TITLE
Chris Perry
18 July 2013

Abstract

1. Introduction

Within the domain of international affairs, the idea of aggregate indexes of national statistics is widely used for cross-country comparisons.[1] These indexes are often weighted combinations of a number of indicators and are compiled at regular intervals in order to chart national progress against specific benchmarks over time. While useful for advocacy and fundraising, their use is limited for strategic forecasting, policy guidance and design, and contingency planning. Further, though international indexes, especially those tailored to conflict and fragility data, are relatively good at showing nations under long-term stress, the underlying data is often highly lagged. Additionally, the high level of aggregation makes it difficult to gain any real insight into conflict dynamics, which are often a complex interplay of local and national trends.

Issues of granularity and timeliness become especially important in the context of conflict prevention and early warning mechanisms, which are much in vogue. For instance, within the United Nations peace and security architecture, the use of methodologies like machine learning[2] derived predictive analytics could have huge implications for strategic forecasting. However, machine learning algorithms are only as good as the quantity and quality of data that they train on.

In order to address the lack of highly granular and timely global statistics, this paper proposes using GIS data processing techniques to generate aggregations of national, subnational and satellite data at the district level. The resulting data is then used as a test case for the application of machine learning algorithms.

2. Problem Definition and Feature Space

For the purposes of this test case, the broad problem can be refined to the following. First, is it possible to use a combination of open source geospatial and national statistics to develop an index of vulnerability to violent events? This index should use prior events to inform the model and should be subnational in nature. Second, is there a way to scale this to incorporate additional feature selection as data becomes available? Relatedly, can the aggregation and modeling process be parallelized for the goal of scalability? A scalable, automated, updated solution would provide decision makers and NGOs alike a valuable resource with which to target interventions and programs.

The data for this test case is pulled from a number of data sources and types. This does a relatively good job of representing the diversity of data that would be necessary for a scalable global aggregated dataset.

3.1 Unit of Observation and Target Class

The dataset takes the district level-year as the unit of observation. Districts were derived using the National Administrative Boundaries GIS dataset provided by Columbia's Global Rural-Urban Mapping Project (GRUMP).[3] The dataset includes 399,747 non-overlapping polygons covering globe. However due to a combination of completeness of corollary data as well as limits of serial processing[4] capacity, the area of study was limited to continental Africa.

---

[1] 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 IPI Catalog of Indices.

[2] Machine learning is a branch of computer science that uses algorithms that perform actions without explicitly being programmed to. A major thread of machine learning research revolves around classification problems, for instance how to classify email as spam based on previously observed patterns in characteristics.

[3] 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).

[4] 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.

The problem dictates a target class that indicates instances of violence. Over the last few years, a number of geospatially tagged datasets of violence have been developed. For the purposes of testing, I chose the Armed Conflict Location and Event Dataset.[5] 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 is cut at 2012. Data was aggregated using QGIS vector join function, which yielded 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.

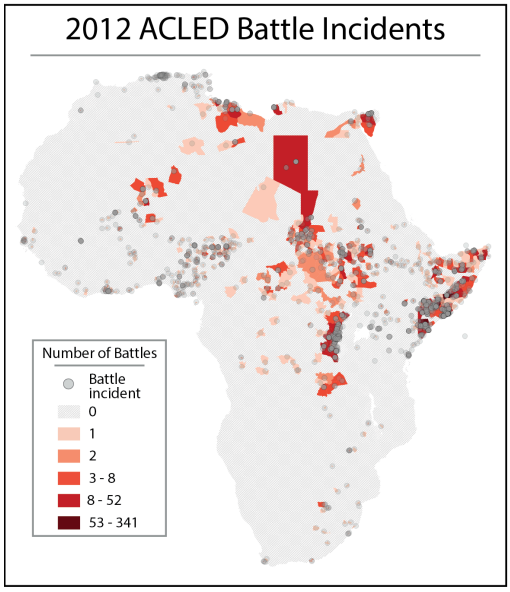


Figure 1: 2012 ACLED battle incidents

While there are a number of ways to derive a **target class, 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 tried, 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 useful in indicating previous levels of violence, which were used a predictor variables. The distributions of all four are provided below. The left side column is the distribution in the full data set and the right side are those after missing values were removed.

3.2 Predictors

Predictor features were processed from a number of levels of granularity, to the district level. These were conceptually inspired by previous econometric studies of the causes of civil war.

- *Ethnic fractionalization*: Supranational ethnic composition is taken from the *Geo-referencing of Ethnic Groups*[6] 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. The QGIS 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 13.[7]

- *Population:* The population feature was taken from Columbia's Gridded Population of the World (GPW)[8] project and uses both the past and future population grid counts. 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 the QGIS zonal statistics function, which derives continuous variables on the sum and mean of values contained in a given polygon.

---

[5] Raleigh, Clionadh, Andrew Linke, Håvard Hegre and Joakim Karlsen. 2010. Introducing ACLED-Armed Conflict Location and Event Data. Journal of Peace Research 47(5) 1-10.

[6] Weidmann, Nils B., Jan Ketil Rød and Lars-Erik Cederman (2010). "Representing Ethnic Groups in Space: A New Dataset". Journal of Peace Research, in press.

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

[8] 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).

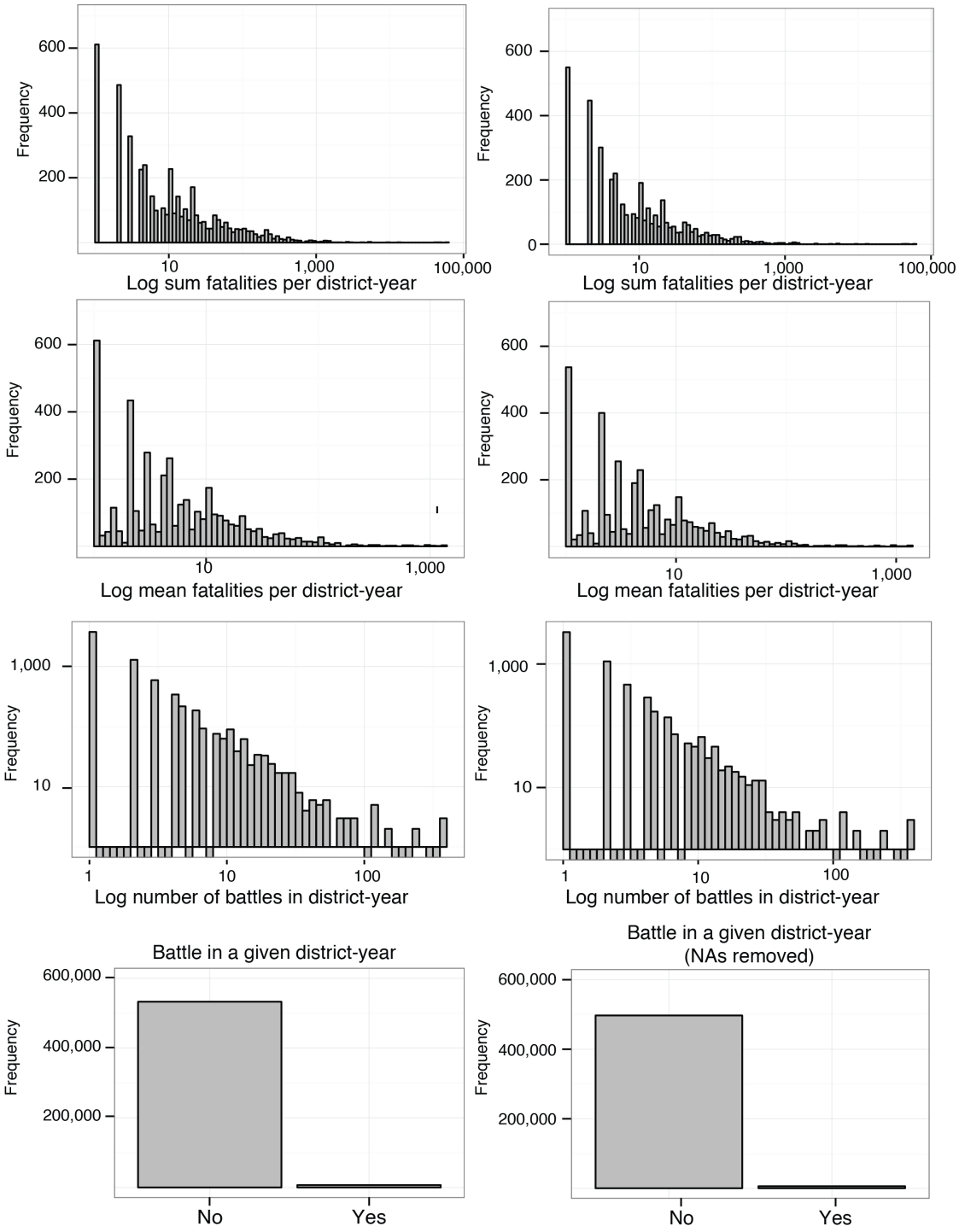


Figure 2: Frequencies of target classes before and after NA's were removed

- *Gross Domestic Product:* GDP data was taken from Columbia's *Global 15X15 Minute Grids of Downscaled GDP Based on the SRES B2 Scenario*.[9] The data consists of 823,680 raster grids indicating an absolute level of GDP at both 1990 and 2025. The data was processed by first taking zonal statistics on both the 1990 and 2025 GDP levels. These end points

[9] Yetman, G., S.R. Gaffin, and X. Xing. 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).

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.

- *Poverty:* Poverty data was comprised of both the *Global Subnational Prevalence of Child Malnutrition* and the *Global Subnational Infant Mortality Rates,*[10] 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. Additionally, the data included a raster file with total number of children under five years of age. This was included, as the data was released in 2000 and children in that age cohort would be between 13 and 18 today, making it a good indicator of current youth bulges. Data was processed using the QGIS zonal statistic function.

- *Hazard:* Hazard data was taken from both the flood and drought components of the *Global Multihazard Frequency and Distribution*[11] from Columbia. Specifically each indicator comes from a 29,652,480 cell raster indicating the frequency of a given hazard on a scale from 1-10, with a zero indicating no or virtually no frequency. Data was aggregated using the QGIS zonal statistics function.

- *Land Usage:* Land usage data was taken from the *Global Agricultural Lands*[12] data from Columbia. 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 the QGIS zonal statistics function.

- *Diamonds:* Diamond data is numeric variable that indicates the number of lootable diamond deposits in a given district. Data comes from the *Diamond Resources*[13] 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 the QGIS merge by location function.

- *Petroleum:* Petroleum data was taken from the PRIO *Petroleum Dataset v. 1.2*.[14] Data consists of shapfiles indicating all know petroleum deposits in the world. The data was processed using the QGIS intersection function to create a binary variable indicating simply whether or not the district contains a petroleum deposit.

- *Governance Indicators:* Finally, governance indicators were taken from the *Database of Political Institutions* (DPI) created by the Development Research Group at the World

---

[10] *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).*

[11] 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).

[12] Ramankutty, N., A.T. Evan, C. Monfreda, and J.A. Foley. 2010. Global Agricultural Lands: Pastures, 2000. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC)

[13] Gilmore, Elisabeth; Nils Petter Gleditsch, Päivi Lujala & Jan Ketil Rød, 2005. 'Conflict Diamonds: A New Dataset', *Conflict Management and Peace Science* 22(3): 257–292

[14] Lujala, Päivi; Jan Ketil Rød & Nadia Thieme, 2007. 'Fighting over Oil: Introducing A New Dataset',*Conflict Management and Peace Science* 24(3), 239-256.

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 feature space includes:

- Target classes
- 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, this feature space serves the useful purpose of testing the feasibility of aggregating these disparate data sources as well as whether machine learning can in fact be applied to the 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 less than the full set of district-years. 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.

3. Analysis

Because of the size of the dataset and the limits imposed by serial as opposed to parallel processing, options were relatively limited. Because of the availability of existing target class data as well as the discrete nature of the target class, this is a supervised learning classification problem.[15] Two types of machine learning algorithms were attempted. First, naïve Bayes[16] provides a relatively simple algorithm that in many cases gives decent performance in terms of both speed and accuracy. Second, the random forest algorithm provides a more sophisticated learner that has gained popularity for performance gains as well as ability to handle large feature spaces. In the case of the random forest learner, data was separated into three partitions. First, 2012 data was split as a final homogenous test case. Data from 2000-2011 was further split into test and training sets randomly with 70 percent allocated for training and 30 percent to testing.

[15] Supervised learning refers to a machine learning task of inferring a function from labeled training data. Classification refers to the problem of identifying which set of categories a new observation belongs based on a training set of observations with known classifications.

[16] The naïve Bayes algorithm is a simple probabilistic classifier. 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.

		Actual	
		No Battles	Battles
Predicted	No Battles	135,499	1,240.25
	Battles	3,657.25	404.5

Table 1: Performance confusion matrix of the averaged binary naïve Bayes classifier (for simplicity only binary model is reported).

		Actual	
		No Battles	Battles
Predicted	No Battles	139,003	195
	Battles	1,328	275

Table 2: Performance confusion matrix of the binary random forest classifier.

The naïve Bayes classifier performed extremely poorly. The algorithm was performed with 4-fold cross validation 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.[17] In both specifications, the algorithm tended to over predict. This over prediction tends to match previous research in predictive and econometric modeling of conflict. The results for the binary classifier are reported in table 1 to the right.

A random forest was used for the second learner. Random forests are a form of ensemble[18] learner that extends the decision tree learner algorithm. 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.

For the purposes of this paper, the randomForest package in R to run a 100-tree forests on the data under two variations: a binary target class indicting existence of battles in a given year and a continuous target class indicting number of battles in a given year. Both versions offered limited improved performance to the naïve Bayes classifier. For the binary target class, the out-of-box error rate 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 from 99% to 99.8% accuracy. In the cases where the model predicted battles, there was a more significant increase in accuracy from 10% to 17.2%. Comparing tables 1 and 2 above demonstrates this. 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.

The performance of both of these models is reported in figure 3 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 till 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.

[17] 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.

[18] Ensemble methods refer to the use of multiple models that are then recombined to increase the predictive performance over any of the single models.

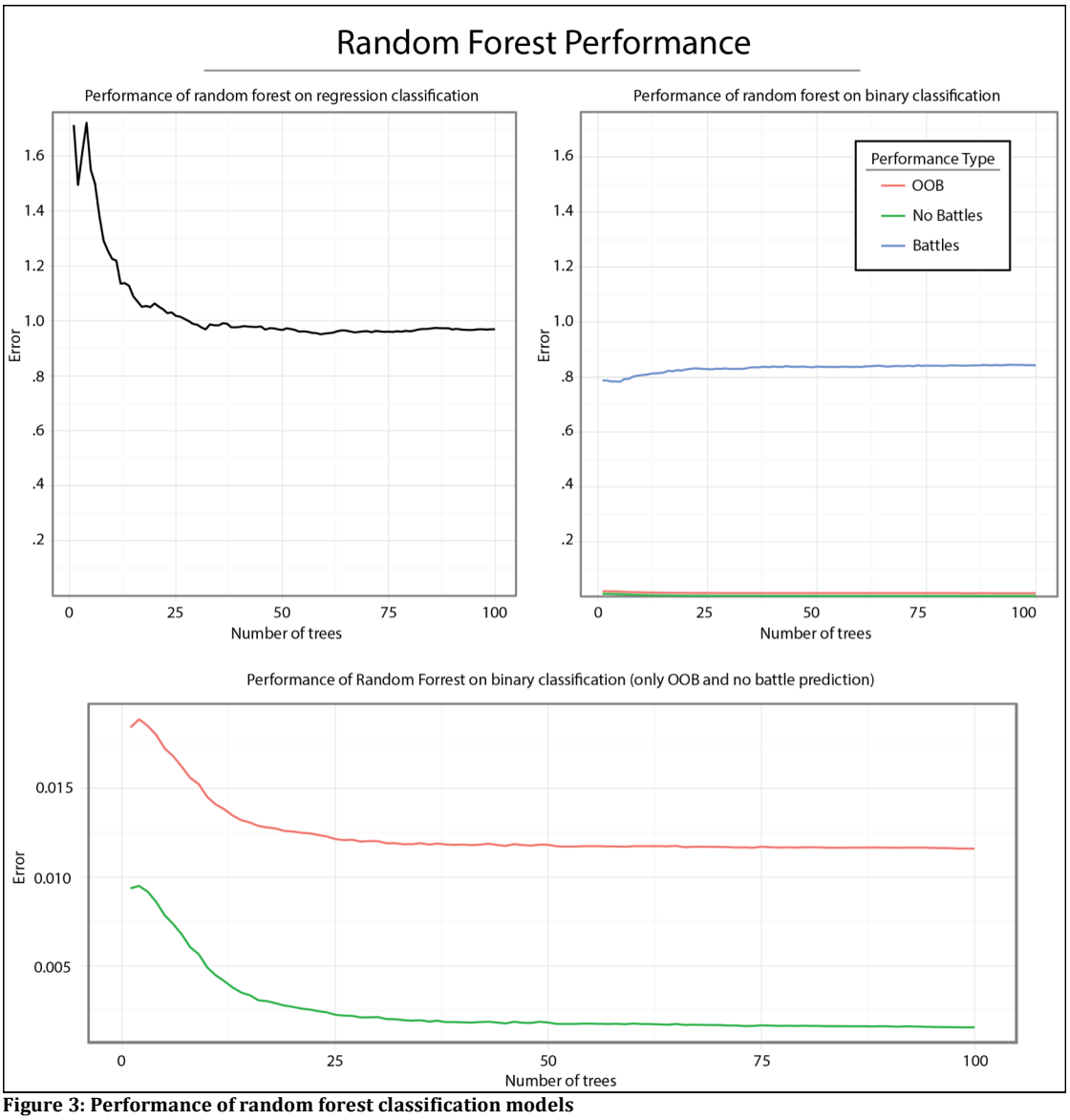


Figure 3: Performance of random forest classification models

An 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 4 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. 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 autocorrelated, future analysis should 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 gave some payoff in terms of increased accuracy. With some processing in future research, hopefully these variables can be further refined to offer more in the way of increased accuracy.

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 against actual instances to give a visual representation of the accuracy of the models. These maps are presented in figure 4 below with the regression forest on the top left, actuals on the top right, and a composite of the

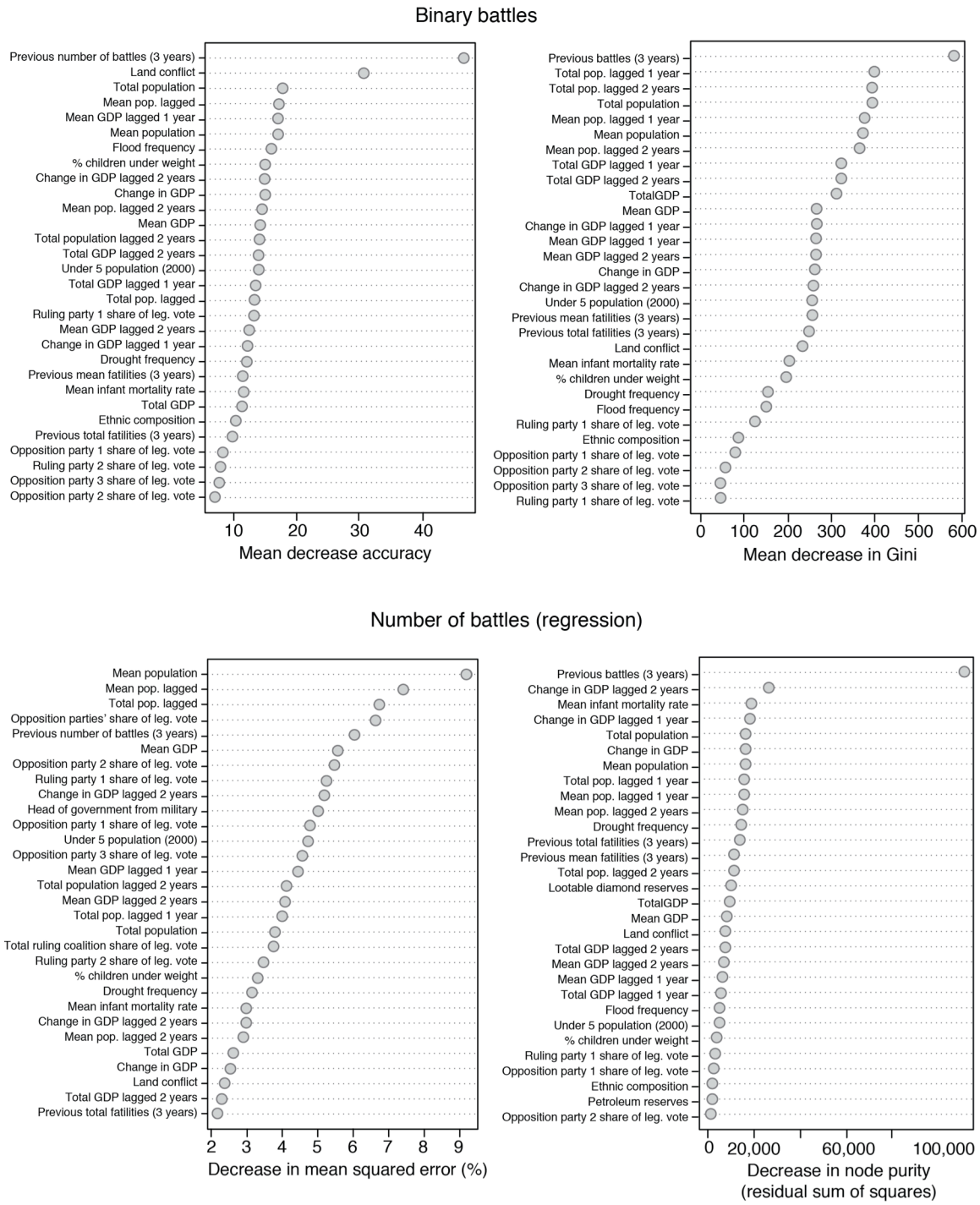
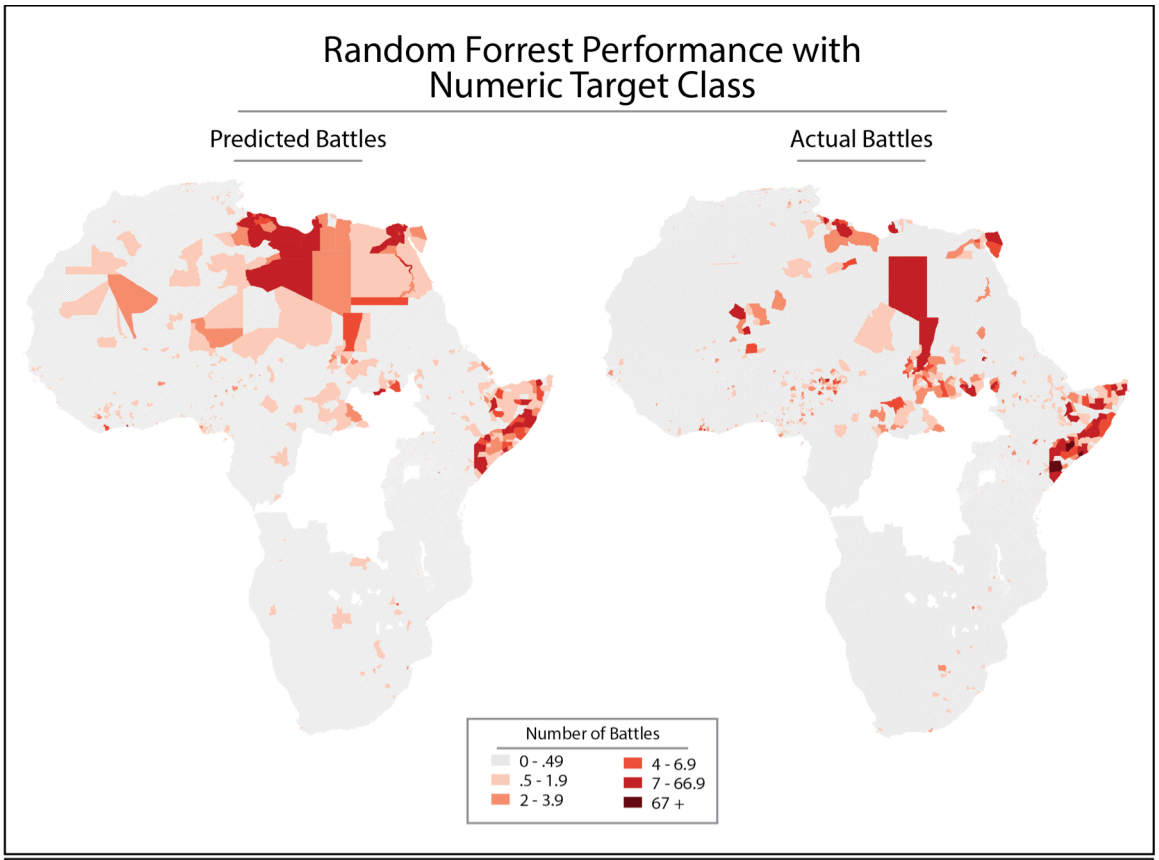


Figure 4: Variable importance for both learners measured in the decrease in accuracy (left) and decrease in overall node purity when variable is removed.

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.



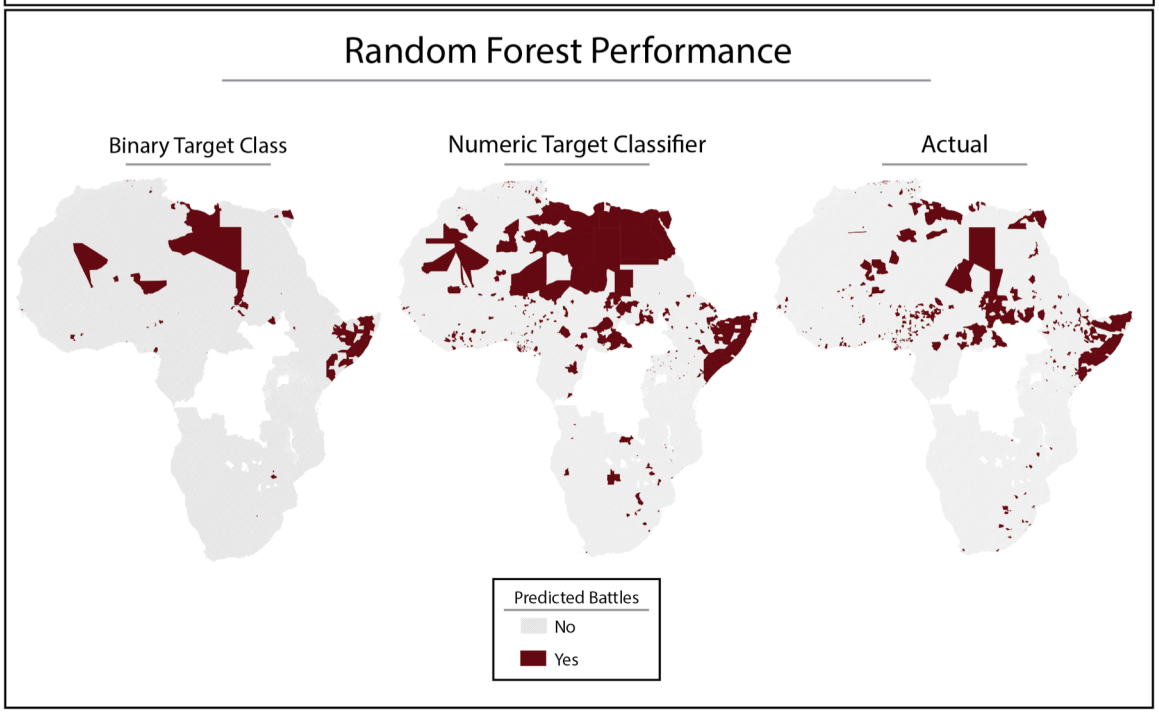


Figure 5: Map of predicted and actual classes for 2012

4. Steps Forward

This paper represents a first baby step towards modeling fragility and vulnerability to conflict at the global level. There are a number of lessons learned that should be incorporated into any future research on the subject.

- First, this there is the case to be made that future research should try to expand the scope globally. In order to do this, the processing of data must be moved towards a parallelized model in the interest of feasibility. Many of the zonal statistics took at least a day to process. Some possible data remained unused because after two to three days the processing had not completed. Fortunately there is a new geoprocessing library, Geotrellis, written in a Java-based language, Scala, which is meant specifically for this task. Future processing can be done Geotrellis on a cluster of machines running in parallel on Amazon's EC2 service for a reasonable cost. Additionally, the machine learning tasks could also be parallelized as all of the learners 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.
- An alternate tract of research should focus on narrowing the scope of analysis to specific conflict regions. For instance, examining the Great Lakes region as a whole could prove useful in gaining tactical and strategic understanding of the current and future conflict environment. A narrower scope would allow for more targeted variable selection as well as better understanding of the specific drivers of conflict in a given context.
- Regardless of the scope of research, future research should address autocorrelation of features by creating a time deteriorating index of temporal data. This project tries to do that with previous conflicts 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 match some of the literature on conflict recidivism indicating that the risk period for resurgence of conflict after the end of hostilities is approximately one decade.
- Finally, any future research must find new and better data. One principle concern for variable selection in this project was expediency. The population data for instance is not ideal as it is really one concrete snapshot and one projection bookending some rudimentary math to deal with the missing data. Likewise, the GDP data. There is a number of new projects aimed at taking yearly snapshots of GDP and population based on satellite imagery that should 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. The target class is also imperfect. 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.
- Finally, 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 brand 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.