IMPROVING AND EXPANDING CONFLICT FORECASTING

A Dissertation in
Political Science
by
Benjamin Fisher

© 2016 Benjamin Fisher

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2016
The dissertation of Benjamin Fisher was reviewed and approved* by the following:

Luke Keele  
Associate Professor of Political Science  
Dissertation Advisor, Chair of Committee

Glenn Palmer  
Professor of Political Science

Christopher Fariss  
Jeffrey L. Hyde and Sharon D. Hyde and Political Science Board of Visitors  
Early Career Professor in Political Science

Guido Cervone  
Associate Professor of Geography

James Piazza  
Associate Professor of Political Science  
Director of Graduate Studies of the Department of Political Science

*Signatures are on file in the Graduate School.
Abstract

Forecasting conflict, for many years the domain of qualitative analysts, pundits, and amateur Nostradamuses, has become increasingly practical thanks to the application of quantitative models. Government groups view these models as key to developing early warning systems for conflict and related violence. As a result, much of the early work in this area was spearheaded government-funded academic research and has only recently begun to spread more generally, even into the private sector. Previous and current models tend to focus on generating probabilistic forecasts at a country-year level. This approach works for some problems, but not everything fits into a country-level format. With a focus on violence against civilians as the outcome of interest, I explore how to improve existing early warning models and develop models to generate forecasts at a subnational and subannual level.
# Table of Contents

List of Figures .................................................. vi

List of Tables .................................................. viii

Acknowledgments ................................................... ix

Chapter 1
Introduction ...................................................... 1

Chapter 2
Improving Country Level Forecasts of Mass Atrocities .......... 6
  2.1 Introduction ................................................. 6
  2.2 A Methods Approach ........................................ 8
    2.2.1 Random Forests ......................................... 9
    2.2.2 Neural Networks ........................................ 10
  2.3 Feature Engineering ......................................... 11
  2.4 Case-Control Sampling ....................................... 12
  2.5 Model Validation ............................................ 14
  2.6 Data .................................................... 16
  2.7 Analysis .................................................. 21
    2.7.1 Model Selection ......................................... 23
    2.7.2 Feature Engineering .................................... 26
    2.7.3 Case Control ........................................... 28
    2.7.4 Inspecting Results ...................................... 30
  2.8 Conclusions ............................................... 31

Chapter 3
Forecasting Atrocities at a Subnational Level ................. 33
  3.1 Introduction ............................................... 33
  3.2 Research Design ........................................... 34
List of Figures

2.1 ROC Curve ................................................................. 16
2.2 ROC Curve with AUC ...................................................... 17
2.3 Onsets by Half Year ...................................................... 18
2.4 Box-and-Whisker Plot of Predicted Probabilities by Model ............ 24
2.5 ROC Curves for Specific Counts Feature Set .......................... 25
2.6 ROC Curves for Feature Sets (raw counts) ............................ 26
2.7 ROC Curves for Feature Sets (differenced counts) ...................... 27
2.8 ROC Curves for Feature Sets (binned counts) ........................ 28
2.9 ROC Curves for Full Sample and Case Control Sample ............... 30
2.10 Top 10 Countries by Risk for June-December 2013 .................... 31

3.1 Africa Grid ................................................................. 36
3.2 ROC Curve ................................................................. 39
3.3 ROC Curve with AUC ...................................................... 39
3.4 Mass Atrocities in Africa 2003-2014 .................................... 42
3.5 Africa 2010-1 Kernel Density Conflict Map ........................... 48
3.6 ROC Curves ................................................................. 50
3.7 Predicted Probability Boxplots .......................................... 52
3.8 Predicted Probabilities for 2014-2 .................................... 53
3.9 Predicted Probabilities for 2014-2 with Events ......................... 54
3.10 Permutation Importance ................................................ 55
3.11 Partial Dependence Plots ............................................... 56

4.1 CRPS example ............................................................ 68
4.2 IS, Rebel, Civilian Material Events ..................................... 72
4.3 IS, Syria, Civilian Material Events .................................... 73
4.4 Syria, Rebel, Civilian Material Events ................................ 74
4.5 RMSE for hyperparameter combinations ............................... 77
4.6 IS, Syria, Civilian Material Forecasts ................................ 78
4.7 Syria, Rebel, Civilian Material Forecasts ............................. 81
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.8</td>
<td>IS, Rebel, Civilian Material Forecasts</td>
<td>82</td>
</tr>
<tr>
<td>.1</td>
<td>Distribution of Drought Data</td>
<td>90</td>
</tr>
<tr>
<td>.2</td>
<td>Excluded Ethnic Groups</td>
<td>91</td>
</tr>
<tr>
<td>.3</td>
<td>MODIS Land Cover Classifications</td>
<td>92</td>
</tr>
<tr>
<td>.4</td>
<td>Mountainous Terrain</td>
<td>93</td>
</tr>
<tr>
<td>.5</td>
<td>Nighttime Light Emissions</td>
<td>94</td>
</tr>
<tr>
<td>.6</td>
<td>Population</td>
<td>95</td>
</tr>
<tr>
<td>.7</td>
<td>Locations of Precious Resources</td>
<td>96</td>
</tr>
</tbody>
</table>
# List of Tables

2.1 Hyperparameter Grid ........................................... 23  
2.2 AUC Scores ....................................................... 25  
2.3 Case Control Regional Distribution .............................. 29  
3.1 Predictions using .0015 threshold .............................. 51  
3.2 Predictions using .014 threshold .............................. 51  
4.1 Hyperparameters of Sims-Zha Reference Prior .................. 64  
4.2 Actor Dyads and Series .......................................... 72  
4.3 Hyperparameter values tested ................................... 76  
4.4 Final Hyperparameter Values by Model .......................... 76  
4.5 IS-Syria Accuracy Metrics ....................................... 79  
4.6 Syria-Rebel Accuracy Metrics .................................... 80  
4.7 IS-Rebel Accuracy Metrics ....................................... 80
Acknowledgments

I would first like to thank Doug Gibler and Phil Schrodt. I would not be where I am today if it were not for their generous help and guidance. I would also like to thank Luke Keele for helping me to shape this dissertation into something presentable, as well as the other members of my committee: Glenn Palmer, Chris Fariss, and Guido Cervone.
Introduction

The past 20 years have seen an increased focus on approaching political science problems from a predictive perspective. Some cases view prediction as an alternative means of model validation and inference (Ward, Greenhill & Bakke 2010). Others apply it in a purely predictive sense to create systems to forecast political outcomes of interest (Goldstone et al. 2010, O’Brien 2010). The range of topics includes everything from US presidential elections, as epitomized by Nate Silver during the 2008 and 2012 elections, to irregular regime change (Beger, Dorff & Ward 2014). The potential of predictive models in political science has drawn interest from all corners. In academia, people like Mike Ward and Phil Schrodt have spearheaded efforts to incorporate predictive models into academic studies. Government groups like IARPA have funded efforts to assess the effectiveness of crowd-sourced predictions. Recently, private sector companies like Predata\(^1\) have emerged marketing tools that predict civil unrest and military conflict for use by businesses with international supply chains.

\(^1\)http://www.predata.com/

The work of the Political Instability Task Force (PITF) is an early example of predictive modeling being applied to political science problems. Originally created in the mid-1990s to assess factors that cause political instability, the focus of the
group soon moved to predicting instability. The first result of this work was a
global model for forecasting political instability, which in this case meant civil war,
ethnic war, adverse regime change, or genocide/politicide (Goldstone et al. 2010).
This model uses a country-year research design, with new predicted probabilities
generated each year for each country. Subsequent PITF models focusing on other
outcomes of interest take the same approach (Fisher & Schrodt 2014, Hazlett 2014).
Academic work looking at forecasting conflict has tended to adhere to the same
country-year approach (Beger, Dorff & Ward 2014).

There are a few reasons that PITF and similar groups have emphasized the
country-year approach. The first is that this was and continues to be how IR re-
searchers typically model international phenomena, especially conflict. This isn’t
to say that these are the only type of conflict forecasting models that have been
developed. However, these remain the types of models actively employed by gov-
ernment groups.\(^2\) Another reason for their ubiquity is ease of maintenance. Once
a model specification has been developed and a statistical method chosen, gener-
ating new forecasts is a routine matter of feeding new data into a trained model
in R or some other statistical software. The only challenge for the end user is
maintaining a consistent source of data. Any sort of model that requires continual
tuning each year is likely to be a nonstarter for many, even those with resident
‘data scientists’. The main reason for the continued reliance on country-year mod-
els is that results are relatively easy to communicate to an audience that has little
to no background in statistics. Results can be communicated as predicted proba-
bilities, which, thanks to daily weather forecasts, the average American should be
able to understand. Alternatively, one could bin countries based on the predicted

\(^2\)The two major government projects using conflict forecasting models are PITF and ICEWS.
Both model their outcomes of interest as binary outcomes. ICEWS has largely remained behind
closed doors since its original publication. It is possible that the group has shifted away from
these models, but the recent public release of ICEWS data suggests that the group continues to
rely on country-year models with a binary dependent variable.
probability of the outcome of interest occurring (e.g. high risk vs. low risk).

The global model is missing two components that could greatly improve the utility of forecasts: temporal and geographic granularity. Much can change over the course of a year, and a year-level model is inherently unable to capture this. Moreover, the yearly structural variables, such as regime type, are largely static, resulting in little change in year to year predicted probabilities. This creates a major problem with false positives in these models. Another issue with this model is the geographic level of aggregation. Using a country-level approach to forecasting substate phenomena is problematic as it misses potentially useful information. For example, someone interested in forecasting attacks on civilians would want to know where within a country violence is likely to occur and who is likely to be the source. These two pieces of information could greatly influence what policy tools governments use to mitigate violence or how aid organizations distribute resources in preparation for dealing with fallout.

This dissertation seeks to improve upon current practices by exploring how to improve current models, as well as alternatives to the country-year approach. Specifically, I focus on developing models that generate subnational forecasts and subannual forecasts. The outcome of interest I seek to predict is the occurrence of mass violence against civilians, or so-called mass atrocities, which have been a focus of recent PITF modeling efforts (Fisher & Schrodt 2014, Hazlett 2014). A mass atrocity in a general sense is defined as an attack on noncombatants by government, rebel, or terrorist groups that results in at least 5 fatalities. I break the approach down in three sections. In the first, I focus on developing a global model to forecast which states are at risk for mass violence against civilians. I will show that a focus on feature engineering yields better results than sampling approaches or new statistical methods. In the second section, I develop a geographic model
by disaggregating the data into a raster grid in order to forecasts which areas of a country are likely to experience mass violence against civilians. Finally, I test the utility of multivariate time series modeling for forecasting short-term civil conflicts, with the Syrian Civil War serving as the subject of analysis.

The first article deals with the standard country level model. Here, I explore three ways of improving the existing approach. The first involves assessing the potential improvement of using newer methods, such as Random Forest and neural networks, in place of the commonly used logistic regression. Research in machine learning indicates that these methods should yield a significant improvement in model accuracy (Caruana & Niculescu-Mizil 2006). The second has to do with correcting imbalances in the data using sampling procedures. I apply case-control sampling, which generates a balanced sample by matching positive observations (i.e. countries that experience a mass atrocity) to negative observations (i.e. countries that do not experience a mass atrocity) based on geographic region. Finally, I assess the extent to which feature engineering, or how data is transformed and incorporated into the model, affects predictive accuracy.

The second article applies a raster format to the data. Appropriate for data with a spatial component, a raster uses a lattice structure to break the observations into grid-squares. This is a common format for geographic data. In this case, I break the continent of Africa into a grid of approximately 50x50 km squares based on latitude and longitude. I then forecast the occurrence of mass atrocity events at a local level using a combination of geographic and political data. The benefits of this approach are the ability to generate forecasts at a subnational level, as well as incorporate spatial data that would typically not be used at national level of analysis.
The final article looks at forecasting multi-actor civil conflicts. I use several time series models to generate forecasts for the Syrian Civil War. The advantage of a time series based approach is the ability to generate subannual forecasts for a heterogeneous set of actors. The main purpose of the article is to assess whether a multivariate Bayesian Vector Autoregression (BVAR) model is more appropriate than simply applying univariate models to individual series. A BVAR is theoretically more appropriate, as the multivariate component accounts for how the actions of one actor affect the actions of another, while the Bayesian component allows me to incorporate prior beliefs about how the actors should behave. The focus on the Syrian conflict provides an additional assessment of how a multivariate model, which is typically prone to overparameterization, performs on a multi-actor civil conflict with a relatively short history.
Improving Country Level Forecasts of Mass Atrocities

2.1 Introduction

In 1994, the Hutu ethnic group in Rwanda, which made up nearly three quarters of the population, launched a genocidal campaign to eliminate the country’s Tutsi minority. Despite the presence of a UN-peacekeeping force, the 100-day campaign resulted in the deaths of over 800,000 Tutsis and moderate Hutus and was only brought to an end with the defeat of the Rwandan government by the Rwandan Patriotic Front, a Tutsi-dominated rebel group. The failure of the international community to anticipate such large scale violence, as well as the failure to end the genocide once it began, put a new focus on developing effective ways of forecasting mass violence.

In response to the Rwandan Genocide and other instances of state failure, the US government formed the Political Instability Task Force (PITF) to better understand the causes of instability events like genocide, coups, and civil wars and develop methods to predict which countries are at risk. The resulting "global model" developed by PITF, which is still in use today, relies on a logistic regres-
sion model to produce forecasts by state of the probability of experiencing a state failure event during the next calendar year (Goldstone et al. 2010). Efforts to model specific types of instability, such as civil war, have followed this same template: a global, country level model with a binary dependent variable (Blair & Sambanis 2015, Chiba & Gleditsch 2015). Similar government-sponsored conflict forecasting programs have used the same approach (O’Brien 2010).

Despite this approach being well-established, there has been little work done on looking at ways of improving model accuracy beyond alternative model specifications. This ignores several potential means of improving the predictive accuracy of models. I seek to remedy this by drawing on ideas from machine learning, a field primarily concerned with developing methods and techniques for predictive analytics. Entire textbooks have been written on best practices for predictive analytics encompassing a range of options (Kuhn & Johnson 2013). Some of these, such as adjusted sampling procedures (Goldstone et al. 2010) and alternative methods (Ulfelder 2012) have been tried, but a systematic assessment of multiple techniques is lacking.

This paper examines three potential avenues of improvement for a global model: alternative methods, feature engineering, and alternative sampling methods. The first option simply applies two well-developed machine learning methods, Random Forests and neural networks, which may be better suited for prediction problems in this field than the typically used logistic regression. The second deals with generating new features, or variables, from existing data. This may involve combining existing variables in some way or creating entirely new ones. I apply this specifically to the creation of variables from event data. Finally, I examine how case control sampling, a resampling procedure used by PITF, compares to using a full sample to create the models in terms of predictive accuracy.
The primary goal of this paper is to explore methods for improving the predictions of a global forecasting model. A secondary goal is to build an effective forecasting model for a substantive problem that has not been extensively covered. Rather than re-explore the general political instability outcome that PITF has typically focused on, I consider a more focused outcome of interest, mass atrocities, which has been one of several more specific topics that PITF has looked at recently (Fisher & Schrodt 2014, Blair & Sambanis 2015, Chiba & Gleditsch 2015). Mass atrocities are instances of mass violence against noncombatants by armed actors, state or non-state, that result in a large number of fatalities. There are a variety of government and non-government groups that would be interested in an effective forecasting model on this topic. As such, I give attention to developing a model that uses data sources that are updated regularly. A forecasting model is of little use if it cannot be used for regular, up-to-date forecasts.

2.2 A Methods Approach

Logistic regression has become the standard for predicting classification problems in political science, and early forecasting work in the field relied on it almost exclusively. The natural issue to begin with then is to address whether gains can be made from using a method other than logistic regression. While performing well in general, logistic regression may not be ideal for the noisy, often nonlinear data dealt with in political science. Fortunately, there are many potential alternatives from the machine learning field, which focuses on the development of better predictive models. Previous work suggests that many of these models outperform logistic regression on predictive tasks (Caruana & Niculescu-Mizil 2006, Ulfelder 2012).
For this study, I compare Random Forest and artificial neural networks, two commonly used, extensively validated machine learning methods to logistic regression using the atrocities data. Ideally, one or both of these methods will be a clear "winner" in comparison to logistic regression in terms of predictive accuracy. Alternatively, it may be that an ensemble approach, where predicted probabilities are averaged across models, is more appropriate. It may be that each method performs well on a certain set of cases, but not others. Averaging the probabilities across models can account for this potential issue. It is often the case that an ensemble of methods is the best approach to take (Opitz & Maclin 1999).

2.2.1 Random Forests

Originally developed in Breiman (2001), the Random Forest algorithm uses multiple decision trees to classify observations. Each tree is constructed by taking a bootstrapped sample of the training data and selects from a random subset of the predictor variables at each split. When a tree is grown, a percentage of the bootstrapped sample is left out. This "out-of-bag" (OOB) data is then run through the tree. This gives an error rate for the tree based on how well it classified the OOB data. This process is repeated for each tree in the forest. Once the forest is constructed, observations are classified by the mode output of trees. The percentage of the trees that vote for a particular outcome can be used as a predicted probability for that observation.

Random Forests have become a standard by which newer methods in the machine learning field and have been used with some success in prediction competitions.\(^1\) The modular nature of the algorithm has resulted in a plethora of variations on Breiman’s original, such as conditional forests (Hothorn, Hornik &

\(^1\)https://www.kaggle.com/wiki/RandomForests
Zeileis 2006), gradient boosted trees (Friedman 2001), and oblique Random Forests (Menze et al. 2011). Each of these corrects for some issue with the original algorithm, and each, of course, claims to deliver superior predictive performance to a standard Random Forest. Because the purpose of this analysis is to determine the utility of common machine learning methods for political predictive problems, rather than adjudicate between machine learning models, I use Breiman’s original algorithm as implemented in the randomForest R package (Liaw & Wiener 2002).

2.2.2 Neural Networks

Artificial neural networks, named for their structural similarity to a biological nervous system, are a machine learning method widely used for prediction tasks. A neural network is composed of multiple layers of ‘neurons’, or units: a layer of input units, which are the independent variables, a layer of output units, which gives predictions, and one or more hidden layers, which are made up of multiple units or ‘neurons’ that process the data. Each unit in a layer is connected to each unit in the next layer by a series of connections, which are represented by a numerical weight. A higher weight indicates a higher degree of influence by one unit on another. If the weight is above a certain threshold, then it ‘activates’ the next unit. The network outputs predictions based on how data on an observation activates units across the hidden layers.

Like Random Forests, neural networks have become a standard in the machine learning field. They have been applied to a diverse range of problems, from image recognition (Lawrence et al. 1997) to rainfall-runoff modeling (Hsu, Gupta & Sorooshian 1995), with a high degree of success. Also like Random Forests, there are a diverse range of implementations that have sprung up in recent years, such as deep neural networks (Schmidhuber 2015), convolutional neural networks
(Krizhevsky, Sutskever & Hinton 2012), and recurrent neural networks (Mandic & Chambers 2001). I again rely on a vanilla implementation, in this case the nnet R package (Venables & Ripley 2002), rather than sift through the numerous variations.

2.3 Feature Engineering

Feature engineering is an informal term that refers to the practice of manually transforming the input data in order to improve the accuracy of a model. A common example of this in political science is taking the natural logarithm of a highly skewed variable, such as a country’s population. Other possibilities include simple transformations like creating a dummy variable from continuous data, or more complex processes like centering and scaling data. Feature engineering plays a major role in predictive modeling. Predictive modeling textbooks devote whole sections to feature engineering (Kuhn & Johnson 2013), and it has proven key to winning predictive modeling competitions like the Knowledge Discovery and Data Mining (KDD) Cup or the website Kaggle. For example, winning teams in the KDD Cup focus heavily on feature engineering instead of tailoring algorithms to the problem (Yu et al. 2010, Wu et al. 2012).

For the purposes of this analysis, the focus will be on how to incorporate event data into the final model. Event data in its raw form is essentially unusable for analysis. Each observation is a record of an event in a who-did-what-to-whom format. Standard practice has been to aggregate events into what are known as quad counts. These refer to the four major types of interactions: verbal cooperation, material cooperation, verbal conflict, and material conflict. Counts are usually recorded for dyadic combinations of actors during a given time period. These can
be between states, for example, or actors within a state, as in this analysis.

I plan to build on the quad count approach by looking at other potential ways of generating features from event data that avoid the pitfalls of the quad count approach. By restricting the data to a few large categories of events, one is forced to give equal weight to very different types of events. For example, the detainment of a suspected member of a rebel group and a military engagement between government forces and rebels receive the same weight in the analysis, despite one being a much more “severe” event than the other. Recent work by Blair & Sambanis (2015) showed that using more granular event counts, in this case of detainments, in place of quad counts improved their ability to accurately forecast the onset of civil wars. My goal is to determine whether focusing in on more specific event counts can improve the accuracy of a mass atrocities model.

2.4 Case-Control Sampling

Like many outcomes of interest in international relations, mass atrocities are a very rare event. The dataset used in this paper has only 128 onsets out of 4711 observations. A lack of positive observations is problematic when developing a predictive model for a couple of reasons. First, model performance becomes highly dependent on the positive observations included in the training set. As fewer positive observations are used to train a model, the weight of potential outliers increases, leading to overfitting. We want the model to fit to the general characteristics of the outcome of interest, rather than the idiosyncrasies of a handful of cases. More generally, it is hard to predict an event that is both rare by nature and occurs in the context of noisy data.
A basic solution to this issue of unbalanced data is to generate a new, balanced dataset using some sort of sampling procedure. Kuhn & Johnson (2013) describe two general methods of sampling imbalanced classes: up-sampling and down-sampling. Up-sampling involves drawing observations from the minority class until there are roughly the same percentage of observations as the majority class. Down-sampling, on the other hand, keeps the original minority observations and samples the majority class, either with or without replacement, until the majority class is the same size as the minority. A model is then trained on the balanced sample, and tested on imbalanced out-of-sample data.

Previous PITF work has used a form of down-sampling known as case-control (Goldstone et al. 2010). Rather than randomly sample from the majority class, case-control matches one or more majority observations to minority observations that are similar in some way. The intuition behind this is to better distinguish why some observations experience an outcome when other, similar observations do not. The PITF global instability model does this by including 3 peace observations for each observation where instability occurs. The three observations are drawn from the same year and geographic region as the positive observation with which they are associated. Additionally, the peace countries used must not have experienced any instability episodes in the two years before or the two years after the year in question.

Goldstone et al. (2010) do not actually use the case-control method for improving predictive performance. Instead, they cite this sampling procedure as a means of obtaining more accurate model inference and do not present any predictive accuracy comparisons between a case-control model and a full sample model. I plan to assess whether this case-control approach adds any accuracy to this type of rare event international relations model by comparing the forecasting accuracy
of a case-control model and a full sample model using the same sampling rules as Goldstone et al. (2010).

### 2.5 Model Validation

A key component of any predictive modeling exercise is assessing the accuracy of the final model. The danger that every predictive model faces is overfitting. A model suffers from overfitting when it predicts the data used to create it with a high degree of accuracy, but performs poorly when used to predict cases that it has not seen. The typical way to avoid incorrectly choosing an overfit model over a better model involves dividing the data into two samples: one as data used to train the model and the other used as testing data that the trained model attempts to predict. The point here is to assess how well the model performs on observations that it has never seen.

After dividing the data into training and testing samples, one must determine what metric by which to compare model accuracy. The obvious choice would be overall accuracy, or the percentage of observations correctly classified. There are a couple of issues that make this metric unsuitable in this case. First, one has to choose a probability threshold at which to divide positive observations from negative observations. There is no easy way to set this threshold in cases where the outcome classes are highly imbalanced. Determining a threshold becomes more problematic when comparing across methods, as some methods may give higher or lower predicted probabilities on average than others. Additionally, it is unclear what qualifies as "good" accuracy in rare events prediction. I could get a 97% accuracy on predicting mass atrocities by naively predicting that there will be no onsets. Since I am more interested in predicting onsets than peace periods, I might...
actually prefer a model with lower accuracy if it predicts more onsets correctly.

Rather than randomly assigning the data to training and testing samples, I divide the sample temporally. The training sample is data from 1996-2009, and the testing sample is data from 2010-2013. This is a form of out-of-sample validation known as out-of-time validation. This form of validation is appropriate when there is a temporal component to the data, and one believes that there is a potential underlying change over time. More intuitively, if we are interested in how our model will do when predicting an event in \( t \), it would be better to test it on \( t - 1 \) rather than \( t - 10 \). We can reasonably expect that \( t \) is much more similar to \( t - 1 \) than to \( t - 10 \). This way, we avoid fitting to some underlying temporal trend that may be present in older data, but has changed for more recent observations.

I rely on the Receiver Operating Characteristic (ROC) curve to compare model accuracy. ROC curves, originally developed to assess the accuracy of air radar (Green & Swets 1966), are a widely used metric in predictive modeling. 'Tournament' papers, which compare the performance of man models, typically use this metric (Caruana & Niculescu-Mizil 2006). Figure 3.2 displays an example of an ROC curve. It plots the true positive rate against the false positive rate across a range of potential probability thresholds. The more accurate a model is, the more the curve will bulge towards the top left corner of the graph. The dashed line in the middle represents what the curve would look like if one were randomly guessing. If the curve falls below this middle line, then the model is actually doing worse than a coin-flip. Comparisons between models are done by calculating the "area-under-the-curve" (AUC), or percentage of the graph that falls below the curve. The shaded area of the graph in Figure 3.3 represents the AUC. An AUC of 1 would indicate that the model perfectly predicts the outcome of interest, while an AUC of .5 or lower would indicate that the model does not better than a coin
flip. The utility of the ROC curve is twofold. First, it allows for comparisons without the need for setting a probability threshold. Second, if one is interested in setting a probability threshold, it shows the performance in terms of false positives vs. true positives at a range of possible cutoffs.

2.6 Data

The data source for the dependent variable is the Worldwide Atrocities Dataset (WAD) (Schrodt 2009). WAD records instances of violence against noncombatants by both state and non-state actors. The sample of countries includes all countries with a population of at least 500,000 with the exception of the United States \(^2\) from 1995 to the present day. WAD collects and codes news stories of incidents that involve 5 or more fatalities, as well as targeted killings such as the assassinations

\(^2\)The dataset is funded by the CIA, which is barred from collecting domestic intelligence
of political officials, journalists and health workers. Unlike many political science datasets, WAD is updated on a monthly basis, making it ideal for the development of real-time forecasting models.

WAD records roughly 11,000 atrocities over a 20-year period. In order to focus in on large-scale occurrences of violence, I set a threshold to define the onset and duration of a mass atrocity event. An onset occurs in month $t$ if there is not already an ongoing event during that month, the time period from $t$ to $t+5$ contains at least 100 deaths, and month $t$ has at least 50 deaths. A mass atrocity event is considered to have terminated during the first month of any three month period where the total number of deaths is below 100. Previous work has shown that this definition strikes a good balance between providing a sufficient number of positive observations for training and testing models, while maintaining a focus on the occurrence of mass, as opposed to isolated, violence (Fisher & Schrodt 2014). Because the goal of this paper is to forecast when mass atrocity events begin, I
drop observations where mass atrocity events are ongoing. Figure 2.3 displays the countries that experience the onset of a mass atrocity by half year. There are 128 onsets total.

The predictor variables fall into two main categories: structural and event. The purpose of the structural variables is to identify aspects of a country that make it predisposed to potential mass violence. These include socioeconomic indicators such as GDP and ethnic composition. However, these indicators are often slow moving, which means a model comprised only of structural variables will produce very similar predicted probabilities for each country year to year. In order to correct for this, a model needs to include data with enough variation to show a spike in risk from one period to another. This is where event data is potentially very useful for forecasting purposes. It provides necessary variation on the RHS of the formula, while maintaining a theoretical basis for inclusion. For example, a spike in events related to government repression may be a precursor to the onset of a mass atrocity.
The model specification relies on a simple set of three structural variables: infant mortality rate, population, and ethnic characteristics of leadership. (Fisher & Schrodt 2014) demonstrate that this specification is robust to more complex alternatives. Data on infant mortality rate comes from the U.S. Census Bureau International Division, via PITF, and measures the number of infants younger than a year dead per 1,000 births each year. Infant mortality rate serves as a proxy indicator of a country’s level of development. Population data comes from World Bank and should capture that a larger population potentially a greater number of opportunities for violence. Data on ethnic characteristics of ruling elites is provided by the Center for Systemic Peace, via PITF. This is a categorical variable indicating the role of ethnicity amongst a country’s ruling elites. There are three categories: elite ethnicity is not salient, elite ethnicity is salient and the political leadership is representative of the largest communal groups or a coalition of groups that together constitute a majority, and elite ethnicity is salient and the political leadership is representative of a minority communal group or coalition of small groups that together constitute less than a majority.

I use the Integrated Conflict Early Warning System (ICEWS) dataset to generate event variables. Event data records interactions between actors in the international system in a who-did-what-to-whom format. For example, an attack by the Egyptian government on Egyptian protestors would be recorded as an event with the Egyptian government as the source, Egyptian protestors as the target, and act of repression as the type of event. ICEWS generates data using a program that parses and codes news reports. Observations are recorded at a daily level, allowing great flexibility in terms of the temporal aggregation of an analysis. ICEWS uses the CAMEO coding scheme (Gerner et al. 2002), which has several dozen unique event categories. Similarity in certain event types makes event aggregation fairly
straightforward. For example, one could combine the four event types 'carry out suicide bombing', "carry out vehicular bombing", "carry out roadside bombing", and "carry out location bombing" into one general "carry out bombing" category. Like WAD, ICEWS is updated on a monthly basis, making it well-suited for the development of real-time forecasting systems.

The fact that event data observations in their raw form are essentially unusable for analysis makes feature engineering a requirement for its inclusion in any model. Event data studies commonly use what are known as quad count variables, which are counts of the number of verbal cooperation, material cooperation, verbal conflict, and material conflict respectively (Yonamine 2011). However, one could theoretically generate any number of event features using different levels of temporal aggregation, monadic vs. dyadic actor combinations, or event combinations. Rather than do an exhaustive search of combinations, which would be computationally infeasible for this project, I generate three different sets of event variables, all centered on measuring the level of material conflict in a state. They are

- **total conflict** - single variable that is a count of the total number of material conflict events between actors in a state

- **dyadic conflict** - counts of material conflict between several actor dyads - government and political opposition, government and rebels, rebels and societal actors ³, government and societal actors, and rebels and political opposition

- **specific counts** - counts of detainments, protests, and violent repression events by the government.

The purpose of these three different sets is basically to understand if relying on a single measure, total conflict, is superior to using a more disaggregated approach.

³Societal actors include groups such as NGOs, businesses, and regular citizens.
The specific counts set and the dyadic set both represent two potential avenues of aggregation, actor aggregation and event aggregation. Essentially, the question is whether who is fighting is a better predictor than what kind of conflict is going on.

In addition, I create two variations on each of these - one where the variables are differenced and one where the variables are binned. A differenced approach may be more appropriate than using raw counts as some countries may have a higher level of conflict events on average than others. As a result, measuring when there is a dramatic increase in conflict events may be more appropriate. Binning variables is a commonly recommended practice in predictive analytics (Kuhn & Johnson 2013). Using binned categorical variables in place of continuous variables addresses potential issues of nonlinearity and overfitting. I bin each variable into three equally sized, ordinal categories (essentially low, medium, and high).

2.7 Analysis

The unit of observation is a country-half year. I set a two year prediction interval on the structural variables and a six month interval on the event variables. The two year interval for the structural variables is due to data availability. For example, infant mortality data for 2014 typically will not be available until some time in 2015. It is necessary then to use data from two years prior to the year I am predicting, rather than the previous year.

Like most machine learning models, Random Forests and neural networks have hyperparameters, or settings associated with a model that must be manually set. One example of this is setting the number variables selected from at each split in a Random Forest model. The hyperparameter values can have a large effect on
the accuracy of the model, so users typically perform hyperparameter tuning to determine which set of values generates the best performance. This is done using grid search, which involves testing predictive performance using all possible combinations of a given set of values across hyperparameters. This requires a holdout sample to predict on. However, we want to avoid fitting to a single holdout sample, and we do not want to use the holdout sample that we plan to test the final models on. So rather than using the holdout sample, or training and predicting on the training data, we use a form of iterative testing known as k-fold cross validation.

Cross-validation is a method for reducing potential overfitting. Rather than choose the hyperparameter values based on performance on one test sample, one chooses values based on average performance on multiple test samples. These multiple test samples are created by dividing the training data into \( k \)-partitions of equal size (Stone 1974). A model is then trained on \( k-1 \) partitions and validated on the left out sample. This process is repeated \( k \) times, with each partition serving as the validation set once. Doing this avoids choosing a set of hyperparameter values that may only have good performance because they overfit the model on a specific test sample. This is particularly a danger when working with only a small sample of positive observations. With a small number of positive observations, the model may fit to the idiosyncrasies of these few cases, rather than the general features of the phenomenon of interest.

For this analysis, I use 5-fold cross validation. There are approximately 75 positive observations in each training partition and 25 positive observations in each testing partition. The Random Forest model has only one hyperparameter to tune: the number of variables selected from at each split (referred to as \texttt{mtry}). One can also adjust the number of trees used to create the model, but Random Forest performance is considered to be a monotonic function of number of trees. There is no
improvement gained from adding more trees after a certain number, but there is no drop in performance either. I use the default of 500 trees set by the caret package for each Random Forest model. The neural network hyperparameters include the number of nodes in the hidden layer, the decay rate of the weights, and whether to use bagging when training the networks. I perform tuning for every combination of model and feature set, excluding logistic regression. The “best” combination of hyperparameter values is determined by the highest average AUC across 5 folds. The grid of values tested is displayed in Table 2.1.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter 1</th>
<th>Parameter 2</th>
<th>Parameter 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>mtry = 1, 3, 5, 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neural Network</td>
<td>size = 5, 7, 9</td>
<td>decay = 0.5, 0.1, 0.01</td>
<td>bag = True, False</td>
</tr>
</tbody>
</table>

2.7.1 Model Selection

I begin by first testing each method on a basic feature set and selecting the best method. Here, I use the dyadic feature set as a starting point, as this format has been used in previous forecasting studies (O’Brien 2010). Before fitting the final neural network and Random Forest models, I tune them using 5-fold cross validation on the training sample. As previously discussed, I use ROC curves and associated AUC scores for the purposes of model comparison. The reasoning behind using ROC curves and AUC scores as accuracy metrics, rather than mean accuracy or classification tables based on a cutoff probability value is clear in Figure 2.4. It displays a box-and-whisker plot of predicted probabilities on the holdout set from the three best models and from an average of probabilities across the three models. Random Forest has a clear tendency to assign lower probabilities

\[ mtry = 3 \text{ for Random Forest and } size = 7, \ decay = 0.01 \text{ and } bag = False \text{ for neural network} \]
on average, but has a larger range in the values assigned. This makes determining a probability cutoff to divide predicted onsets from no predicted onsets to use as a comparison problematic. Another approach is necessary in cases where models deliver varying ranges of predicted probabilities. The focus should really be then on determining which models do the best job of assigning higher probabilities to positive observations and lower probabilities to negative observations. The ROC curve, which compares the rate of true positives to the rate of false positives at multiple cutoff points, is ideal for measuring this.

Figure 2.5 displays the ROC curves for each model and their associated AUC. In this case, logistic regression outperforms the two machine learning methods, a surprising result. The logistic regression’s AUC of .87 is significantly better than the other two methods.

Since this result runs contrary to what previous studies of method performance
have taught us to expect, I do a robustness check by applying the methods to the other two general feature sets, total conflict and specific accounts. This should account for the possibility of the logit’s performance being an idiosyncratic result due to the feature set. Table 2.2 shows the results of this test. Again, the logit model performs better overall, with the neural network model having a very slim advantage over it for the total conflict feature set. Another interesting result is that there does not appear to be one best feature set across the methods. For example, the specific counts feature set performs the best overall when used in conjunction with the logit model, but performs worse when used with a neural network model compared to the total conflict feature set.

Table 2.2: AUC Scores

<table>
<thead>
<tr>
<th></th>
<th>Logit</th>
<th>Random Forest</th>
<th>Neural Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>total conflict</td>
<td>.840</td>
<td>.799</td>
<td>.844</td>
</tr>
<tr>
<td>dyadic conflict</td>
<td>.830</td>
<td>.820</td>
<td>.816</td>
</tr>
</tbody>
</table>

Figure 2.5: ROC Curves for Specific Counts Feature Set
2.7.2 Feature Engineering

Having determined that a logistic regression is the optimal method to use for this problem, I proceed to test the various engineered feature sets using logit models. I continue to use ROC and AUC as a means of comparison. I begin by fitting models to the three different feature sets, total conflict, dyadic conflict, and specific counts in their plain form - raw counts. Figure 2.6 displays ROC curves and associated AUC scores for each of these models. The specific counts feature set outperforms the other two with an AUC of .87.

Next, I look at differenced versions of the three feature sets. It is likely that looking at the change in the amount of conflict occurring in a state is a better means of predicting the onset of a mass atrocity. Figure 2.7 presents results for these three models. Again, the specific counts feature set outperforms the other two feature sets. However, the models perform worse than each of their coun-
terparts in the raw count models. Results from the binned feature sets, presented in Figure 2.8, similarly underperform compared to the raw count versions of the feature sets. It may be that at this level of temporal aggregation, there is too much noise for transformations like differencing to gain much leverage. It is also possible that a different number of bins or different cutoffs might improve the performance of the binned variables, but the lack of any gains in their current form suggests that pursuing a binned approach would not be worth the investment.

Overall, these results demonstrate that feature engineering can have a drastic effect on model performance. The best feature set, raw specific counts, yields an AUC score of .87, while the worst feature set, differenced dyadic conflict, has an AUC score of .76. This is a significant difference, illustrating the need to test multiple sets of engineered features in order to optimize the performance of a predictive model.
2.7.3 Case Control

I use the same criteria as Goldstone et al. (2010) when developing a case control sample. For each onset in the training sample, I match three observations from the peace observations in the training sample. The three peace observations are drawn from the same year and geographic region as the onset observation with which they are associated. There are seven unique geographic regions: Africa, East Asia, Europe, former Soviet states, Latin America, Near East, and North America. Additionally, these observations must not have experienced any mass atrocity episodes in the two years before and after the period in question. This generates a total of 396 observations in the training data, with 99 onsets and 297 with no onsets.

Another way of looking at the case control sample approach is that it creates a biased sample of only at-risk countries. For example, there are no cases drawn
from North America, and only 2 positive observations and 6 negative observations matched to them for Europe. It is a way of admitting that the probability of a mass atrocity occurring in developed regions is almost nonexistent. Instead, case control provides a sample of at-risk countries, some of which experience mass atrocities, but most of which do not. By training only on these countries, the model might be better able to identify why a country experiences a mass atrocity, while its neighbors remain at peace.

Table 2.3: Case Control Regional Distribution

<table>
<thead>
<tr>
<th>Region</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>200</td>
</tr>
<tr>
<td>East Asia</td>
<td>44</td>
</tr>
<tr>
<td>Europe</td>
<td>8</td>
</tr>
<tr>
<td>Former Soviet</td>
<td>16</td>
</tr>
<tr>
<td>Latin America</td>
<td>32</td>
</tr>
<tr>
<td>Near East</td>
<td>96</td>
</tr>
</tbody>
</table>

Rather than repeat the full model selection and feature engineering process for the case control sample, I only fit a model for the best feature set and method - logistic regression and the specific counts feature set, respectively. I fit the model on the case control sample and predict on the same holdout set used for the full sample models. Figure 2.9 displays a ROC curve for a logistic regression model fitted on the case control sample compared to one fitted on the full sample. The case control model performs similarly to the full sample model. With no clear gains in accuracy, resampling using a case control approach when developing a forecasting model is not worth the extra steps required.
2.7.4 Inspecting Results

In order to get a better idea of what the results look like on closer inspection, I map the 10 countries with the highest predicted probability of experiencing a mass atrocity onset in the second half of 2013, presented in Figure 2.10. The 10 countries are highlighted in red. Of these 10 countries, three, the Central African Republic, Somalia, and Egypt, experienced the onset of a mass atrocity during this period, the first half of 2013. The remainder are all countries that have experienced mass atrocities in the past. This implies that, even though the models produce a large number of false positives, many of these "make sense" intuitively, in that they are countries we would expect to be at risk given a history of mass violence. Given the noisy nature of the process behind the outcome being predicted, it is unsurprising that there are going to be a number of false positives. These are acceptable as long as they are countries that no one would reasonably expect to be at risk for mass atrocities (e.g. Western European countries).
2.8 Conclusions

There are several important takeaways from the results of this analysis. The first is that alternative methods do not yield gains over standard logistic regression. Logistic regression actually does better on average across the feature set combinations than either neural networks or Random Forests. This is unexpected given the both of these models outperform logistic regression in methods tournaments in other fields. It could be that logistic regression is performing best on this particular sample of cases and performance may drop over multiple test sets. With this in mind, the best model in the long run is likely an ensemble of results from two or more of these methods. Scholars should not expect to find a silver-bullet method that will always outperform alternatives.
Feature engineering can have a dramatic effect on the accuracy of a model. The feature sets used all measure the same general concept, conflict, but the accuracy varies widely. The best performing feature set has an AUC of .87, while the lowest has an AUC of .75. Feature engineering matters. When presented with data sources that must be transformed before use, like event data, time should be spent developing and testing different transformations and aggregations. Additionally, the ranking of the feature sets varies by method used. As previously suggested, the best approach is to ensemble results across the best combinations of models and feature sets to account for potential variation in performance on test samples.

The case control approach does not provide any improvement over using the full sample of cases. Though there may be inference-related reasons to apply this type of sampling, it does not help for prediction tasks. The fact that case control is a particular form of the general downsampling procedures recommended by machine learning literature suggests that downsampling in general is not appropriate for these problems.

Ultimately, developing a good forecasting model for mass violence, and political unrest in general, has to involve a comprehensive, iterative approach. Political scientist interested in developing forecasting models should be systematic in their approach to exploring and testing the possible avenues of improving predictive accuracy. As evidenced by the results of the methods section and the feature engineering section, finding the best model requires iterating over several possible options, with the likely outcome being an ensemble of multiple models. There is no silver-bullet for creating the perfect model, but combining existing approaches and methods yields strong results.
3 Forecasting Atrocities at a Sub-national Level

3.1 Introduction

For the past 15 years, Nigeria has experienced a continual problem with atrocities perpetrated by nonstate actors. One of the features of the violence has been its geographic isolation. The major problem area at first was the Niger delta, where armed groups used violence against the military and civilian population in order to assert control over oil wealth in the region. The violence there subsided after a general amnesty in 2009. Though the situation in the Niger delta had improved, the country began experiencing a continuation of mass violence in the northeastern part of the country with the emergence of the Boko Haram insurgency, which soon became notorious for targeting civilians.

Other examples of countries with isolated violence include Iraq with its so-called Sunni Triangle, Myanmar, whose northern region is home to multiple ethnic rebels groups, and China and its Xianjiang region. These examples together suggest that a country-level forecasting model only scratches the surface of the problem at hand. The natural step then is to develop a model that forecasts at a subnational level.
Assuming that there is some sort of geographic component to violence within a country, this model should be appropriate. From a policymaking perspective, spatially specific forecasts can provide better information on how to allocate resources, as well as which groups may potentially be the source of violence.

In this paper, I develop a forecasting model of mass violence based on the underlying assumption that risk is not uniformly distributed across a state. This is not the first model to do so. Raleigh (2010), for example, use a raster grid to analyze how geography affects the likelihood of conflict, as does Kucera et al. (2011). The former from a scholarly, inference-based approach, while the latter is more concerned with policy implications. Both focus only on a handful of sub-Saharan African countries within limited periods of time. Perry (2013) forecasts civil conflict at a subnational level for all of Africa, but uses second and third order subnational administrative units as the units of observation instead of raster cells. I build on previous work by expanding the analysis to a raster grid of the African continent.

3.2 Research Design

I use a raster structure for the data. In the field of geography, a raster is a term for an array of x and y coordinates of equal size, which can contain one or more attribute values for each square, often referred to as bands. Visually speaking, it is a simple grid structure, with each cell of the grid a unique observation assuming no temporal component. One of the major advantages of a raster format is the ease with which one can aggregate or disaggregate to different resolutions. This has made it widely used as a format for satellite imagery data, many of which come in varying resolutions. For example, the Moderate-Resolution Imaging Spectrorad-
diometer (MODIS) Land Cover dataset, which uses satellite imagery to determine land cover type (e.g. grasslands) (Friedl et al. 2002). The dataset uses a 0.05x0.05 decimal degree resolution, meaning the cells of the grid are 0.05 degrees long on each side. Each cell is assigned a value based on the predominant type of land-cover within it according the satellite imagery. One could easily aggregate this to a lower resolution (i.e. larger cells) by assigning the new cell the modal value of its constituent cells. Conversely, one could divide the cells into smaller cells in order to merge with a higher resolution dataset. It is a natural choice of data structure for any study using geographic data.

In this case, I overlay a raster on the continent of Africa. Figure 3.1 shows a visual representation of the raster. I use a 0.5x0.5 decimal degree resolution. This translates to cells that are approximately 55x55 kilometers in size. Choosing a resolution is, to some extent, an arbitrary process that largely comes down to balancing too little detail versus too much detail. Obviously, I want to be more specific than a national level, but it is highly unlikely that I will get enough leverage on the data to accurately predict square mile by square mile. In order to adhere to a semblance of precedent, and for the sake of convenience, I base my choice of resolution on the one used by the Peace Research Institute Oslo’s (PRIO) GRID\textsuperscript{1} project (Tollefsen, Strand & Buhaug 2012). PRIO-GRID provides both a grid structure, as well as a large amount of accompanying data, some of which I use in this paper. Using the PRIO-GRID resolution should provide a good balance on the level of detail and make this comparable to any future studies in political science that rely on PRIO-GRID. Figure 3.1 displays a 0.5x0.5 decimal degree resolution grid overlaid on Africa.

A key component of any predictive modeling exercise is assessing the accuracy

\textsuperscript{1}The all capital spelling is apparently an aesthetic choice and not because GRID is an acronym
of the final model. The danger that every predictive model faces is overfitting. A model suffers from overfitting when it predicts the data used to create it with a high degree of accuracy, but performs poorly when used to predict cases that it has not seen. The typical way to avoid incorrectly choosing an overfit model over a better model involves dividing the data into two samples: one as data used to train the model and the other used as holdout data that the trained model attempts to predict. The point here is to assess how well the model performs on observations that it has never seen.

After dividing the data into training and testing samples, one must determine
what metric by which to compare model accuracy. The obvious choice would be overall accuracy, or the percentage of observations correctly classified. There are a couple of issues that make this metric unsuitable in this case. First, one has to choose a probability threshold at which to divide positive observations from negative observations. There is no easy way to set this threshold in cases where the outcome classes are highly imbalanced. Determining a threshold becomes more problematic when comparing across methods, as some methods may give higher or lower predicted probabilities on average than others. Additionally, it is unclear what qualifies as "good" accuracy in rare events prediction. I could get a 97% accuracy on predicting mass atrocities by naively predicting that there will be no onsets. Since I am more interested in predicting onsets than peace periods, I might actually prefer a model with lower accuracy if it predicts more onsets correctly.

Rather than randomly assigning the data to training and testing samples, I divide the sample temporally. The training sample is data from 2003-2011, and the testing sample is data from 2012-2014. This is a form of out-of-sample validation known as out-of-time validation. This form of validation is appropriate when there is a temporal component to the data, and one believes that there is a potential underlying change over time. More intuitively, if we are interested in how our model will do when predicting an event in $t$, it would be better to test it on $t - 1$ rather than $t - 10$. We can reasonably expect that $t$ is much more similar to $t - 1$ than to $t - 10$. This way, we avoid fitting to some underlying temporal trend that may be present in older data, but has changed for more recent observations.

I rely on the Receiver Operating Characteristic (ROC) curve to compare model accuracy. ROC curves, originally developed to assess the accuracy of air radar (Green & Swets 1966), are a widely used metric in predictive modeling. "Tournament" papers, which compare the performance of multiple models, typically use
This metric (Caruana & Niculescu-Mizil 2006). Figure 3.2 displays an example of an ROC curve. It plots the true positive rate against the false positive rate across a range of potential probability thresholds. The more accurate a model is, the more the curve will bulge towards the top left corner of the graph. The dashed line in the middle represents what the curve would look like if one were randomly guessing. If the curve falls below this middle line, then the model is actually doing worse than a coin-flip. Comparisons between models are done by calculating the "area-under-the-curve" (AUC), or percentage of the graph that falls below the curve. The shaded area of the graph in Figure 3.3 represents the AUC. An AUC of 1 would indicate that the model perfectly predicts the outcome of interest, while an AUC of .5 or lower would indicate that the model does not better than a coin flip. The utility of the ROC curve is twofold. First, it allows for comparisons without the need for setting a probability threshold. Second, if one is interested in setting a probability threshold, it shows the performance in terms of false positives vs. true positives at a range of possible cutoffs.

3.3 Methodology

Originally developed in Breiman (2001), the Random Forest algorithm is a supervised learning method that can be used for both regression and classification problems. The algorithm is considered an ensemble method, as it uses multiple decision tree models to classify observations. Each tree is constructed by taking a bootstrapped sample of the training data and selects from a random subset of the predictor variables at each split. When a tree is grown, a percentage of the bootstrapped sample is left out. This "out-of-bag" (OOB) data is then run through the tree. This gives an error rate for the tree based on how well it classified the OOB data. This process is repeated for each tree in the forest. Once the forest
Figure 3.2: ROC Curve

Figure 3.3: ROC Curve with AUC
is constructed, observations are classified by the mode output of trees. The per-
centage of the trees that vote for a particular outcome can be used as a predicted
probability for that observation. While the modular nature of the algorithm has re-
sulted in a plethora of variations on Breiman’s original, such as conditional forests
(Hothorn, Hornik & Zeileis 2006), gradient boosted trees (Friedman 2001), and
oblique Random Forests (Menze et al. 2011), I use Breiman’s original algorithm as
implemented in the randomForest R package (Liaw & Wiener 2002).

In addition to prediction, Random Forests provide two means of looking at the
effects of individual variables. The first is permutation importance. Permutation
importance is a method of assessing the predictive importance of each variable
in the model. This is done by randomly permuting the value of a variable of in-
terest for each observation in the OOB data, then recomputing the OOB error
rate using the permuted data. The difference between the original error rate and
the permuted error rate is used as the importance score. Essentially, we would
expect that changing the value of a variable that has high predictive importance
would hurt the accuracy of the model at classifying observations, while accuracy
should be unaffected by changing the value of an important variable. The score
is typically either the mean decrease in accuracy, or the mean decrease in Gini
impurity, a metric measuring the proportion of positive to negative observations
within a terminal node of a decision tree. While permutation importance gives an
idea of the absolute predictive importance of a variable, it does not show whether
a variable increases or decreases the probability of a given outcome occurring.

Random Forests include a procedure known as partial dependence to calculate
the directional effect of a variable on the predicted probability of witnessing an
outcome of interest. Like permutation importance, partial dependence makes use
of the OOB data to calculate effects. Given a set of unique values \{j_1, j_2, ..., j_i\} for
variable $X$, it sets the value of $X$ in all OOB observations to $j_1$. It then generates an average predicted probability associated with that value by reclassifying each of the OOB observations. The process is then repeated for the next value, $j_2$, and continues for every possible value or a predetermined number of values. The resulting predicted probabilities are then plotted to create a partial dependence plot.

3.4 Data

The format of the data is a 0.5x0.5 degree raster covering the continent of Africa from 2003 to 2014. The data is drawn from several sources, with the recently updated PRIO-GRID project serving as a source for many of the independent variables. Non-PRIO data was transformed to fit a 0.5x0.5 degree raster using the software program QGIS. Maps of the distribution of each variable are included in the appendix.

3.4.1 Dependent Variable

The data source for the dependent variable is the Worldwide Atrocities Dataset (WAD) (Schrodt 2009). WAD records instances of violence against noncombatants by both state and non-state actors. The sample of countries includes all countries with a population of at least 500,000 with the exception of the United States\(^2\) from 1995 to the present day. WAD collects and codes news stories of incidents that involve 5 or more fatalities, as well as targeted killings such as the assassinations of political officials, journalists and health workers. Unlike many political science datasets, WAD is updated on a monthly basis, making it ideal for the development

\(^2\)The dataset is funded by the CIA, which is technically barred from collecting domestic intelligence
of real-time forecasting models.

Most importantly for this study, the latitude and longitude of the location where each event occurred is recorded. I made adjustments for obvious errors (e.g. an event listed as occurring in South Africa, but with latitude coordinates in the northern hemisphere) and assigned each event to the cell in which it occurred. The final dependent variable is dichotomous: 0 if no event occurred in a cell during a given period of time, or 1 if at least 1 event occurred. There is no differentiation between types of events or intensity.
3.4.2 Distance

Previous research suggests that location within a country, plays an important role in the occurrence of civil conflict (Raleigh 2010). State or non-state forces may resort to violence in order to maintain control in areas farther from the capital. Border regions may host rebel groups supported by a neighboring state, or people displaced by conflict in a bordering country. One example of this is eastern Democratic Republic of the Congo, where the region bordering Rwanda has been the site of much violence over the years by government and rebel troops, often supported by Rwanda, in a struggle to assert control over the region.

To reflect this, I use two distance measures: distance to capital and distance to the nearest international border. Both are taken from the PRIO Grid dataset. The first is the spherical distance in kilometers from the cell centroid to the capital of the country in which the cell lies. Spherical distance, also known as great-circle distance, calculates the shortest distance between two points on a sphere, in contrast to geometric distance, which assumes that the points fall on a flat surface. The second is spherical distance in kilometers from the cell centroid to the nearest international border. In the case of Madagascar, an island nation, this would include the distance across the Mozambique Channel.

3.4.3 Economic Health and Development

Many of the atrocity events recorded in the WAD dataset that take place in Africa involve violence by herders against farmers or vice versa as a result of competition for land resources. To account for conditions that might exacerbate this competition, I use the occurrence of a drought as a measure of periods when tensions between these two groups would be high. The drought measure comes from PRIO
based on data from the International Research Institute for Climate and Society at Columbia University. Severity is measured using the Standardized Precipitation Index (SPI) (Guttman 1999). This variable is calculated at a yearly level. It is the highest number of consecutive months where the SPI value falls below $-1.5$. For example, if the SPI value fell below $-1.5$ each month from January through April, and again in September and October, the drought value for that year would be $4/12 = .33$, since the longest streak of drought was 4 months.

As a proxy for economic development, I use data on nighttime emissions. Higher emissions indicate the presence of electric light, which then indicates that it is a developed area. The measure of nighttime light emission comes from PRIORI based on data from DMSP-OLS Nighttime Light Time Series dataset by NOAA, which collects data on nighttime light emissions using satellites. The original data is collected at a yearly level with a .05 degree resolution, with each pixel assigned a value based on the brightness of the area at night. A higher value indicates greater brightness. Because the original dataset uses a higher resolution than the PRIORI Grid dataset, I use the average value of the original cells that fall within a cell in my dataset.

### 3.4.4 Population and Demographics

Population can be an important predictor of where violence is unlikely to occur, in that there are few or no civilians to kill in an unpopulated region. I use the Landscan dataset to measure the population in each cell (Dobson et al. 2000). Landscan employs an algorithm to generate population data at a 1km level resolution using a combination of census data, geographic data, and satellite imaging. As with some of the other datasets, I aggregate the cells to match the 0.5x0.5 decimal degree resolution I use. The population variable is simply the total population within a
cell according to the Landscan dataset.

Regions with ethnic cleavages may have a higher probability of violence than homogenous regions, or those where all or most ethnic groups are included in the political process (Collier & Hoeffler 2004). In order to measure this, I look at the number of ethnic groups excluded from the political process in some way within each cell. Data on excluded ethnic groups comes from the Geographic Ethnic Power Relations (GeoEPR) dataset via PRIO. GeoEPR codes the location and status of politically relevant ethnic groups (Vogt et al. 2015), as well as three different levels of power access: the group rules alone, the group shares power, or the group is excluded. To measure potential ethnic tensions, PRIO-GRID operationalizes this data by creating a variable counting the number of excluded ethnic groups in each cell.

3.4.5 Geography

Assuming that physical geography may have an affect on conflict and violence, I include measures for land cover type, as well as mountainous terrain. Data on land cover comes from the Moderate-Resolution Imaging Spectroradiometer (MODIS) Land Cover dataset (Friedl et al. 2002), which is based on satellite imaging from USGS and NASA. MODIS collects data on land cover type at a yearly level with a .05 degree resolution. Land cover types are divided into 16 categories. Because the original data is at a higher resolution than my grid, I assign each cell a value based on the modal category of the original MODIS cells that fall within its boundaries. This MODIS data is currently only available through 2012.

For measuring mountainous terrain, I use a measure from the PRIO-GRID dataset based on data from UNEP’s Mountain Watch Report. The original UNEP
dataset is a basic binary metric indicating whether a pixel is mountainous or not. Because the PRIO’s dataset uses a lower resolution, the PRIO version of the variable is the proportion of the cell that would be considered mountainous according to the UNEP data.

### 3.4.6 Resources

Competing armed groups are likely to assert control over mineral or energy resources in order to generate funding. Violence may result from groups competing for these resources or coercing the civilian population into aiding in their extraction. I create two separate resource variables: one for petroleum resources and one for precious mineral resources. Data for these variables were obtained via the PRIO-GRID project, which bases its codings on a variety of data sources (Balestri 2012, Lujala, Rød & Thieme 2007, Lujala 2009). The petroleum variable is binary, indicating whether a petroleum deposit is present in the cell. The other variables is also binary, indicating whether a there is a deposit of either diamond, gold, or gems within a cell.

### 3.4.7 Conflict

Violence against civilians often occurs within the context of a wider civil or interstate conflict. I use the Armed Conflict Location Event Data (ACLED) project’s Africa dataset as a source of data on conflict events in Africa (Raleigh et al. 2010). ACLED records instances of political violence, including the geographic coordinates of where an event occurred, as well as the actors involved. The data covers all African countries from 1997 to the present day. In order to avoid including instances of violence against civilians on the right hand side of the equation, I omit all events where the target actor was coded as civilian.
Rather than creating a dummy variable indicating whether a violent event occurred in a cell, I use kernel density estimation to account for the possibility that conflict in one cell increases the likelihood of violence occurring in neighboring cells. Statistically, kernel density estimation is a method for estimating the probability density function for a random variable. In a GIS setting, kernel density works by applying a density function to fit a smoothly curved surface over a point. This creates a circular surface around a point, with surface values, generally on a 0 to 1 scale, highest at the point and diminishing as one moves away from the center. There are two main parameter settings to consider. The first is the radius, or how far out from the point the kernel surface will extend. The second is the decay ratio, which determines the rate at which surface values decrease moving away from the center. A decay ratio of 0 causes high values to be concentrated at the center, while a value of 1 distributes values uniformly across the surface.

I use QGIS’s Heatmap plugin to generate a density raster for the ACLED conflict events. I set the radius at 0.5 degrees so that the effects of a conflict event extend into neighboring cells. I set the decay ratio at 0.2. This means that the values on the edge of the surface will be approximately 20% of the values at the center. Figure 3.5 displays a kernel density map of ACLED conflict events in the first half of 2010. The final variable is the sum of the surface values within a cell. Because surfaces are allowed to overlap, the values can stack and add up to a value much greater than 1 for cells with a high number of conflict events in or around them.
3.5 Results

The unit of observation is a cell half-year. Except for the conflict variable, all variables have a two year lead time due to data availability. For example, new Landscan data on 2014 may not be available until some time in 2015. The conflict variable is set to a six month lead time. The training sample contains just under 150,000 observations, 473 of which are observations where an atrocity occurred. The testing sample has approximately 48,000 observations, 517 of which are positive.
Like many machine learning methods, Random Forests have a hyperparameter, in this case the number of candidate variables chosen from at each split in a tree, that must be manually set. This particular parameter is typically referred to as $mtry$. Because the value of a hyperparameter can have a significant effect model accuracy, it is standard practice to perform hyperparameter tuning to determine which value generates the best performance by training and testing the model across a set of possible values to determine which provides the highest accuracy. This requires a holdout sample to predict on. However, we want to avoid fitting to a single holdout sample, and we do not want to use the holdout sample that we plan to test the final models on. So rather than using the holdout sample, or training and predicting on the training data, we use a form of iterative testing known as $k$-fold cross validation.

Cross-validation is a method for reducing potential overfitting. Rather than choose the hyperparameter values based on performance on one test sample, one chooses values based on average performance on multiple test samples. These multiple test samples are created by dividing the training data into $k$-partitions of equal size (Stone 1974). A model is then trained on $k-1$ partitions and validated on the left out sample. This process is repeated $k$ times, with each partition serving as the validation set once. Doing this avoids choosing a set of hyperparameter values that may only have good performance because they overfit the model on a specific test sample. This is particularly a danger when working with only a small sample of positive observations. With a small number of positive observations, the model may fit to the idiosyncrasies of these few cases, rather than the general features of the phenomenon of interest.

For this analysis, I use 5-fold cross validation to determine the optimal value for $mtry$. I test with values of 1, 3, 5, 7, and 9. There are roughly 378 and 95
positive observations in each training and testing partition respectively. The "best" value is determined by the highest average AUC across the 5 folds. In this case, the best performance occurs when \textit{mtry} is set to 9. This means that the decision trees will select from a subset of 9 of the 11 predictors at each split.

Figure 3.6 displays results from predicting the holdout data using a Random Forest model trained on the training data. As a baseline comparison, I include results for predicting atrocities using only a variable indicating whether there was an atrocity in that cell one period prior. This serves as a good "better than a coin flip" style comparison for rare events conflict data. The AUC for the full model is 0.82, while the AUC for the lagged model is .63, meaning the model is a significant improvement over a model based on prior violence alone. The results were similar if the training and testing samples were flipped, with the model trained on the smaller test sample and tested on the larger training sample.
Table 3.1: Predictions using .0015 threshold

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No Event</td>
<td>34502</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>Event</td>
<td>13177</td>
<td>426</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: Predictions using .014 threshold

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No Event</td>
<td>43436</td>
<td>260</td>
<td></td>
</tr>
<tr>
<td>Event</td>
<td>4243</td>
<td>257</td>
<td></td>
</tr>
</tbody>
</table>

One issue with using ROC curves to assess rare events models is that, while the do well for model comparison, they do not give a good idea of how well a model is doing in absolute terms (Davis & Goadrich 2006). Assuming a usable model would be one with a recall of .75 (i.e. correctly predicts 75% of the events), I determine a probability threshold to achieve this rate, in this case .0015, and run a cross tabulation presented in Table 3.1. Of particular interest is the false positive rate. Any predictive model can achieve perfect recall with a low enough probability threshold, but this comes with an associated cost in terms of an increasing number of false positives. This fact is particularly evident here. Achieving a recall of .75 comes with an associated ratio of roughly 31 false positives to each true positive. Reducing the target recall to .5 reduces the ratio to 16 as shown in Table 3.2. There is unfortunately no set rule for adjudicating this tradeoff.

One of the main motivators behind this paper is that mass violence risk is not evenly distributed across a state, but clusters in geographic subregions. I assess this in a couple of ways. First, I do a simple boxplot, displayed in Figure 3.7, of predicted probabilities of cells within a country for a five countries that have experienced mass violence. We would expect predicted probabilities to be relatively equal across the cells of a country if there was no geographic component to risk.
Instead, there are clearly cells within each country that are much higher risk than others. Next, I map the predicted probabilities for the latest period, the second half of 2014. Figure 3.8 is a heatmap of the predicted probabilities by cell, while Figure 3.9 displays the same map with points overlaid indicating where atrocities occurred during that period. Dark orange and red indicate cells with a higher probability of violence occurring, while the blank white areas are where observations were dropped due to data missingness. Overall, cells with higher risk appear to cluster geographically.

Figure 3.7: Predicted Probability Boxplots
Figure 3.8: Predicted Probabilities for 2014-2
Finally, I look at variable importance using the permutation importance and partial dependence features of the Random Forest algorithm. The importance scores in Figure 3.10 show that civil conflict is the strongest predictor of mass atrocities, with an average decrease in accuracy of over 4% when the value of the variable is randomly altered. The two distance measures, distance to capital and distance to international border, are also important, as is the measure of night time light emissions. To get a better idea of how these predictors affect the likelihood of violence, I generate partial dependence plots for the top six variables, shown in Figure 3.11. The results for most of the variables are as expected. There is a greater likelihood of a mass atrocity occurring in populated areas with a higher degree of
development (as measured by night time light emissions). A higher number of excluded ethnic groups and a history of civil conflict also increase the probability of a mass atrocity happening. The two distance variables, on the other hand, appear to have a U-shaped relationship, with higher probabilities associated with either being very close to a capital or border, or very far away. Distances in between have relatively low probabilities.

3.6 Conclusion

This analysis demonstrates that a subnational forecasting model based on a raster grid is feasible. By leveraging satellite imagery data in conjunction with more traditional political conflict data, it’s possible to generate reasonably accurate forecasts at a 0.5 degree resolution, with an AUC measure of 0.82. In most cases this would be a great score, but it masks the class imbalance and the resulting poor precision.
Figure 3.11: Partial Dependence Plots

(a) Civil Conflict

(b) Population

(c) Excluded Ethnic Groups

(d) Distance to International Border

(e) Nighttime Light Emissions

(f) Distance to Capital City
of the model. Ultimately, what one considers "reasonably accurate" largely depends on what one considers to be an acceptable number of false positives. Improving the precision of this type of model should be the focus of future work in this area.

In terms of model specification, the presence of civil conflict continues to be a strong predictor of violence against civilians. Some of the variables based on satellite imagery data, which would not typically be used in a country level model, have strong predictive power. Considering that the satellite imagery measures were created using basic aggregations, additional feature engineering on the satellite imagery data could possibly improve the results. Overall, the results indicate that a combination of geographic, demographic, and political data is key to developing a working model.

This paper is really just a start. There are several possibilities going forward, the main one being to expand the model to a global level. The main restriction on this front is a lack of regularly updated, geocoded conflict data. Improvements in automated geocoding for event datasets like Phoenix may solve this issue, as well as allow for the inclusion of other event types beyond violent conflict events. Besides that, additional exploration in determining the ideal resolution when using a raster approach, or comparing a raster approach to one using administrative districts as the unit of observation would also be useful. There are many possibilities that remain for how a geographically disaggregated approach can be applied.
4
Forecasting Multi-Actor Civil Conflicts

4.1 Introduction

In the spring of 2011, protests against the Syrian government broke out in Syria in conjunction with the wider Arab Spring movement taking place across the Middle East and Northern Africa. The protests in Syria quickly morphed into armed opposition in the face of violent government crackdowns and the country eventually plunged into a full-blown civil war. One of the great difficulties for outside parties seeking to bring an end to the war has been the complicated nature of the fighting, with multiple armed groups fighting both the government and each other. This also creates difficulties for researchers interested in forecasting the direction of the conflict with quantitative models.

The ability to model and forecast a conflict like the one in Syria is important for policymakers, NGOs, and others who have a stake in managing or resolving the conflict. Large numbers of competing actors complicate any negotiated settlement. The longer the violence goes on, the more the humanitarian situation worsens. As a result of the Syrian conflict, for example, nearly 4 million refugees have fled the
country, creating a humanitarian crisis. While improved conflict forecasting will not resolve these problems, it can help decision makers. Advance warning of an increase in fighting would help aid groups prepare for a potential increase in refugees, while knowledge of whether the conflict is expected to intensify or quiet may help mediators in deciding when to press harder for a negotiated settlement. Conflict forecasts have great potential as a tool for conflict management and mediation.

Beyond giving a general year-to-year risk of civil conflict, country level models are not much use here. They do not provide a way to model the dynamics of a conflict. Neither are they well-suited for handling heterogeneous actors. Conflicts like the one in Syria require a model that can generate subannual forecasts for a set of heterogeneous actors. There is some precedent for this. Early work in this area dealt primarily with the conflicts in the former-Yugoslavia states. Pevehouse & Goldstein (1999), for example, use Vector Autoregression (VAR) to predict that the NATO bombing of Serbia would not change Serbian policy towards Kosovo (a prediction that turned out to be incorrect). The authors do not perform explicit forecasting, but base their predictions on significance tests on the relationship between international actions against Serbia and Serbian actions against Kosovo. Work by Phil Schrodt applied Hidden Markov models to forecast "high conflict" and "low conflict" weeks in the Balkans (Schrodt 2006), as well as the Levant (Schrodt 1997).

In this paper, I focus on applying time series methods, specifically Bayesian Vector Autoregression (BVAR), to conflict forecasting. I build on previous work that has used BVAR for conflict forecasting in several ways. First, these studies have considered only cases of interstate interactions involving two actors, such as Israel in Palestine by Brandt, Freeman & Schrodt (2011) and China and Taiwan by Brandt et al. (2013). With cases of civil conflict involving multiple actors becom-
ing, such as Libya and Syria, increasingly becoming of interest to policymakers, it is important to examine the utility of methods like BVAR for modeling these conflicts. Unlike the previous analyses of interstate tensions, these conflicts often involve more than two actors. Second, previous studies only consider conflicts with a very long history. Very few conflicts are going to have multiple decades worth of data, as the Israeli-Palestine and China-Taiwan conflicts do. Practically speaking, it would be best to begin generating forecasts of a conflict as soon as possible after it begins. Policymakers may not have the luxury of waiting a few years for a sufficient number of observations to fit a model. Currently, it is unclear how accurate this type of model may be when applied to a recent, ongoing conflict with a limited number of observations to draw from. In this case, a multivariate time series model like BVAR may produce forecasts that are no more accurate than univariate forecasts due to the short history, or, in the worst case scenario, there may not be enough data to even fit a model. Finally, the baseline comparisons in previous work have been against basic univariate autoregressive (AR) models, ignoring the fact that a univariate autoregressive integrated moving average (ARIMA) model is more appropriate to test. I incorporate ARIMA models in the analysis to serve as a baseline comparison.

To demonstrate the applicability of this method on multi-actor civil conflicts with a short time span, I generate forecasts for the ongoing Syrian Civil War. This is an area well-suited for an BVAR model. First of all, the conflict is relatively recent, ongoing, and of high degree of interest to a number of government and nongovernment groups for a variety of reasons. Additionally, the conflict involves several competing actors whose actions are contingent on the actions of the other actors. How different governments, IGOs, and NGOs act to mitigate the violence is dependent on both the intensity of the violence and who the perpetrators are. The model developed in this section will generate forecasts with a focus on predicting
conflict intensity between armed actors, both state and non-state, involved in the Syrian Civil War, as well as forecasts of the intensity of violence against civilian targets by these armed actors. If successful, this approach should be applicable to other regions experiencing ongoing violence involving multiple actors.

4.2 Vector Autoregression

Vector autoregression is a multiequation model in which the number of equations are equal to the number of variables, and the current value of each variable is explained by lagged values of both it and the other variables in the model. This type of model is best-suited for situations where there is reason to believe that there are linear interdependencies between multiple time series (Karlsson 2013). A structural VAR, which accounts for contemporaneous relationships among the dependent variables, can be written as

\[
A_0 y_t = c_0 + y_{t-1} A_1 + \ldots y_{t-p} A_p + u_t, \quad t = 1, 2, \ldots, T \quad (4.1)
\]

where \(A_0\) is an \(m \times m\) contemporaneous coefficient matrix, \(y_t\) is a \(k \times m\) vector of observations of the dependent variable, \(A_i\) (for \(i = 1, \ldots, p\) with \(p\) being the maximum lag) is a \(k \times k\) matrix of coefficients, and \(u_t\) is a \(k \times 1\) vector of error terms, or structural shocks in this case. I use a reduced form of this equation by premultiplying \(A_0^{-1}\) throughout to get

\[
y_t = c + y_{t-1} A_1 + \ldots y_{t-p} A_p + u_t, \quad t = 1, 2, \ldots, T \quad (4.2)
\]

where \(c\) is a \(k \times 1\) vector of intercepts, \(y_t\) is a \(1 \times m\) vector of observations at time \(t\), in this case for the behavior of the actor dyads in the system, \(A_t\) is the \(m \times m\) coefficient matrix for the \(lth\) lag, \(p\) is the maximum number of lags, and \(u_t\)
are the reduced form residuals. This is essentially an ordinary least squares (OLS) regression estimated using the lagged values of each variable, including the one being considered.

This type of model is potentially an ideal fit for forecasting the behavior of actors in the international system. We might expect that the behavior of one actor towards another is affected by prior behavior of the second actor, which in turn may be affected by prior behavior of the original actor. For example, Syrian rebels may stage an attack on government forces in Damascus in response to government bombings of rebel positions, which may in turn result in more bombings by the government in response. This type of behavior is important to take into account. By taking into account the behavior of two or more actors in a group whose behaviors we expect to be interdependent in some way, we can produce better forecasts.

Unfortunately, producing forecasts with VAR is problematic. The multiple equations used by the model generates a large number of parameters, making it susceptible to overparameterization. The number of parameters is equal to $n \times n \times p + n$, where $n$ is the number of variables and $p$ is the number of lags. As an example, a study of directed dyadic behavior involving Syria and Syrian rebels, along with their actions towards civilian actors, might involve 6 unique variables. Using a four-lag VAR, the model has to estimate 150 parameters. This makes it difficult to generate forecasts with any degree of certainty. Early work applying VAR to international conflict typically relied on Granger causality tests alone due to issues with overparameterization (Goldstein & Pevehouse 1997, Pevehouse & Goldstein 1999). Granger causality is a simple test of whether some series $X$ is a better predictor of $Y$ than past values of $Y$. As such, no actual predictions were made beyond assumptions that an increase in conflict by one actor would affect the behavior of another actor in some way based on the sign and statistical significance...
of the associated coefficient.

As suggested by Brandt & Freeman (2006) and Brandt & Freeman (2009), employing a Bayesian approach can resolve the issue of overparameterization. A properly applied prior can reduce the uncertainty of the estimates. In this case, I use a prior developed by Sims & Zha (1998), which as been used in previous studies on forecasting international conflict (Brandt, Freeman & Schrodt 2011, Brandt et al. 2013). The Sims-Zha prior uses a set of inexact restrictions to shrink lagged coefficients towards zero. Specifically, it assumes that the prior mean of the higher order lags is close to 0, giving greater weight to the lower order lags. Essentially, the prior is a good fit for any situation where one expects recent behavior to be the best explainer of current behavior. This intuitively fits with the expectation that an actor’s actions will follow from their own recent behavior and take into account the recent behavior of other actors involved in the system.

The Sims-Zha prior can be described as follows. We first place the elements of $A_0$ into $a_0$, where the columns of $A_0$ are stacked in column-major order for each equation. The remaining parameters pertaining to the lag dynamics are denoted as $A_+$, an $(m^2p+1)$ matrix stacking the lag coefficients and the constants for each equation. The columns of $A_+$ are then placed into a vector $a_+$ in column major order. The prior $\pi(a)$ is then

$$\pi(a) = \pi(a_0)\Phi(\tilde{a}_+, \Psi), \quad (4.3)$$

where $\tilde{a}_+$ indicates the mean parameters of the prior density for $a_+$, and $\Phi(\tilde{a}_+, \Psi)$ is a multivariate normal with mean $\tilde{a}_+$ and covariance $\Psi$.

The prior in this case is set on the structural parameters of the model with an eye towards scaling the standard deviations of the regression coefficients. The
hyperparameters $\lambda_0$, $\lambda_1$, and $\lambda_3$, for example, serve to scale the prior covariance matrix. Table 4.1 provides information on the associated hyperparameters. $\lambda_4$ is intended to handle the prior variance of the constant, while $\lambda_5$ does the same for the exogenous variables. $\mu_5$ and $\mu_6$ are intended to handle two respective sets of dummy observations added to the model to handle unit roots and trends.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Range</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_0$</td>
<td>[0,1]</td>
<td>Overall scale of the error covariance matrix</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>$&gt;$ 0</td>
<td>SD about $A_1$ (persistence)</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>$=$ 1</td>
<td>Weight of own lag versus other lags</td>
</tr>
<tr>
<td>$\lambda_3$</td>
<td>$&gt;$ 0</td>
<td>Lag decay</td>
</tr>
<tr>
<td>$\lambda_4$</td>
<td>$\geq$ 0</td>
<td>Scale of SD of intercept</td>
</tr>
<tr>
<td>$\lambda_5$</td>
<td>$\geq$ 0</td>
<td>Scale of SD of exogenous variables coefficients</td>
</tr>
<tr>
<td>$\mu_5$</td>
<td>$\geq$ 0</td>
<td>Sum of autoregressive coefficients component</td>
</tr>
<tr>
<td>$\mu_6$</td>
<td>$\geq$ 0</td>
<td>Correlation of coefficients/initial condition component</td>
</tr>
</tbody>
</table>

### 4.3 Research Design

This paper considers weekly time series of event counts for actors involved in the Syrian Civil War. While previous work has typically considered time series at a monthly level, I focus on a weekly level primarily due to data constraints. As of now, the number of observations at a monthly level of aggregation are not sufficient to fit a BVAR model. The purpose of these series is to capture the level of conflict between combatants, as well as combatants and noncombatants. I develop three models to focus on three major groupings of actors. The first looks at interactions between the Syrian government, Syrian rebels, and noncombatant civilians, the second at interactions between the Syrian government, the Islamic State (IS), and noncombatant civilians, and the third at interactions between Syrian rebels, the Islamic State, and noncombatants civilians. The decision to model
these three grouping separately is due to potential overparameterization from including so many series into one VAR and the uncertainty of any resulting forecasts.

For each grouping, I do three types of models: BVAR with an informed prior, BVAR with a flat prior, and an autoregressive integrated moving average (ARIMA) model. The latter two models serve as baselines. The BVAR model with an informed prior should perform better than one with a flat prior assuming the prior is suited to the problem at hand. The ARIMA model serves as a baseline to compare the multivariate time series against more generally. I perform tuning for each method to optimize their performance. For BVAR, this involves determining the best values of the hyperparameters, while, for ARIMA, this means determining the best values for the values for $p$, $d$, and $q$, parameters which govern the autoregressive, integrated, and moving average components of the model respectively. The specification of the model is typically represented as ARIMA$(p, d, q)$. If neither of the multivariate models outperform an ARIMA model, then one is better off using a more straightforward and computationally cheaper univariate model.

A key component of any predictive modeling exercise is assessing the accuracy of the final model. The danger that every predictive model faces is overfitting. A model suffers from overfitting when it predicts the data used to create it with a high degree of accuracy, but performs poorly when used to predict cases that it has not seen. In order to determine how well the model generalizes to unseen data, I use a type of assessment known as cross-validation (Stone 1974). The basic form of cross-validation involves dividing the data into two samples: one as data used to train the model and the other as holdout data that the trained model attempts to predict. This can be extended using $k$-fold cross-validation, which entails dividing the data into $k$ partitions and then training and testing $k - 1$ models using a different partition as the holdout sample each time.
Data for cross-validation is typically split randomly. Because of the serial nature of time series modeling though, I split the data using a date as a cutpoint. In this case, I use the last 10 weeks worth of data, roughly the end of October 2015 to the end of December 2015, as the testing sample, and the remainder of the data going back to mid-June 2014 as the training sample. Rather than predict all 10 weeks at once, I use rolling predictions, an approach previously used in both economics (Bhardwaj & Swanson 2006) and political science (Brandt et al. 2013), to get a more robust assessment of how the models are performing. A rolling prediction in this case means that I begin by fitting the model on the training data, then testing it on the first week of the testing data. I then refit the model, this time including the first week of the testing data when fitting the model, and predict on the second week of the testing data. This is done iteratively until forecasts have been produced for all 10 weeks of the testing data.

The main advantage of this approach is that it provides a more robust assessment of how the model is performing. The model is fit multiple times on slightly different training samples, and multiple week-to-week predictions are made. This, of course, only tells us how well the model does at predicting one week in advance. Predicting multiple weeks in advance would require either using a different level of aggregation, or using forecast values of intervening weeks. For example, predicting week $y_{t+2}$ would require us to first forecast the value of $y_{t+1}$, then incorporate that value into the model to forecast $y_{t+2}$. If the forecast for $y_{t+1}$ is poor, then the forecast for $y_{t+2}$ is likely to be poor as well. We may assume then that a model that predicts well one week in advance, will likely outperform a model that performs poorly one week in advance when applied to forecasting multiple weeks ahead.

Because there are a variety of options for assessing the accuracy of models
that produce continuous predictions, I use multiple methods as a robustness check against metrics that may favor one model over another. The first two are root mean square error (RMSE) and mean average error (MAE). These are both very popular metrics for assessing out-of-sample accuracy for models that produce point forecasts. However, as point metrics, they are unable to take into account the forecast densities produced by BVAR beyond assessing the mean or median forecast. Additionally, previous research has shown them to be sensitive to be sensitive to outliers (Armstrong & Collopy 1992).

To better measure the accuracy of the forecast densities produced by the BVAR models, I use the Continuous Rank Probability Score (CRPS) metric as described by Hersbach (2000), which defines it as follows: let $x$ denote the value of the forecast for a variable, while $x_a$ denotes the observed value, and their PDFs denoted by $P(x)$ and $P_a(x)$ respectively for

$$CRPS(P, x_a) = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 dx. \quad (4.4)$$

with $P$ and $P_a$ being the cumulative distributions

$$P(x) = \int_{-\infty}^{x} \rho(y)dy \quad (4.5)$$

and

$$P_a(x) = H(x - x_a) \quad (4.6)$$

where $H$ is the Heaviside function

$$H(x) = \begin{cases} 
0 & \text{for } x < 0 \\
1 & \text{for } x \geq 0 
\end{cases}. \quad (4.7)$$

The CRPS is then the difference between the totals of the forecast and observed cumulative distributions. Because we want the distributions to match as closely
as possible, lower scores suggest more accurate forecasts.

The CRPS provides a useful means of assessing the forecast density both in terms of how well it matches the observations and how concentrated the forecast distribution is, or the level of uncertainty of the forecast in other words. Figure 4.1 illustrates the CRPS. The graph on the left shows the distribution of predictions, in this case of temperature, in relation to the observed value. The CRPS is then the difference between the two shaded areas in the right hand graph. It is ultimately measuring how centered the distribution of predictions is around the observed value, so lower values indicate better performance.

The CRPS is typically averaged over a number of forecasts, \( k \), with weights \( w_k \) set by the forecaster:

\[
\text{CRPS} = \sum_k w_k \text{CRPS}(P_k, x^k). \quad (4.8)
\]

In addition to these standard measures, I include a simpler measure of whether the model correctly forecast an increase or a decrease in the level of conflict for the
next period. I assess this as mean accuracy. I do this for a combination of reasons. First of all, policymakers may not necessarily be interested in the total number of conflict events occurring, but the direction in which the conflict is headed. Because the data is based on news reports, potential underreporting, a real possibility in conflict zones, means that we may not be capturing the actual number of events occurring. There is a deal of uncertainty to these forecasts that makes point estimates unreliable. The CRPS does a better job than RMSE or MAE of assessing this uncertainty, but may be difficult to communicate to a government or NGO audience from a qualitative background. Calculating how well the model did on average at predicting an increase or decrease in the level of conflict is intuitively easier to understand for a non-technical audience.

4.4 Data

This analysis uses event data. Event data records events that occur in the international system. These events are defined as interactions between two politically relevant actors, such as states, militaries, or NGOs. Events are recorded at a daily resolution in a who-did-what-to-whom format. For example, an attack by the Syrian government on Syrian protestors would be recorded as an event with the Syrian government as the source, Syrian protestors as the target, and act of repression as the type of event. The key pieces of information are the source of the event, the target of the event, and the type of event. Additional information is often included on the location of the event and more general typological categories for the actors and the event type.

I use event data collected by the Phoenix Data Project for this study (Schrodt, Beiler & Idris 2014). Phoenix generates data using an automated coding pro-
gram known as PETRARCH (Norris 2016) to parse and code text of news reports scraped from RSS feeds of a white list of news sources. Because events are typically reported by multiple sources, the project uses a one-a-day filter that permits only one event of a given type between two actors in a specific geographic area per day. For example, multiple reports on the Syrian government bombing rebel position in Aleppo and Homs would generate two events for that day, one for the bombing of Aleppo and one for the bombing of Homs. Phoenix uses the CAMEO coding scheme (Gerner et al. 2002), which has several dozen unique categories related to conflict and mediation events. The actor dictionaries used to code relevant actors are regularly updated and take into account the emergence of new, politically relevant actors like the Islamic State group in the Middle East and Northern Africa. The resulting dataset is a series of dyad-event observations at a daily level.

Ease of aggregation is one of the main advantages of using event data. Rather than using each of the several dozen event categories, I use a more general categorization provided by Phoenix known as pentaclass. The pentaclass variable groups event types into five general categories: neutral\(^1\), verbal cooperation, material cooperation, verbal conflict, and material conflict. Based on these categories, I generate weekly counts for dyads of interest, excluding the neutral category. This level of aggregation is standard in studies using event data for forecasting purposes (Yonamine 2011).

I look at four types of actors for this analysis. The first are Syrian civilians, which cover any events in Syria where the target type is coded as CVL. Actors related to the Syrian government, such as the Syrian military or police, are grouped together as a unique category. I also include Syrian allies providing direct military support in the conflict, such as Hezbollah and more recently Russia, in the Syrian

\(^1\)these are typically general public statements
government grouping. The remaining categories are related to non-government militarized actors in Syria. One is a grouping of rebel groups in Syria, coded based on events where REB is specified as the source or target actor type. The dataset does not identify specific groups, so this includes both Islamist and secular rebels. The final actor category identifies events involving the Islamic State that take place in Syria.

I generate weekly directed time series for counts of verbal cooperation, material cooperation, verbal conflict, and material conflict for each combination of actors, with the exception of dyads involving civilians. In this case, I generate material cooperation and material conflict series for events where civilians are the target of one of the other three actor types. To reduce the number of series that I have to work with, I subtract the material conflict series from its associated material cooperation series for each dyad to generate a series measuring net cooperation. I do the same for the verbal conflict and verbal cooperation series. In the resulting series, positive values indicate that there was more cooperation occurring in a given week, while negative values indicate that there was more conflict overall. Table 4.2 provides a summary of the dyads of interest and their associated series, while Figures 4.2 - 4.4 display material plots for the relevant dyads. There are a total of 15 series, each extending from mid-June of 2014 to the end of 2015. For the analysis, I consider data at a weekly level of aggregation. This generates a total of 80 observations in each series, 70 of which make up the training data and 10 of which make up the testing data.
Table 4.2: Actor Dyads and Series

<table>
<thead>
<tr>
<th>Dyad</th>
<th>Identifier</th>
<th>Series</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syria -&gt; Rebel</td>
<td>S2R</td>
<td>Verbal, Material</td>
</tr>
<tr>
<td>Rebel -&gt; Syria</td>
<td>R2S</td>
<td>Verbal, Material</td>
</tr>
<tr>
<td>Syria -&gt; IS</td>
<td>S2I</td>
<td>Verbal, Material</td>
</tr>
<tr>
<td>IS -&gt; Syria</td>
<td>I2S</td>
<td>Verbal, Material</td>
</tr>
<tr>
<td>Rebel -&gt; IS</td>
<td>R2I</td>
<td>Verbal, Material</td>
</tr>
<tr>
<td>IS -&gt; Rebel</td>
<td>I2R</td>
<td>Verbal, Material</td>
</tr>
<tr>
<td>Syria -&gt; Civilian</td>
<td>S2C</td>
<td>Material</td>
</tr>
<tr>
<td>Rebel -&gt; Civilian</td>
<td>R2C</td>
<td>Material</td>
</tr>
<tr>
<td>IS -&gt; Civilian</td>
<td>I2C</td>
<td>Material</td>
</tr>
</tbody>
</table>

Figure 4.2: IS, Rebel, Civilian Material Events
Figure 4.3: IS, Syria, Civilian Material Events
4.5 Analysis and Results

I fit univariate models, as well as Bayesian models with and without an informed prior on three separate sets of data: interactions between Syria and IS, IS and non-IS rebels, and Syria and non-IS rebels, as well as their actions towards civilians. This means each model is built on 6 time series. 4 series account for directed material events and directed verbal events between armed actors and an additional
For the Bayesian models, I begin by tuning the hyperparameters associated with the Sims-Zha prior in order to determine the set of hyperparameter values for each model that yields the best accuracy. I begin by generating a sets of values to test for each hyperparameter, which are displayed in Table 4.3. I then fit models for each combination of hyperparameter values. The models are fit on the training data. I then generate in-sample fit measures for each model based on log likelihood scores, log marginal density scores, and RMSE. I determine the final set of hyperparameter values for each model based on the set that achieves the best overall accuracy across these in-sample fit measures. I use a lag length of 3 when performing these tests. Figure 4.5 displays the RMSE for a model of the IS-Syria data using different combinations of the hyperparameters $\lambda_0$, $\lambda_1$, $\lambda_3$, and $\lambda_4$, with the remaining hyperparameters held constant. In this case, the best model has a $\lambda_4$ value of 0.5 and a $\lambda_0$ value of 1.0, while the values for $\lambda_1$ and $\lambda_3$ do not appear to make much of a difference. Each model is fit with a lag value of 3. There were issues fitting the posterior when testing with lag values higher than 3. This suggests that there may be issues with overparameterization despite the use of a Bayesian prior.

After determining the best set of hyperparameters for each model, displayed in Table 4.4, I proceed to fit informed BVAR and flat BVAR models for each actor grouping, using a lag length of 3 in both cases. The flat BVAR models serve as a check as to whether an informed prior improves a multivariate model. In addition, I fit univariate ARIMA models for each series, with lag lengths determined by the AIC. The ARIMA models serve as a check as to whether multivariate models are a better forecasting tool than univariate models. For the ARIMA models, I test
values for $p$ and $q$, the autoregressive and moving average terms respectively, from 0 to a max of 5. I set $d$, the order of differencing, based on a KPSS test that determines whether the data is stationary. If non-stationary, one order of differencing is applied. Another KPSS test is applied to determine if the data is then stationary. If not, the next order is tested and so on until the KPSS test does not reject the null hypothesis of stationarity. Similar to tuning the hyperparameters in the BVAR model, I fit a model for each combination of parameters values (e.g. ARIMA(1, 0, 0), ARIMA(1, 0, 1), ARIMA(1, 0, 2), etc.). The best model for each series is chosen based on the lowest AIC. The final parameter values vary by series.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>IS-Syria</th>
<th>Rebel-Syria</th>
<th>IS-Rebel</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_0$</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\lambda_3$</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\lambda_4$</td>
<td>0.5</td>
<td>0.15</td>
<td>0.5</td>
</tr>
<tr>
<td>$\lambda_5$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\mu_5$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\mu_6$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
used to generate the forecast density. Similar forecast densities are created using 5000 unconditional draws from the univariate ARIMA models. Figures 4.6–4.8 display plots of the forecasts compared to the observed data. The forecasts include 68% credible intervals (corresponding to one standard deviation) to account for uncertainty. Tables 4.5–4.7 show accuracy metrics for each model by series, with the best results in bold (except in cases where there are ties). With the exception of the CRPS scores, the accuracy metrics were generated based on the mean forecasts. Negative values on the y-axes of the plots indicate net conflict, while positive values indicate net cooperation.

The forecasts for the Syria-IS models and their associated accuracy metrics are
<table>
<thead>
<tr>
<th>Year-Week</th>
<th>Net Cooperation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-43</td>
<td>10</td>
</tr>
<tr>
<td>2015-45</td>
<td>10</td>
</tr>
<tr>
<td>2015-47</td>
<td>10</td>
</tr>
<tr>
<td>2015-49</td>
<td>10</td>
</tr>
<tr>
<td>2015-51</td>
<td>10</td>
</tr>
</tbody>
</table>

**Figure 4.6: IS, Syria, Civilian Material Forecasts**

displayed in Figure 4.6 and Table 4.5 respectively. The results are mixed across all metrics, with no model being best overall. The univariate models actually perform best overall when judged solely by the CRPS metric, which is ostensibly more appropriate when judging forecast densities. Differences in the univariate and Bayesian models are quite apparent in Figure 4.6. The univariate models tend towards more stable, consistent forecasts across time, while the BVAR forecasts are prone to sharp increases and decreases. These changes in the BVAR forecasts,
Table 4.5: IS-Syria Accuracy Metrics

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Direction</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>I2S-M</td>
<td>S2I-M</td>
<td>S2C-M</td>
</tr>
<tr>
<td>BVAR informed</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.70</td>
</tr>
<tr>
<td>BVAR flat</td>
<td><strong>0.90</strong></td>
<td>0.60</td>
<td>0.40</td>
<td><strong>0.90</strong></td>
</tr>
<tr>
<td>Univariate</td>
<td>0.50</td>
<td>0.70</td>
<td><strong>0.80</strong></td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BVAR informed</td>
<td><strong>8.73</strong></td>
<td>15.36</td>
<td><strong>0.68</strong></td>
<td>3.65</td>
</tr>
<tr>
<td>BVAR flat</td>
<td>9.20</td>
<td>15.14</td>
<td>1.06</td>
<td>4.92</td>
</tr>
<tr>
<td>Univariate</td>
<td>9.50</td>
<td><strong>13.02</strong></td>
<td>0.60</td>
<td><strong>3.03</strong></td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BVAR informed</td>
<td><strong>6.78</strong></td>
<td>11.74</td>
<td><strong>0.51</strong></td>
<td>2.72</td>
</tr>
<tr>
<td>BVAR flat</td>
<td>7.79</td>
<td>11.65</td>
<td>0.82</td>
<td>3.87</td>
</tr>
<tr>
<td>Univariate</td>
<td>8.13</td>
<td><strong>9.29</strong></td>
<td>0.64</td>
<td><strong>2.11</strong></td>
</tr>
<tr>
<td></td>
<td>CRPS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BVAR informed</td>
<td><strong>5.31</strong></td>
<td>8.60</td>
<td>0.60</td>
<td>1.96</td>
</tr>
<tr>
<td>BVAR flat</td>
<td>5.35</td>
<td>8.71</td>
<td>0.66</td>
<td>3.08</td>
</tr>
<tr>
<td>Univariate</td>
<td>5.51</td>
<td><strong>6.90</strong></td>
<td><strong>0.52</strong></td>
<td><strong>1.60</strong></td>
</tr>
</tbody>
</table>

particularly for the models with an informed prior, tend to chase the trend of the data, missing sharp upturns and downturns by one period.

Figure 4.7 and Table 4.6 display forecasts for interactions involving Syria and non-IS rebels, as well as the forecasts’ associated accuracy metrics. It is important to note in this case that there are no instances of attacks by non-IS rebels against civilian targets during the test period. For the rest of the series, the univariate models deliver the best performance across all metrics except the directional shift of the forecasts, in which case it does well for the civilian series, but poorly for the series involving Syria and rebels. The univariate models perform worse than a coin-flip when attempting to predict the overall direction the series will go in the next period for the Rebel-Syria and Syria-Rebel directed dyads.

The results, displayed in Figure 4.8 and Table 4.7 for the final set of series involving IS and non-IS rebels, are not particularly illuminating given that there
### Table 4.6: Syria-Rebel Accuracy Metrics

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Direction</th>
<th>R2S-M</th>
<th>S2R-M</th>
<th>R2C-M</th>
<th>S2C-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVAR informed</td>
<td>0.80</td>
<td>0.50</td>
<td>0.60</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>BVAR flat</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>Univariate</td>
<td>0.70</td>
<td>0.70</td>
<td>1.00</td>
<td>0.70</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RMSE</th>
<th>BVAR informed</th>
<th>2.93</th>
<th>6.13</th>
<th>0.19</th>
<th>3.76</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVAR flat</td>
<td>2.38</td>
<td>5.69</td>
<td>0.72</td>
<td>4.47</td>
<td></td>
</tr>
<tr>
<td>Univariate</td>
<td>2.09</td>
<td>6.02</td>
<td>0.13</td>
<td>3.06</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MAE</th>
<th>BVAR informed</th>
<th>2.19</th>
<th>4.82</th>
<th>0.17</th>
<th>2.43</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVAR flat</td>
<td>2.15</td>
<td>4.03</td>
<td>0.55</td>
<td>3.40</td>
<td></td>
</tr>
<tr>
<td>Univariate</td>
<td>1.77</td>
<td>4.97</td>
<td>0.13</td>
<td>2.20</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CRPS</th>
<th>BVAR informed</th>
<th>1.69</th>
<th>3.60</th>
<th>0.28</th>
<th>1.93</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVAR flat</td>
<td>1.39</td>
<td>3.21</td>
<td>0.40</td>
<td>2.59</td>
<td></td>
</tr>
<tr>
<td>Univariate</td>
<td>1.23</td>
<td>3.76</td>
<td>0.17</td>
<td>1.64</td>
<td></td>
</tr>
</tbody>
</table>

### Table 4.7: IS-Rebel Accuracy Metrics

<table>
<thead>
<tr>
<th>Model</th>
<th>I2R-M</th>
<th>R2I-M</th>
<th>I2C-M</th>
<th>R2C-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVAR informed</td>
<td>0.90</td>
<td>0.60</td>
<td>0.50</td>
<td>0.60</td>
</tr>
<tr>
<td>BVAR flat</td>
<td>0.30</td>
<td>0.30</td>
<td>0.60</td>
<td>0.50</td>
</tr>
<tr>
<td>Univariate</td>
<td>0.80</td>
<td>0.50</td>
<td>0.80</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RMSE</th>
<th>BVAR informed</th>
<th>0.81</th>
<th>1.54</th>
<th>0.86</th>
<th>0.11</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVAR flat</td>
<td>1.11</td>
<td>1.44</td>
<td>0.89</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>Univariate</td>
<td>0.92</td>
<td>2.16</td>
<td>0.69</td>
<td>0.13</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MAE</th>
<th>BVAR informed</th>
<th>0.63</th>
<th>1.24</th>
<th>0.55</th>
<th>0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVAR flat</td>
<td>0.97</td>
<td>1.19</td>
<td>0.76</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Univariate</td>
<td>0.85</td>
<td>1.82</td>
<td>0.64</td>
<td>0.13</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CRPS</th>
<th>BVAR informed</th>
<th>0.69</th>
<th>0.92</th>
<th>0.62</th>
<th>0.23</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVAR flat</td>
<td>0.69</td>
<td>0.81</td>
<td>0.55</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>Univariate</td>
<td>0.61</td>
<td>1.26</td>
<td>0.52</td>
<td>0.17</td>
<td></td>
</tr>
</tbody>
</table>
were little to no events involving these actors during the test period. However, one interesting aspect of the forecasts is that the BVAR models generate much tighter forecast densities than the univariate models when the series is relatively flat.

The major question is of course whether a Bayesian multivariate modeling approach is better than relying on simple univariate models. Judging by multiple accuracy metrics across multiple time series, the Bayesian approach rarely out-
performs a univariate approach. The results are too mixed to justify using an exclusively multivariate approach. There is some consistency across metrics that particular models are better for particular series. For example, a BVAR with an informed prior is the best forecaster of material events in the IS-Syria directed dyad when judged by RMSE, MAE, and CRPS. The same is true for univariate models when forecasting the associated Syria-IS directed dyad. This lack of accuracy, coupled with the fact that Bayesian models were unable to fit when using lag
orders greater than 3, suggests that BVAR models are not a superior alternative to univariate analysis when forecasting multi-actor conflicts with a short history.

4.6 Conclusion

Overall, there is little evidence to suggest that BVAR models are clearly superior in this case. Instead, the results are mixed. While the BVAR models perform marginally better than univariate counterparts for some series, they perform worse for others. Additionally, these models suffer from overparameterization despite applying Bayesian methods to remedy this issue. This limits the level of aggregation to a weekly level, which may be less appropriate than a monthly level in some cases. Complex BVAR models are also much more expensive in computational terms than univariate models, requiring more processing time to fit and generate forecasts, in addition to the need to properly tune the hyperparameters of the prior beforehand.

If basic ARIMA models perform comparatively well to complex multivariate models, this suggests that the focus should be on univariate methods. For long term conflicts at least, variations of BVAR appear to be the best avenue of improving forecasts (Brandt et al. 2013), but future work on forecasting recent civil conflicts involving multiple actors should explore existing univariate methods, such as ARFIMA (Baillie, Chung & Tieslau 1996, Bhardwaj & Swanson 2006), or a combination of univariate methods for some series and multivariate for others. The trends of conflicts like the Syrian Civil War may be difficult, or ultimately impossible to accurately predict using statistical models, but the real world costs of these conflicts makes continued research in this area important.
In general, a predictive modeling exercise can be divided into a few stages. First, one defines the problem: what are we predicting? From there, the exercise proceeds to choosing predictors, then collecting data. Next comes assembling the model itself and generating predictions. In the final stage, one optimizes the model by looking for different ways to improve model accuracy through things like hyperparameter tuning or feature engineering.

If one were to apply these stages of development to predictive modeling as a subfield within political science, we are somewhere between the main modeling stage and the optimization stage. We now have a decent grasp of the main tenets of predictive modeling based on work in other fields like meteorology and machine learning that have a much longer history with prediction. Adopting these best practices is still a work in progress though. For example, I reviewed a paper last year that applied predictive modeling to transnational terrorism risk. While the paper had a clear idea of the problem and a strong set of predictors, it made several mistakes in the modeling stage, mistakes that would be considered fairly basic in a field like machine learning. On the other hand, political scientists who have been at the forefront of predictive modeling in our field are well into exploring how to better optimize and tailor models to both old and new problems. Mike Ward,
for example, has done extensive work on applying ensemble methods to improve forecasts of existing models (Montgomery, Hollenbach & Ward 2012, Beger, Dorff & Ward 2014). We are not at the point where predictive modeling is a standard practice in political science, but it is heading in that direction.

Early work in predictive modeling in political science grew out of groups like the Political Instability Task Force (PITF), whose work was primarily applied rather than academic. The outcomes of interest were generally major international events, like coups and civil wars. This was with good reason. These events have a major destabilizing effect on the states that experience them and create a multitude of problems, such as refugees fleeing into neighboring countries, that policymakers, NGOs, and others have to deal with. Early warning systems are potentially valuable tools that allow them to anticipate and prepare for these destabilizing events. The private sector has begun to realize the potential of these tools as well. Mastercard, for example, recently filed a patent application describing a system to forecast political events using data on credit card transactions (Unser, Bernard & Malgatti 2014). Work on forecasting political events will continue regardless of whether academia maintains its interest.

This dissertation falls in the optimization stage. It takes as a given that predictive modeling has been established as a practice in political science and explores ways to improve and expand the existing models we have with a focus on developing models that can be used in a real world setting. In the first article, I consider the existing country level model approach as popularized in work by PITF. With mass atrocities as the outcome of interest, I explore ways of improving this approach. Next, I disaggregate the model spatially. Instead of predicting mass atrocities at a state level, I break the data into a raster, or a grid of longitude and latitude coordinates. This allows for subnational forecasts of violence and incorporation
of geographic data. In the last article, I explore the use of time series methods, particularly multivariate Bayesian Vector Autoregression (BVAR), for modeling conflict dynamics in the Syrian Civil War.

The first article focuses on how best to improve a country-year style predictive model. The outcome in this case is the onset of a mass atrocity within a country. I consider three ways of improving the model. The first is to apply some new method to the problem. In this case, I compare a commonly used logistic regression to two methods borrowed from machine learning, Random Forest and neural networks, both of which are generally considered better for prediction than logistic regression. I find that there is no real improvement using either method. Next, I apply a sampling method known as case-control to determine whether a more balanced sample of training data improves predictive accuracy. As with the new methods, there is no discernible improvement in model accuracy. The final avenue of improvement I explore is feature engineering. Feature engineering is the practice of transforming existing data in some way to create a new predictor. I apply feature engineering to event data to determine to what extent the level of aggregation affects results. I find that the manner in which event data is aggregated as a predictor can have a significant effect on accuracy, suggesting that, of all the potential avenues of improvement, this is one that we should continue to explore.

There are several implications of the first article. The first, and one of the more surprising, is that newer methods do not significantly improve accuracy. This is contrary to how these methods are viewed in machine learning. Logistic regression is typically used as the punching bag when demonstrating how well some new method performs. The reason for this divergence in findings is likely a result of the data itself. Methods like neural networks are very adept at accounting for com-
plex, nonlinear relationships, such as those encountered when dealing with text data. The issue in this case though is not one of nonlinearity or complexity, but of noise. The relationship between predictors like infant mortality or ethnic characteristics of leadership, and an outcome like mass violence is noisy. There is no clear value of these predictors at which mass violence is definitely going to happen. This noisiness, coupled with the fact that these events are very rare, makes it difficult for any method to get leverage on the problem. This same problem is likely why case-control sampling yields no improvement either. The focus should instead be on improving existing predictors using feature engineering. The results of the analysis on feature engineering with event data demonstrates that model accuracy can be highly dependent on the format of the predictors. Better tailoring data to the problem at hand is a strong way of improving our ability to forecast political events.

In the second article, I explore a raster approach as a potential means of generating subnational forecasts. I forecast mass atrocities in Africa at an approximately 50x50 kilometer level by dividing the continent up into a grid based on latitude and longitude. A spatially disaggregated approach allows me to incorporate geographic data as predictors in addition to predictors used in the preceding article to forecast mass violence against civilians at a national level. I find that it is possible to generate relatively accurate forecasts of violence at a subnational level. Additionally, the predictors based on geographic data prove to be quite useful for prediction.

The major strength of this sort of subnational approach is that it is very easy to either expand or narrow. One could apply a raster to the entire globe or focus in on a single country, or even region of a country. The level of resolution can be tailored as well, from the broad resolution I use to a much narrower kilometer level. This is all dependent on available data. This analysis, for example, could easily be expanded to a global level give a source of conflict data that covers most
or all countries. The ability to adapt a raster-based model to the problem at hand makes it very practical for applied work where some groups are interested in very general models, while others may want to focus in on particular regions of the world. Additionally, the ability to better incorporate geolocated data in this sort of model opens up the door to a host of new data sources, such satellite imagery. There is a lot of potential in this type of subnational model.

The third and final article looks at using time series method for modeling short-term civil conflict. I apply both multivariate methods, in this case BVAR, as well as univariate ARIMA methods to forecast interactions between actors involved in the ongoing Syrian Civil War. The objective here is to evaluate how well a multivariate BVAR, both with and without an informed prior, performs compared to simpler univariate methods. Previous research found that BVAR performed well for prediction, but only applied it to conflicts with a long history and used very basic autoregressive models as a baseline. I find that BVAR yields no meaningful improvement over univariate methods when applied to conflicts with a short history and compared to univariate methods that have undergone some tuning.

The results suggest that a multivariate approach is not particularly useful for civil conflicts with a short history. There is no clear advantage over the univariate models. When accounting for the increased computational cost associated with fitting and tuning a BVAR model, there is even less reason to prefer it over a univariate approach. Additionally, relying on a BVAR model may restrict the level of aggregation available to a researcher. Here I was forced to fit the models on weekly data due to an insufficient number of observations to fit a BVAR model at a monthly level of aggregation. This could create problems in an applied setting where policymakers and other may be interested in forecasts at varying levels of aggregation. Future work in this area should explore more advanced univariate
methods as an alternative to existing systems.

This study is by no means exhaustive, but it does show the potential of alternative approaches. There are a wide range of directions in which conflict forecasting can go. Forecasters need not restrict themselves to the usual probabilistic country level forecasts that characterized early research in this area. Attention should be paid to tailoring the forecasting approach to the problem at hand. As I find in the second and third articles, breaking down forecasts both subnationally and subannually is both possible and preferable in some cases. For applied work on developing early warning systems, the best approach may be to use multiple types of models. Take the problem of forecasting mass atrocities. A country level model could serve as a means of identifying the countries most at risk. A raster-based model can then be used to identify which regions of these at risk countries are most likely to experience violence, while a time series style model can help to forecast who is likely to be involved in the violence. The building blocks of such a system are there. It is only a matter of putting them together.
Appendix

Maps of Geographic Data

Figure 1.1: Distribution of Drought Data
Figure 2: Excluded Ethnic Groups
Figure 3: MODIS Land Cover Classifications
Figure 4: Mountainous Terrain
Figure 5: Nighttime Light Emissions
Figure 6: Population
Figure 7: Locations of Precious Resources


Kucera, Jan, Mayeul Kauffman, Ana-Maria Duta, Ivette Tarrida Soler, Patrizia Tenerelli, Giovanna Trianni, Catherine Hale, Lauren Rizzo & Stefano Ferri. 2011. “Armed Conflict and Natural Resources.”


Ulfelder, Jay. 2012. “Forecasting Political Instability: Results from a Tournament of Methods.” *Available at SSRN 2156234*.


Wu, Kuan-Wei, Chia-Hua Ho, An-Chun Liang, Chun-Heng Huang, Wei-Yuan Shen, Jyun-Yu Jiang, Ming-Hao Yang, Ting-Wei Lin, Ching-Pei Lee et al. 2012. “A two-stage ensemble of diverse models for advertisement ranking in KDD Cup 2012.” *KDD Cup*.


101
Benjamin Fisher

521 Toftrees Avenue #210
State College, PA 16803
251-472-6256
bensfisher@gmail.com
bensfisher.com

EDUCATION

PhD., Political Science, 2016
Master of Arts, Political Science, 2014
Pennsylvania State University, State College, PA
Major: Quantitative Methods
Minor: International Relations

Bachelor of Arts, International Relations, 2012
University of Alabama, Tuscaloosa, AL
Major: International Relations
Minor: French, German

EXPERIENCE

Quantitative Social Scientist
Verisk Maplecroft
December 2015 - Present
- Develop predictive models for political events such as state instability and civil unrest using a mix of machine learning and GLM models.

Research Assistant
Professor Bruce Desmerais
August 2015 - December 2015
- Developed command line and Python tools to process text and extract citation information from government documents as part of an NSF-funded project.

Analytics Intern
Verisk Innovative Analytics
June 2015 - September 2015
- Developed models to forecast the risk of maritime piracy at global and regional levels with varying lead times, as well as modeling work for insurance claims. Final deliverables were an ensemble of GLM and machine learning methods.

Statistical Consultant
Leidos, Warfighter and Mission Services Division
May 2014 - February 2015
- Developed statistical models for the Political Instability Task Force using machine learning and time series methods as part of an effort to create an early warning system for mass violence against civilians.

TECHNICAL SKILLS

Languages: Python, SQL
Programs: R, Stata
Statistics: Regression Analysis, Maximum Likelihood Estimation, Time Series, Machine Learning, Network Analysis
Other: Amazon Web Services (EC2, Elastic MapReduce, Hive, S3), Natural Language Processing (Stanford CoreNLP, spaCy)