Poverty has been posited often as a driver of conflict through both the greed and grievance models of conflict with a variety of causal mechanisms posited.[10] GDP data was taken from Columbia's Global 15X15 Minute Grids of Downscaled GDP Based on the SRES B2 Scenario (Yetman et al. 2004). The data consists of 823,680 raster grids indicating an absolute level of GDP at 1990 and a projection of 2025 levels. The data was processed by first taking zonal statistics on both the 1990 and 2025 GDP levels. These end points were used to calculate both the yearly increment and the yearly percent change necessary to get from 1990's level to 2025's level. Yearly national GDP estimates were then taken from the World Bank to calculate a given district's deviation from the long term projected average. This deviation was applied to the calculated incremental change and applied beginning in 1990 to calculate a district's new yearly trajectory and account for changes since the data was created in 2000. This data was used to create mean (the average GDP of a raster cell in a district), change (the change in GDP from the previous year) and sum (the total GDP of a given district) values. These values where in turn used to create values for the current year as well as values for one and two years lagged. The sum data ranged from 0 to 733.0744 with a mean of 1.5950.

An alternate proxy is comprised of both the Global Subnational Prevalence of Child Malnutrition and the Global Subnational Infant Mortality Rates (CIESIN 2005), both from Columbia. The data consists of 823,680 raster grids indicating the percentage of children under five years of age that are underweight and the grid's child mortality rate per 10,000 live births. Data was processed using a zonal statistic function. Infant mortality rates ranged from 0 to 2031 with a mean of 950. Underweight values ranged form 0 to .54 with a mean of .29.

Acute hazards can place stress on areas that can lead ultimately to conflict. This is especially the case for hazards that have long-term implications for livelihoods. Hazard data was taken from both the flood and drought components of the Global Multihazard Frequency and Distribution (CHRR et al. 2005) from Columbia. Specifically each

indicator comes from a 29,652,480-cell raster indicating the frequency of a given hazard on a scale from 0-10, with a zero indicating no or virtually no frequency. Data was aggregated using a zonal statistics function. Frequency values ranged from 0 to 10 with a mean flood value of 4.899 and a mean drought value of 5.956.

In some cases, conflicts can arise over land use between pastoralists and farmers. For this paper, land usage data was taken from the Global Agricultural Lands data from Columbia (Ramankutty et al. 2010). The data includes two raster files consisting of 9,331,200 cells indicating percentage of land used for pasture and crops. These were processed by first using the QGIS raster calculator to multiply the two percentages together and multiplied by four. This creates and index where a score of one indicates 100 percent of the land being used for pastures and crops, with fifty percent devoted to each use. A lower score can indicate either a shift in this proportion or a percentage of the land used for other purposes. The resulting raster was aggregated via a zonal statistics function. The values range from 0 to 1 with a mean of .1968.

Natural resources have been debated ad-infinitum as drivers of conflict. For this study we focus on the availability of two resources. Diamonds tend to better represent lootable resources that are often used as illicit financing instruments, while the availability of petroleum can be an incentive for state capture. Diamond data is numeric variable that indicates the number of lootable diamond deposits in a given district. Data comes from the Diamond Resources (Gilmore et al. 2005) dataset from the Peace Research Institute Oslo (PRIO), which is geo-referenced and disaggregated by type of deposit. Lootable diamonds are those that are extracted relatively easily using artisanal methods and are often alluvial or surface deposits. Non-lootable diamonds, in contrast, require heavy mining equipment and substantial extraction infrastructure and tend to be less easily used for financing armed non-state actors. These points were aggregated to district polygons using a GIS merge by location function. Values ranged from 0 to 3 with a mean of .00453. Petroleum data was taken from the PRIO Petroleum Dataset v. 1.2 (Päivi et al. 2007). Data consists of shapefiles indicating all know petroleum deposits in the world. The data was processed using an intersection function to create a binary variable indicating simply whether or not the district contains a petroleum deposit, which was the case in two percent of the districts.

Finally, institutions are a key factor to a functioning society, effective rule of law, and peaceful conflict resolution mechanisms. Governance indicators were taken from the Database of Political Institutions (DPI) created by the Development Research Group at the World Bank. The DPI contains 125 variables, mainly measuring aspects of the political system and electoral rules and is aggregated at the country-year. SQLite was used to propagate this data through subnational levels.

The final list of variables considered at this stage is:

- Ethnic composition
- Land conflict index
- Mean flood and drought frequency
- Number of lootable diamond deposits

- Existence of petroleum
- Population of children under five years of age in 2000
- Mean percentage of children under five years of age underweight
- Mean infant mortality rate
- Mean GDP of the current and previous two years
- One year change in GDP for the three previous years
- Mean population for the current year and the previous two years
- Sum and mean of battle related fatalities and number of battle events from the previous three years calculated by $y_1 + (.5 * y_2) + (.25 * y_3)$
- Whether the government is head of government is from the military
- The vote share of government and opposition coalitions in the legislative branch

While imperfect, these data represent a useful first step in testing the feasibility of aggregating disparate data sources as well as to test whether the application of machine learning has any added value for this problem domain. Governance data was especially problematic as the features selection came down to choosing most complete variables. Even taking this into account, the final dataset was 37,153 observations short of the full set of district-years due to missing data. The distribution of target classes is included above on the right and there seems to be no discernable systemic loss of data as far as distribution. However it should be noted that in the final data set, the Democratic Republic of the Congo and South Sudan fall out completely. Future research will address this important shortcoming.

Analysis

For the purposes of this paper, two learning algorithms were applied: naïve Bayes and random forest. As a baseline, both classifiers were tested using only prior conflict data as a predictor. This allows a test to see if machine learning offers any improvement in predictive power over simply knowing that a given district had experienced conflict in prior years. The naïve Bayes classifier provides a relatively simple algorithm that offers a tolerable base case in terms of speed and accuracy. The random forest algorithm provides a more sophisticated learner and offers the possibility of performance gains between models. The predictor features and target classes of the full model are outlined in the section above. Following machine learning convention, the data was divided randomly with 70 percent of observations allocated for a training set and 30 percent for a test set. As a final test of accuracy, the data for 2012 was first separated as a final homogenous test set. This means that the models were trained and tested on data spanning 2000-2011, with a final accuracy test reported for the full set of 2012 data.

Figure 3: Performance of baseline, naïve Bayes and random forest machine learning algorithms

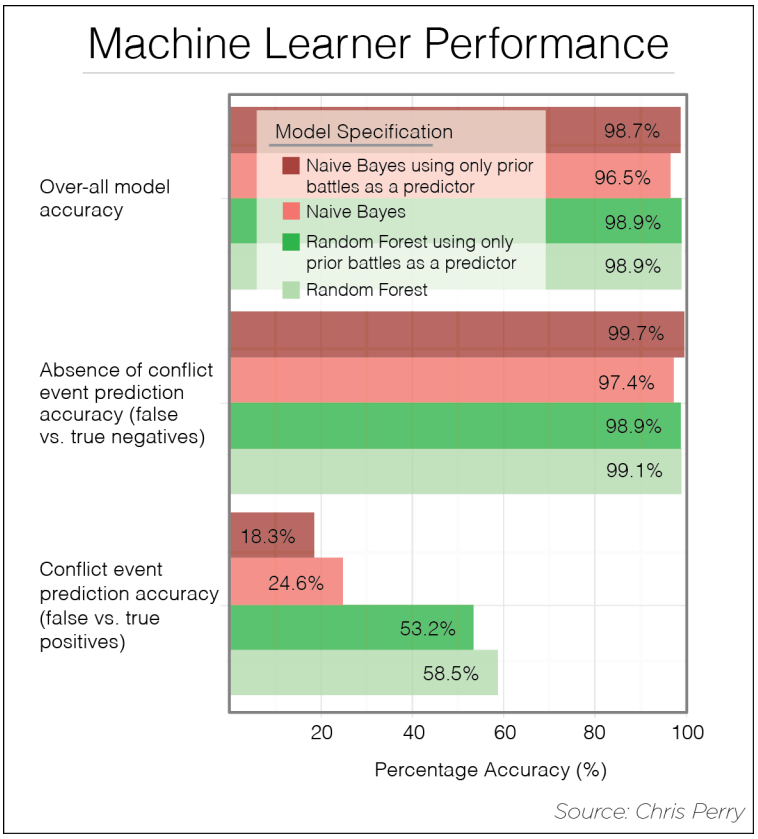
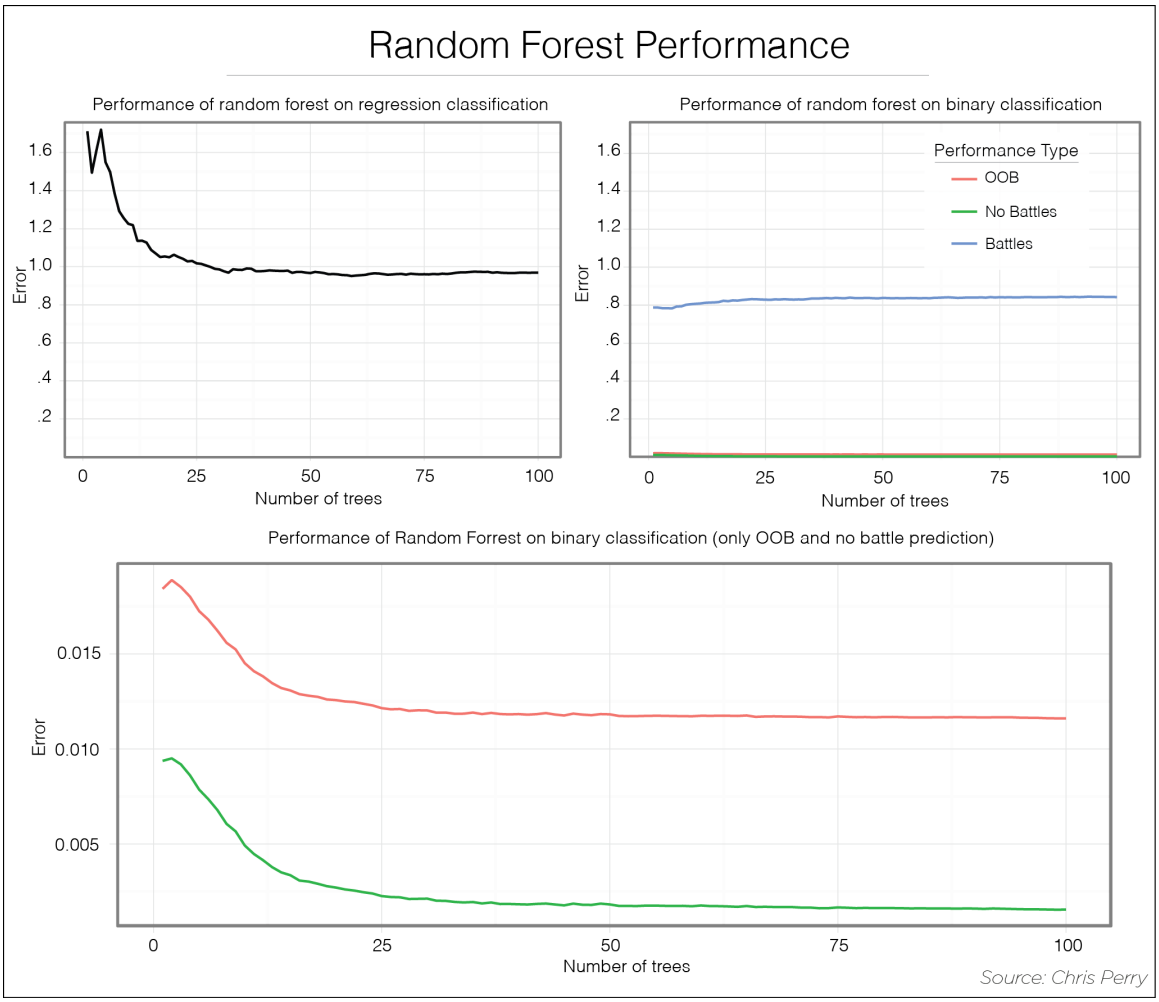


Figure 3 shows the accuracy performance of the algorithms. There was not a large variation in terms of total accuracy across models. This is not surprising as prior research indicates that machine learning tends to over predict instances of conflict and districts experiencing outbreaks of violence are exceedingly rare. Therefore, large swings in the accuracy of predicting violence (as opposed to predicting a lack of violence) would have relatively little effect on the overall predictive accuracy. Indeed, that is born out by the algorithm's performance against the baseline. However, while the overall variation in accuracy is minimal, there are large differences between the models when looking at the ratio of false to true positives. That is, the naïve Bayes algorithm is gains six and a half percentage points of accuracy when predicting an outbreak of violence while only loosing a two percentage points of accuracy in predicting the absence of violence over the baseline. The algorithm was performed with 4-fold cross validation[11] on two variants of the target class: a binary indicator of battles in a given year and a twelve level, binned indicator of the number of battles in a given year.[12]

The random forest algorithm offers drastically better results than the naïve Bayes algorithm, both for the baseline and for the full model. The model is slightly more accurate overall than the naïve Bayes baseline. The full model also looses no accuracy when moving from the baseline to the full model while gaining over five percentage points in accuracy when predicting outbreaks of violence. For the purposes of this paper, the R randomForest package was used to run a 100-tree forest on the data. The model used two variants: a binary target class indicting simply the existence of battles in a given year and a continuous target class indicting number of battles in a given year. Both versions offered significant improved performance to the naïve Bayes classifier. For the binary target class, the out-of-bag error rate[13] estimate was only 1.16%. However, nearly all of this error was cases of false positives. In the cases where the model predicted no battles there was a slight increase in accuracy from 97.4% to 98.9%. In the cases where the model predicted battles, there was a more significant increase in accuracy from 24.6% to 58.5%, indicating a decrease in false positives. In the case of the regression classifier,

comparison between the random forest and naïve Bayes models is more difficult. The model does estimate that 26.9% of the inherent variance is explained by the model. But still in both cases, the number of false positives far outnumbers false negatives.

Figure 4: Performance of learners and random forest classification models

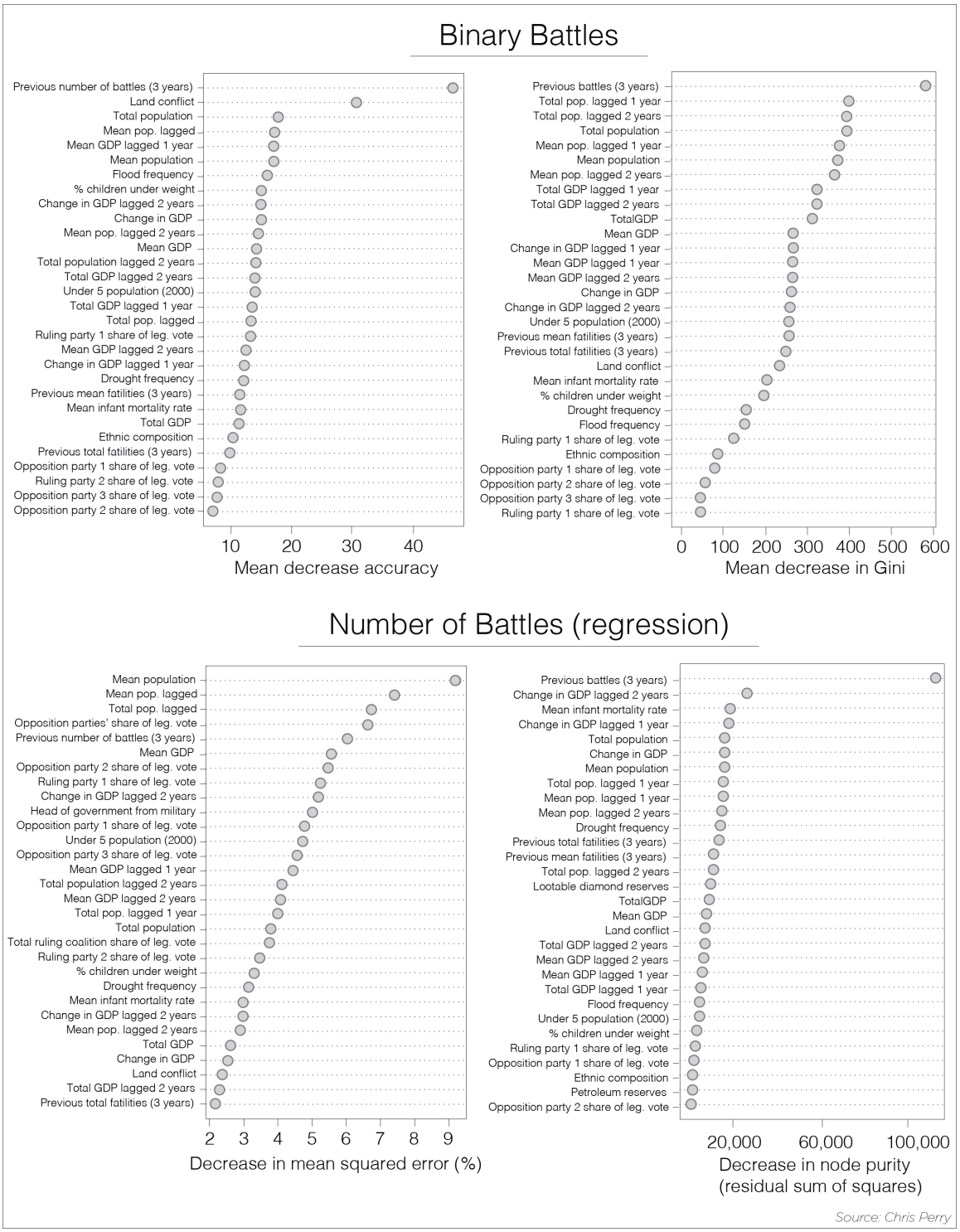


The performance of both of the random forest models is reported in Figure 4 below. In the case of the regression model, accuracy of the model seems to peak at around the 60th tree. Somewhat surprisingly, the binary model has the opposite trend in that accuracy peaks at a very early iteration (around 10) and gets worse until leveling at about the 40th tree. This is very slight increase in error, however, and is offset by a corollary decrease in error for both the out-of-box and 'no battle' prediction which again seems to peak around the 50-60th tree iteration.

One advantage of the random forest is the ability to determine variable importance based on each variables contribution to an increase in purity and a decrease in error. These are reported in Figure 5 below. The charts on the left measure the decrease in accuracy of the model when a given variable is removed. The charts on the right measure the aggregate mean decrease in purity of the leaf elements of trees when a given variable is removed. The top thirty most important features are reported for each chart.

Chris Perry
Deleted: F

Figure 5: Variable importance for both learners measured in the decrease in accuracy (left) and decrease in overall node purity when variable is removed.

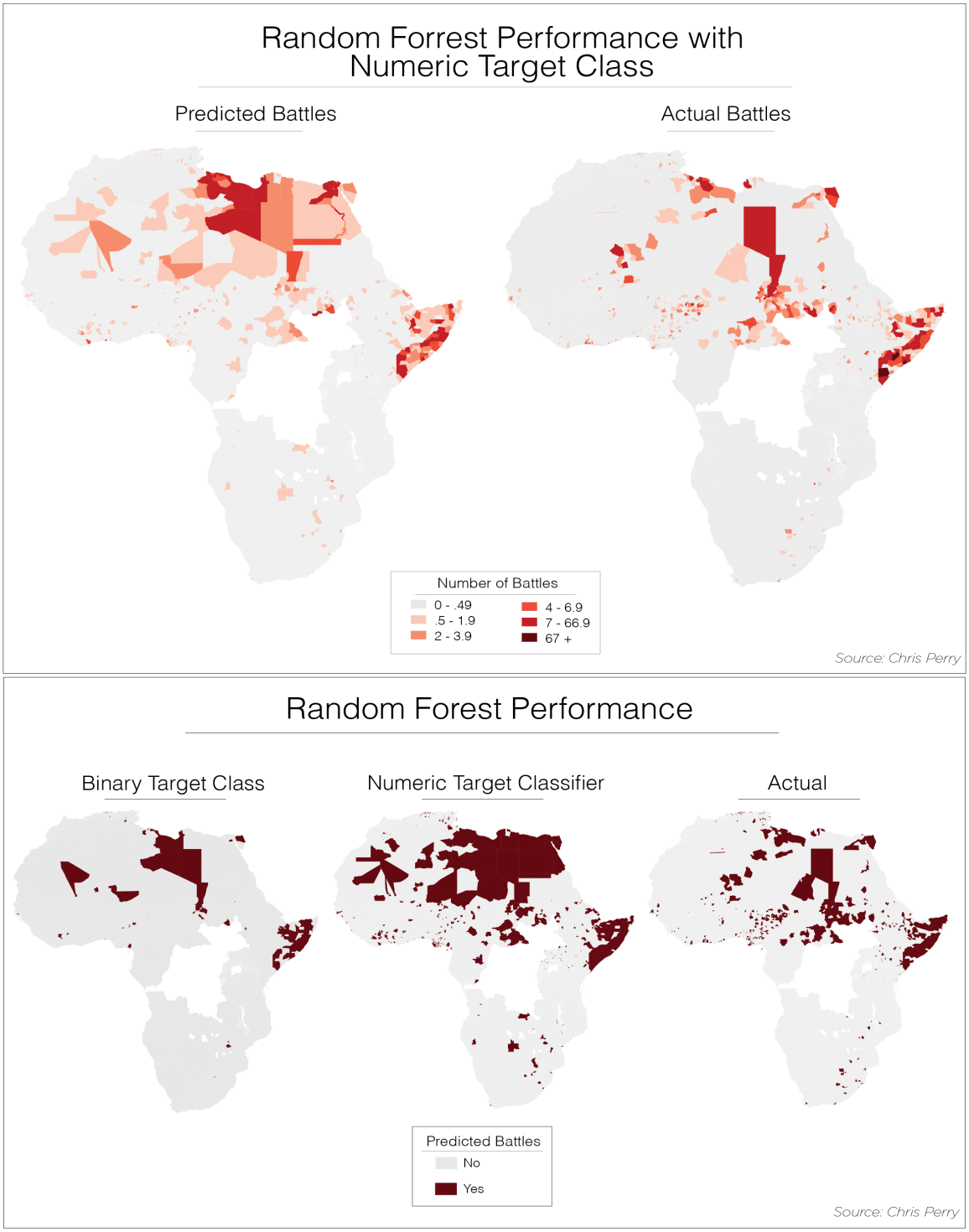


Unsurprisingly, the number of battles in the previous three years figures highly as does population. Due to the fact that many of these component variables are auto correlated, future analysis will consider aggregating lagged measures into a single variable by using time decomposition aggregation. Promisingly, variables like ethnic composition, drought

and flood frequency, and some of the governance indicators offered a payoff in terms of increased accuracy. Future research will focus on refining these indicators as well as identifying new and better ones.

As a final test of the learners, the derived models were applied to 2012's data, which was initially partitioned from the rest of the dataset. The predictions were then mapped

Figure 6: Map of predicted and actual classes for 2012



against actual instances to give a visual representation of the accuracy of the models. These maps are presented in Figure 6 above with the regression forest on the top left, actuals on the top right, and a composite of the binary forest, regression forest converted to a binary class and the actual binary occurrences on the bottom. These maps give a good visual representation of the accuracy of both learners. Interestingly, the binary random forest learner actually heavily under-predicts for 2012. What this also shows is that while the regression forest learner tends to over-predict; it does not do so randomly. Over-prediction values tend to be geographically clustered in the right place, indicating that there may be some promise going forward as the data is expanded upon and refined.

Chris Perry
Deleted: F

Conclusion

This paper represents an initial step towards using machine-learning methodology to model fragility and vulnerability to conflict at the global level. It is true that these initial gains from the application of machine learning seem small and may leave much to be desired for policy makers hoping for big gains in actionable information. It needs to be stressed however, that this paper only represents an initial feasibility study. Next steps will focus on refining the models to improve accuracy as well as creating scalable applications to process incoming data more quickly and efficiently. That said, there are a number of lessons learned that should be incorporated into future research.

First, there is the question of using a broader or narrower geographic scope. On the one hand, a more narrow regional or country based scope could be helpful for a few reasons. Limiting the geographic area could help to better create a rough hierarchy of risk factors in a specific context. The missing data issues would likely be less an issue in a narrower scope, as it would preclude the need for data with global coverage.

On the other hand, there are a number of advantages to developing models with a global scope. For strategic planners, those working at multilateral organizations for instance, it can be imperative to have a global picture of risk. This is one of the reasons that within the domain of international affairs aggregated national indexes of fragility and risk are so widely utilized.[14] Using machine learning techniques to tie these types of risk indexes to actual risk of conflict or violence would help make them more effective. Global scoping of subnational data aggregation also would allow researchers to capture risk and vulnerability in areas that may seem relatively stable to outside observers.

A second lesson learned is the need for better data management and processing tools. There are a number of methods to parallelize the processing of large datasets. The most feasible and scalable solution would be to move processing tasks to a cloud-based architecture. Cloud services like Amazon's Elastic Cloud Computing lets users rent processing bandwidth and storage on a per-use basis. This would allow future research iterations to expand processing power as the need arises. And this need tended to arise often, especially when dealing with zonal statistics. Many of the zonal statistic operations took at least a day to process. Indeed, some data was left out of the model due to barriers on reasonable processing time. Fortunately, there exists a geoprocessing library written in a Java-based language called Geotrellis. It is meant specifically for the task of processing geographical data and contains a number of features for both performing zonal statistics

and running the processes in parallel. Future processing should be done using Geotrellis on a cluster of machines running in parallel on Amazon's EC2. Additionally, the machine learning tasks could also be parallelized using the same architecture. Finally, parallelization would allow an application of machine learning the option to harness disparate sources of data in real time if such data was deemed useful. All of the learning operations took upwards of six hours to complete in serial. Parallelizing the learning tasks could open up possibilities of utilizing different learning algorithms as well experimenting more with variations on the random forest paradigm. Finally, using parallel processing techniques opens up the possibility of performing real-time or near real-time analysis in the future.

The third take-away is that future research will need to address autocorrelation of some of the data. This could be done by creating a time deteriorating index of temporal data for certain categories of indicators. This project makes a first attempt in that direction by counting number of fatalities of battles in a previous year as full, two years previous as half and three years prior as one-quarter. A useful extension of this might be to change the deterioration equation to extend out ten full years. This would mirror some of the literature on conflict recidivism, which indicates that the risk period for resurgence of conflict after the end of hostilities is approximately one decade.

A fourth lesson is that the success of modeling violence going forward will be determined by finding new and better sources of data. One principle concern for variable selection at this stage of the project was expediency. The population data for instance is not ideal as the data used were three snapshots at five-year increments for a baseline and three projections at five-year increments used to interpolate missing values. Likewise with the GDP data. There are a number of new projects aimed at taking yearly snapshots of GDP and population based on satellite imagery that could greatly improve any modeling accuracy. Likewise, the governance data was problematic. Since the DRC and South Sudan were omitted due to missing data, and these are cases in which there were a number of battles, there is likely a large subsequent drag on predictive power. This needs to be rectified in future research. An interesting model is the Failed States Index with uses a combination of natural language processing on news and open source documents as well as other data to create national aggregations of component indicators. They have had some success in targeting their data aggregation techniques to the subnational level in specific cases, which could prove a useful addition if they were so inclined.

Fifth, the target class used at this stage is imperfect. Ideally going forward, models would use an expanded definition of violence that would offer a much richer target class. One line of research could extend analysis and modeling across the additional component classes of the ACLED data, for instance on rioting or violence against civilians. Another interesting tract is to use a new dataset called the Global Data on Events, Location and Tone (GDELT). This data includes more than 200-million geo-located events with global coverage from 1979 to the present day. This could prove useful in developing better target classes, especially target classes with global scope.

Notes:

[1] *See* for example Schrodt (1999).

[2] Serial processing is a programming paradigm in which computational tasks are conducted one at a time in sequence. This is in contrast to parallel processing, in which processing tasks are delegated to multiple CPUs or cores to run concurrently.

[3] Data science refers to a branch of multidisciplinary applied quantitative research that applies methodologies from mathematics, statistics, advanced computing, data modeling, data visualization and hacking as well as specific domain expertise.

[4] *See* for example Tikuisis et al. (2013); and Avra et al. (2013). It should be noted that both of these articles are based on the data of the International Conflict Early Warning System (ICEWS) and later Global Database of Events, Language and Tone (GDELT). While this data is indeed well suited to machine learning methodology, it was unfortunately not available when the research in the current article was undertaken. Future iterations of IPI's work on conflict prediction will seek to utilize GDELT data.

[5] It is important to note a key factor that distinguishes supervised machine learning from the sorts of quantitative analysis, such as econometrics, that are more often used in policy research. Namely, hypothesis testing versus prediction. In the types of statistical methods more widely used in the policy arena (and in conflict research more specifically), research begins with a hypothesis as to why a certain event occurs or an actor engages in a certain behavior. The researcher then uses statistical methods to test whether this hypothesis holds and adjusts the hypothesis accordingly. The goal of this methodology is to develop a more and better systematic understanding of the causes of something like conflict. Predictive analysis on the other hand is less interested developing an explanatory model of behavior, and rather more interested in finding patterns in historical or incoming data to predict trends or behavior. While predictions are often focused on future events, they can also be applied to past events (for instance 'predicting' who committed a crime) and ongoing events (for instance identifying credit card fraud in real time). In short, this distinction is important for conflict-focused policy makers to keep in mind when thinking about the application of machine learning. Hypothesis testing is most useful for making decision as to what sorts of responses to take. Predictive analytics and machine learning are most useful for modeling where and when hot spots will occur.

[6] For an excellent meta-analysis of these various threads of research as well as a commentary on the state of the art, *see* Blattman and Miguel (2010).

[7] It should be noted here that there are a number of limitations in ACLED's data collective methodology. As Kristine Eck surmised in her recent article 'those interested in sub-national analyses of conflict should be ware of ACLED's data due to quality-control issues which can result in biased findings if left unchecked by the researcher,' (Eck 2013). The author acknowledges these limitation and future research will utilize GDELT data, which was only made available after this paper was drafted.

[8] *See for example* Fearon and Laitin (2003).

[9] A value of zero indicates either and area largely unpopulated (the Sahara desert for instance) or where no information exists.

[10] *See* Blattman and Miguel (2010).

[11] Cross validation is a method to make models derived from machine learning algorithms more generalizable. When data is divided into test and training sets, a model's accuracy could be an artifact of the particular way the data was divided. When

performing k-fold cross validation, each step of the validation partitions the data anew, creating k unique 'test-training set' combinations.

[12] To create the bins, the number of battles in a given year was separated into no battles, ten equal intervals for 1-100 battles and a final bin of number of battles exceeding 100.

[13] The out-of-bag (OOB) error is a method to get an unbiased estimate of the test set error. In the random forest model, each tree is constructed using a unique bootstrapped sample from the original data. The OOB measures the proportion of the times that the predicted class of an unsampled case in a given tree is incorrect. For a more detailed description of random forest methodology *see* Breiman and Cutler.

[14] See for instance the Global Peace Index or the Failed States Index for good examples related to state fragility. For a somewhat comprehensive listing and visualization of international indices, see the International Peace Institute, *IPI Catalog of Indices* (2012).

References

Avra, B, Beieler, J, Fisher, B, Gustavo, L, Schrodt, P A, Song, W, Sowell, M, and **Stehle, S** 2013 Improving Forecasts of International Events of Interest. In: the 3rd Annual Meeting of the European Political Science Association, Barcelona, Spain in June 2013. Available at: http://eventdata.psu.edu/papers.dir/Arva.etal_EPSA_13.pdf [last accessed 15 October 2013].

Blattman, C and **Miguel, T** 2010 Civil War. *Journal of Economic Literature,* 48(1): 3-57. DOI: http://dx.doi.org/10.1257/jel.48.1.3.

Breiman, L and **Cutler, A** Random Forests. Available at http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm [Last accessed 12 September 2013].

Center for International Earth Science Information Network (CIESIN)/Columbia University, United Nations Food and Agriculture Programme (FAO), and **Centro Internacional de Agricultura Tropical (CIAT)** 2005 *Gridded Population of the World, Version 3 (GPWv3): Population Count Grid, Future Estimates.* Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Available at: http://sedac.ciesin.columbia.edu/data/collection/gpw-v3 [last accessed 11 July 2013].

Center for International Earth Science Information Network (CIESIN)/Columbia University, International Food Policy Research Institute (IFPRI), The World Bank, and **Centro Internacional de Agricultura Tropical (CIAT)** 2011 *Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Population Density Grid.* Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Available at: http://sedac.ciesin.columbia.edu/data/collection/grump-v1 [last accessed 11 July 2013].

Center for International Earth Science Information Network (CIESIN)/Columbia University 2005 *Poverty Mapping Project: Global Subnational Infant Mortality Rates.* Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC).

Available at: http://sedac.ciesin.columbia.edu/data/set/povmap-global-subnational-infant-mortality-rates [last accessed 11 July 2013].

Center for Hazards and Risk Research (CHRR)/Columbia University, Center for International Earth Science Information Network (CIESIN)/Columbia University, and International Bank for Reconstruction and **Development/The World Bank** 2005 *Global Multihazard Frequency and Distribution.* Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Available at: http://sedac.ciesin.columbia.edu/data/set/ndh-multihazard-frequency-distribution [last accessed 11 July 2013].

Eck, K 2013 In Data We Trust? A Comparison of UCDP GED and ACLED Conflict Events Datasets. *Cooperation and Conflict,* 47(1): 124-141. DOI: http://dx.doi.org/10.1177/0010836711434463.

Fearon, J D, and **Laitin, D D** 2003 Ethnicity, Insurgency, And Civil War. *American Political Science Review,* 97(1): 75-90. DOI: http://dx.doi.or/10.1017/S0003055403000534.

Gilmore, E, Gleditsch, N P, Lujala, P, and **Rød, K** 2005 Conflict Diamonds: A New Dataset. *Conflict Management and Peace Science,* 22(3): 257–292. DOI: http://dx.doi.org/10.1080/07388940500201003.

Gustavo, L, Schrodt, P A, Song, W, Sowell, M, and **Stehle, S** 2013 Improving Forecasts of International Events of Interest. In: the 3rd Annual Meeting of the European Political Science Association, Barcelona, Spain in June 2013. Available at http://eventdata.psu.edu/papers.dir/Arva.etal_EPSA_13.pdf [last accessed 15 October 2013].

International Peace Institute 2012 *IPI Catalog of Indices* 19 September 2012. Available at http://www.theglobalobservatory.org/indices.html [Last accessed 19 October 2013].

Kilcullen, D and **Courtney, A** Big Data, Small Wars, Local Insight: Designing for Development with Conflict-Affected Communities. McKinsey on Society. Available at http://voices.mckinseyonsociety.com/big-data-small-wars-local-insights-designing-for-development-with-conflict-affected-communities/ [last accessed 11 September 2013].

Päivi, L, Rød, J K, and **Thieme, N** 2007 Fighting over Oil: Introducing A New Dataset. *Conflict Management and Peace Science,* 24(3): 239-256. DOI: http://dx.doi.org/10.1080/07388940701468526**.**

Raleigh, C, Linke, A, Hegre, H, and **Karlsen, J** 2010 Introducing ACLED-Armed Conflict Location and Event Data. *Journal of Peace Research,* 47(5): 1-10. DOI: http://dx.doi.org/10.1177/0022343310378914.

Ramankutty, N, Evan, A T, Monfreda, C, and **Foley, J A** 2010 Global Agricultural Lands V.1, 2000. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Available at: http://sedac.ciesin.columbia.edu/data/collection/aglands/sets/browse [last accessed 11 July 2013].

Schrodt, P A 1999 Early Warning of Conflict in Southern Lebanon Using Hidden Markov Models. In: Starr, H *The Understanding and Management of Global Violence*. New York: St. Martin's Press. pp. 131-162.

Smith, C, Mashhadi, A, and **Capra, L** 2013 Ubiquitous Sensing for Mapping Poverty in Developing Countries. Submitted to: D4D NetMob 2013 Third Conference on the Analysis of Mobile Phone Datasets, Cambridge, MA on 1-3 May 2013. Available at: http://www0.cs.ucl.ac.uk/staff/l.capra/publications/d4d.pdf [last accessed 11 September 2013].

Soto, V, Frais-Martinez, V, Virseda, V, and **Frais-Martinez, E** 2011 Prediction of Socioeconomic Level Using Cell Phone Records. In: 19th International Conference, UMAP 2011, Girona, Spain on 11-15 July 2011, pp. 377-388. Available at: http://www.vanessafriasmartinez.org/uploads/umap2011.pdf [last accessed 11 September 2013].

Tikuisis, P, Carment, D, and **Samy, Y** 2013 Prediction of Intrastate Conflict Using State Structural Factors and Events Data. *Journal of Conflict Resolution*, 57(3): 410-444. DOI: http://dx.doi.org/10.1177/0022002712446129.

UN Security Council 1992 An Agenda for Peace: Preventive Diplomacy, Peacemaking and Peacekeeping: report of the Secretary-General pursuant to the statement adopted by the summit meeting of the Security Council on 31 January 1992. New York: United Nations (UN Doc. A/47/277 - S/24111).

UN Security Council 2011a Preventive diplomacy: Delivering results: report of the Secretary-General, 26 August 2011. New York: United Nations (UN Doc. S/2011/552).

UN Security Council 2011b Security Council Pledges Strengthened UN Effectiveness in Preventing Conflict, Including Through the Use of Early Warning, Preventative Deployment, Mediation. Department of Public Information, 22 September 2011. New York: United Nations (UN Doc. SC/10392).

Warden, P 2011 *Big Data Glossary*. California: O'Reilly Media. p. 31.

Weidmann, N B, Ketil Rød, K, and **Cederman, L E** 2010 Representing Ethnic Groups in Space: A New Dataset. *Journal of Peace Research,* 47(4): 491-499. DOI: http://dx.doi.org/10.1177/0022343310368352.

Yetman, G, Gaffin, S R, and **Xing, X** 2004 Global 15 x 15 Minute Grids of the Downscaled GDP Based on the SRES B2 Scenario, 1990 and 2025. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Available at: http://sedac.ciesin.columbia.edu/data/set/sdp-downscaled-gdp-grid-b2-1990-2025 [last accessed 11 July 2013].