FLORIDA STATE UNIVERSITY

COLLEGE OF SOCIAL SCIENCES AND PUBLIC POLICY

APPLIED PREDICTIVE MODELING FOR MEASUREMENT AND INFERENCE IN INTERNATIONAL CONFLICT AND POLITICAL VIOLENCE

By

PHIL HENRICKSON

A Dissertation submitted to the
Department of Political Science
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2018

Phil Henrickson defended this dissertation on November 16, 2018.
The members of the supervisory committee were:

Mark Souva

Professor Directing Dissertation

Jonathan Grant

University Representative

Robert J. Carroll

Committee Member

Sean Ehrlich

Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

To my parents, my siblings, and my wife, for their unwavering support and love.

ACKNOWLEDGMENTS

I am forever indebted to the political science department at Florida State University for many reasons.

First, I thank my professors at Florida State: Mark Souva for his instruction throughout my time as a graduate student (as well as rescuing me at the airport during my first visit to FSU); Rob Caroll, for his patience, mentorship, and investment in both my professional and personal development; Sean Ehrlich, for his relentless support and encouragement for my entire cohort; Will Moore, for directing my gaze towards the world of political violence and human rights.

Second, I thank Dennis Langley, Kevin Fahey, John Griffis, Scott Meachum, and Patrick Scott for many silly and wonderful memories playing board games during our time at FSU, without which I could scarcely have survived graduate school. Here I owe an additional thanks to the fine people of Shut Up and Sit Down for helping to foster our love of games, as well as providing many podcasts for my hours spent writing and coding.

Finally, I am most thankful for another FSU political science graduate student, Sydney Gann, who is now my wife. Of all my time spent at FSU in the last five years, I am most grateful for our hikes together in Tallahasee's beautiful parks and forests; our hours spent chatting over coffee at Catalina Cafe; our weekly happy hours with friends at Gaines Street Pies; our cross country road trips with our cats to visit family (and flee from hurricanes). I am thankful for these memories, for Sydney, and for all the moments we will share in the years to come.

TABLE OF CONTENTS

List of Tables . . . vii

List of Figures . . . ix

Abstract . . . xiii

1 Predicting the Costs of War **1**

1.1 The Expected Costs of Conflict . . . 5

1.1.1 Bargaining and the Costs of Fighting . . . 5

1.1.2 Expected Outcome vs. Expected Costs . . . 8

1.1.3 Measuring the Costs of War . . . 9

1.2 Learning Cost Expectations . . . 10

1.2.1 Setting up the Data . . . 10

1.2.2 Average, Aggregate, or Strongest Opponent? . . . 11

1.2.3 The Outcome . . . 12

1.2.4 The Predictors . . . 14

1.2.5 The Predictive Criterion . . . 16

1.2.6 Methods . . . 17

1.3 Results . . . 19

1.3.1 Predicting Battle Deaths . . . 19

1.3.2 Explaining Battle Deaths . . . 21

1.3.3 The New Measure: Dispute Casualty Expectations (DiCE) . . . 28

1.3.4 Substantive Examples . . . 29

1.4 Conclusion . . . 32

2 A New Measure of Foreign Threat **34**

2.1 Conceptualizing Threat . . . 38

2.1.1 Uses in the Literature . . . 38

2.1.2 Threat Expectations . . . 41

2.1.3 Measuring Threat . . . 43

2.2 Towards a New Measure of Threat . . . 46

2.2.1 The Approach . . . 46

2.2.2 Data . . . 48

2.2.3 Predictive Criterion . . . 51

2.2.4 Classifiers . . . 52

2.2.5 Candidate Classifiers . . . 53

2.3 Results . . . 54

2.3.1 Ensembling . . . 56

2.3.2 Test Set . . . 59

2.3.3 Comparing Models . . . 60

2.3.4 What Matters for Predicting MIDs? . . . 63

2.3.5 How Have MIDs Changed Over Time? . . . 70

2.4 The New Measure . . . 74

2.5 Conclusion and Advice for Researchers . . . 82

3 Foreign Threats and State Repression **84**
3.1 When Do States Repress? . . . 87
3.1.1 Domestic Factors . . . 88
3.1.2 International Factors . . . 89
3.1.3 Linking Foreign Threats to Repression . . . 91
3.2 Methodology . . . 94
3.3 Data . . . 95
3.4 Results . . . 101
3.4.1 OLS Variable Permutations . . . 101
3.4.2 Random Forest Variable Importance . . . 103
3.4.3 Inference . . . 114
3.5 Conclusion . . . 119

Appendices

A Predicting the Costs of War **123**
A.1 The Ensemble Model . . . 123
A.2 Alliances and War Costs . . . 125
A.3 Political Institutions and War Costs . . . 127
A.4 Lasso and Ridge Regressions . . . 129
A.5 Tuning Parameters . . . 129

B A New Measure of Foreign Threat **134**
B.1 Logistic Regression Coefficient Plot . . . 134
B.2 Modeling with Lagged Inputs . . . 134
B.3 Modeling with Incremental Years . . . 135
B.4 Examining Additional Year Comparisons . . . 136
B.5 Tuning Parameters . . . 138

C Foreign Threats and Repression **140**
C.1 Correlation Heatplot . . . 140
C.2 Additional OLS Variable Permutations . . . 141
C.3 Additional Conditional Random Forest and Ranger Variable Permutations . . . 145
C.4 Lasso, Ridge, and OLS Coefficients . . . 147

Bibliography . . . 150

Biographical Sketch . . . 161

LIST OF TABLES

1.1 Out of sample performance estimated using nested 5-fold cross validation, tuned to minimize RMSE. Results averaged across 10 imputations. 19

1.2 Proportional reduction in loss (PRL) from each of the candidate models compared to the null model. Results averaged across 10 imputations. 20

1.3 Results of linear models using all predictors with a cubic polynomial for year, with and without country fixed effects. 95% confidence interval reported from 1000 bootstraps. Results averaged across 10 imputations. 27

2.1 Data structure for modeling MID onset with Iraq and US as a motivating example. Note that the year itself (1990) is also included as a separate feature. 48

2.2 Predictors of MID onset - country level 49

2.3 Predictors of MID onset - dyad level 50

2.4 Predictors of MID onset - system level 51

2.5 Results from cross validation on training set. N = 610794 with 1208 observed instances of states being the target of fatal MIDs. Log likelihood and area under the receiver operating curve estimated using 10 fold cross validation. Training time in minutes reported for each method. 55

2.6 Results from validation set using tuned classifiers. N=135732 with 258 observed instances of fatal MIDs. Classification performance assessed using log likelihood and area under the receiver operating curve. 56

2.7 Results from the test set using each tuned classifier along with various ensemble models. Classification performance assessed using log likelihood and area under the receiver operating curve. 60

3.1 List of outcome variables for state repression and human rights. Listed with description of the variable and its source. 97

3.2 Predictor variables used in models of state repression and human rights and source of the data. 97

3.3 Linear models of Fariss' dynamic latent score, using a reduced and full set of predictors, with bootstrapped standard errors. 120

3.4 Linear models of the Political Terror Scale, using a reduced and full set of predictors, with bootstrapped standard errors. 121

3.5 Linear models of CIRI's Physical Integrity Index, using a reduced and full set of predictors, with bootstrapped standard errors. 122

A.1 Results for each candidate model with a weight greater than 0.001 in the ensemble model. The performance of the ensemble in terms of RMSE and proportional reduction in loss from the null model for comparison to each of the candidate models. 124

A.2 Test performance of the candidate models having incorporated ally features more directly. 127

A.3 Coefficient estimates from an OLS compared to shrunken coefficients from the lasso and ridge. Predictors centered and scaled. 131

B.1 Results from cross validation on training set using lagged inputs 135

C.1 Coefficients from lasso, ridge, and ordinary least squares regression models of Fariss' dynamic latent score using all predictors without a lagged dependent variable. 147

C.2 Coefficients from lasso, ridge, and ordinary least squares regression models of the Political Terror Scale using all predictors without a lagged dependent variable. 148

C.3 Coefficients from lasso, ridge, and ordinary least squares regression models of CIRI's Physical Integrity Index using all predictors without a lagged dependent variable. . . 149

LIST OF FIGURES

1.1 Battle deaths: Untransformed 15

1.2 Battle deaths: Inverse hyperbolic sine 15

1.3 Scatter plots of model performance across each of the candidate models. A model which predicts perfectly would see all observations fitting along a 45 degree line from the bottom left to top right. 22

1.4 Variable importance plots from linear models, Random Forests, and Cubist, fit to the aggregate data. Bootstrapped 1000 times to produce the resulting confidence interval. 23

1.5 Partial dependence plots from each of the variables in a random forest fit to the entirety of the aggregated data, with and without year included as a predictor (with seven randomly selected predictors and 500 trees). 25

1.6 Expected battle deaths from a random forest fit using the aggregate opponent approach, with LOESS lines and breaks at 1826, 1919, and 1945. Results averaged across 10 imputations. 28

1.7 DiCE: US-Russia 30

1.8 DiCE: China-Japan 30

1.9 DiCE: Brazil-Argentina 31

1.10 DiCE: North Korea - South Korea 31

1.11 Expected battle deaths for conflicts post 1990 compared to observed battle deaths, bootstrapped 100 times. 33

2.1 Weighted ensembling 58

2.2 Stacked ensembling 59

2.3 Predicted probabilities for the full dataset from a logistic regression. Observations with a fatal MID onset are highlighted in red while observations without a conflict are in black. 61

2.4 Predicted probabilities for the full dataset from the stacked ensemble with extreme gradient boosted trees. Observations with a fatal MID onset are highlighted in red while observations without a conflict are in black. 61

2.5 Comparing predictions for the best performing ensemble and logit for the years 1937-1941. Observations in which B attacked are highlighted in red; jitter used to help distinguish between observations. 63

2.6 Comparing the ensemble model and a logistic regression. For both plots, the y axis is the difference between test set probabilities from the ensemble and a logistic regression. On the left are observations in which conflict took place; on the right are observations in which no conflict occurred. The ensemble records wins (plotted in blue) when it assigns a higher probability and conflict occured, as well as when it assigns a lower probability and conflict did not occur. The ensemble ''oses' (plotted in blue) when it assigns a higher probability and conflict occurred, as well as when it assigns a lower probability and conflict did not occur 64

2.7 Permuted variable importance scores from Ranger. Five hundred trees grown with two randomly selected predictors. 66

2.8 Variable importance scores computed using the Boruta package for Ranger. 68

2.9 Predictive performance for all classifiers from 1870-2001 assessed using log-loss. . . . 70

2.10 Ranger fit to years 1870-1913. N = 62,983 with 153 occurrences of Fatal MIDs 72

2.11 Ranger fit to years 1914-1945. N= 106,157 with 613 occurrences of Fatal MIDs 72

2.12 Ranger fit to years 1946-1990. N = 591,650 with 997 occurrences of Fatal MIDs . . . 73

2.13 Ranger fit to years 1991-2001. N = 256,337 with 249 occurrences of Fatal MIDs . . . 73

2.14 The country-level threat measure shown for all countries over the time period 1870-2001. 74

2.15 Measure of foreign threats displayed for 1920. Top 20 countries on measure in year shown on right. 77

2.16 Measure of foreign threats displayed for 1950. Top 20 countries on measure shown on right. 77

2.17 Measure of foreign threats displayed for 1970. Top 20 countries on measure shown on right. 78

2.18 Measure of foreign threats displayed for 1995. Top 20 countries on measure shown on right. 78

2.19 US foreign threat level, 1870-2001 80

2.20 China foreign threat level, 1870-2001 80

2.21 Russia foreign threat level, 1870-2001 81

2.22 Iraq foreign threat level, 1870-2001 81

3.1 Cross validated variable importance scores for the Dynamic Latent Score from Fariss (2014). Variable importance estimated using linear regression models with 10 fold

cross validation, iteratively adding each variable to a baseline specification including logged GDP per capita and logged population. Results bootstrapped 1000 times. . . . 98

3.2 Cross validated variable importance scores for the Political Terror Scale Variable importance estimated using linear regression models with 10 fold cross validation, iteratively adding each variable to a baseline specification including logged GDP per capita and logged population. Results bootstrapped 1000 times. 99

3.3 Cross validated variable importance scores for CIRI's Physical Integrity Index. Variable importance estimated using linear regression models with 10 fold cross validation, iteratively adding each variable to a baseline specification including logged GDP per capita and logged population. Results bootstrapped 1000 times. 100

3.4 Variable permutation scores from a conditional random forest for the Dynamic Latent Score from Fariss (2014). Forest grown with 500 trees and 10 randomly selected predictors. 103

3.5 Variable permutation scores from a conditional random forest for the Political Terror Scale. Forest grown with 500 trees and 10 randomly selected predictors. 104

3.6 Variable permutation scores from a conditional random forest for CIRI's Physical Integrity Index. Forest grown with 500 trees and 10 randomly selected predictors. . . 105

3.7 Boruta variable importance for Fariss' Dynamic Latent Score 109

3.8 Boruta variable importance plots for the Political Terror Scale 110

3.9 Boruta variable importance plots for CIRI's Physical Integrity Index 111

3.10 Scatter plot of Fariss' dynamic latent score and the measure of foreign threats. LOESS line fit with 95% confidence interval 116

3.11 Scatter plots of the Political Terror Scale and the measure of foreign threats (left) and CIRI's Physical Integrity Index and the measure of foreign threats (right). LOESS line fit with 95% confidence interval 117

A.1 Out of sample performance of the ensemble model. The scatter plot shows the ensemble model's out of sample predictions regressed against the observed values with a LOESS line with a 95% confidence interval. 124

A.2 Observed values of battle deaths sorted from least to greatest against the model's predictions for each of these observations, LOESS line with a 95% confidence interval added. 125

A.3 Expected battle deaths by dyad Polity scores 128

A.4 Lasso variable trace plots for all predictors used in modeling battle deaths 130

A.5 Ridge regression variable trace plots for all predictors used in modeling battle deaths 130

B.1 Logistic regression using all predictors for the full dataset. Standardized coefficients to facilitate comparison; 95% robust standard errors reported around each point estimate. 134

B.2 Training and test results using log likelihood and area under the receiver operating curve for incremental runs of a logistic regression 136

B.3 Comparing predictions for the best performing ensemble and logit for the years 1970-1974. Observations in which B attacked are highlighted in red; jitter used to help distinguish between observations. 137

B.4 Comparing predictions for the best performing ensemble and logit for the years 1990-1994. Observations in which B attacked are highlighted in red; jitter used to help distinguish between observations. 137

C.1 Correlation plot for all predictors and outcomes. 140

C.2 Cross validated variable importance scores for CIRI's political prisoners measure. Variable importance estimated using linear regression models with 10 fold cross validation, iteratively adding each variable to a baseline specification including logged GDP per capita and logged population. Results bootstrapped 1000 times. 141

C.3 Cross validated variable importance scores for CIRI's political disappearances measure. Variable importance estimated using linear regression models with 10 fold cross validation, iteratively adding each variable to a baseline specification including logged GDP per capita and logged population. Results bootstrapped 1000 times. 142

C.4 Cross validated variable importance scores for CIRI's torture measure. Variable importance estimated using linear regression models with 10 fold cross validation, iteratively adding each variable to a baseline specification including logged GDP per capita and logged population. Results bootstrapped 1000 times. 143

C.5 Cross validated variable importance scores for CIRI's political killings measure. Variable importance estimated using linear regression models with 10 fold cross validation, iteratively adding each variable to a baseline specification including logged GDP per capita and logged population. Results bootstrapped 1000 times. 144

C.6 Variable permutation scores from a conditional random forest for CIRI's political prisoners measure. Forest grown with 500 trees and 10 randomly selected predictors. . 145

C.7 Variable permutation scores from a conditional random forest for CIRI's political disappearances measure. Forest grown with 500 trees and 10 randomly selected predictors. 145

C.8 Variable permutation scores from a conditional random forest for CIRI's torture measure. Forest grown with 500 trees and 10 randomly selected predictors. 146

C.9 Variable permutation scores from a conditional random forest for CIRI's Physical Integrity Index. Forest grown with 500 trees and 10 randomly selected predictors. . . 146

ABSTRACT

Advances in computing and machine learning have enabled researchers to use many different tools to learn from data. This dissertation is devoted to using predictive modeling to learn from existing data in international conflict studies with the aim of offering new measures and insights for applied researchers in international relations.

In the first chapter, I explore the expected cost of war, which is a foundational concept in the study of international conflict. However, the field currently lacks a measure of the expected costs of war, and thereby any measure of the bargaining range. I develop a proxy for the expected costs of war by focusing on one aspect of war costs - battle deaths. I train a variety of machine learning algorithms on battle deaths for all countries participating in fatal military disputes and interstate wars between 1816-2007 in order to maximize out of sample predictive performance. The best performing model (random forest) improves performance over that of a null model by 25% and a linear model with all predictors by 9%. I apply the random forest to all interstate dyads in the Correlates of War dataverse from 1816-2007 in order to produce an estimate of the expected costs of war for all existing country pairs in the international system. The resulting measure, which I refer to as Dispute Casualty Expectations (DiCE), can be used to fully explore the implications of the bargaining model of war, as well as allow applied researchers to develop and test new theories in the study of international relations.

In the second chapter, I use these expected costs of war to explore another foundational concept in international relations: foreign threats. Researchers commonly theorize about the impact of a state's international security environment - that is the extent to which a state is threatened by other states - yet the field currently lacks a measure which can effectively proxy for expectations

of conflict. In order to create a new measure of threat, I train a number of machine learning algorithms on fatal militarized disputes over the years 1870-2001. I aggregate the predictions from these models at the country level to create a new measure of international conflict expectations for all states. In so doing, I am able to revisit the causes of international conflict via a data-driven approach, as well as provide a new measure of foreign threat for applied researchers.

Finally, in the third chapter, I make use of this new measure to assess how international security affects a state's human rights behavior. International relations scholars have increasingly relied on domestic institutions to explain international conflict but less work has focused on reversing the arrow. To this point, political violence scholars have principally relied on domestic factors to explain the conditions under which leaders use coercive means to maintain power. But, political leaders do not exist in a vacuum; their decision making is informed by international and domestic factors. Therefore, I rely on both a predictive and inferential approach to assess whether foreign threats matter for state repression. The measure of foreign threats does emerge as an important variable in predicting state repression, which suggests that there is a meaningful relationship between international security and human rights behavior. Additionally, I find some (limited) evidence that the measure is negatively related to human rights behavior: states with high levels of foreign threat are associated with higher levels of state repression. But this finding is sensitive to model specification and merits further inspection.

CHAPTER 1

PREDICTING THE COSTS OF WAR

One of the central puzzles of international relations continues to be the puzzle of war: war is tremendously costly for states, yet nonetheless they choose to fight. Why do states go to war? Though the answers to this question are many, one prominent answer lies in the introduction of the bargaining model in the seminal work of Fearon (1995). His explanation begins with a simple premise: because war is costly, states would be better off reaching a mutually preferred bargain rather than fighting. The question of why states fight is thus really the question of why states fail to bargain. Rationalist explanations of war have been built on this premise, as researchers continue to explain conflict's occurrence by focusing on the factors which prevent states from reaching a negotiated settlement.[1]

Central to the rationalist explanations of war is the assumption that war is costly, where the range of mutually preferred bargains is set by the expected costs of war for both states. That is, before fighting, both states must assess how costly conflict would be in the event that they choose to fight, and use these expectations to determine their preferences for a peaceful settlement. That a bargaining range exists in Fearon's model is due to the assumption that states expect to pay some cost in order to fight. But what are these expectations of war costs? How costly will a war be for a state? Will the bargaining range be small or large? Will the war be more costly for one state than other? The rationalist explanations of war are built on the premise that states

[1]The canonical rationalist explanations from Fearon being information asymmetries with incentives to misrepresent, credible commitment problems, and indivisible goods; See Jackson and Morelli (2011) for a review of the literature of explanations for war.

estimate war costs routinely in their interactions with other states as they must always evaluate the costs and benefits of conflict. Yet for all of the theoretical and empirical work that has been done in using the bargaining model to explain international conflict, we have not yet developed any means of *measuring* states' expectations for the costs of war. At present, we do not have a readily available means of answering the following question: if two states were to fight, how costly would their conflict be? As such, applied researchers currently lack a measure of one of the most important theoretical concepts in international relations.

This is not to suggest that the field has overlooked a simple measurement task. Like many theoretical concepts in international relations and political science, the cost of war is not easy to measure directly. In the bargaining model, the cost of war represents losses incurred by choosing to fight, which can incorporate a wide variety of outcomes: loss of life, loss of territory, value of the war to the nation or leader, state resources spent training, mobilizing, and deploying military forces, and the opportunity costs from a loss of trade or from a transformation of a nation's workforce. While researchers have agreed that the background concept of war costs involves any sort of loss incurred by choosing to fight rather than bargain, the concept is broad enough to make the task of measurement difficult.[2] Any effort to estimate the expected costs of war must make decisions about what can and cannot be incorporated. I focus on one aspect of war costs: battle deaths. While it is certainly the case that "a variety of human, material, and psychic losses go into the costs of war" (Bueno de Mesquita, 1983, 353), one of the most pressing costs in war remains the loss of human life due to fighting. Indeed, the conflict literature often points to calculations leaders make with regards to expected fatalities in the war, with democracies commonly thought to be

[2]See Adcock (2001) for a careful discussion of measurement validity for political scientists. For an example of measuring the economic costs of conflict, see Stiglitz and Bilmes (2008) for a thorough discussion of the task of estimating the economic cost of the 2003 Iraq War.

more sensitive to war costs than autocracies (Kant and Reiss, 1970; Reiter and Stam, 2002; Filson and Werner, 2007, 2004). If we wish to examine the costs of war, an easy starting point is the loss of life from fighting. I therefore seek to develop a measure of expected battle deaths for all possible interstate dyads in the international system with the aim of proxying for cost expectations. While imperfect, this endeavor will provide researchers with a measure that the field is currently lacking. Moreover, my approach is not intended to be the final word on measuring the expected costs of war; instead, it represents the first effort to produce a measure of this important theoretical concept in international relations.

The task at hand is one of prediction: we wish to know the number of battle deaths which would have occurred in conflicts which did not actually take place. What this amounts to is an out of sample problem: we need to train a model on the observed cases of battle deaths and be confident in its predictions for new data. Fortunately, advances in computing and machine learning have made this task not only feasible but relatively straightforward. I therefore use a variety of methods with the aim of maximizing out of sample performance on battle deaths from all fatal military disputes and interstate wars in the Correlates of War dataset over the period of 1816-2007. The first task of this paper is to evaluate the predictive performance of these models. The question motivating this paper is simple: can we predict battle deaths for interstate conflicts? I find that the answer is yes: the best performing model (random forest) improves out of sample predictions of conflict battle deaths over that of a null model by 25% and over a standard practice model (a linear model with all predictors) by 9%. This finding is also encouraging given that I use a relatively limited set of predictors, relying primarily on on country-level predictors from the Correlates of War National Material Capabilities and the Polity IV project. It is easy to imagine that future work can improve upon the effort here by using the same methodology but with additional features

from the international system. After first establishing that we can use country-level predictors to predict battle deaths, the second task of this paper is to extend the model to all hypothetical disputes which could have taken place in the international system. Having trained a random forest on observed cases of battle deaths, I apply it to all interstate dyads in order to estimate the expected battle deaths for all pairings in the Correlates of War dataverse. This amounts to over 1.5 million estimates, covering all possible country pairings over the years 1816-2007. I refer to these estimates as Dispute Casualty Expectations (DiCE), which I argue can serve as the best existing proxy for the expected costs of war.[3]

The central contribution of this paper is illustrating the utility of predictive modeling for measurement in international relations. In predicting battle deaths I am additionally able to contribute to the literature on how observable factors at the outset of conflict affect the dynamics of war. The question motivating this paper is whether observable factors at the outset of war can be used to predict, and thereby inform, our understanding of war costs. Though the explicit task of this paper is to predict battle deaths from interstate conflict, I'm also able to speak to *how* observable country characteristics predict battle deaths. As I will show, I find that features thought to proxy for state power - CINC scores, military personnel, and state energy consumption - emerge as the most important variables for out of sample prediction. While not surprising, this finding revisits and contrasts with the work of Maoz (1983), who found that observable capabilities did not affect militarized dispute outcomes. What is perhaps more surprising is that the lone variable thought to capture institutional effects - a state's Polity 2 score - offers little improvement in predictive performance conditional on all other variables in the model. Notably, I find that all of these predictors

[3]My approach mirrors that of Carroll and Kenkel (2016), who run an ensemble learner and use cross validation to construct the Dispute Outcome Expectations (DOE) variable, which is an estimate of the probability that a state would win a hypothetical dispute. My approach differs in that rather than estimating the probability of winning or losing a dispute, I am seeking to estimate the costs from fighting in a war, a fundamentally different outcome than winning or losing.

are conditional on time: the year of the conflict emerges as the most important predictor across all of my models, as expected battle deaths have steadily decreased over time in the international system since 1950.

1.1 The Expected Costs of Conflict

1.1.1 Bargaining and the Costs of Fighting

The costs of war are foundational in the study of interstate conflict and the onset of war. Consider the canonical example from Fearon (1995) where two states, A and B, are in a dispute over a good. As unitary rational actors seeking to obtain as much of the good as possible, the states know if they fight there is some probability p which determines who will win the war and receive the good. In this case, the decision to go to war represents a costly lottery, where A will win the war and receive its desired outcome with probability p. If the states choose to fight rather than bargain, they each will have to pay some cost, c. The terms c_A and c_B represent each state's ex ante cost of war; it is what they will expect to pay only in the event that they choose to fight.[4] Because c_A and c_B are assumed to be positive, fighting is costly. From this, choosing to fight is always inefficient ex post; states should prefer to reach a peaceful ex ante bargain in the range of $[p - c_A, p + c_B]$ rather than fight and be forced to pay the costs of conflict. Thus there always exists a range of mutually preferable bargains to war. This key insight in Fearon's work underpins the rationalist explanations for war which have been fundamental to conflict studies for the last two decades.[5]

[4] Fearon notes: "in this formulation the terms c_A and c_B capture not only the states' values for the costs of war but also the value they place on winning or losing on the issues at stake. That is, c_A reflects state A's costs for war, relative to any possible benefits... if two states see little to gain from winning a war against each other, then c_A and c_B would be large even if neither side expected to suffer much damage in a war". (387). This is important to note, because my measure will *not* incorporate the value states place on the desired outcome, but only the costs in expected fatalities which would take place in the event of conflict.

[5] The model does not represent the final word on studies of international conflict, as it has been explored and extended over the years. Researchers have explored whether war can be thought of as a costly lottery (Wagner,

It is crucial to note here that the model invokes *expectations* regarding the cost of war. That a range of mutually preferable bargains exists relies on the notion that states or leaders are able to form these ex ante expectations of costs before fighting begins. But how do states or leaders develop these expectations? There is a great deal of work on how states develop beliefs and expectations about war outcomes, with various arguments stressing the role of psychology and mis-perceptions about capabilities (Levy, 1983; Blainey, 1988; Kaufmann, 2004; Johnson and Tierney, 2011), political institutions and domestic politics (Allison, 1999; Reiter and Stam, 2002), and rivalries (Goertz and Diehl, 1995).[6] One recurrent theme in all of this work is the inherent uncertainty of war and the difficulty in predicting hypothetical outcome: "because war is an uncertain process... the leaders of two countries must each form expectations about the results of a conflict to guide their decision making" (Fey and Ramsay, 2007, 738).

For the purpose of this paper, I make a simplifying assumption that states develop cost expectations by observing outcomes in the international system. This is similar to the work of Crescenzi, who studies the effect of reputation on international conflict and writes that in the absence of complete information, "states are forced to generate expectations about the behavior of [other states]... one possible learning schema for generating these expectations is to *textobserve how other states behave in similar situations and use this observations as a precedent, or prior, for the current situation* (emphasis added)" (2007, 388). Thus if states seek to develop expectations about war costs, they must look to instances in the international system in which fighting took place. In order

2000), how the future affects bargaining in the present (Powell, 2006), and how bargaining affects conflict termination (Reiter, 2009). But the model remains a key pillar in the study of international conflict, and the conception of war costs as first posited by Fearon has been carried through in future work.

[6]One prominent implication of the democratic victory argument is rooted in the notion that democratic states have greater access to information, allowing them to form better estimates of war outcomes and select wars which they are more likely to win (Reiter and Stam, 2002).

to estimate the states' expected costs of fighting, we can likewise look to the universe of realized conflicts with the aim of training a model which can predict this particular outcome.[7]

Before proceeding to is important to discuss the conceptualization of 'war' with regards to the bargaining model and costs of fighting. Though international conflict scholars make a distinction between disputes and wars by using a fatality threshold (typically 1000 deaths), Fearon's model makes no arguments relating to the scale of the ensuing conflict or the form of the war. In the model, states have the ability to reach a bargain or fight in a costly conflict, which is conceived of as a lottery where bargaining ends and the war resolves with one winner who gets to decide the outcome. A strict reading of the model could lead one to infer that because bargaining stops and the conflict only ends when one side has won a decisive victory, the war is 'absolute'. But as Clausewitz writes, "war can be thought of in two different ways - its absolute form *or one of the variant forms that it actually takes* (emphasis added)" (1976, 582). Because most wars culminate in a negotiated settlement rather than a decisive victory (Reiter, 2009), the form war more commonly takes is that of limited, rather than absolute, war. From this, in seeking to estimate war costs, I argue that it is better to focus on the costs states expect to pay in the event of a limited war. I therefore include all conflicts in which fighting and fatalities have taken place, with the aim of letting the data speak to the expected severity of a hypothetical conflict rather than imposing an assumption a priori.

[7]Here it is important to explicitly note that the task at hand is developing the best possible prediction for a state's battle deaths given observable factors in the event that it chooses to engage in conflict with another state. That is, I am not able to speak to a leader's perception or mis-perception with regard to their own expectations of war costs. Though leaders are often mistaken in how they expect a war to unfold, I assume here that they make decisions with the best possible estimate of their own expected war costs. For specific examples of how leaders have made prewar decisions , see Downes (2009) for a discussion of the Johnson administration's decision to begin the Vietnam War and Kaufmann (2004) for a discussion of the Bush administration in the 2003 Iraq War.

1.1.2 Expected Outcome vs. Expected Costs

One immediate thought might be that state expectations regarding the costs of fighting are the same as that of expectations about power and state capabilities. Indeed, the field has long explored conflict theoretically by focusing on power and how it relates to war outcomes. (Blainey, 1988), in his comprehensive study of the causes of war, was principally interested in whether state's formed similar expectations about the distribution of power: "if two nations are deep in disagreement on a vital issue, and if both expect that they will easily win a war, then war is highly likely. If neither nation is confident of victory, if they expect victory to come only after long fighting, then war is unlikely". Similarly, expected utility theories of conflict in the 1980s focused on expectations about the relative strength of the opponent: "the probability of gaining or losing in a conflict is directly related to the relative ability of the antagonists to bring power to bear in the conflict" (Bueno de Mesquita, 1980, 919). Smith and Stam (2004) develop a model of war in which states form heterogenous beliefs about their distribution of capabilities, and these beliefs "shape nations' expectations of the duration of conflict and which nation is likely to be the eventual winner if the war is fought to a decisive conclusion" (787). The crux of their model is that nations repeatedly fight battles until either one side is decisively defeated or both nations' beliefs on the true state distribution of capabilities, and therefore the eventual outcome, converge.

These existing explanations of war outcomes have focused on how power and capabilities are related to the outcome of conflict in the sense of which state wins or loses. But, as Filson and Werner (2007) argue, there are really two outcomes involved in fighting a war, that of winning the conflict and the costs which are paid in reaching that outcome. Existing work which focuses only on the the success or failure in conflict thus misses a crucial point because winning or losing conflict does not tell us anything about how *costly* a conflict would be. For instance, Carroll and Kenkel

(2016) develop a measure of Dispute Outcome Expectations (DOE) which gives the probability that states will succeed or fail in a conflict. But this measure gives no indication of the costs of fighting for each state. There is surely a key difference between a state which wins a dispute while paying little cost and a state which wins a dispute but only after costly fighting. Currently we do not have any means of distinguishing between these two potential outcomes. Estimating only the eventual outcome of the dispute overlooks the process by which that outcome is produced. This is not to criticize existing measures or suggest that they do not well for their intended purpose of proxying for power. Instead, I argue that while we have a measure of power, we do not at present know how well power maps to the expected costs states would pay in the event of conflict. How capabilities and observable factors relate to war costs is an empirical question, rather than one that can be assumed. In terms of the bargaining model, the field now has a measure of p, but at present we do not have any measure of c. Because of this I contend that we need to investigate another outcome of war beyond winning and losing and instead devote careful attention to identifying the costs states expect to pay in the event of conflict.

1.1.3 Measuring the Costs of War

The field thus stands to benefit from the enterprise of measuring the expected costs of conflict. In order to construct such a measure, I use existing data on the number of battle deaths which have taken place in fatal military disputes and interstate wars during the time period 1816-2007. The loss of human life remains the most damaging consequence of war, and is likely the most correlated with other forms of war costs such as economic and material losses from fighting (Beger, working, 1). There are also readily available estimates of battle deaths within each interstate military dispute and war which can be used to train a model of war costs.

Ideally, a measure of war costs would be able to incorporate the economic costs of waging war. Indeed, it is easy to conceive of arguments by which states might differ when it comes to paying economic vs human costs in war. Such an approach would involve first estimating the economic impact of fighting on states in interstate conflicts, then training a model on observable factors to predict the economic impact. This would mirror the methodology in this paper with the outcome simply being shifted from battle deaths to monetary costs. But this approach is currently infeasible, as at present such estimates do not exist. Instead, I opt for battle deaths because these have been gathered for all conflicts and, despite difficulties in gathering data from conflicts, remain an objective measure of war cost.

I thus seek to estimate the expected costs of with the understanding that a good proxy for war costs should be able to predict battle deaths from interstate conflict well. The task at hand is therefore explicitly a problem of prediction, for which I rely on tools from machine learning in order to maximize predictive performance. As it stands, to my knowledge, in the literature of international conflict such a proxy does not exist. In the following section I will detail my methodology for predicting battle deaths so as to fill this gap in the literature.

1.2 Learning Cost Expectations

1.2.1 Setting up the Data

Before I discuss the characteristics of the data I plan to use in estimating battle deaths, I describe the process of setting up the data. Theoretically, I am seeking to develop a measure of the expected cost of fighting between two states. That is, I am addressing the question: if two states were to engage in a costly dispute, what is their expected cost of fighting (in terms of battle deaths)? In order to estimate this, I rely on the universe of fatal military disputes and interstate

wars which have taken place. I model battle deaths as a function of country characteristics involved in the disputes and use cross validation to assess the out of sample predictive performance of these models. I then use the best performing model to make predictions about the costs of hypothetical disputes between all dyads in the international system.

1.2.2 Average, Aggregate, or Strongest Opponent?

The most useful measure of war costs for researchers is dyadic, reflecting the different ex ante costs of fighting for states should they choose to fight. The problem with the dyadic approach is that the majority of interstate conflicts are multilateral, and this presents an issue for modeling battle deaths using directed dyads.[8] To illustrate this, consider the hypothetical scenario where states A and B fought a war with state C. If I wish to model the battle deaths for states A and B, I can simply include the features from each respective state compared with their opponent C. But this leads to question of what do with modeling state C's battle deaths. I use three different approaches. First, I take the average of A and B's capabilities and pair C with this average. The problem with this approach is that it can punish sides with multiple participants rather than reflect that alliances are stronger than an individual state. If a weaker state sides with a stronger state, the data will treat the observation as weaker than if the strong state had fought on its own. Alternatively, I could pair C with the aggregate of A and B's capabilities and simply add each state's military and national features, taking the lowest Polity score of the two. Finally, I could pair C with the strongest opponent it faced in the war, using raw capabilities in order to determine which is the strongest opponent. I would then simply match C with whoever had the highest

[8] See Poast (2010) for a discussion of problems which arise from modeling complex outcomes in international relations using an assumed dyadic approach.

military capabilities score between A and B. In the event that both sides in the war have multiple participants, I would then pair each participant with the strongest opponent from each side.

Average Approach:
C Battle Deaths = f(C Features, avg(A+B Features), Dyad Features, Year)

Aggregate Approach:
C Battle Deaths = f(C Features, (A+B Features), Dyad Features, Year)

Strongest Opponent Approach:
C Battle Deaths = f(C Features, max(A+B) Features, Dyad Features, Year)

To preview the results, these three different modeling approaches all lead to slightly different predictions, though predictive performance remains relatively stable across each setup with the aggregate approach performing better in most circumstances.[9]

1.2.3 The Outcome

With the set up of the data in mind, the task of predicting interstate battle deaths begins with a discussion of the data that is available. The outcome of interest is battle deaths from interstate conflicts, including both fatal military disputes and wars.[10] For this task, scholars in international relations have largely relied on the Militarized Interstate Dispute and Correlates of War 4.1 datasets, which has data on "the number of battle-connected fatalities among military personnel" (Sarkees and Schafer, 2000, 128) for all participants in fatal disputes interstate wars between 1816 and 2007. These datasets are appealing because they cover the largest period of time

[9] In order to potentially gain from each approach to the data, I combined predictions from models run on each of these approaches in an ensemble model, which resulted in a small increase in predictive performance. Details on the ensemble model can be found in the appendix. I also briefly discuss the effects of alliances on war costs more directly in the appendix, though I aim to more directly incorporate alliances in future work.

[10] By itself, estimating the number of deaths which have taken place within a conflict is a difficult task. A large literature is devoted to the task of estimating the number of deaths incurred as a result of war. In this paper I will not address the various methodologies used to produce estimates of battle deaths, but will instead use the figures which commonly been analyzed in the literature.

amongst all available datasets, which not only offers the more observations for training the models but can also address substantive questions of how the costs of war may have changed over time.[11]

One drawback of these datasets is they estimate the number of combatant deaths by participant, but do not disaggregate the data annually. Instead, they only provide an estimate for battle deaths for the entire war. This should not be a problem for the task at hand, as I am principally interested in estimating the ex ante cost of conflict, meaning the expected cost of the entire war, but it does limit the ability to develop a more fine grained estimate. Another drawback is these datasets seek to measure deaths from battle, but it does not record civilian fatalities, or nonviolent deaths of any kind.[12] Thus the measure potentially understates the true cost of war as it does not explicitly seek to account for civilian fatalities related to the war.[13]

While the MID dataset deliberately does not include instances of war, it does include observations in which there were a high number of fatalities. As the measurement task at hand is predicting battle deaths for all conflicts in which fighting took place, I include all MIDs in which there was at least one fatality on at least one side of the dispute, taking care not to overlap any

[11]Including conflicts as far back as 1816 might provide reason for caution, as Jenke and Gelpi (2017) find there to be significant temporal variation in international conflict: their results suggest that causes of international conflict are substantially different in the Cold War era than in all other historical eras. They ultimately recommend that quantitative scholars in international conflict be sensitive to the temporal generalize ability of their results, and fully explore the impact of time. I seek to follow this advice by privileging out of sample prediction and using flexible models with 'year' as a predictor to capture the effects of time via a data-driven exercise. By including all conflicts from 1816-2007, my work is ultimately similar to that of Jenke and Gelpi (2017), in that I find similar temporal breaks in the international system with regards to the costs of conflict as they find for the onset of conflict.

[12]Lacina and Gleditsch (2005) demonstrate that the COW data varies between recording combatant, battle, and war deaths in codings for intrastate wars, but is unclear whether this extends to the interstate COW battle deaths data. While the COW project has a "total deaths" measure for civil and extrasystemic war, the interstate war data has no such measure

[13]Another possibility comes from the UCDP/PRIO dataset, which has compiled a dataset of all battle deaths in interstate war from 1945-2001. This dataset is notable because it is aggregated at the country-year level and distinguishes between the types of fatalities. This dataset has a total number of war deaths, but makes the distinction between battle deaths and non-battle deaths. This dataset also provides upper and lower bounds on the number of deaths in each war to represent the uncertainty in counting fatalities. While all of these features are appealing, the biggest issue with this dataset is the number of observations. By limiting the dataset to just conflicts post 1945, there are fewer number of interstate wars with which to train the model. This might not immediately appear to be a big issue if the data itself is of higher quality, but this sample cuts the number of observations for training in half. Additionally, it would limit the ability to examine expected costs over time, which I find to be very important in using data from the Correlates of War project

disputes with that of the interstate war data.[14] I then combined the fatal MIDs and interstate wars, giving the resulting dataset 1189 directed dyads. I next transformed the data to reduce the impact of skewness on model performance. Normally, a log-transformation would be the standard approach, but because there are a number of zeroes in the dependent variable. I instead use the inverse hyperbolic sine transformation.[15]

1.2.4 The Predictors

As the task of this paper is predictive, the selection of predictor variables is guided by data availability rather than by theory. I would prefer to have many variables available for modeling and then lean on preprocessing and feature selection to determine what is relevant for modeling our outcome. Unfortunately, as the outcome variable in question ranges from 1816-2007, the number of covariates for which we have full rank is going to be relatively limited. I principally rely on the National Material Capabilities dataset from the Correlates of War Project which has annual data on six aspects of a country's military capability: military expenditures, military personnel, iron and steel production, primary energy consumption, total population, and urban population.[16] The Polity IV dataset covers the entire time period, so I can include each country's Polity score as a feature in the model to determine how political institutions affect the costs of war. I additionally include the salience of the dyad from the Issue Correlates of War Project (Hensel, 2009). Finally,

[14]The MID dataset codes fatalities from fighting in ordered categories; for instance, the variable is coded as 1 if there were between 12-25 battle deaths and coded as 5 if there were 500-999 battle deaths. Since I am seeking to combine the fatal MIDs with the Correlates of War interstate conflict data, I need the fatalities from MIDs to be on the same scale. For each MID I randomly sampled from the appropriate interval and set this as the fatality for that particular MID.

[15]There are a number of zeroes present in the dataset because of conflicts in which one nation in the militarized dispute did not experience any battle deaths (UK-Albania in 1946 and Greece-Bulgaria 1952). I estimated the appropriate theta by selecting the value which minimizes the Kolmogorov-Smirnov test statistic against a normal distribution in order to select the distribution which is approximately normally distributed.

[16]Following Carroll and Kenkel (2016), I apply an hyperbolic sine transformation (Burbidge, Magee and Robb, 1988) to each of these variables as they are all right-skewed.

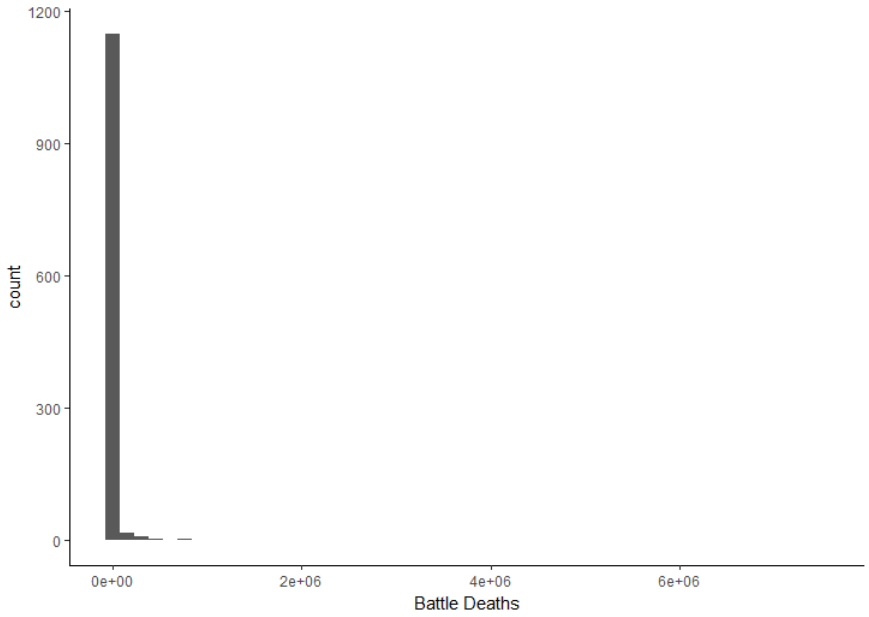

Figure 1.1: Battle deaths: Untransformed

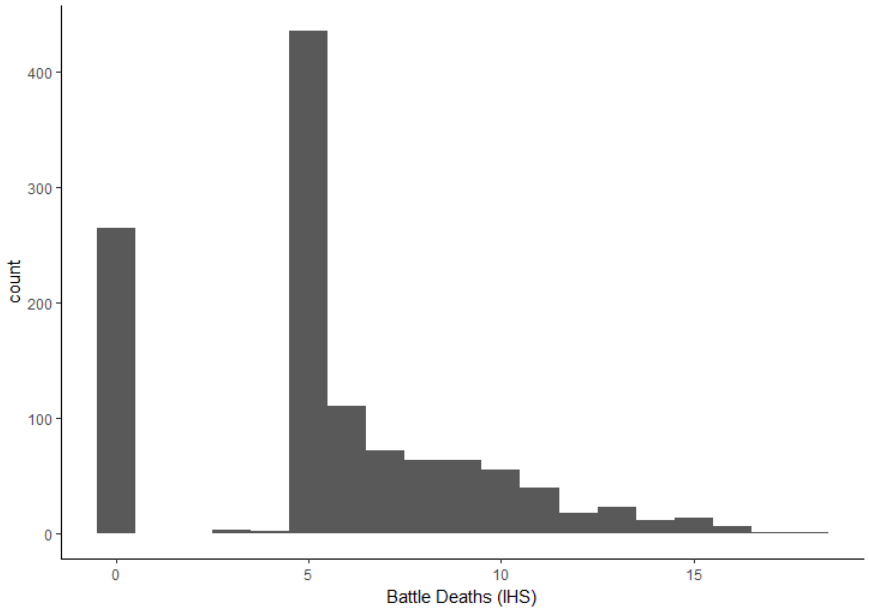

Figure 1.2: Battle deaths: Inverse hyperbolic sine

I include each dyad's contiguity from COW, and I include the year of the dispute to capture time effects.

1.2.5 The Predictive Criterion

With the data in hand, I now turn to the task of predicting battle deaths. The most common measure of predictive power in the regression setting is the root mean squared error (RMSE), where $\hat{f}$ is a model and $\hat{f}(x_i)$ is the prediction that a model gives for the ith observation, while y_i is the actual outcome for that observation (Hastie, Tibshirani and Friedman, 2009).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}$$

The RMSE will be small when the predictions of the model are close to the observed outcomes, and will be large if some of predictions differ substantially from the truth. I care about the generalized test error of the models - how well they would perform in predicting new data that was not used in fitting the model. Ideally, I would be able to randomly split the data into a training set and a validation set. I could then use the training set to fit the model, and then assess its performance on the validation set. This validation set approach of splitting the data would be be feasible with a large sample, but given that I have a relatively low number of observation this approach is not suitable. Instead, I rely on k-fold cross validation. In this approach, the data are randomly divided into k groups or 'folds', usually 5 or 10, of approximately equal size. The first k fold is then left out as the validation set, while the model is fit on the remaining $k-1$ folds. This process is repeated k times where each time a different k fold is treated as the validation set. This

process produces k estimates of the test error, and the final k-fold CV estimate is the average of these values.[17]

1.2.6 Methods

To maximize out of sample performance, I rely on tools from machine learning which are designed to predict well without making strong assumptions about the structure of the data. As there are numerous methods dedicated to this task in machine learning, I rely on the advice of Wu et al. (2007) and Fernández-Delgado et al. (2014), who have identified some of the the best performing algorithms for data mining and prediction. Guided by their recommendations as well as advice from Kuhn and Johnson (2013), I use the following methods via the caret package in R:

1. An intercept-only model to serve as the baseline for predictive performance.
2. A linear model using a subset of predictors
3. A linear model using all predictors (LM).
4. A penalized linear model with L1 and L2 regularization via the elastic net (Zou and Hastie, 2005)
5. Partial least squares (Wold, 1985).
6. Multivariate adaptive regressive splines (Friedman, 1991).
7. K-nearest neighbors (Cover and Hart, 1967).
8. Classification and regression trees (Breiman et al., 1984).
9. Random forest (Breiman, 2001).

[17] Another method for estimating the generalized test error would be to use the 632+ bootstrap from Efron and Tibshirani (1997), which reduces the variance of cross validated test error but has more bias, particularly in small samples.

10. Stochastic gradient boosted trees (Friedman, Hastie and Tibshirani, 2001; Elith, Leathwick and Hastie, 2008).

11. Cubist (Kuhn et al., 2012) which is an extension of the M5 tree algorithm citepquinlan1992learning for the regression setting.

12. Support vector machines with a radial kernel (Scholkopf et al., 1997).

13. Averaged neural networks (Ripley, 1996; Hastie, Tibshirani and Friedman, 2009).

For each of these methods, I use cross-validation to estimate how well they perform in predicting out of sample. For the methods which rely on tuning parameters, I first use cross validation to estimate the appropriate values for these parameters. That is, I use cross validation in an inner loop to select the appropriate values for the tuning parameters, and then use cross validation in an outer loop to estimate their test error. This nested cross validation is important, as Varma and Simon (2006) demonstrate that cross-validation can otherwise be too generous in estimating the out of sample performance of models which rely on tuning parameters.

To summarize, I split the data into five folds. I designate one fold as the test set and use the remaining folds as the training set. I then perform repeated 5-fold cross validation on the training set in order to estimate the appropriate values for tuning parameters. In this case, I select the values which minimize the RMSE of model.[18] I then evaluate the tuned model on the test set to estimate the model's true out of sample performance. I repeat this process five times, setting each fold as the test set in turn. The final estimate of the performance of the model is the average of the test error across each of the five folds. I repeat this process for all methods using the same set of predictors, using each of the different dyadic data approaches in turn.

[18]For the more computationally intensive models, such as support vector machines and neural networks, I select the tuning parameters which are within 1 standard deviation of the minimum RMSE from tuning, as recommended by Hastie, Tibshirani and Friedman (2009)

Table 1.1: Out of sample performance estimated using nested 5-fold cross validation, tuned to minimize RMSE. Results averaged across 10 imputations.

		Strong		Average		Aggregate	
Method	Predictors	RMSE	SD	RMSE	SD	RMSE	SD
Null	Intercept	3.816	0.116	3.816	0.116	3.816	0.116
LM	CINC, Year	3.453	0.077	3.482	0.084	3.434	0.077
PLS	All	3.279	0.149	3.250	0.118	3.125	0.080
LM	All	3.277	0.126	3.258	0.112	3.112	0.083
LM - Elastic Net	All	3.271	0.124	3.250	0.116	3.110	0.097
CART	All	3.207	0.105	3.217	0.102	3.236	0.063
KNN	All	3.158	0.067	3.152	0.115	3.156	0.100
Neural Nets	All	3.130	0.103	3.122	0.074	3.092	0.067
MARS	All	3.078	0.085	3.137	0.128	3.070	0.068
SVM - Radial	All	3.056	0.068	3.072	0.090	3.091	0.099
Boosted Trees	All	2.964	0.076	2.970	0.048	2.938	0.065
Cubist	All	2.964	0.101	2.950	0.093	2.907	0.065
Random Forest	All	2.840	0.082	2.845	0.067	2.846	0.063

1.3 Results

1.3.1 Predicting Battle Deaths

In this section I briefly discuss the out of sample performance of the candidate models. Across the three modeling approaches for dyads, the aggregated approach generally yields the best test performance. Within each of these dyadic modeling strategies, the tree-based, ensemble models - random forests, boosted trees, and Cubist - routinely show the best out of sample performance. Random forests consistently produce the lowest test error in cross validation across all three settings. Table 1.1 displays all of this information numerically, highlighting the three methods which show the best performance.[19]

To give a point of reference for the performance of these models, I now compare them to the baseline, intercept-only model. This model simply predicts the mean number of transformed battle

[19]I additionally ran each of these models without 'Year' as a predictor but do not include these results here. I found degraded performance across all of the models, but the results remained similar with still random forests showing a 20% improvement over the null.

Table 1.2: Proportional reduction in loss (PRL) from each of the candidate models compared to the null model. Results averaged across 10 imputations.

		PRL		
Method	Predictors	Strong	Average	Aggregate
LM	CINC, Year	0.095	0.087	0.100
PLS	All	0.141	0.148	0.181
LM	All	0.141	0.146	0.184
LM - Elastic Net	All	0.143	0.148	0.185
CART	All	0.160	0.157	0.152
KNN	All	0.172	0.174	0.173
Neural Nets	All	0.180	0.182	0.190
MARS	All	0.193	0.178	0.195
SVM Radial	All	0.199	0.195	0.190
Boosted Trees	All	0.223	0.222	0.230
Cubist	All	0.223	0.227	0.238
Random Forest	All	0.256	0.254	0.254

deaths (5.53) for all conflicts. Table 1.2 shows the reduction in test error for each of the candidate models over the null model. If we were to see no meaningful improvement over that of a null model, the task of predicting battle deaths might not be feasible given the current set of predictors. Happily, I found that all models substantially improve over the null, with random forests achieving a 25% reduction in test error when using the strongest opponent approach. This indicates that the models are managing to learn about the outcome from the available predictors and thereby provide a meaningful improvement when asked to predict new data. But this improvement, while drastic, is perhaps overstated. A more worthy comparison would be that of linear models using common predictors in international conflict. The random forest achieves a 15% improvement over a linear model with CINC scores and 'year' as predictors, and a 7-11% improvement over a linear model with all predictors.

There are two main takeaways at this point. First, we can reasonably conclude that it is possible to improve upon our predictions of battle deaths using the standard (and fairly limited)

set of predictors from the Correlates of War National Materials. Observable country factors do in fact offer us some information about war costs. Second, the results demonstrate that we can improve our predictions by using more flexible algorithms. As evidenced by the scatter plots, models from the linear family under predict at higher values of the outcome variable, while the tree based ensemble methods - random forests, boosted trees, and Cubist - perform well in expectation across the entire range of the outcome variable. Random forests in particular seem to be well suited to the task at hand. Their improvement over parametric models is likely because of their ability to easily detect nonlinearities and interactions present in the data generating process which would not captured unless specifically assumed by the researcher. Additionally, random forests implicitly conduct feature selection and mitigate the impact of irrelevant predictors. For these reasons random forests have seen increased use for predictive problems in political science (Hill and Jones, 2014; Barrilleaux and Rainey, 2014; Carroll and Kenkel, 2016; Muchlinski et al., 2016) and are appealing because they perform well in prediction while also providing results which can easily be interpreted, as I detail in the followings section.

1.3.2 Explaining Battle Deaths

It would be understandable to be hesitant about 'black-box' methods to improve our predictions if the gains are minimal and the results are difficult to understand (De Marchi, Gelpi and Grynaviski, 2004). The appeal of linear modeling is that it offers interpretable results, cleanly explaining the relationship between X and Y. While many methods are opaque - neural nets in particular - this is not the case for the methods which have performed well with this dataset. I am able to examine the relationship between country characteristics and battle deaths using random forests while also improving our predictions. Namely, what matters for predicting battle deaths, and what is the relationship between the predictors and the outcome?

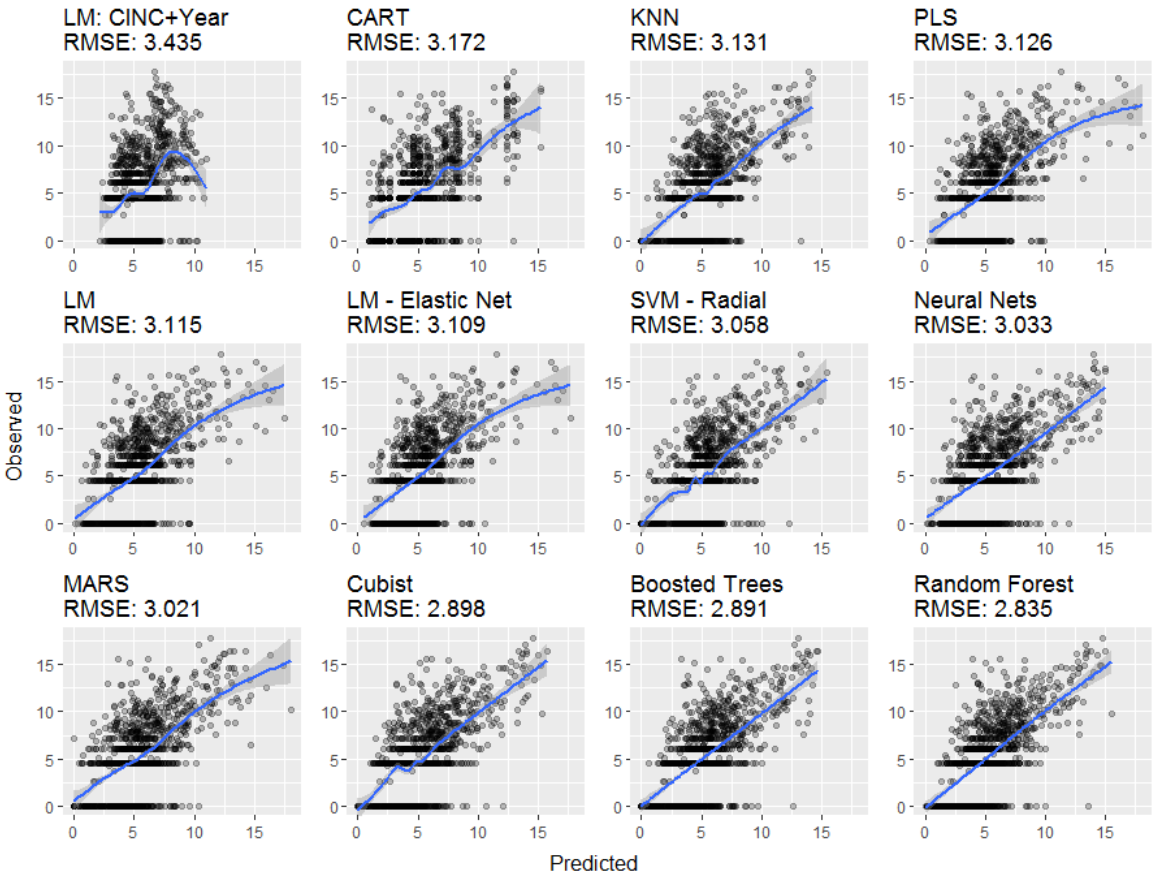

Figure 1.3: Scatter plots of model performance across each of the candidate models. A model which predicts perfectly would see all observations fitting along a 45 degree line from the bottom left to top right.

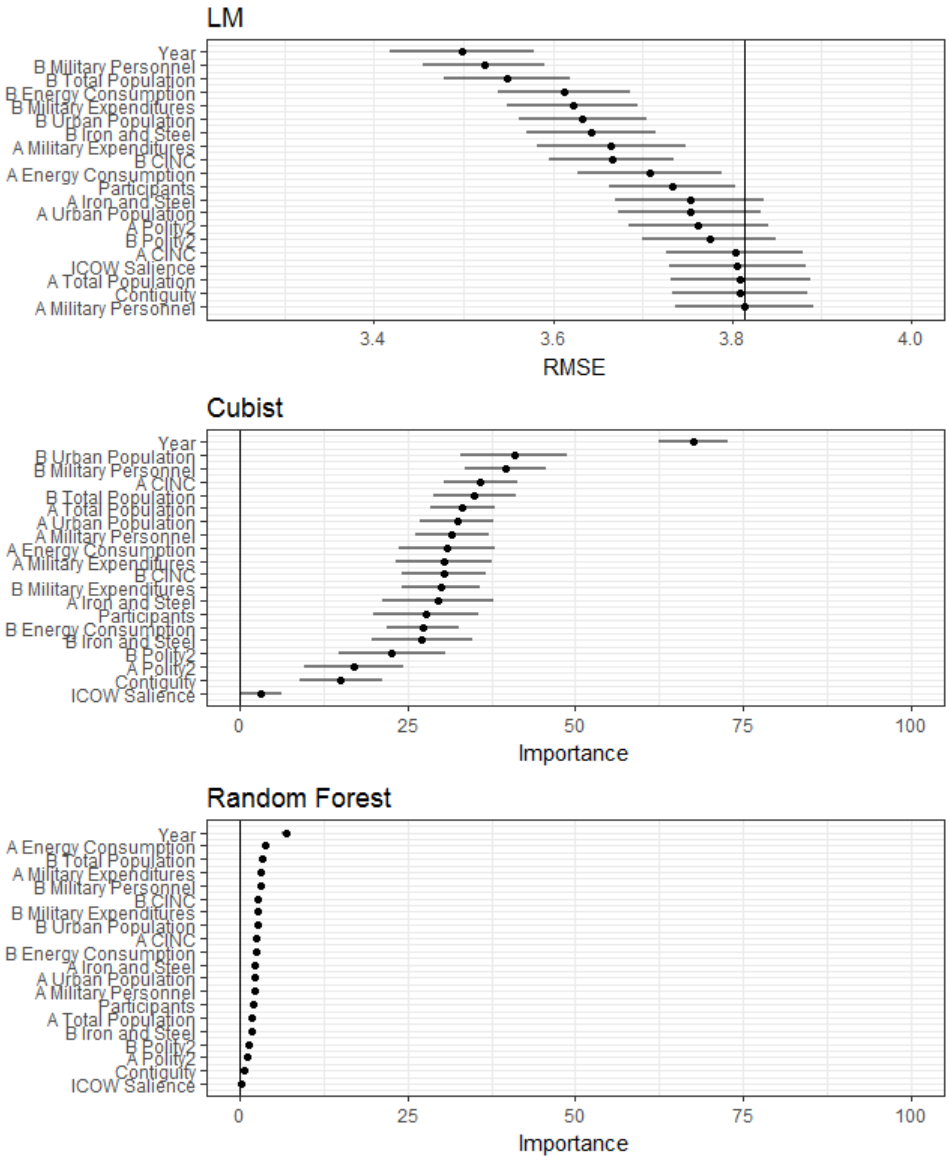

Figure 1.4: Variable importance plots from linear models, Random Forests, and Cubist, fit to the aggregate data. Bootstrapped 1000 times to produce the resulting confidence interval.

The first way I assess individual predictors is by using variable importance scores from linear models, Cubist, and random forests, following (Hill and Jones, 2014). The intuition behind these variable importance score is to capture the mean decrease in test error which results from randomly leaving out a predictor variable. If a predictor is strongly related to the outcome, then leaving out that predictor will result in decreased performance for the model. If a predictor has no relationship with the outcome, then we would expect no meaningful decrease in performance. For a linear model, I assess the predictive performance by individually including each variable and using cross validation to estimate its test error. Cubist shows the percentage of times each variable was used in its terminal node after conducting internal feature selection. Finally, the random forest variable importance scores are computed by averaging the amount of change in test set performance when each variable is permuted from the forest.

Figure 1.4 displays these variable importance scores. 'Year' emerges as the most important predictor, indicating the importance of time in modeling conflict in the international system, consistent with the work of Jenke and Gelpi (2017). As will be seen in the following sections, there are a number of structural breaks in expected battle deaths over time, with costs decreasing generally in the second half of the 20th century. Beyond just time effects, the general trend seems to be that variables capturing the raw military capabilities of the states involved are the most important in predicting the number of battle deaths. Military personnel, population, and energy consumption consistently emerge as the most important predictors across each of these three models.

A more interesting finding is that the institutional variables - state A and B's Polity scores - offer little in the way of additional predictive power with all other variables in the model. One reading of this might be that, for all of the work that is devoted to the role of political institutions and how they relate to conflict, raw capabilities in the form of military, energy, and population

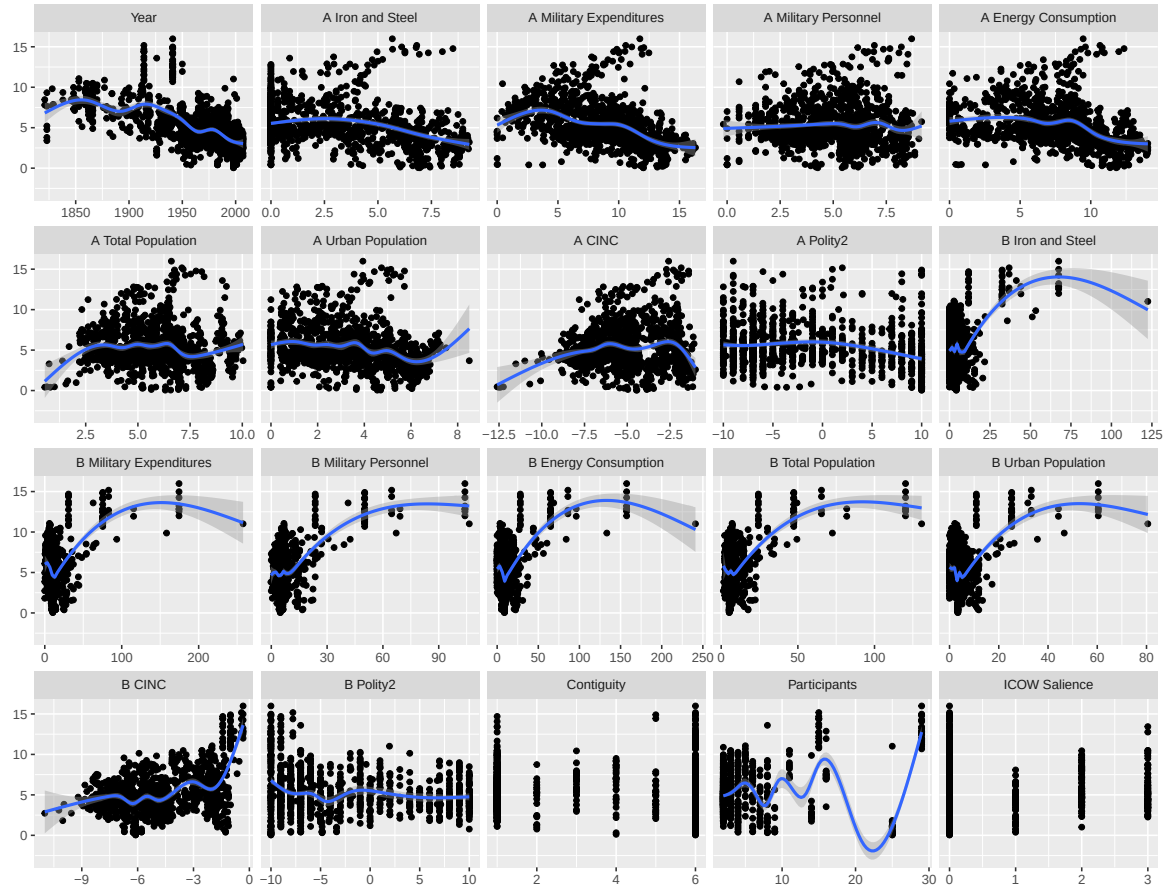

Figure 1.5: Partial dependence plots from each of the variables in a random forest fit to the entirety of the aggregated data, with and without year included as a predictor (with seven randomly selected predictors and 500 trees).

are the most important thing for predicting a war outcome given that conflict has started. That is, while costly signaling and diplomacy may matter for the onset of conflict, once it is underway material capabilities are ultimately what matter for predicting how that conflict will unfold. Even this reading of the results must be met with some hesitation, as the effects of institutions are likely felt in a state's economic and military development. Additionally, variable importance scores in and of themselves must be read with some caution, as Strobl et al. (2007) finds that importance scores from random forests are prone to bias in the presence of correlated predictors.

While variable importance scores might give us some idea of what is useful for prediction, they do not give us any sense of the relationships in the dataset. One way of exploring the relationship between X and Y is by using partial dependence plots from the random forest. These provide a means of assessing the marginal change in the outcome variable (State A's battle deaths) based on changes in a predictor while holding all other predictors at their observed values. These are useful for identifying nonlinearities in the relationship between predictors and the outcome variable, though they may be misleading due to hidden interactions in the data. Figure 1.5 shows the partial dependence of battle deaths on each of the variables in the aggregate data, with each point being the predictions of the forest at that particular value of the predictor. We can see here the relationship between time and predicted battle deaths, with a general decrease in battle deaths over time. We can also see the linear relationships we might expect in state A's military expenditures, iron and steel, and energy consumption: militarized, industrious states have lower predicted battle deaths, on average. Additionally of note is that the predictions do not differ meaningfully based on either state A or B's polity score.

One immediate question might be how these results differ from that of standard linear modeling. To check this, I regressed state A's battle deaths on the same set of predictors, with and without country fixed effects, including a cubic polynomial for time. These models offer somewhat inconsistent findings relative to that of the random forest. While the variable importance scores indicate that A's energy consumption and B's population offer the biggest improvements for out of sample prediction, the linear model does not find a significant relationship for either of these variables, nor does it find any evidence of an effect for A's military expenditures. The linear models do identify some similar findings, as B's military expenditures and personnel are positive and significant while A's iron and steel is negative and significant. Contiguity is also negative and significant, while the

random forest indicates that it offers little value for out of sample prediction. The overall lesson is that while linear modeling can uncover some of the same findings as that of the other methods, relying strictly on statistical significance can lead us to overlook interesting relationships and patterns in the data.

Table 1.3: Results of linear models using all predictors with a cubic polynomial for year, with and without country fixed effects. 95% confidence interval reported from 1000 bootstraps. Results averaged across 10 imputations.

	Linear Model 1			Linear Model 2		
		95% CI			95% CI	
Variable	Coef	LB	UB	Coef	LB	UB
(Intercept)	0.4820	−4.5715	5.1757	−2.6765	−11.9084	7.2667
A Iron and Steel	−0.2160	−0.3246	−0.1054	−0.1342	−0.3132	0.0468
A Military Expenditures	0.0150	−0.1204	0.1565	0.1364	−0.0640	0.3173
A Military Personnel	0.3647	0.1478	0.6025	0.4796	0.1408	0.8966
A Energy Consumption	−0.0611	−0.1680	0.0471	−0.0424	−0.2062	0.1273
A Total Population	0.6335	0.3500	0.9184	0.4309	−0.5665	1.3385
A Urban Population	−0.2650	−0.5526	0.0303	−0.1967	−0.7877	0.3527
A CINC	−0.3136	−0.7724	0.1017	−0.6046	−1.3812	0.1350
A Polity2	0.0007	−0.0232	0.0244	0.0074	−0.0385	0.0511
B Iron and Steel	−0.0524	−0.1287	0.0275	−0.0598	−0.1573	0.0250
B Military Expenditures	0.1393	0.0573	0.2103	0.1144	0.0226	0.2001
B Military Personnel	0.3046	0.2024	0.4049	0.3963	0.2723	0.5315
B Energy Consumption	−0.0742	−0.1412	0.0028	−0.0691	−0.1486	0.0345
B Total Population	−0.0779	−0.1897	0.0454	−0.1313	−0.2700	0.0136
B Urban Population	−0.3374	−0.5506	−0.1415	−0.3377	−0.6028	−0.1235
B CINC	0.1817	0.0389	0.3156	0.1074	−0.0500	0.2785
B Polity2	−0.0180	−0.0422	0.0068	−0.0017	−0.0296	0.0304
Contiguity	−0.2204	−0.2966	−0.1430	−0.2444	−0.3515	−0.1578
Participants	−0.0158	−0.0502	0.0214	0.0223	−0.0192	0.0810
ICOW Salience	−0.0355	−0.2425	0.1726	0.1281	−0.2305	0.4170
N		1189			1189	
Country Fixed Effects?		No			Yes	
Year Polynomial?		Yes			Yes	

1.3.3 The New Measure: Dispute Casualty Expectations (DiCE)

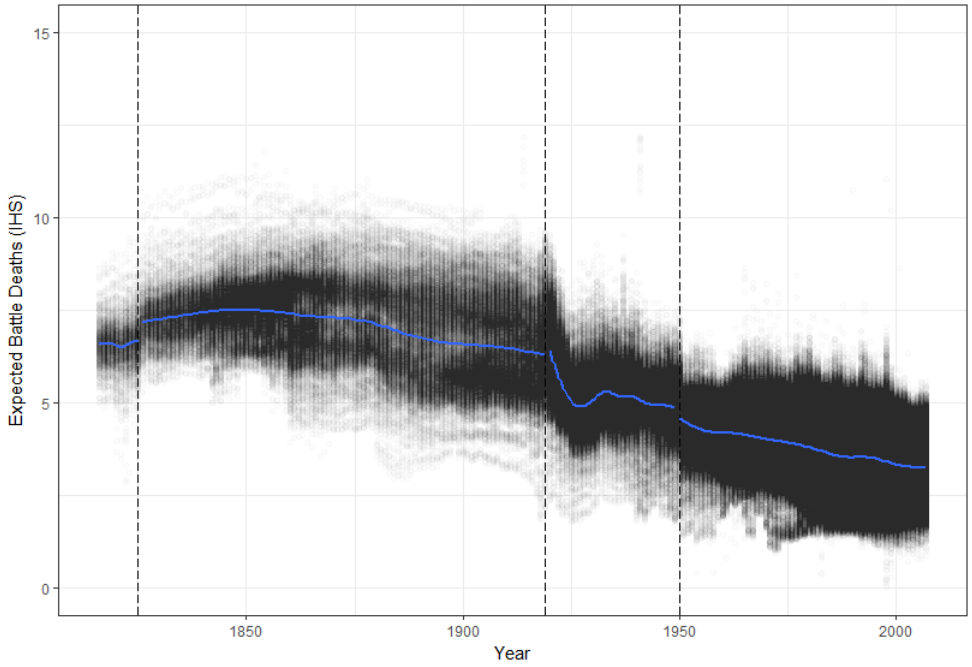

Figure 1.6: Expected battle deaths from a random forest fit using the aggregate opponent approach, with LOESS lines and breaks at 1826, 1919, and 1945. Results averaged across 10 imputations.

The main task of this paper is to produce the best estimates for battle deaths in hypothetical disputes in the international system. I use a random forest to predict all directed dyads in the Correlates of War dataverse between 1816 and 2007. This yields an estimate of battle deaths for hypothetical bilateral military disputes between members of the international system, which I argue serves as the best existing proxy for the expected costs of war. I call the resulting measure DiCE (Dispute Casualty Expectations) which is visualized in Figure 1.6. The figure illustrates the general pattern of battle deaths from interstate conflict over the last two centuries. There are a number of 'breaks' in these estimates, occurring around the years 1826, 1919, and 1945. In particular, the

general reduction in battle deaths from interstate conflict after 1950 has been a source of academic interest (Braumoeller, 2013; Pinker, 2011). [20]

At this stage in the paper, I originally sought to replicate existing models in international conflict and show how a measure of war costs can be used to improve model fit and inform existing theory in the literature on conflict. But I have generally found that there is very little existing empirical work which explicitly invokes variation on war costs for hypothesis testing. This is likely due to the fact that we have no existing measure of expected war costs. In my estimation, the field has engaged in theorizing and testing the role of national capabilities because we have had indicators, however crude, of national capabilities for many years. I hope that the field can, with measure of expected war costs, develop theory which explicitly invokes expectations about the costs of war.

1.3.4 Substantive Examples

To gain some sense of what the measure looks like, Figures 1.7-1.10 show expected casualties for specific pairings of dyads in the international system. At first glance these results might not map well to our own expectations of war costs. Indeed, we would expect a conflict between the US and Russia in 1980 to have been much more costly than a conflict between Brazil and Argentina. This hypothetical indicates the results here must be interpreted with care, as they merely reflect the best predictions based on observed conflicts in the international system. The model is making low predictions for conflicts between major powers after 1950 because we have not observed a major interstate war in this time period.

[20] Though these estimates are bilateral, the methodology here is flexible and can easily extend to itself to hypothetical multilateral conflicts - I discuss this in the appendix when I discuss the impact of alliances.

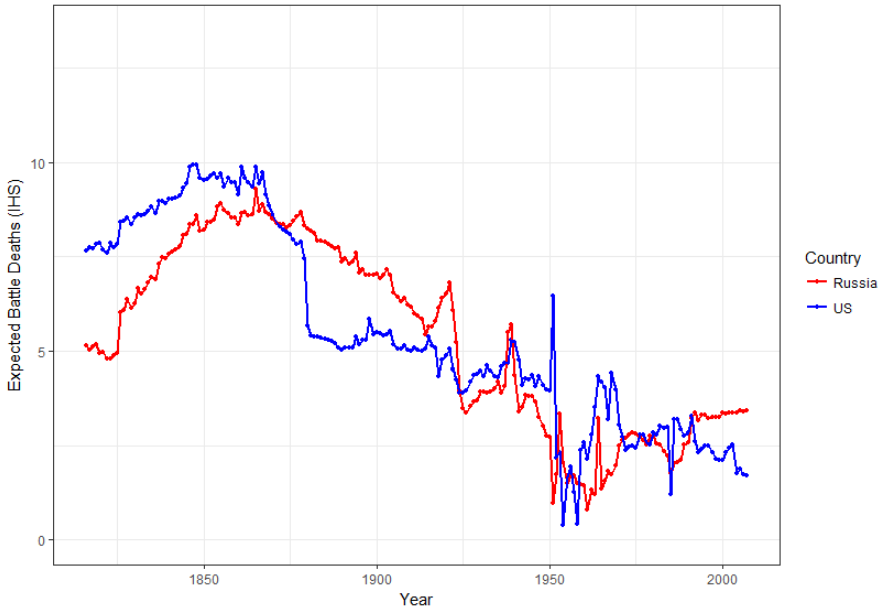

Figure 1.7: DiCE: US-Russia

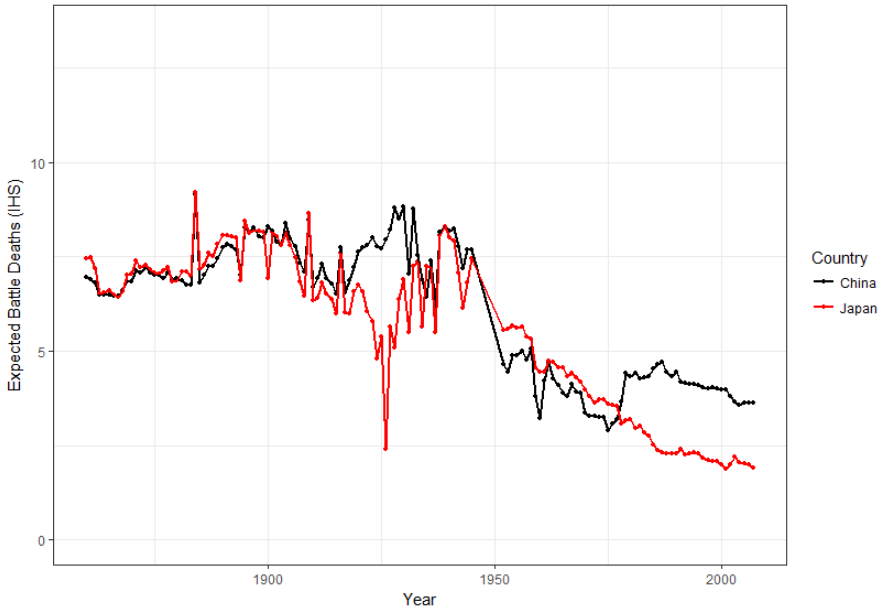

Figure 1.8: DiCE: China-Japan

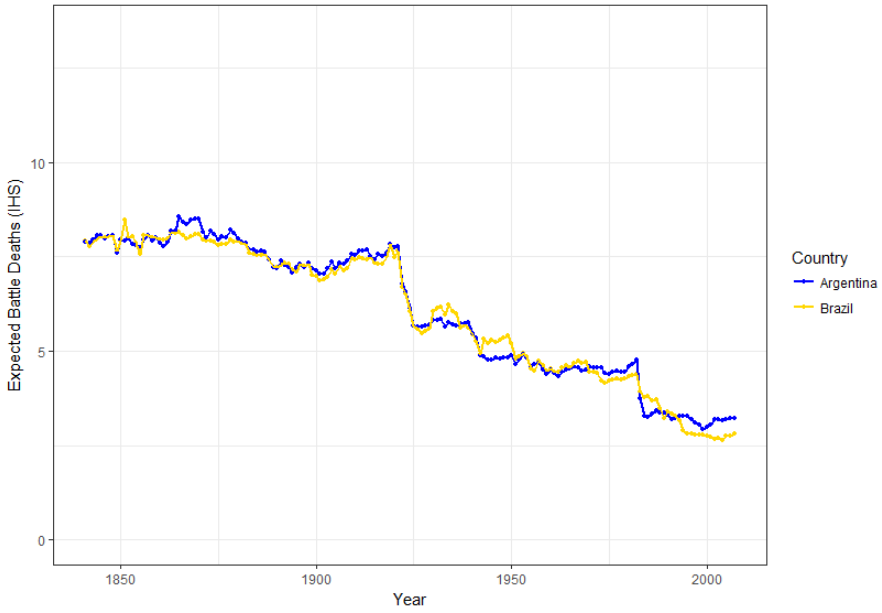

Figure 1.9: DiCE: Brazil-Argentina

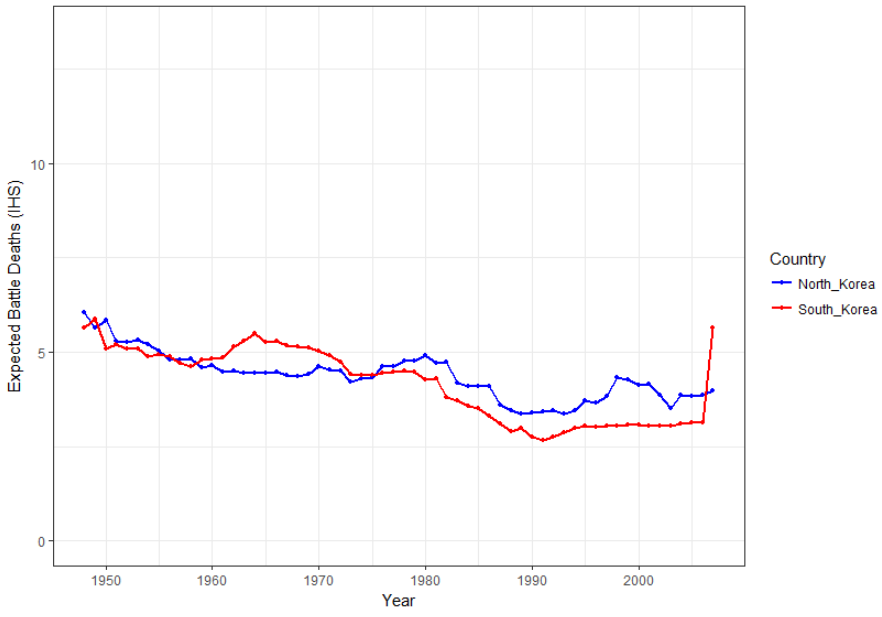

Figure 1.10: DiCE: North Korea - South Korea

Though there is some variation after 1950, the expected costs for hypothetical conflicts between these countries have generally been decreasing over time. The general pattern of costs decreasing over time is not only because there are few instances of disputes between major powers, but also because when two large powers do engage in military conflict, it is generally resolved before it escalates into a major war. If the pattern of interstate peace were to stop and war was to become more prevalent in the international system, we would expect the model to promptly update its predictions. Because of this, the estimates presented here should be thought of in Clausewitzian terms, as the expected costs of *limited* rather than *absolute* war.

An immediate way to check the performance of the model is in looking at particular instances of conflict. Figure 1.11 shows the expected battle deaths for conflicts which occurred after 1980 against the models bootstrapped predictions. Here we can see that the model is clearly imperfect, as it fails to capture the magnitude of the 1980 Iran-Iraq conflict. Similarly, the model understated the battle deaths of the 1982 Israel-Syria conflict, as well as the US's battle deaths in the two Iraq Wars. However, the model does a reasonable job in capturing the relative difference in costs between the two states in each of these conflicts, and performs admirably in predicting the asymmetric costs for both states involved in the conflicts between the US and Yugoslavia and Afghanistan, respectively.

1.4 Conclusion

This paper is intended to be a targeted exploration at predicting the costs of war. This, I argue, fills a gap in the literature for applied researchers of international conflict. More importantly, I believe this measure can be used to build and develop novel theories which explicitly incorporate expectations of war costs, which will aid applied researchers in the study of all aspects of international relations.

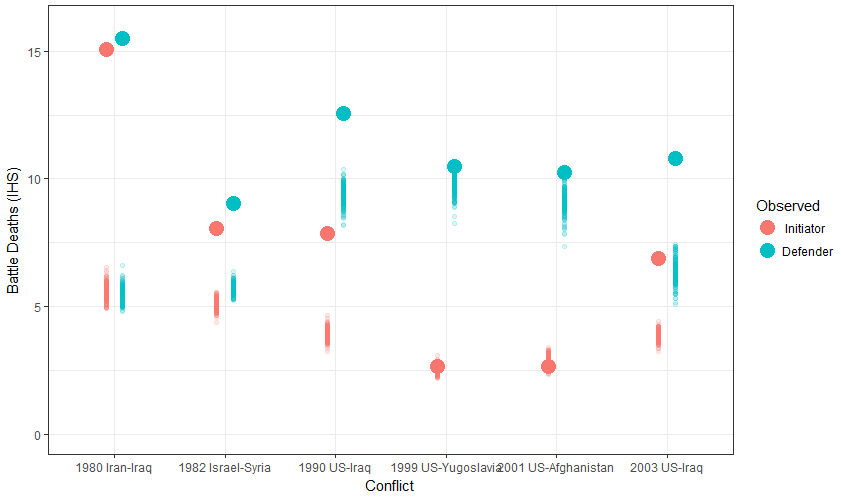

Figure 1.11: Expected battle deaths for conflicts post 1990 compared to observed battle deaths, bootstrapped 100 times.

There are a number of limitations with my approach, the first of which is that I am only able to speak to the expected costs of war in terms of military battle deaths. To say nothing of civilian deaths, there are a wide variety of costs associated with fighting that are of interest to states. Future work must continue to explore and expand on the approach here in using other forms of war outcomes to proxy for war costs. Second, I use a limited number of predictors in order to model cost expectations for the entirety of the time period of 1816-2007. The methodology here is intended to be the first cut at predicting the costs of war. The results here can be seen as the baseline model, and researchers can easily incorporate additional predictors in an effort to improve upon these predictions.

In many ways the question raised by this paper is more important than the results I find or the measure I eventually produce. This is only a first step towards developing a measure of war costs. But in so doing I hope to show that by using a different set of tools we can improve the study of conflict in ways not yet fully realized.

CHAPTER 2

A NEW MEASURE OF FOREIGN THREAT

Foreign threats are critical to the study of international relations and conflict. Waltz's (1979) foundational work points to the importance of the international environment in shaping the behavior of states, as anarchy creates an arena in which states are motivated by a desire for survival.[1] But as important as external threats may be in the lead up to war, the impact of threats extends far beyond international conflict. A state's security environment affects not only the possibility of conflict but is fundamental to their very construction. Waltz posited that states resort to developing military capabilities and forming alliances in order to ward off external foes, and others have argued this decision to mobilize resources for the possibility of war is vital to a state's construction. Tilly points to preparation for war as a key factor for the consolidation of state power in European history as states responded to a threatening security environment: "the largest and most persistent stimulus to increases or changes in national fiscal burdens over the great period of European state-making was the the effort to build armed forces and wage war" (1975, 54). This viewpoint is echoed by Thompson (1996), who argues that a state's geopolitical safety is critical in determining a state's choice of political system, whether autocratic or democratic.[2] This second-image-reversed emphasis on the impact of the international environment calls into question fundamental theories

[1] Organski (1958) examined this concept in power transition theory, arguing that conflict will most likely emerge when there is a change in the distribution of power in the international system - when the landscape of the international environment begins to shift and states are forced to re-assess the potential for conflict. Power transitions have been a key instance in which expectations of foreign threats have fomented conflict. Persia and Rome's conflicts were marked by their fear of each other's capabilities and the threat each posed to the other (Frankopan, 2015); Germany's decision to wage war in 1939 was in direct response to the threat posed by the potential power of the rising Soviet Union (Copeland, 2000, 119).

[2] This perspective goes back to Hintze Hintze, Gilbert and Berdahl (1975) and his historical essays on the organization of the state.

in international relations, as it implies that democratization may result from, rather than produce, peace.[3]

When do states view each other as threats? Hobbesian and Waltzian logic implies that in a system characterized by anarchy all states view each other as potential threats. But it stands to reason that some states are more threatening than others, whether through geography, alliances, or past realizations of conflict. With this in mind the field has used a number of different indicators of the core theoretical concept of external threats: militarized disputes, alliances, strategic rivalries, protracted conflict, and territorial challenges have all seen use as measures of a state's international security environment. Researchers have used these measurers to examine the conceptual importance of external threats in a wide variety of settings: state building (Tilly et al., 1992; Thompson, 1996; Desch, 1996; Lu and Thies, 2013), tax and spending policies (Lamborn, 1985), democratic transitions and democratic survival (Gibler and Wolford, 2006; Gibler and Sewell, 2006), civil-military relations (Desch, 1998; Feaver, 1999), military spending (Nordhaus, O'Neal and Russett, 2012), political participation and political trust (Hutchison, 2011*a*,*b*), state repression and diversionary conflict (Poe and Tate, 1994; Poe, Tate and Keith, 1999; Enterline and Gleditsch, 2000; Davies, 2016), international conflict initiation (Ghatak, Gold and Prins, 2017), and coup initiation (Arbatli and Arbatli, 2016), among many others. In short, external threat remains a prominent theoretical concept for researchers in international relations.

The importance of this concept to researchers demands a re-examination of both the conceptualization of foreign threats and the operationalization of measures used in the literature (Adcock, 2001). Towards this end, there are a number of concerns with the current state of measurement for this key theoretical concept. Critically, many of the measures currently used (militarized disputes,

[3]See Gibler (2007) and Gibler and Tir (2010) for prominent empirical examinations of this argument.

protracted conflict, territorial challenges) are indicators of conflict only when it is already taking place. To illustrate, it is common practice for a researcher to invoke the idea of external threats in influencing a state's policy decisions, and then use the presence of a militarized dispute to proxy for the presence of a threatening security environment. If a state is in a militarized dispute, it is said to be under threat; if a state is not in a militarized dispute, it is not. Here it is vital to make a conceptual point: it is not only the realization of conflict which should influence state behavior, but the *expectation* of foreign conflict. Sparta militarized in response to rise of Athena and the expectation of conflict to to come; Germany decided to wage war in the present because of the expectation of future conflict with Russia. Conceptually, a state's external threat environment can best be conceived of as a latent variable which we should expect to influence behavior in many cases *before conflict actually takes place*. This conception of threats is best described by Schelling, who writes:

> it is the threat of damage, or of more damage to come, that can make someone yield or comply. It is latent violence that violence that can influence someone's choice - violence can still be withheld or inflicted, or that a victim believes can be withheld or inflicted... *it is the expectation of more violence that gets the wanted behavior*, if the power to hurt can get it at all" (1966, 3).[4]

From this point of view some measures - such as alliances and rivalry - may be better served to proxy for a state's security environment than the presence of a militarized dispute. By making conflict more costly allies might reduce the expectation of being attacked; states who having fought in the past may be more likely to fight in the present. But both of these indicators capture only one aspect of external threats and neither can wholly capture the picture of when conflict is most

[4]It is important here to note that Schelling is primarily speaking to crisis bargaining and the issuance of compellant threats, such as those measured by Sechser. But this same idea applies to the conception of a state's security environment - it is the power to hurt before conflict takes place which should influence state behavior.

likely to occur. If a state possesses allies but has a strategic rival in close proximity, we would expect that to be different than a state with only allies. Similarly, a strategic rivalry may be of lesser importance if a state is in a position of economic and military superiority. These existing measures can surely aid us in our search for a proxy of external threat, but ultimately we would like a measure that has been produced with a wide variety of information about states beyond the presence of alliances or rivalries.

In order to measure the expectation of external threats for states, we need to answer the following question: how likely is a state to be attacked by another state in a given year? To operationalize this theoretical concept, I argue that we can instead ask, how likely is the state to be the target of a fatal militarized dispute by all other states in a given year? Given the existing data in the Correlates of War universe, scholars have often selected fatal MIDs as the outcome of interest in order to study international conflict behavior. I similarly can use this dataset and approach it from the perspective of prediction: given information about two states in a given year, we want to estimate the probability of conflict (even in settings where conflict did not occur) and be reasonably confident in our predictions.

One prominent paper in the literature already adopts such an approach, as (Nordhaus, O'Neal and Russett, 2012) make a novel contribution to the measure of foreign threats by directly estimating the probability of conflict between states and constructing a state-level measure of the probability a state becomes involved in conflict in a given year. While this effort speaks more closely to the theoretical concept than others, their measure extends only from 1950-2000 and assumes an a priori model specification for external threats. For the purpose of this paper, it is vital to assess the predictive accuracy of the model in a rigorous way in order to trust the probabilities stemming from our model. If our model is a poor fit to the data, the resulting measure will fail to

accurately reflect expectations of conflict. Given that this is a predictive problem, I aim to improve on the efforts of Nordhaus, O'Neal, and Russet by using flexible models from machine learning which are noted for their predictive performance while imposing few assumptions on the structure of the data. Namely, I create a stacked ensemble using extreme gradient boosted trees as a meta learner on an ensemble of candidate learners (neural networks, random forests, logistic regression) to maximize out of sample performance in predicting fatal MID initiation over the years 1870-2001. I then apply the results of the ensemble model to all interstate dyads to estimate the probability of conflict initiation between all states. I sum these probabilities for each country in each year to create a state-level measure of the international security environment for each state. This, I argue, better proxies for foreign threats than current approaches, offering a new measure of an important concept for researchers in the study of international relations.

This paper proceeds as follows: I first examine notable IR theory on the concept of foreign threats, defining and clarifying the measure I aim to create. Next, I outline limitations with existing proxies and detail my methodology for creating a new measure. I then discuss the results of my predictive approach and seek to unpack and compare the models used for prediction. Finally, I present the new measure and conclude with advice for researchers.

2.1 Conceptualizing Threat

2.1.1 Uses in the Literature

The word threat has a wide variety of uses in the study of international relations and international conflict. For the purpose of this paper, I emphasize the core concept as originally posited by Hobbes: the prospect of being attacked by another state. No sovereign actor exists who can enforce cooperation between states, and as a result, states exist in an environment which is by default said

to be threatening. Anarchy ensures that 'each state must guarantee its own survival since no other actor will provide its security. *All other states are potential threats*, and no international institution is capable of enforcing order or punishing powerful aggressors [emphasis added] (Mearsheimer, 1990, 12)".[5]

With this in mind, the key theoretical concept of threat involves a state's prospective chances of being targeted by another power in militarized conflict. In his famous discussion of the security dilemma, Jervis (1978) notes that the most important part of a state's security environment is 'the perception of threat (that is, the estimate of whether the other [state] will cooperate).' (176). Many different issues may prevent states from reaching mutually preferred bargains, but he writes 'more frequently, the concern is with *direct attack* [emphasis added]' (169). It is this possibility of attack which ultimately produces the security dilemma and the potential for conflict. The security dilemma arises because of the possibility of direct attack - states monitor potential threats and adjust their behavior accordingly, and this often produces conflict.

Similarly, studies on deterrence and alliance formation also focused precisely on the possibility of being attacked by states or aggregations of states in the formation of alliances. In assessing the decision to bandwagon or balance via alliances, Walt (1985) argues that states do not mechanically ally against the most dominant power, but instead against the state they perceive to be the most threatening:

> it is more accurate to say that states will ally with or against the most *threatening* power. For example, states may balance by allying with other strong states, if a weaker power is more dangerous for other reasons. Thus the coalitions that defeated Germany

[5]Though this view is traditionally held by realists, this point is not without contention in the literature; see Wendt (1992) discussion of an anarchy for a constructivist critique of this argument.

in World Wars I and II were vastly superior in total resources, but united by their common recognition that German expansionism posed the greater danger' (8-9).

The emphasis in each of these cases on threat involves a state's baseline expectations of being attacked. In discussing the impacts of international security on state behavior, theories implicitly assume that states monitor and adjust their behavior in response to the potential of future conflict.[6] Though the core concept might be thought of as state specific (how likely a state is to be attacked?), scholars often speak of threat by referring to the level of security in the international system as a whole. Certain periods of history are thought to have been more stable than others, such as end of the Napoleonic Wars until the Crimean War, or the 'Long Peace' created by the presence of nuclear weapons in the aftermath of World War II. In focusing on the macro level of threat, scholars often reach conflicting conclusions and predictions. For instance, Mearsheimer (1990) argued that the end of the Cold War would induce a period of heightened threat throughout the international system as the world adjusted from bipolarity to multipolarity. In contrast, Desch (1996) believed "the end of the Cold War may represent a "threat trough" - a period of significantly reduced international security competition (237)" as states shift away from traditional alliances. Differences aside, such arguments make predictions about the level of threat in the international system as a whole. But in order to measure this quantity of interest, the first step is to focus on state specific interactions. As Gilpin (1988) writes in his discussion of hegemonic war, "The relations among individual states can be conceived as a system; the behavior of states is determined in large part by their strategic

[6]Here it is important to note that there is another frequent use of the term threat in the literature. At a more granular level, crisis bargaining is concerned with the stages and results of explicitly delivered threats, whether to compel or deter, on the eve of conflict. Audience cost theory has its origins in this area, as Fearon (1994, 1997), Schultz (1999) and Sechser (2010) are concerned principally with how states respond to demands made by other countries. In these cases, a threat is delivered in a demand to obtain a bargain, where conflict will occur in the event that a bargain is not reached. In this case, the threat of conflict is very directly speaking to the possibility of being attacked.

interaction" (592). Because of this, in order to measure the security environment of the international system, we can aggregate up from dyadic interactions of states.

2.1.2 Threat Expectations

Given the discussion so far about the use of the term in the literature, it is important to make a theoretical clarification. Fundamental to the core concept of threat is the possibility, or expectation, of being attacked. It is therefore this *expectation* which influences behavior, rather than simply the *realization* of conflict. While it is certainly true that states behave differently while at war, we should expect to see states strategically adapt to their environment before conflict takes place. To limit the role of international conflict to its realization suggests that leaders consider international factors only after they have already involved themselves in war. This does not fit with our conception of political leaders as strategically minded actors who select policies in anticipation of decisions made by other actors. As Bueno de Mesquita and Siverson (1995) write, "Leaders, of course, recognize the existence of opposition and the designs of others on the office they hold. They consequently select policies to minimize the opportunities available to those seeking to remove them from power" (842). Instead, if leaders are forward thinking, they should factor in the threat of foreign conflict into their decision making often before conflict is realized. By this logic, it is not the actual realization of the war which matters so much as the expectation of foreign conflict.

From this perspective, it is insufficient to study the importance of foreign threats by focusing only on events in which conflict occurred. There are numerous instances throughout history of states altering their behavior in the anticipation of conflict even when conflict did not ultimately come to pass. The emergence of the Ottoman Empire following its capture of Egypt in 1517 and military successes against Hungary in 1526 led European powers to brace for the prospect of a war

within central Europe.[7] As Jervis notes in his 1976 discussion of the spiral model for war, states often act on the basis of threats which are not realized:

> ...each state is protected only by its own strength. Furthermore, statesmen realize that, even if others currently harbor no aggressive designs, there is nothing to guarantee that they will not later develop them. In the 1920s Canada's only war plan "held that the principal external threat to the security of Canada lay in the possibility of armed invasion by the forces of the United States" - leading the director of military operations to engage in reconnaissance missions in the Pacific Northwest. (162)

Jervis points to this notion by contrasting the policies of Britain and Austria after the Napeolonic Wars. The key difference between these two states was their international vulnerability - the possibility of being attacked by another power. Because of its geographical isolation, Britain was able to adopt a less centralized political system with little intention of interfering in the internal affairs of other states. Austria, surrounded by strong powers, adopted a centralized political system which defended its own right to engage in policy which would affect other state's security. For Austria, the security dilemma was present, whereas in Britain it was not. Similarly, in the lead up to World War II, Britain and France differed with respect to their foreign policy in their assessment of Germany. For France, Germany's rise in power appeared highly threatening; to its interests the construction of the Maginot Line was an effort to protect itself against foreign threats. Britain, believing itself geographically isolated, did not make such preparations and believed conciliation would succeed in preventing future conflict. In each of these cases, policies undertaken were in response to the state's initial security environment and its expectations of being attacked.

[7] Pope Leo X stated at the time: "Now that the most atrocious Turk has captured Egypt and Alexandria and the whole of the eastern Roman empire, he will covet not just Sicily and Italy but the whole world". Quoted in the Frankopan 2015.

The emphasis on expectations here is nothing new, as theories of international conflict often invoke expectations of conflict. Powell (2006) focuses on capability changes and points to the possibility of expected conflict in the future creating conflict in present, hearkening back to the concept presented by Thucydides. Similarly, studies which focus on the role of executives routinely assume that state leaders are strategic and adjust their behavior in anticipation of what will happen in the future (Ritter, 2014). Therefore, in order to be consistent with our assumptions of strategic behavior, we should be assessing the *possibility* or *expectation* of conflict in order to measure the presence of threat. Conceptually, the appropriate measure of external threats should reflect expectations of conflict - more specifically, expectations of being attacked. In summary, for the purpose of international relations theory, much has been written about the effects of the security environment at both the international and state level. I argue that in order to accurately measure the concept of threat for each state, we need to estimate the possibility of conflict for each specific state.

2.1.3 Measuring Threat

Due to its conceptual importance, there have been a number of different approaches taken to measuring foreign threats in the literature. One prominent effort to measure this concept exists in the discussion of state rivalries. As many conflicts in history have been recurring between specific pairs of countries, researchers have aimed to identify pairs of countries which have experienced high levels of conflict in the past, classifying these pairs of rivals. Goertz and Diehl (1995) and Bennett (1997) examined the density of disputes between states and classified states as rivals - which we may think of as a specific instance of external threat - when x number of disputes occurred in y

number of years.[8] By using observed conflict these scholars aimed to develop a measure which indicates when pairs of states will be more likely to escalate a conflict.

But there some limitations with this approach. First, the measure explicitly requires the onset of conflict in the past, limiting its ability to predict the possibility of conflict in settings where conflict has not yet occurred. While in many cases a rivalry will reflect that a state is about to experience conflict, it will fail to classify instances in which conflict occured from a different state. This is a crucial component to the concept of foreign threats, and a measure of this concept should be able to identify when conflict is likely to occur from a wide variety of states. Second, the dispute density approach does not necessarily categorize rivalries as well as one would expect. Thompson (1995) demonstrates that this approach fails to identify rivalries in dyads we would expect (Germany-Russia) while classifying rivalry in dyads we would not expect (United States-Ecuador). To improve on these efforts, Thompson (2001) and Colaresi and Thompson (2002) aimed to identify rivalry through historical research, defining strategic rivalry as a competitive relationship between independent states where both states identify the other as an enemy and an explicit threat. While an improvement, this measure still does not fully capture the conceptual idea of a state's security environment. Rivalry may be an important feature in assessing whether a state is likely to be attacked - a relationship I examine later in this paper - but by itself it does not capture the entirety of a state's security environment. Instead, as I will show, it is important to know a wider range of information (in addition to rivalry) about a pair of states when predicting their probability of conflict.

The closest approach to fully measuring the presence of foreign threats comes from Nordhaus, O'Neal and Russett (2012). These authors estimate the probability of a fatal militarized interstate

[8]This approach is itself similar to the work done on protracted conflict citepbrecher1997study; see Colaresi and Thompson (2002) for a full discussion.

dispute (MID) for all pairs of states in the international system, then aggregate these probabilities to produce a state-level measure of the estimated probability of foreign conflict. While an improvement over the dispute density approach, there are a number of limitations with the measure they ultimately introduce. The first issue is methodological. Though they use a classifier (logistic regression) to produce probabilities of fatal MIDs, Nordhaus, O'Neal and Russett (2012) do not evaluate how well their model does in predicting fatal MID onset. This is a critical step in validating the measure, as they are using probabilities from the model to proxy for the presence of foreign threats. The extent to which these probabilities succeed as a proxy depends on how well these probabilities accurately predict MIDs *out of sample* - if it was to make predictions on a new collection of data. Though the authors include a suite of variables from the liberal and realist schools of conflict, the most important criterion is not the inclusion of features which are thought to matter or the in-sample fit of the model, but the ability of the model to predict well when exposed to new data. If the model is overfit on their training data, it will perform poorly in generalizing to new data and its probabilities will be unreliable as a measure for the latent concept. Conversely, if the model predicts well when exposed to new data, we can be confident in using the model's predictions as a proxy. It is therefore important to spend a considerable amount of time assessing the predictive validity of the model, using techniques designed to maximize out of sample performance as I demonstrate later.

If this was the only issue with Nordhaus, O'Neal and Russett (2012), a quick fix would be to replicate their work and validate the predictive capabilities of their model. But the second issue with Nordhaus, O'Neal and Russett (2012) is theoretical. The authors estimated a model in order to proxy for the presence of foreign threats - how likely a state is to be attacked. But in so doing the authors structured their data using a nondirected approach. Theoretically, the appropriate

measure is directed, as a foreign threat entails the prospect of the opponent initiating conflict. If a state is the initiator of the conflict, it would be incorrect to automatically infer that they were being threatened. To illustrate this, consider if there are three states in the international system, A, B, and C. If we want to know the presence of foreign threats to state A, we wish to know the probability that countries B and C will initiate a MID with A. If we adopt the nondirected approach, we could estimate the probability that A becomes involved in a conflict with B and C, but this allows for the possibility that A is the one initiating the conflict. If we really want to know the threat environment facing A, we need to focus on when they will be targeted by B and C. This conceptual concept can be learned using data on directed conflict initiation, which I detail in the following section.

2.2 Towards a New Measure of Threat

2.2.1 The Approach

In order to develop a state-level measure of threat, I focus on predicting directed fatal MID onset over the years 1870-2001. In so doing, I utilize an ensemble of machine learning algorithms in order to maximize predictive performance when generalizing to new data. After examining a variety of candidate methods for classification and estimating their out of sample performance, I select the models which perform best on a separate hold out set. I then apply the results of the best performing models to all interstate dyads to estimate the probability of conflict initiation between all states. I sum these probabilities for each country in each year to create a state-level measure of the international security environment for each state. This, I argue, better proxies for foreign threats than current approaches, offering a new measure of an important concept for researchers in the study of international relations.

This paper's main contribution is the development of a new measure of a state's international security environment. But due to the predictive nature of this task I am able to make additional contributions to the study of militarized disputes and origins of international conflict. Though the algorithmic approach adopted here may be criticized for relying on 'black-box' methods in order to predict outcomes - and this criticism is often true for many opaque methods such as neural networks, support vector machines, or ensemble models - I am able to use a number of tools to speak to the relative predictive importance of features in the onset of fatal MIDs.[9]

While the field has traditionally relied on tests of statistical significance to determine whether a variable is an important determinant of conflict, this approach ignores the ability of a model to actually *predict* conflict, as variables which are statistically significant do not necessarily increase the ability of a model to predict an outcome of interest (Ward, Greenhill and Bakke, 2010). By assessing the *predictive* importance of traditional variables in the study of conflict initiation I am able to re examine what the field has come to believe 'matters' for the onset of conflict in search of a model that is externally valid (Fariss and Jones 2017). The emphasis on prediction in this paper is shared by recent work in political science which has highlighted the utility of predictive models for monitoring political events (Beger, Dorff and Ward, 2016) as well as validating the models from which we draw inferences (Montgomery, Hollenbach and Ward, 2015) (Hindman, 2015). In examining variable importance in this project, for instance, I am able to speak to whether conflict has changed over time by examining fluctuations in predictive performance, contributing to the work of Jenke and Gelpi (2017) in examining systemic changes in international conflict over the last century and a half.

[9]Even so, the criticism of algorithmic modeling for relying on black boxes is less applicable as research in machine learning has started to focus on the interpretability of its models. See Ribeiro, Singh and Guestrin (2016) for a discussion of innovations in explaining predictions in the classification setting.

Table 2.1: Data structure for modeling MID onset with Iraq and US as a motivating example. Note that the year itself (1990) is also included as a separate feature.

Dyad (A-B)	B Attacked	A Features	B Features	Dyad Features	System Features
US-Iraq	No	US Features	Iraq Features	US-Iraq Features	1990 Features
Iraq-US	Yes	Iraq Features	US Features	US-Iraq Features	1990 Features

2.2.2 Data

I now turn to a discussion of the data. I use directed data on fatal MID onset from the newly released set of MIDs from Gibler, Miller and Little (2016). The unit of analysis is at the dyad level, and the outcome to be classified is a binary variable for whether the first state in the dyad was targeted by the second state in a military dispute. Following Nordhaus, O'Neal and Russett (2012), I include only MIDs in which at least one participant experienced a fatality. This amounts to 2,012 instances of directed fatal MID onset among 1,017,988 total observations over the years 1870-2001. To clearly illustrate how the dataset is constructed, take the example of the conflict between the US and Iraq in 1990. This pair of countries will have two observations in the dataset, one in which the Iraq is classified as the target of a fatal MID initiated by the United States. The probability of fatal MID initiation is modeled as a function of features of each state (capabilities, overall trade, institutions), dyad features (alliances, trade), system features (number of states, number of democracies, and year.

Table 2.2: Predictors of MID onset - country level

Features	Description	Source
A&B Polity	Polity score representing level of democracy. Scores range from 10 (full democracy) to -10 (full autocracy)	Polity IV
A&B Population	Total population for each country	CoW National Material Capabilities 4.0
A&B Urban population	Urban population (total population living in cities with population greater than 100,000) for each country	CoW National Material Capabilities 4.0
A&B Iron and Steel	Iron and steel production for each country	CoW National Material Capabilities 4.0
A&B Energy	Energy production for each country	CoW National Material Capabilities 4.0
A&B Military Personnel	Total military personnel for each country	CoW National Material Capabilities 4.0
A&B Military expenditures	Total military expenditures for each country	CoW National Material Capabilities 4.0
A&B Age	Length of state tenure in the international system in years.	CoW State System Membership 2016
A&B Major Power Status	Dummy variable indicating whether each country is a major power	CoW State System Membership 2016
A&B Ongoing Civil War	Dummy variable indicating whether each country is experiencing a civil war.	CoW War Data
A&B Instability	Dummy variable indicating whether each country is experiencing political instability	Polity IV
A&B Occupied	Dummy variable indicating whether each country is occupied by a foreign power	Polity IV
A&B Ongoing MIDs	Dummy variable indicating whether each country is involved in an interstate conflict	MIDB 4.01
A&B Number MIDs	Count of ongoing interstate conflicts for each state	MIDB 4.01
A&B Alliances	Outside alliance memberships for each country	ATOP 3.0
A&B Total Imports	All trade imports for each country	CoW National Trade 3.0
A&B Total Exports	All trade exports for each country	CoW National Trade 3.0

Table 2.3: Predictors of MID onset - dyad level

Features	Description	Source
A&B Victory	Dispute Outcome Expectations (DOE); each country's probability of winning, losing, or tying in a potential military dispute between the two countries	Carroll and Kenkel (2016)
A&B Costs	Dispute Casualty Expectations (DiCE); each country's expected battle deaths for a potential conflict between countries the two countries	Henrickson (working)
A&B Imports	Total trade imports for each country from the other in the dyad	CoW Dyadic Trade 3.0
A&B Exports	Total trade exports for each country from the other in the dyad	CoW Dyadic Trade 3.0
A-B Rivalry	Dummy variable indicating whether the countries are considered to be strategic rivals	Thompson (2001)
A-B Joint Democracy	Dummy variable indicating whether both countries are democracies; coded as 1 if both states are 7 or greater on Polity	Polity IV
A-B Contiguity	Direct contiguity between the two countries	CoW Direct Contiguity 3.2
A-B Distance	Distance between the two countries capitals	CoW Direct Contiguity 3.2
A-B Territorial Disputes	Dummy variable indicating whether the two countries are involved in a territorial dispute	The Issue Correlates of War Project
A-B MIDinLast10	Dummy variable indicating whether the dyad has experienced a fatal MID in the last ten years	Gibler, Miller and Little (2016)
A-B PeaceYears	Number of years since last fatal dispute between the two countries	(Gibler, Miller and Little, 2016)

Table 2.4: Predictors of MID onset - system level

Features	Description	Source
Number of States	Number of states in the international system	CoW State System Membership 2016
Number of Democracies	Number of democracies in the international system (Polity >= 7)	CoW State System Membership 2016 & Polity IV
Percent of Democracies	Percent of total states in the international system which are democratic (states classified as Polity >= 7)	CoW State System Membership 2016& Polity IV

As the emphasis of this project is on prediction, the selection of features for modeling the response is guided by availability rather than theoretical expectations. Rather than assuming a model specification a priori, I include a wide variety of state, dyad, and system level features in addition to the year of the observation. Given the number of features I include, it is hard to discuss each individual predictor in detail. However, the inclusion of 'year' as a predictor is important to defend from a theoretical perspective. The object of this paper is to estimate a state's expected probability of foreign conflict in a given year. It is not an effort to *forecast* this probability ahead of time. Instead, my approach is to answer the following question: given information about two states in a particular year, what was the most reasonable expectation of foreign conflict? As such, year is included as a feature to capture time-varying effects in the onset of fatal MIDs. Its eventual importance as a feature is an indication of changes in the structure of conflict in the international system over time.

2.2.3 Predictive Criterion

In the classification setting, a common metric for evaluating predictive accuracy is the percent correctly predicted from the model. Namely, each observation has an estimated probability from the model of a MID occurring. Observations with a probability above some threshold, usually

50%, are classified as predicting the onset of a MID, while the others predict that a MID will not occur. These predictions are then compared to the actual values of the outcome variable and the resulting confusion matrix can be used to determine the percent correctly predicted. While intuitive, this approach is inappropriate for the setting of this fatal MIDs because of the severe class imbalance in the outcome variable. As most country pairs in a given year do not engage in a military dispute, a model will perform well on this metric by simply predicting that no MIDs will ever occur. Instead, in order to train a model to perform well in classifying onsets, I rely on the log-loss function for training (Hastie, Tibshirani and Friedman, 2009, 221), which is the negative of the average log-likelihood, and also report the area under the receiver operating curve (AUC) to assess the performance of the models as both of these metrics focus on the ability of the classifier to correctly separate the classes in the dataset. To illustrate the log-loss, let $\hat{f}$ be a model and $\hat{f}(X_i)$ be the prediction that the model gives for the ith observation, while Y_i is the actual outcome for that observation. This metric quantifies accuracy by penalizing a model for making false classifications. Lower values of the log loss equate to better predictions as a model which perfectly classifies the outcome would have the value of 0.

$$\ell(\hat{f}, X, Y) = -\frac{1}{N}\sum_{i=1}^{N} Y_i log\hat{f}(X_i) + (1 - Y_i)log(1 - \hat{f}(X_i))$$

2.2.4 Classifiers

The task of estimating the probability of foreign conflict is at its core an out of sample problem: we wish to know the probability of foreign conflict between all states, including states that did not participate in MIDs. To accomplish this, I have to train a model using the the cases in which MIDs did occur, then extrapolate the results of this model to all interstate dyads to yield the probability

of MID onset between states which did not actually enter into a dispute. This yields estimated probabilities of MID initiation which I argue can proxy for the presence of foreign threats. To maximize out of sample performance, I rely on tools from machine learning which are designed to predict well without making strong assumptions about the structure of the data. As before, I rely on the advice of Wu et al. (2007) and Fernández-Delgado et al. (2014), who have identified some of the the best performing algorithms for this task. Guided by their recommendations I select the algorithms which are suited to the data at hand.[10]

2.2.5 Candidate Classifiers

- A constant-only logistic regression which always predicts the proportion of each class in the dataset. As fatal MIDs are relatively rare events, this null model will skew heavily towards predicting no MIDs between countries and achieve relatively decent predictive performance as a result.
- Logistic regression with all predictors, as would be standard practice in the literature of international conflict, mirroring the approach of Nordhaus, O'Neal and Russett (2012).
- Boosted logistic regression Friedman et al. (2000).
- Penalized logistic regression model using an elastic net for regularization Zou and Hastie (2005).
- Multivariate adaptive regression splines Friedman (1991).
- Classification and regression trees (Breiman et al., 1984).
- Ranger, which is an adaptation of random forests Breiman (2001).
- Support vector machines with a weighted radial basis function (Cortes and Vapnik, 1995)
- Averaged neural networks Ripley (1996)

[10] I also ran KNN, C5.0, and Naive Bayes in addition to the models listed, but due to poor performance I proceeded with the remaining models to minimize computational time.

2.3 Results

In order to accurately estimate the test error of the models, I take the full dataset (N=1,017,988) and split it into three separate datasets: training (60%) validation (15%), and test (25%) maintaining the same proportions of 1s to 0s in each partition. The quantity of interest is the generalized test error of the models - how well they perform in predicting new data that was not used in fitting the model. To estimate this quantity, I develop models on the training set, using 10-fold cross validation to select the appropriate values for models which rely on tuning parameters to optimize their performance. In this approach, the data are randomly divided into 10 groups or 'folds' of approximately equal size. The first k fold is then left out as the validation set, while the model is fit on the remaining $k-1$ folds. This process is repeated k times where each time a different k fold is treated as the validation set. This process produces k estimates of the test error, and the final k-fold CV estimate is the average of these values.

Table 2.5 reports the cross validated estimates of model performance on the training set. The goal for each of these candidate classifiers is to offer an improvement relative to that of a null model which simply predicts the average proportion of 1s to 0s in the dataset for every observation. The results of training show immediately that each candidate learner is able to use the available data to dramatically improve over that of the null model. Ranger offers the best performance in cross validation, though its training time is prohibitively longer than most of the other classifiers (save for support vector machines).[11]

[11] In fact, the cross validated performance of each classifier at this stage was so strong that it potentially signaled an issue with how the dataset was constructed. I investigated whether this was the case by inspecting whether there were particular features which stood out as problematic, such as peace years or participation in another MID. This wasn't the case, as I will show later, so I then tried lagging the inputs, removing the world wars from the dataset, and iteratively training on five year chunks and estimating performance in the following two years over the entire time period. In each of these cases the models still showed strong performance, leaving me to conclude that the available predictors with the selected classifiers are able to perform well in recovering the data generating process for fatal MID onset. See the appendix for additional details.

Table 2.5: Results from cross validation on training set. N = 610794 with 1208 observed instances of states being the target of fatal MIDs. Log likelihood and area under the receiver operating curve estimated using 10 fold cross validation. Training time in minutes reported for each method.

Training Set			
	LL	AUC	Minutes
Ranger	0.007	0.980	5483.110
avNNet	0.009	0.956	509.489
GLM	0.009	0.962	13.408
enet	0.009	0.962	161.378
MARS	0.010	0.931	46.834
CART	0.011	0.896	20.783
logitBoost	0.011	0.904	14.100
SVM Radial	0.012	0.849	9398.503
Null	0.014	0.500	0.417

The results at this stage are encouraging, but it would be potentially problematic to report these as accurate estimates of the generalizable error of the classifiers. For models which use cross validation to select the appropriate tuning parameters, cross validation can lead to a biased estimate of their actual performance on unseen data Varma and Simon (2006). Therefore, I relied on the validation set in order to gain a more accurate assessment of each classifier's performance. Table 2.6 reports these results. The performance of each classifier is slightly reduced on this dataset due to the decrease in N, but the results are happily very similar to cross validation on the training set. Ranger continues to show the strongest performance of the available classifiers, offering a 42% reduction in loss over that of the null model and a 20% reduction in loss over a standard practice logistic regression model. The strong performance of Ranger, which uses ensembles of decision trees, is likely due to its ability to conduct feature selection as well as identify interactions and nonlinearities present in the data which would not be detected by a less flexible model. For this reason random forests have increasingly seen use in political science (Hill and Jones, 2014; Barrilleaux and Rainey,

2014) and are an excellent tool for a researcher seeking an off-the-shelf predictive algorithm in most applications. Given the performance at this stage it would be reasonable to proceed to the test set with the classifiers which have been run so far. However, I explore ensembling in the following section in order to seek improvements in classification performance.

Table 2.6: Results from validation set using tuned classifiers. N=135732 with 258 observed instances of fatal MIDs. Classification performance assessed using log likelihood and area under the receiver operating curve.

Validation Set			
	LL	AUC	PRL
Ranger	0.008	0.975	0.429
avNNet	0.009	0.952	0.357
GLM	0.010	0.955	0.286
enet	0.010	0.955	0.286
CART	0.011	0.894	0.214
MARS	0.011	0.923	0.214
logitBoost	0.011	0.842	0.214
SVM Radial	0.013	0.790	0.071
Null	0.014	0.500	–

2.3.1 Ensembling

In most applied predictive modeling projects it is vital to balance performance with computational time. Given the strong performance of ranger and averaged neural networks on the validation set, I could reasonably proceed with the averaged neural networks based on their speed and performance. However, for the purpose of this paper, the goal is to achieve the best possible out of sample predictions for fatal MID onset with less of an emphasis on the timeliness of producing those predictions. I therefore seek to investigate whether performance can be improved via ensembling. For the purpose of predictive modeling, an ensemble is a group of classifiers whose predictions are combined with the aim of achieving better performance from the group than from one individual model. Random forests demonstrate the intuition of ensembling, as predictions averaged across

hundreds of decisions trees outperform predictions from just one tree. This improvement comes from the bias-variance tradeoff. For instance, a single decision tree is a low-bias, high-variance estimator; by averaging across hundreds of different decision trees, a random forest is able to reduce the variance of these classifiers while still minimizing bias. For a technical discussion see section 7.3 in (Friedman, Hastie and Tibshirani, 2001), but there is a large body of research demonstrating that techniques such as bagging (Breiman, 1996), boosting (Schapire, 2003) and stacking (Wolpert, 1992) allow ensemble models to outperform individual classifiers (Caruana, Munson and Niculescu-Mizil, 2006; Whalen and Pandey, 2013).

In an effort to outperform Ranger - a method which itself performs well because of its reliance on ensembling via bagged decision trees - I test two different ensemble approaches which make use of all of the individual classifiers I have run so far. First, I use a weighted ensemble, which seeks to improve on an individual classifier by taking a weighted average of a wide variety of models. In order to construct a weighted ensemble, I take the cross validated predictions from the training set for each candidate model and bind them into a matrix.[12] I then use Y.Yes general nonlinear augmented Lagrange multiplier method solver (Ye, 1987; Ghalanos and Theussl, 2015) to estimate the optimal weights for minimizing the loss function, in this case the log-loss, on the outcome in the validation set. I then predict the final test set using each individual classifier and then weight these predictions using the weights estimated on the validation set.

[12]Here I use only the candidate models at the optimal tuning parameter selected via cross validation. It is common in ensembling to use many classifiers at various levels of tuning parameters, but due to computation complexity I have chosen to narrow my focus to only the models which performed well during training.

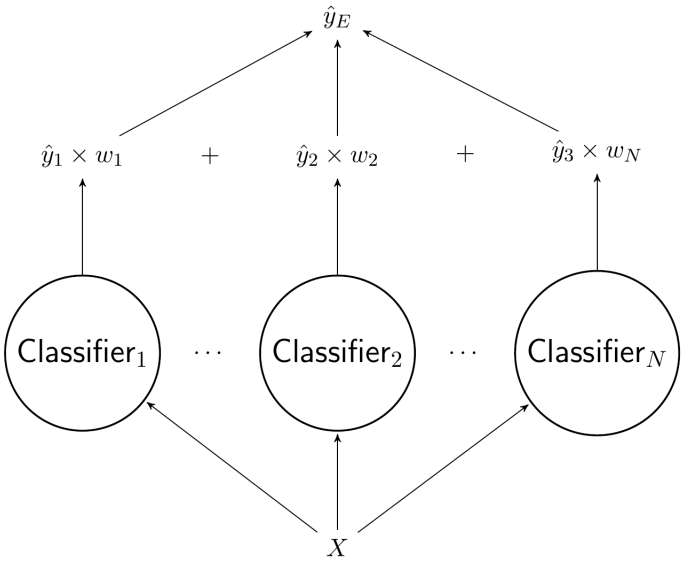

Figure 2.1: Weighted ensembling

Another common approach to ensembling is stacked generalization, or stacking. First introduced by Wolpert (1992), in a stacked ensemble the predictions from individual classifiers are used as features in a higher level classifier. For example, a CART and Naive Bayes may be used to predict the validation set. These predictions are then used as features in another classifier such as a logistic regression. By using information from a wide variety of classifiers, the meta classifier can achieve better performance than from one individual classifier (Whalen and Pandey, 2013, 3). I used four different methods for the higher level classifier - logistic regression, penalized logistic regression via elastic net regularization, averaged neural networks, and extreme gradient boosted trees - which I trained on predictions for the validation set. To be precise, I used the prior classifiers to make predictions on the validation set. I then used these predictions as features in training the ensemble models. I was then able to predict the test set in the same way: first making predictions using each individual classifier, then feeding those predictions into the ensemble models.

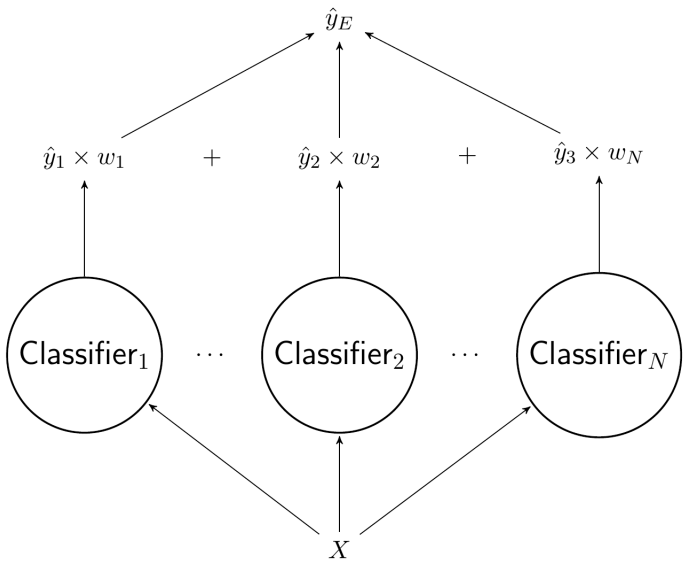

Figure 2.2: Stacked ensembling

2.3.2 Test Set

Having developed the ensemble models, I now turn to the final assessment of classifier performance. As noted earlier, I set aside roughly 25% of the data which was never touched in the training process as the true test of each model's out sample performance. Table 2.7 shows the classification results for each of the classifiers discussed so far, including the ensembles. Ranger and averaged neural networks continue to perform well, but the ensembles in general show strong performance across the board (the stacked ensemble with logit as the meta classifier is a the lone exception here, performing worse than a standard logit). Notably, the stacked model with extreme gradient boosted trees as as the meta classifier offers a slight improvement over that of Ranger, achieving a final proportional reduction of loss of nearly 48%. Though the improvement is marginal, the task here is one of predicting the outcome well with less attention to computational speed. Applied researchers in this area would likely be well served proceeding with averaged neural

networks or Ranger given their performance and speed, but for the purpose of this paper I proceed to constructing the measure of threat by using the ensemble.

Table 2.7: Results from the test set using each tuned classifier along with various ensemble models. Classification performance assessed using log likelihood and area under the receiver operating curve.

Test Set			
	LL	AUC	PRL
Ensemble - xgbTrees	0.007	0.981	0.478
Ranger	0.008	0.980	0.466
Ensemble - Weights	0.008	0.981	0.466
Ensemble - Neural Nets	0.009	0.981	0.380
Neural Nets	0.009	0.953	0.339
GLM	0.010	0.958	0.310
Elastic Net	0.010	0.958	0.309
Ensemble - GLM	0.010	0.974	0.288
CART	0.011	0.882	0.259
MARS	0.011	0.925	0.228
Boosted Logit	0.011	0.924	0.208
SVM Radial	0.012	0.866	0.189
Null	0.014	0.500	0.000

,

2.3.3 Comparing Models

The results so far have demonstrated that a number of classifiers achieve better performance than logistic regression, the common method of choice for conflict researchers. But where are more complex models achieving better performance? To gain a visual sense of how the classifiers are performing, I compare a 'standard practice' model to that of the best performing classifier (stacked ensemble with extreme gradient boosted trees). Figures 2.3 and 2.4 display out of sample predicted probabilities for the entire dataset. Both models show the uptick in the probability of experiencing fatal MIDs surrounding the world wars, as well as a heightened probability of conflict post 1950.

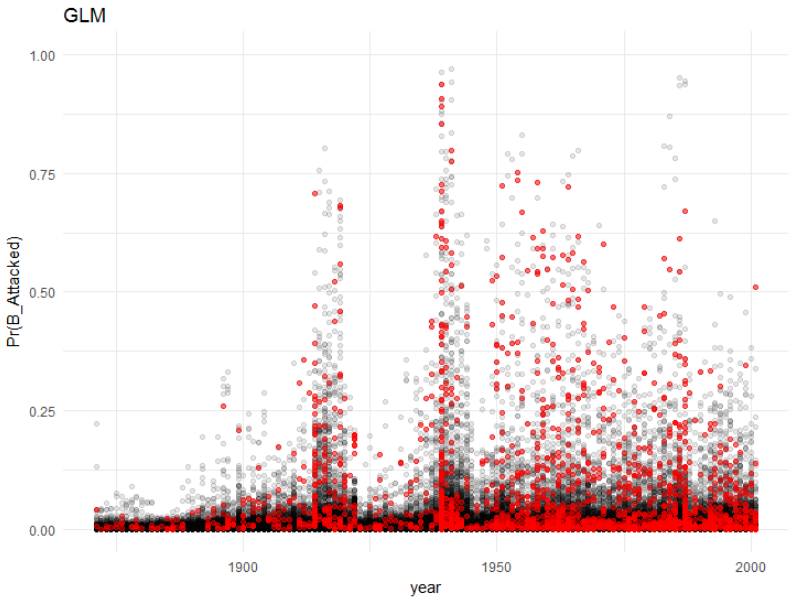

Figure 2.3: Predicted probabilities for the full dataset from a logistic regression. Observations with a fatal MID onset are highlighted in red while observations without a conflict are in black.

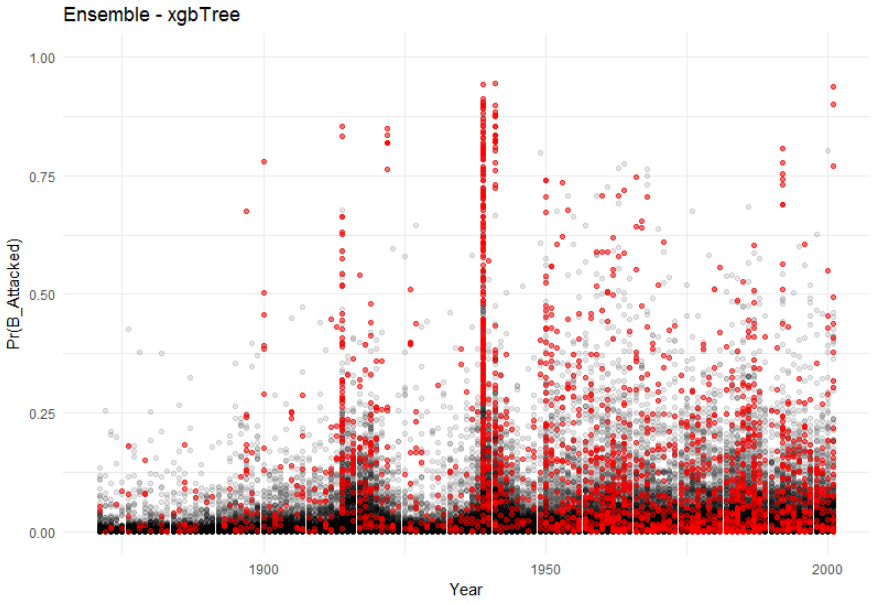

Figure 2.4: Predicted probabilities for the full dataset from the stacked ensemble with extreme gradient boosted trees. Observations with a fatal MID onset are highlighted in red while observations without a conflict are in black.

Though the models are broadly similar when assessing the probability of conflict over time, the models differ in their ability to accurately detect the realized instances of conflict. Here it is important to remember that the eventual purpose of these models is to use the probability of fatal MID onset as a means to proxy for the presence of foreign threats; it is the *probabilities* that we care about from the model rather than the classification. Because of this, in comparing two models for the purpose of measurement, the better model will assign a higher probability of conflict if conflict occurred; if conflict did not occur, the better model will assign a lower probability of conflict.

Figures 2.5 and 2.6 explore this comparison between logit and the ensemble. In order to better understand the output of the models, I first inspect a small subsample of years and look at 1937-1940 as a motivating example.[13] As can be seen in looking especially at 1939 and 1941, the ensemble assigns higher probabilities to conflicts in a conflict did occur. This is not without some cost, as it also increases its rate of false positives in these years, but logit also suffers from this problem. In the case of logit, it starts to produce false positives from 1939-1941 while still failing to identify many dyads which experienced conflict.

In Figure 2.6, I next separate observations in which conflict took place from those in which there was no conflict. I then plot the difference between probabilities from the ensemble and logit. When B attacked, if the ensemble predicted a higher probability of conflict (the difference being positive), the ensemble records a 'win'. When B attacked, if the ensemble predicted a lower probability of conflict (the difference being negative), the ensemble records a 'loss'. With this setup, the ensemble 'wins' 76% of the time compared to logit when conflict took place, being more accurate in 1516 of the 2012 observations in which a fatal MID took place. It is notable where the ensemble picks up losses, however, as it has a tendency to be bold when it does predict conflicts, especially during the

[13] To see two other selection of years (early 1970s, early 1990s), see the appendix.

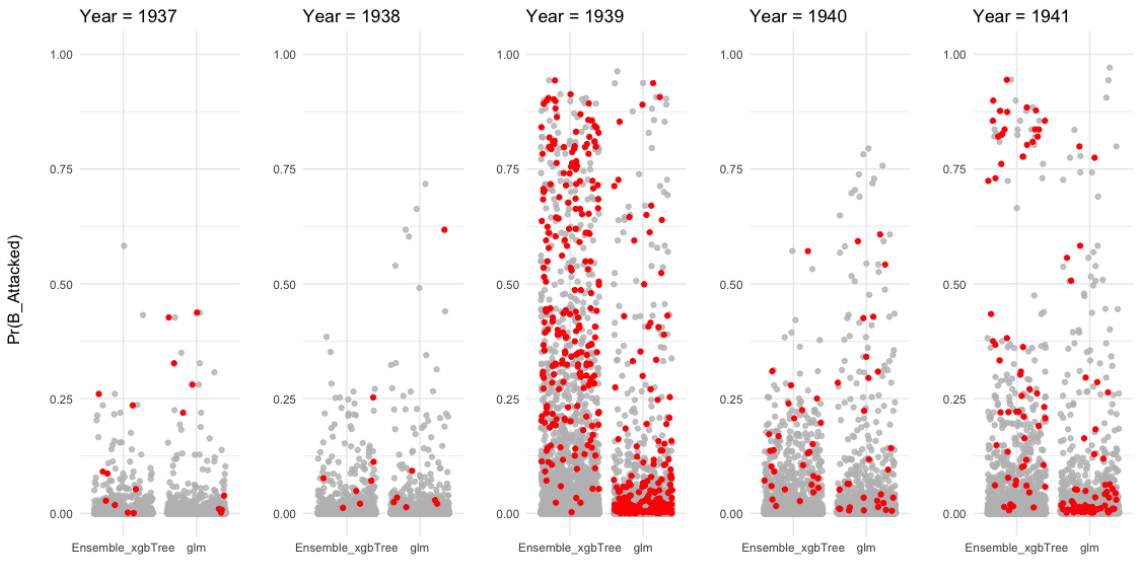

Figure 2.5: Comparing predictions for the best performing ensemble and logit for the years 1937-1941. Observations in which B attacked are highlighted in red; jitter used to help distinguish between observations.

=

world war years. This would be an issue if the ensemble model routinely opted to predict conflicts and failed in identifying true negatives, but it still manages to outperform a logistic regression when a MID did not take place, winning 93% of these 1,015,976 observations. That the ensemble performs better in both settings is not surprising given the results so far, but it is important to note the for the purpose of measurement. In many cases the measure of threat will be developed for use in years in which conflicts did not take place, so it is vital to have reliable predictions in these years.

2.3.4 What Matters for Predicting MIDs?

The aim of this paper was to accurately predict fatal MID onsets in the international system in order to produce a new measure of threat for researchers. In some sense, we would be pleased with a black box which was able to accomplish this task even if we were unable to understand how

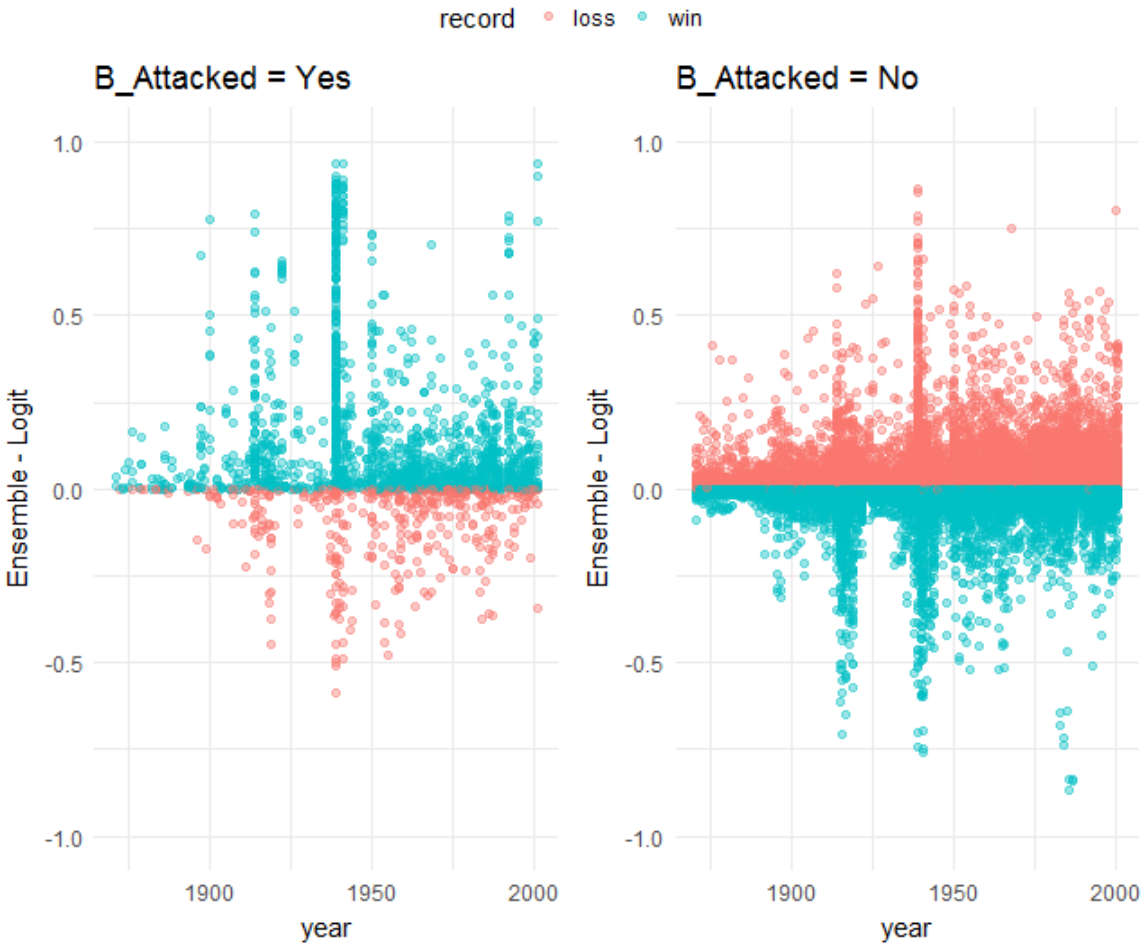

Figure 2.6: Comparing the ensemble model and a logistic regression. For both plots, the y axis is the difference between test set probabilities from the ensemble and a logistic regression. On the left are observations in which conflict took place; on the right are observations in which no conflict occurred. The ensemble records wins (plotted in blue) when it assigns a higher probability and conflict occured, as well as when it assigns a lower probability and conflict did not occur. The ensemble ''oses' (plotted in blue) when it assigns a higher probability and conflict occurred, as well as when it assigns a lower probability and conflict did not occur

these predictions were being made. However, the algorithmic approach to predicting MIDs does offer an additional way to assess the determinants of conflict in the international system. Given a model that was trained to predict well, which predictors mattered the most in correctly predicting MIDs?

A natural way of assessing the importance of a predictor is to assess how much worse the model would perform if that variable was not included. If a predictor is strongly related to the outcome, then leaving out that predictor will result in decreased performance for the model. If a predictor has no relationship with the outcome, then we would expect no meaningful decrease in performance. Ranger offers an immediate way of assessing variable importance by permuting each predictor across all of its trees and estimating the resulting decrease in test performance. Figure 2.7 displays the results of permutation importance scores for every predictor in the dataset. Using these scores, 'Year' emerges as the most importance variable, indicating (as we know) that there is variation in the pattern of conflict over time. This result is followed closely by expected battle deaths for each side as estimated previously. This result seems intuitive, as we would naturally expect that the expected costs of conflict would be important in estimating the onset of conflict. While this relationship may well play a role (and lend some face validity to the measures computed in my previous paper!), it is important to note that these specific variables are also closely related to time. As will be seen, there are large 'jumps' in the probability of conflict around the occurrence of the two world wars, with a general increase in the probability of fatal MIDs after 1950 until 1990. The variables which perform well in prediction seem to be those best situated to picking up changes over time, such as what we might think of as the 'structural' variables of the international system - the number of states, number of democracies, and percentage of democracies in the international system. These features are followed in importance by measures of state trade and power, such

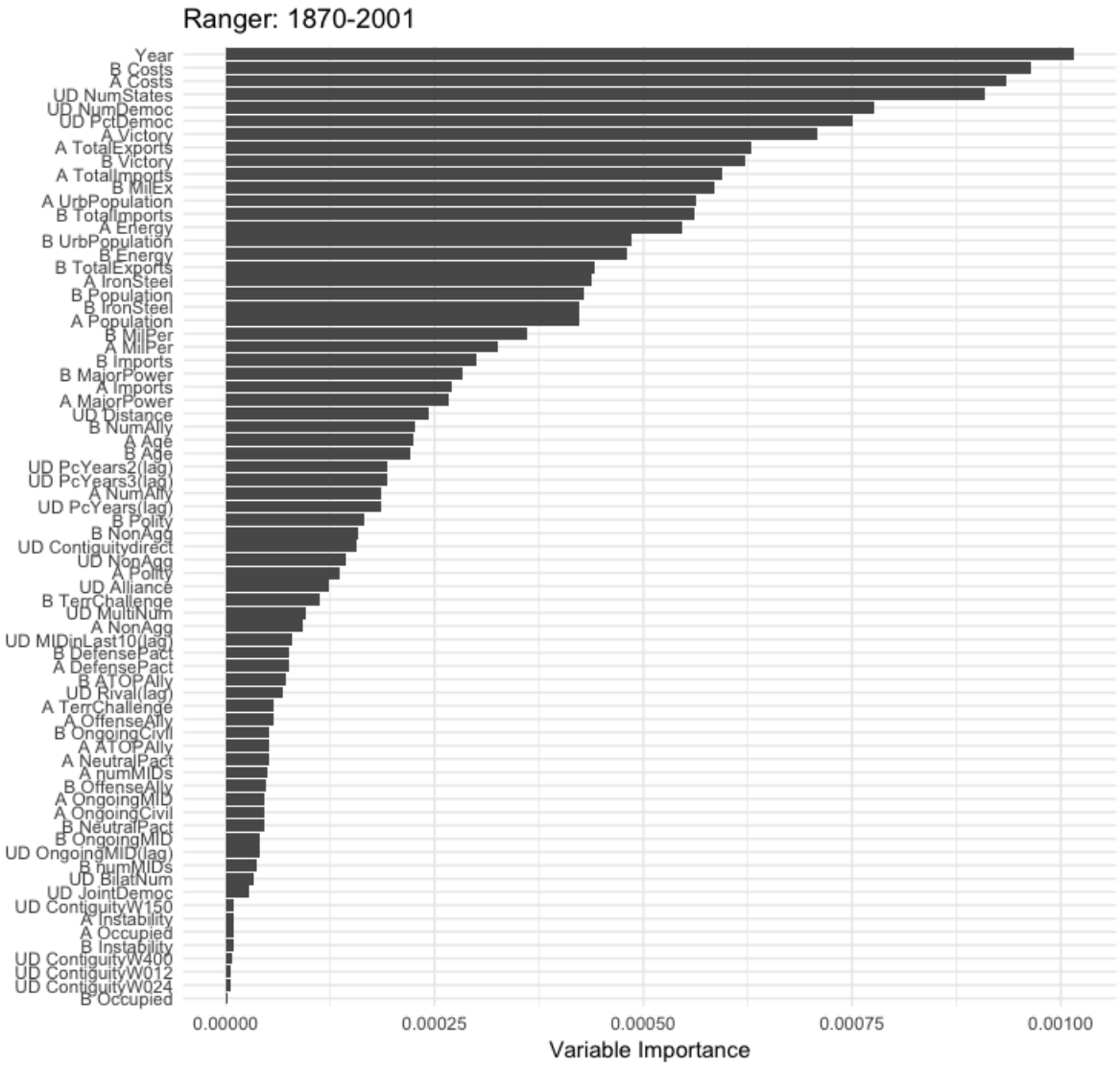

Figure 2.7: Permuted variable importance scores from Ranger. Five hundred trees grown with two randomly selected predictors.

as the estimated winner of a conflict (A, B victory) from Carroll and Kenkel (2016) as well as population and energy. Though unsurprising, these results lend some credence to both the liberal and realist schools of international conflict, as traditional indicators of state power state power and trade emerge as importance features in modeling conflict.

Certain variables also emerge quite low in terms of importance, such as contiguity, rivalry, and the presence of allies. But it's important to note here that permuted importance scores must be read with some caution, as these scores can be prone to bias in the presence of correlated variables (Strobl et al., 2007) as well as variables with higher variance (as the trees will be better able to find splits in these variables). Additionally, it is difficult to use permutation importance scores to rule out a variable as 'mattering'. Though a large decrease in predictive importance can be used to justify the inclusion of a variable, it is difficult to use the opposite to rule out a variable. It is also insufficient to filter individual features based on their relationship with the outcome, as a feature may only be important when used in conjunction with others (Guyon and Elisseeff, 2003).

For these reasons, I used one other method to identify features important in modeling fatal MIDs, using the Boruta algorithm (Kursa, Rudnicki et al., 2010) as a wrapper around Ranger. In this method, each predictor is shuffled to remove its correlation with the response, generating a random 'shadow' variable with a similar distribution for comparison with the actual variable. As the shadow variable is by definition randomly associated with the response, it represents a baseline for no predictive importance. By repeatedly comparing the true variable with randomly generated shadow variables, it is possible to directly assess the importance of an individual variable. The main purpose of this test is to identify predictors which are needed for modeling the outcome with the aim of trimming irrelevant predictors. But in iteratively repeating the random forest, the algorithm also offers a means of assessing the stability of the importance of individual variables. Variables

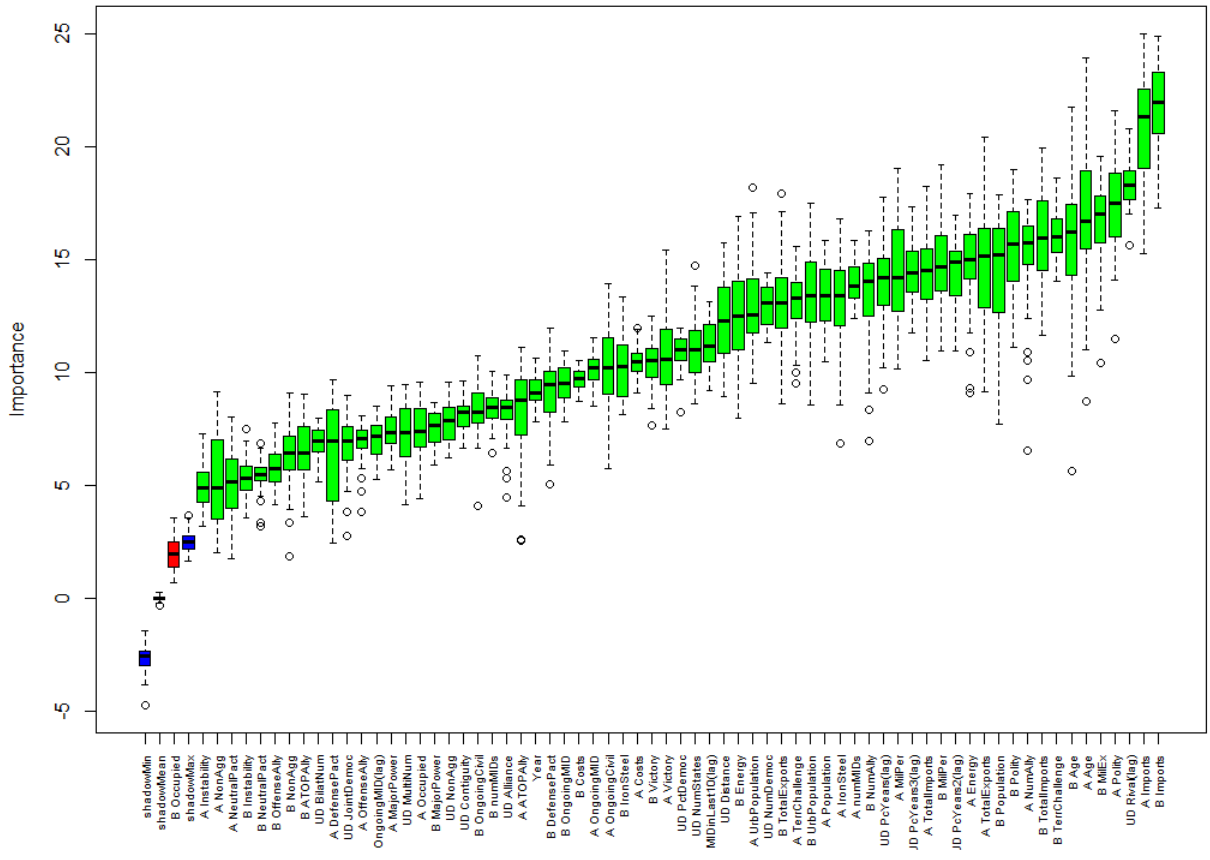

Figure 2.8: Variable importance scores computed using the Boruta package for Ranger.

which emerge as consistently important can be thought of as the most important in predicting the outcome. Figure 2.8 displays the result of Boruta variable importance using all features in the dataset.

There are two main takeaways from this assessment of variable importance. First, virtually no variables from the CoW universe emerge as unimportant in predicting fatal MID onset. Of the 60+ features, only one emerges as demonstrably unimportant - whether country B was occupied by a foreign power. From a predictive modeling perspective this exercise fails in demonstrating one of

the main uses of Boruta variable importance scores, as irrelevant features can be omitted and the random forest re run in order to improve both computational (by reducing the number of features) and predictive performance (by reducing the variance across individual decision trees). But from the perspective of international conflict, these results can be seen as encouraging - the variables we have been using to study conflict are well associated with the outcome we study. Though this might be unsurprising, it nonetheless represents predictive validation for many of the CoW variables used by researchers.

Second, the importance scores themselves do differ from the raw permutation scores from Ranger. While Boruta scores may be seen as a more refined estimate of variable importance, it is important to re iterate that they do not imply causal importance. Indeed, as Breiman et al. (2001) notes, it is quite often the case that two models which both predict an outcome well may identify different predictors as important. Different means of assessing variable importance will similarly offer competing results. For instance, trade between countries emerges as the most important variable in this setting, followed by Rivalry and Polity. The expected costs of conflict variables are less important using Boruta, possibly hinting that their predictive power in permutation importance is in part coming from the high degree of variance present in these measures. But, critically, the overall reading of these results should not be to pick out individual variables that have been deemed as important with the purpose of including only a select few as control variables in future studies. Instead, the overall lesson is that many variables are reasonably important in modeling international conflict outcomes. The nature of international conflict is complex and dynamic. Researchers will be better served by starting with the common stable of predictors and relying on flexible methods which can combat overfitting and identify complex interactions rather than assuming a functional form a priori.

2.3.5 How Have MIDs Changed Over Time?

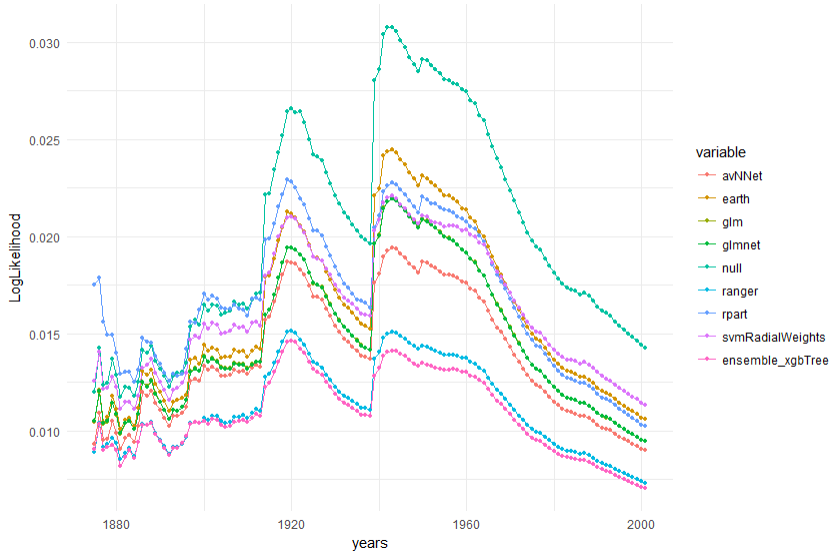

Figure 2.9: Predictive performance for all classifiers from 1870-2001 assessed using log-loss.

The results so far have assessed predictive performance and importance of individual features using the entire sample from 1870-2001. But how have MIDs changed over time? I first assessed each classifier's predictions over each year of the dataset, the result of which can be seen in Figure 2.9. There is variation in predicting MIDs over this entire time period corresponding to key events in the 20th century. Prior to World War I, there were relatively fewer number of MIDs taking place and the performance of each classifier is relatively stable. By the start of the first World War, however, all classifiers see degraded performance in accounting for the presence of conflict from 1914-1919 before stabilizing in the interwar period. The same is true for World War II, as each classifier performs less well in accounting for the large spikes in conflict which occured on the precipice of the world wars.[14]

[14]This result comes from treating time naively: an observation in 1890 may be classified with a model that was trained using observations in 1975. For the aim of recovering threat expectations, this approach is appropriate: the

This poses a potential problem for researchers of international conflict, as we would hope that the models we use to predict conflict would be able to recover the system wide shock presented by the two world wars. If our models cannot predict the presence of widespread conflict in these years, the inferences we draw from modeling these years may be flawed. This concern is somewhat lessened by noting that each classifier still outperforms the null, and the degradation in performance is less severe in the top performing classifiers. Notably, Ranger and the stacked ensemble strictly outperform all other other classifiers over the entire time period of this dataset. One imperfect solution to addressing this structural variation in conflict would be to drop the world wars from an analysis. Otherwise, researchers hoping to address this structural variation in the international system would be well served to use flexible methods which can recover the nonlinear and interactive effects of time.

Second, do variables shift in importance over different eras of conflict? I re examined variable importance from Ranger for four different periods of time: 1870-1913, 1914-1945, 1946-1990, and 1990-2001. It is important to note that the number of observations as well as the number of Fatal MIDs is different in each of these four eras. As before, these scores must also be interpreted with some caution. But, given these caveats, the results themselves are largely unsurprising. While 'Year', 'A/B Costs' , and the structural variables (number of states, number of democracies) were the most important variables for the full dataset, once we partition the dataset these variables

emphasis is only on finding the relationship between a set of inputs and the output. But as an additional check on how MIDs may have changed over time, I aimed to deal with time more directly. Starting in 1875, I trained a logistic regression on all years prior to t and classified observations in years $t+1,2$. I proceeded in this way until 2001, classifying observations using only data in the years preceding to train the models. In order to reduce the computational strain of this task, I first downsampled the dataset to roughly 100 observations without conflict (B Attacked=0) for every observation with conflict (B Attacked = 1), having first estimated that this proportion best balanced computational and predictive performance. Though it takes a few decades for the model to stabilize (as there are few observations prior to 1900), the result is generally the same. The world wars create a degradation in test performance while the model continues to otherwise steadily improve over time. See the appendix for plots of these results.

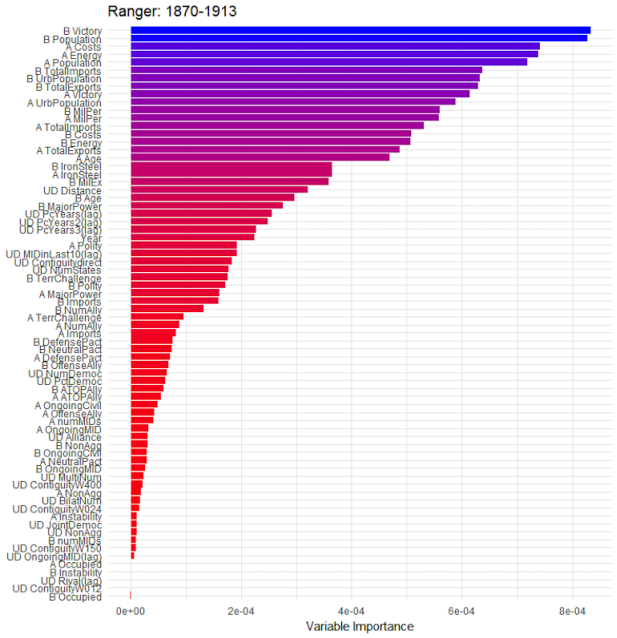

Figure 2.10: Ranger fit to years 1870-1913. N = 62,983 with 153 occurrences of Fatal MIDs

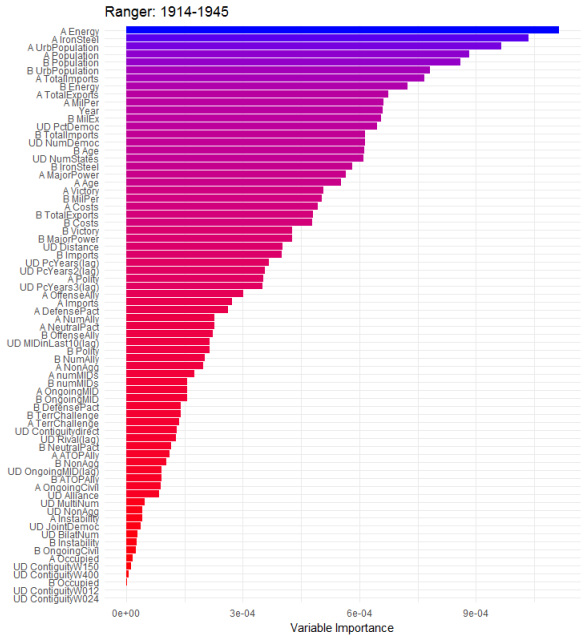

Figure 2.11: Ranger fit to years 1914-1945. N= 106,157 with 613 occurrences of Fatal MIDs

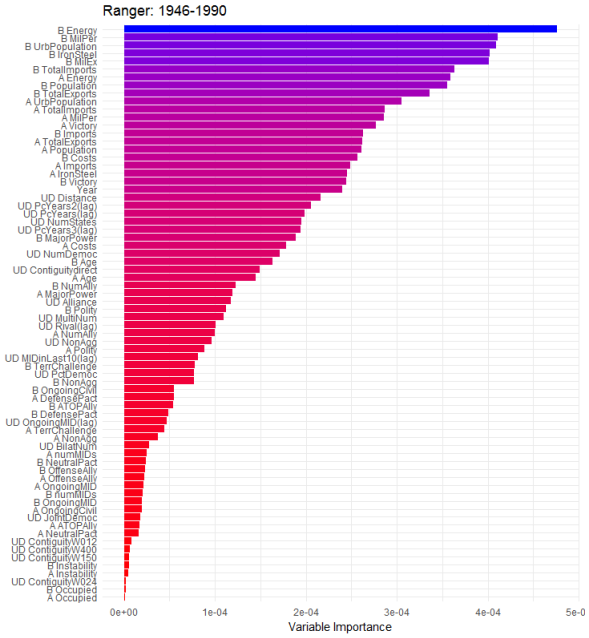

Figure 2.12: Ranger fit to years 1946-1990. N = 591,650 with 997 occurrences of Fatal MIDs

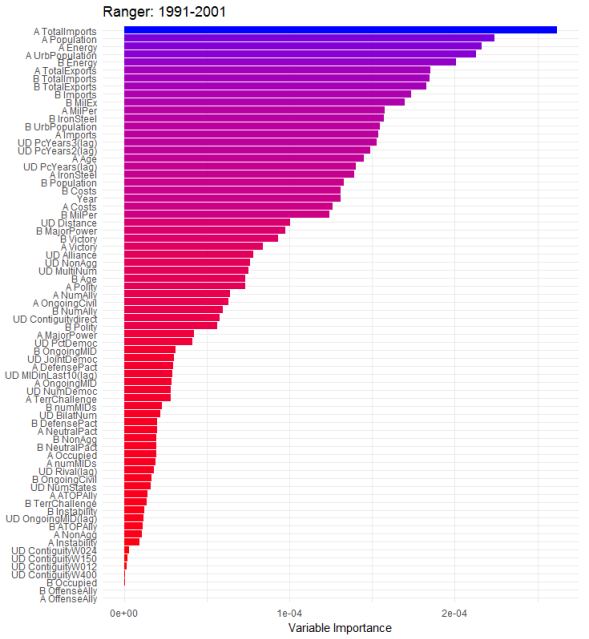

Figure 2.13: Ranger fit to years 1991-2001. N = 256,337 with 249 occurrences of Fatal MIDs

become less important. Instead, variables corresponding to each country's raw material capabilities such as population and energy prove to be among the most important in predicting conflict for each era of conflict. Patterns of trade also continue to be towards the top, with trade increasing in predictive importanenvironmentce in the period of 1991-2001. Otherwise, on the whole there does not appear to be a large shift in the process of predicting international conflict over time. A wide variety of features continue to be important in each time period, with variables proxying for a state's raw capabilities and trade towards the top at every turn.

2.4 The New Measure

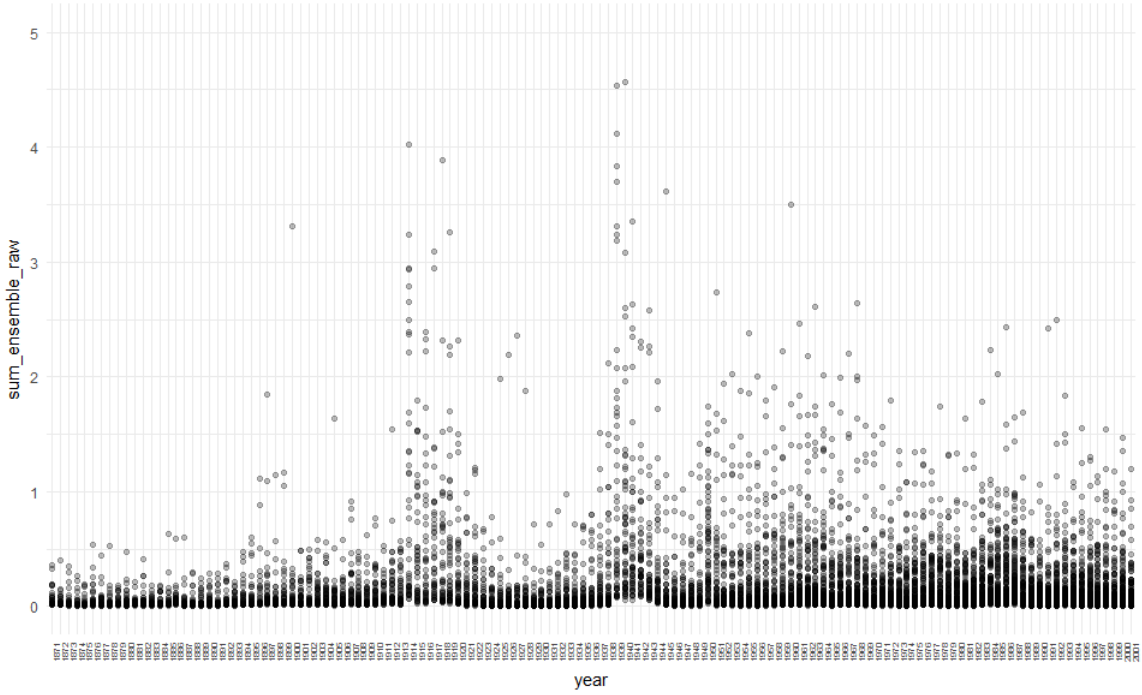

Figure 2.14: The country-level threat measure shown for all countries over the time period 1870-2001.

Finally, I turn to the task of creating the new measure of foreign threat for each country for each year using the classifiers discussed previously. The predictive modeling task outlined so far

amounted to estimating the probability that each country would be targeted by all other countries in a given year. For instance, if we wanted to know how much France was threatened in 1890, we have so far estimated its probability of being targeted by all other countries in that year. In order to create the measure of threat for France in 1890, we need to aggregate these probabilities into one final measure. While we could reasonably take the average or median of these probabilities, I focus on summing. The quantity of interest we care about is the overall level of threat for a country - how each potential source of conflict (all other countries) contributes to its security environment. It is natural then that we include all potential sources of threats and not reduce the overall information we have for each country. While one might be concerned about the impact of 'irrelevant' dyads from this approach, we can lean on the strength of the predictive model. If other states are truly irrelevant, their constitutive probabilities will be low and the measure will be unaffected by their inclusion in measuring a state's security environment.

Figure 2.14 displays the threat measure for every country for every year from 1870-2011. While I present the sum of the raw probabilities here, I also scaled the probabilities to account for each country's relative power based on DOE scores from Carroll and Kenkel (2016). The results are largely unchanged save for a shift in scale. This view can best be thought of as the level of threat in the international system in a given year. Relatively stable prior prior to 1900, there are two large upticks corresponding to the world wars, along with a general elevation in threat during the cold war era before a slight reduction post 1990. The increase in conflict in the international system post-1950 relative to the 1800s is of some interest. This may in part be a function of the data: it is easier to observe and document conflicts between countries in this time period relative to conflicts in the latter half of the 19th century. A better explanation may simply be that with the increase of states post 1950, there was greater potential for the onset of fatal conflict in the international

system. Though the scale of these conflicts may not have reached the heights of previous conflicts, the measure is constructed by looking for the prospect of any fatal conflict.

To gain a sense of variation between countries, I took snapshots of the measure at four periods in time: 1920, 1950, 1970, and 1995, seen in Figures 2.15-2.18. In general, there is one notable pattern from these snapshots. Despite some year to year variation, it is usually the case that major powers appear towards the top of the most threatened states on a given year. In 1920, the balance of threat was greatest in Europe and Asia, as Russia, Germany, Turkey, and the UK are among the most threatened states. Post World War II this pattern starts to change, as the United States, Israel, and China move upwards on the measure in 1950, though the UK is somewhat surprisingly listed as the most threatened state in this year.[15] The beginning of the Cold War sees the US and Russia appear towards the top of the measure alongside Israel. By 1995, Iraq - having just been attacked by an international coalition in 1990 - becomes the most threatened state in the world.

This pattern poses a question. Though the measure indicates that major powers are more likely to be targeted by a fatal MID, does this mean they are actually more threatened? Namely, does the measure actually reflect the background concept? From one perspective this might appear to not be the case: we might expect a smaller state to be more vulnerable than a larger state, possessing fewer raw capabilities for defense.

[15]This is likely a function of its colonies, as the UK will be coded as being a target if any of its areas of influence were under attack.

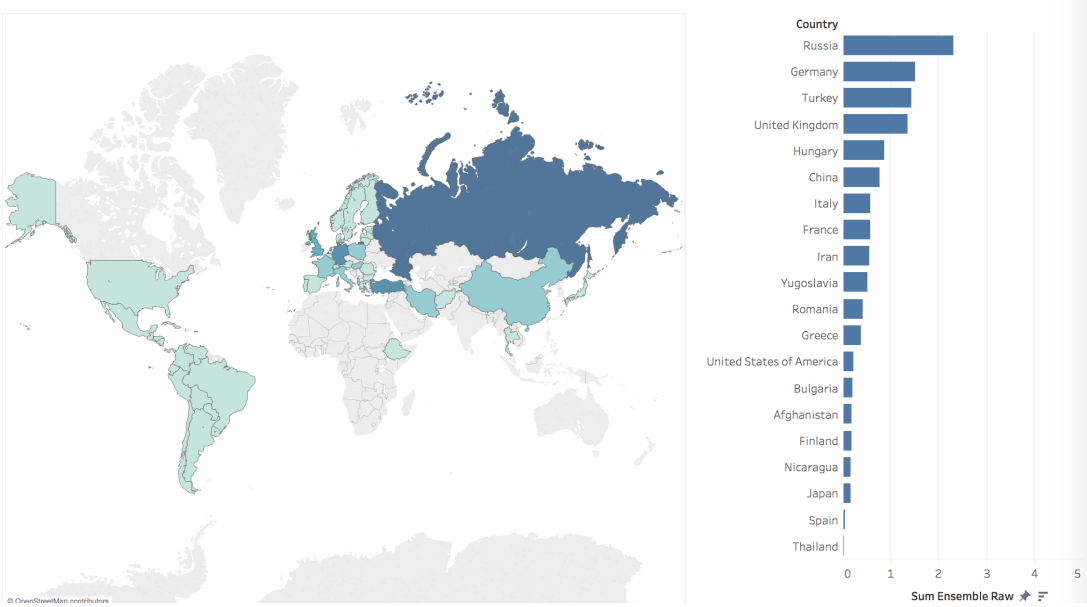

Figure 2.15: Measure of foreign threats displayed for 1920. Top 20 countries on measure in year shown on right.

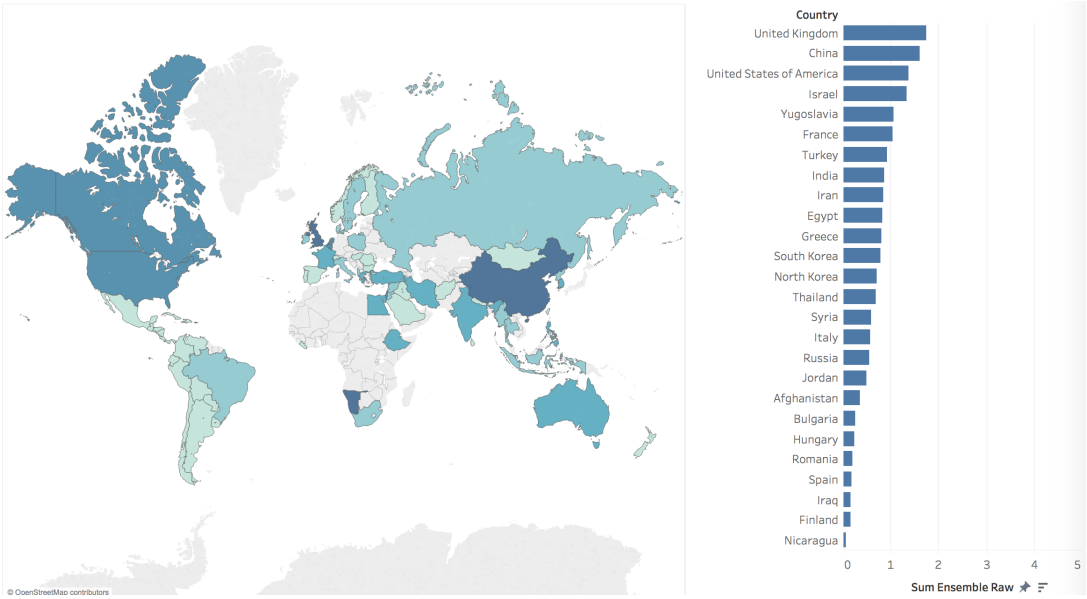

Figure 2.16: Measure of foreign threats displayed for 1950. Top 20 countries on measure shown on right.

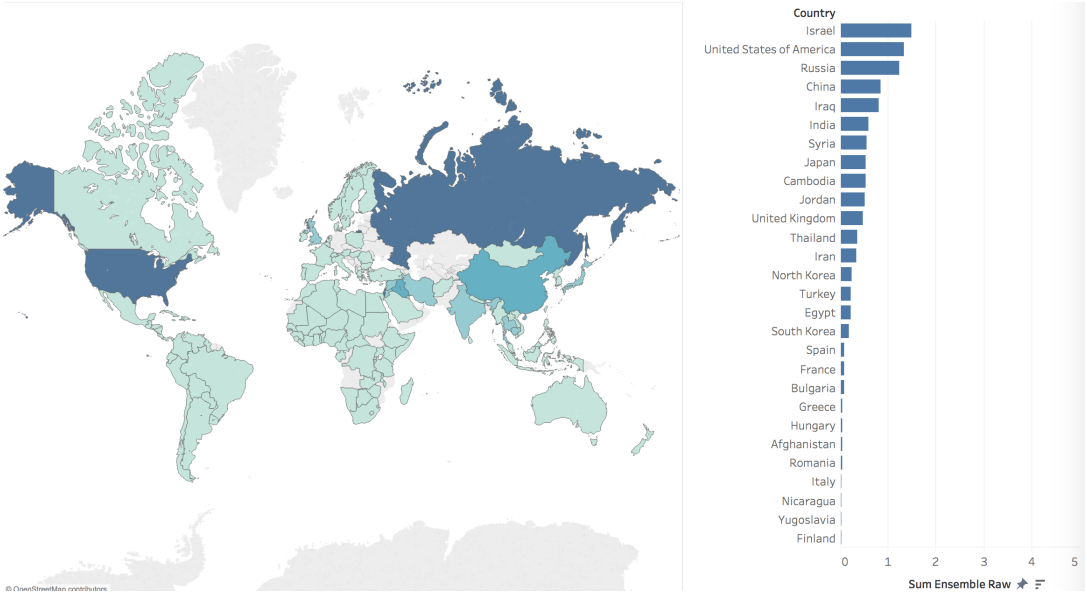

Figure 2.17: Measure of foreign threats displayed for 1970. Top 20 countries on measure shown on right.

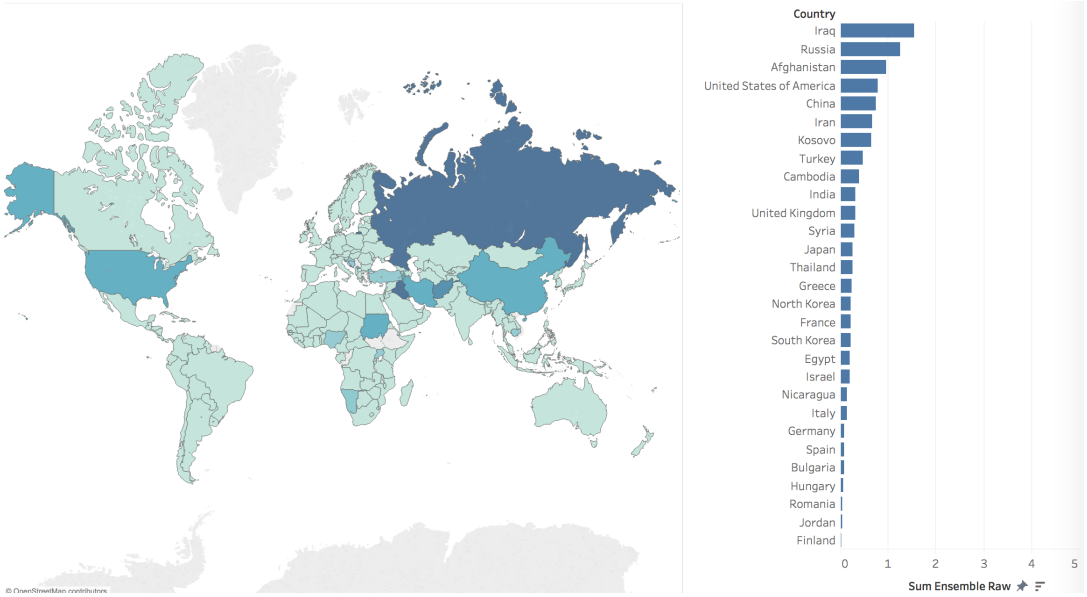

Figure 2.18: Measure of foreign threats displayed for 1995. Top 20 countries on measure shown on right.

But a state's international security environment is not just its capabilities. The extent to which a state is threatened is also a function of its nterests. Jervis writes in his discussion of the security dilemma:

> Defending the status quo often means protecting more than territory. *Nonterritorial interests, norms, and the structure of the international system must be maintained.* If all status quo powers agree on these values and interpret them in compatible ways, problems will be minimized. But the potential for conflict is great, and the policies followed are likely to exacerbate the security dilemma. *The greater the range of interests that have to be protected, the more likely it is that national efforts to maintain the status quo will clash.* (emphasis added) (1978, 185)

From this perspective, we *should* expect to see the major powers as being more threatened: they have a greater domain of interests to protect than other, smaller states. For instance, the Eisenhower doctrine represented a clear statement to the international system about the status quo desired by the United States; naturally, this created situations in which the US would potentially be involved in military disputes.

Indeed, in examining variation in threat over time in Figures 2.19 and 2.21, it is evident that the US and Russia were at a high level of threat throughout the entire period of the Cold War, as we would expect. Each state faced the possibility of being targeted by each other's allies. As a result, the US sees a reduction in its level of threat with the dissolution of the Soviet Union, while Russia maintains or increases in threat during this time. China likewise sees a high level of threat throughout the Cold War period before decreasing post 1990. To gain some sense of contrast between a major power and a smaller state, I also display Iraq's variation in threat over this time. The measure places Iraq as being the most threatened state in the international system in 1995, having steadily increased from 1975 until its conflict in the early 1990s.

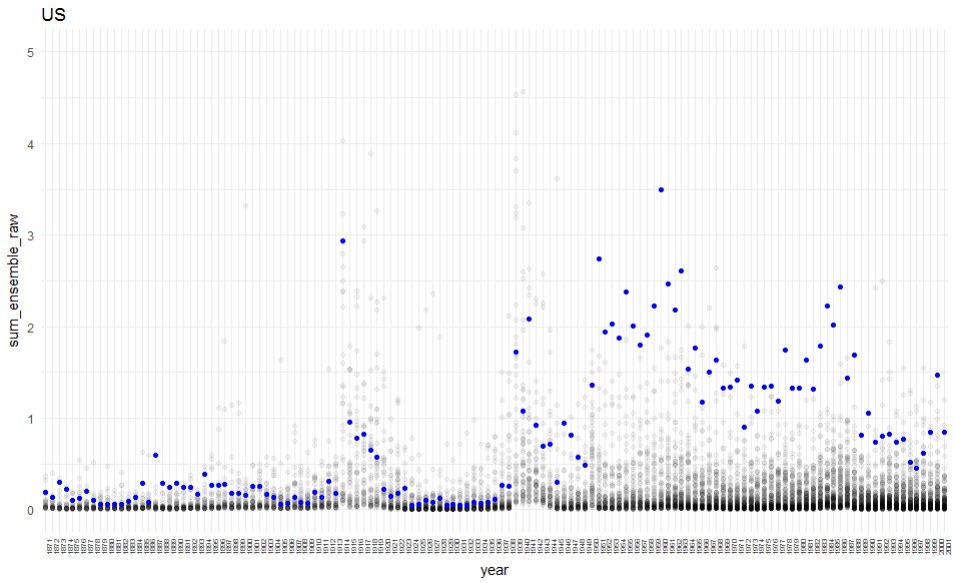

Figure 2.19: US foreign threat level, 1870-2001

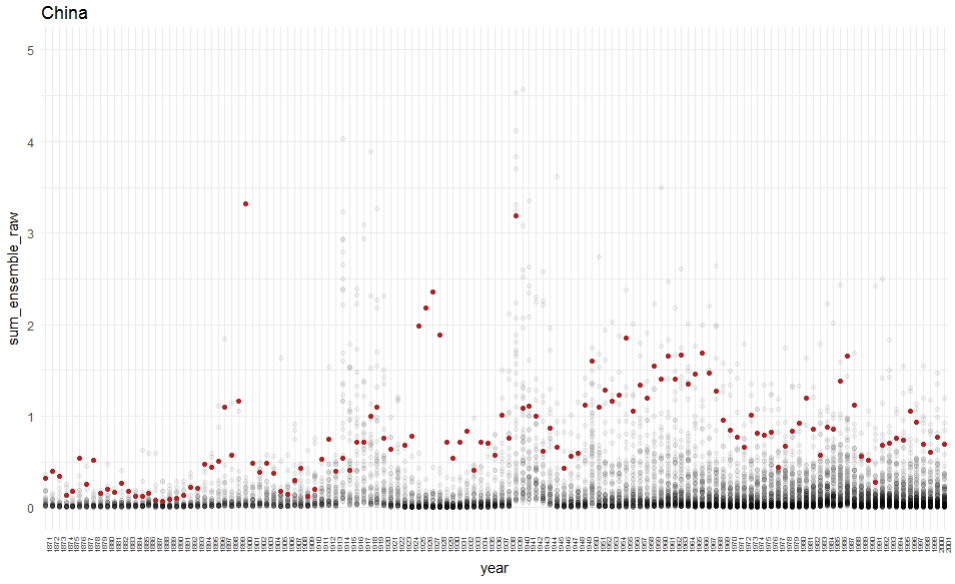

Figure 2.20: China foreign threat level, 1870-2001

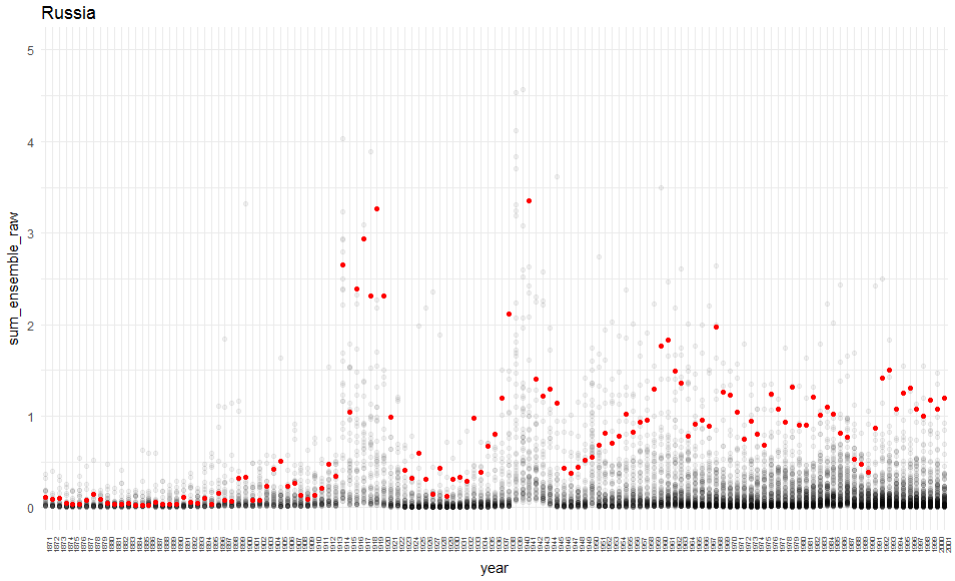

Figure 2.21: Russia foreign threat level, 1870-2001

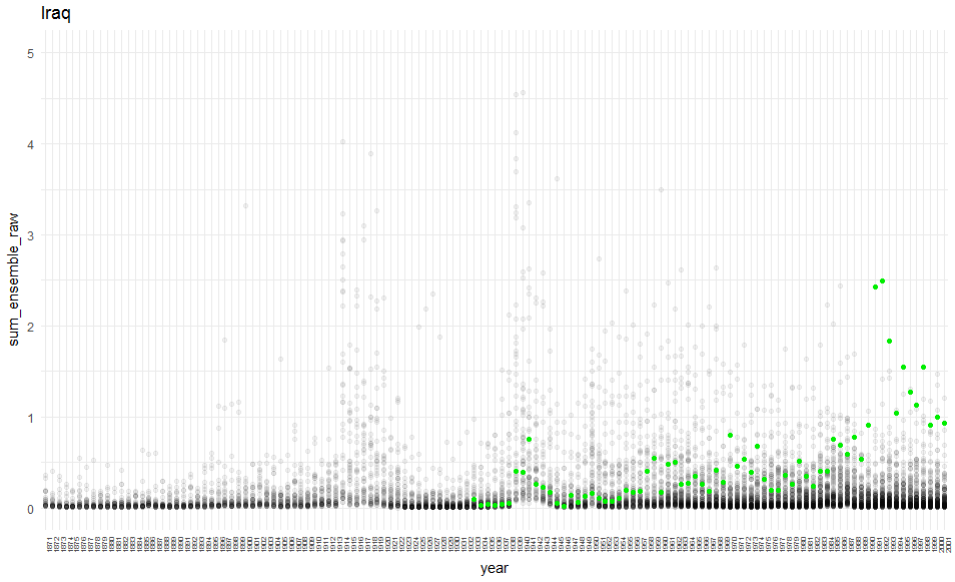

Figure 2.22: Iraq foreign threat level, 1870-2001

2.5 Conclusion and Advice for Researchers

This paper sought to create a new measure for researchers in international relations. Much has been written about the impact of the international security environment on state behavior, but the field has lacked a measure which can precisely measure a key feature of international relations: the probability of being attacked by other states. By developing a predictive model of fatal MIDs, I aimed to fill this gap by producing a new measure of foreign threat.

With this in mind, the applied researcher also must be careful in examining specific outcomes. For instance, it is important to note that the measure cannot be used in models of conflict onset or severity, as it was trained using data on realized conflict. Additionally, as I used a wide variety of predictors, the researcher must be careful before using the measure to be sure that the same variable does not appear on each side of the equation. For instance, one might wish to test the hypothesis that a state's international security environment will affect the formation of alliances. But as a number of ATOP variables were included in the model, the researcher can not immediately apply the measure to these very same variables.

Lagging the measure may be one solution to this obstacle and may be theoretically appropriate depending on the outcome of study. Military spending, as originally studied by Nordhaus, O'Neal and Russett (2012), in year t may be produced as a function the state's threat level in year *t-1*. Another option would to re run the predictive algorithm while purging the dataset of the outcome sought by the researcher. While this might result in a loss in predictive performance, the flexibility of Ranger allows it to perform even with a narrower field of predictors.[16]

[16]Another possibility here is for me to create a reduced version of the measure. Instead of pulling from all variables, I can use a number of methods (filter, lasso/ridge) to identify the minimal set of predictors needed to achieve performance within one standard error of the best performing model. This would allow researchers to use a 'reduced' measure if there is a concern about endogoneity.

Along with these caveats, the measure should be used conceptually to reflect *expectations* of conflict. It is therefore best served in studying how the expectation of conflict might affect other aspects of state behavior: military spending, repression, and alliance formation, among many others. For this purpose, the measure will offer researchers the ability to re-visit old hypotheses in the fields of both international relations and comparative politics.

CHAPTER 3

FOREIGN THREATS AND STATE REPRESSION

How does the international environment affect a state's domestic politics? Classic treatises on the second image reversed have speculated about various impacts of the international system on states. Many studies have examined the effects of globalization and economic interdependence on a state's conflict behavior (Gartzke and Li, 2003), political institutions (Rudra, 2005; Milner and Mukherjee, 2009), and human rights records (Hafner-Burton, 2005*a*; Apodaca, 2001). Despite the effort to explore the effects of trade and economic interdependence, one aspect of the international system has remained relatively unexamined in modern empirical work: the effect of a state's international security environment. One of the foundational tenets of international relations is that states exist in a system characterized by threats.[1] Classical realist theory posits that states are motivated primarily by their desire to survive. Under anarchy, states exist in an environment in which they are threatened by other states, and as a consequence states must take actions to prepare for the possibility of conflict. This poses a question: how does the possibility of foreign conflict affect state behavior?

International relations scholars have increasingly relied on domestic institutions to explain international conflict (Fearon, 1994; Weeks, 2008, 2012), but less work has focused on reversing the arrow. In as much as Putnam (1988) illustrated the importance of domestic politics for international outcomes, scholars risk neglecting the second level of the game if they do not account for the influence of the international system on domestic outcomes. It is particularly vital to consider the

[1]See Wendt (1992) for a discussion and criticism of this concept in international relations.

possibility of conflict in the international system, as a state's security environment is fundamental to its construction and thereby its efforts to maintain power. Waltz (1979) posited that states resort to developing military capabilities and forming alliances in order to ward off external foes, and others have argued this decision to mobilize resources for the possibility of war is necessary for the development of the state. Tilly and Ardant (1975) points to preparation for war as a key factor for the consolidation of state power in European history as states responded to a threatening security environment: "the largest and most persistent stimulus to increases or changes in national fiscal burdens over the great period of European state-making was the effort to build armed forces and wage war" (54).

With this background in mind, I return to a question asked by researchers in the field of political violence: under what conditions do leaders use repression to maintain control of their regime? Dating back to Machieavelli and Hobbes, it has been argued that leaders use repression in response to domestic threats: leaders repress in order to suppress challenges to their grip on power. With this in mind, the field of political violence has devoted significant theoretical and empirical attention to understanding the conditions which threaten a leader's hold on power and influence the decision to repress. To date, these explanations have generally privileged factors internal to the state: democracy, political and economic instability, leader security, and demographics have emerged as some of the most important correlates in explaining state repression (Davenport and Armstrong, 2004; Davenport, 2007; Ritter, 2014). In stressing the role of domestic characteristics, the influence of the international system on state repression has been limited to human rights treaties, nongovernmental organizations, and the effects of globalization. While these certainly play a role, I contend the international system influences states on a more fundamental level by shaping the security environment of the state and influencing a leader's political security. Just

as leaders face threats of removal from domestic opposition, they also must contend with foreign adversaries. In seeking to maintain control of power, leaders play a two-level game where decisions made at the domestic level influence the international level, and vice versa. The threat or realization of foreign conflict can disrupt a state's internal politics and influence a leader's hold on power and thus the calculation to repress. From this I argue that studies of state repression should examine the international security environment of the state.

In this paper, I examine whether expectations of foreign conflict influence observed levels of state repression. International war appears as a standard correlate in cross-national studies of state repression, examining the relationship between repression and international conflict that is already underway. But, critically, it is not only the realization of conflict which should affect outcomes in domestic politics, but the *expectation* of conflict. If states and state leaders leaders are strategic, they should anticipate and account for the possibility of foreign conflict in their decision making. Because of this, it is inappropriate to proxy for the presence of foreign threats by relying on indicators of whether conflict is underway. Instead, in my previous work I develop a measure which proxies for conflict *expectations* and the presence of foreign threats.

Using this measure, I examine whether foreign threats affect state repression. First, I examine the predictive importance of a measure of foreign threats alongside other concepts said to matter in explanations of state repression. While the field has traditionally relied on null hypothesis tests of statistical significance to determine whether a variable is an important determinant of repression, this approach ignores the ability of a model to actually *predict* state repression, as variables which are statistically significant do not necessarily increase the ability of a model to predict an outcome of interest (Ward, Greenhill and Bakke, 2010).[2] I therefore rely on cross-validation with linear models

[2]Hill and Jones (2014) show that many of the variables identified in the literature of state repression provide no substantive improvement to prediction and are likely the result of over fitting

and variable importance scores from random forests to assess the predictive validity of foreign threats to the study of state repression. Second, having first found some suggestive evidence that the measure of threats matters in predicting repression, I further explore the relationship in a way that is more common in the literature. Building off the work of Poe and Tate (1994); Poe, Tate and Keith (1999); Keith, Tate and Poe (2009), I use linear regression to examine the in-sample effect of threats on different measures of repression and human rights practices. To preview these results, I find mixed evidence in this setting. Generally, I find some evidence of a negative relationship between the measure of foreign threats and human rights practices: states facing greater expectations of conflict are associated with higher levels of repression. But this finding is fragile, not robust to model specification or the selection of the outcome variable. Further work is necessary to unpack the relationship between foreign conflict expectations and repression.

This paper proceeds as follows: I first define repression and examine the literature's current work on explaining where and when states protect human rights. While the literature has largely privileged domestic factors, I examine existing work on how international factors might affect a leader's calculus to repress. After examining competing implications for the effects of foreign threats, I dive into the data and explore the predictive validity of the measure in models of repression before evaluating the results of linear models. I conclude with a discussion of the results and implications for future work in this area.

3.1 When Do States Repress?

Under what conditions do states repress? The subfield of political violence often defines state repression as "a wide variety of coercive efforts employed by political authorities to influence those within their territorial jurisdiction: overt and covert; violent and nonviolent; state, state-sponsored

(e.g., militias), and state-affiliated (e.g., death squads); successful and unsuccessful" (Davenport, 2007, 3).[3] It has been argued since Machiavelli and Hobbes that states repress in response to threats, and this relationship has been repeatedly supported in the empirical literature on state repression (Poe, Tate and Keith, 1999; Davenport, 1995, 1996; Davenport and Armstrong, 2004). The consistency of this relationship has led Davenport (2007) to deem it the Law of Coercive Responsiveness: "when challenges to the status quo take place, authorities generally employ some form of repressive action to counter or eliminate the behavioral threat" (7). Critical to this relationship is a leader's political security, as this determines when a leader feels threatened and must resort to coercive means in order to retain power. Ritter (2014), for instance, models the interaction between a leader and a dissenting group and shows that repression and dissent hinge on the political security of the leader. Similarly, Conrad and Ritter (2013) show that compliance with human rights treaties is also a function of leader political security.

3.1.1 Domestic Factors

Studies of repression have generally focused on domestic factors which influence a leader's political security, and from this the mechanisms by which a leader might wield repression. In practice this research implicitly adopts a decision-theoretic approach which examines the structural conditions under which repression becomes more or less costly for political leaders. One strand of the literature focuses on the role of political institutions. In line with the view of democratization scholars, the domestic democratic peace posits that democratic leaders are less likely to use coercive means to retain power and this has generally been observed.[4] Other work focuses on distilling the

[3]These generally constitute the use of force by the state to violate First-Amendment style rights of citizens, such as due process, peaceful assembly, and freedom of speech.

[4]Davenport and Armstrong (2004) and Bueno de Mesquita et al. (2005) find that state coercive behavior is only influenced at the highest levels of democratization, and Carey, Colaresi and Mitchell (2015) also finds that weak democracies are more likely to form ties with pro government militias in an effort to evade accountability, which are associated with worse human rights records (Mitchell, Ring and Spellman, 2013)

institutional mechanism which constrains the use of repression, as scholars have examined the effect of domestic legal institutions on a state's decision to violate rights (Davenport, 1996; Powell and Staton, 2009). Other domestic factors which are often explored in the literature include GDP per capita, population size, and the occurrence of civil wars (Poe and Tate, 1994; Poe, Tate and Keith, 1999). In sum, the literature has explored a wide variety of domestic explanations for repression, and it appears that this emphasis has not been without merit. In a recent examination of competing explanations for state repression, Hill and Jones (2014) find that domestic features of a state outperform international features in predicting state repression. They conclude: "the contrast between results for domestic/international factors suggests that the institutional (political) and legal constraints that exist at the domestic level are more important for the decision to repress than any international constraints arising from treaties, NGO activity, or a state's situation in the global economy" (674).

3.1.2 International Factors

Political violence scholars have devoted far less time to exploring the presence of international conflict and external threats. This is surprising, as it is often noted that leaders face not only domestic but international challenges to their hold on power. Poe, Tate and Keith (1999) discuss the relationship between foreign threats and state repression: "When faced with threats, leaders will consider responding to them with repression... threats, which existing theory and research suggest are most important, stem from domestic and international political conflict" (293). Though the field has considered the theoretical role of international threats, empirical work on repression has only examined the impact of international war.

Pulling from Putnam (1988), Poe and Tate (1994) acknowledge that leaders play two simultaneous games, one in the domestic political arena, one in the international system. Poe and Tate

note that this consideration was important to the study of human rights and state repression, as they argued that "leaders' actions in the domestic political realm will likely be affected when their nations are a direct participant in an international crisis situation" (859). To accommodate this expectation, they included a binary indicator of international war as a correlate of state repression. The presence of international war has been included as a correlate in studies of state repression ever since, and the literature has generally found that international war increases repression (Davenport and Armstrong, 2004). This finding has recently been questioned by Hill and Jones (2014), however, who find that international war adds little predictive validity to models of state repression.

One study which does consider expectations of future conflict comes from McMahon and Slantchev (2015), who examine the interplay between a state and its military in the presence of foreign threats. They begin with the premise that "rulers govern in an environment characterized by foreign and domestic threats, and must provide for their security if they are to survive in power" (McMahon and Slantchev, 2015, 297). In their model, the international security environment affects the decision making of the military, who might otherwise wish to seize power from the leader. This is because armed forces which seek to intervene in the politics of the state face not one but two challenges: first, they must successfully complete a coup against the state, but crucially they must also fend off foreign challengers. Because armed forces are playing this two-level game, McMahon and Slantchev argue that foreign threats can induce military loyalty to the regime: "sufficiently grave external threats can discipline even a potentially disloyal general and deter him from executing a coup, a sort of 'circling the wagons'" (301). This is similar to the argument of Desch (1998), who contends that external threats can enable political leaders to control their own armed forces and prevent internal instability. Agüero (1995) contends as well that the presence of

an external threat helps with the process of democratization, as foreign threats enabled civilian leaders to focus on defense policy and smooth the transition from authoritarianism to democracy.

3.1.3 Linking Foreign Threats to Repression

What does this logic imply for levels of repression? There are two viewpoints in the literature which yield conflicting implications. From one view, as states consolidate their militaries and focus on external rather than internal opposition, they will become less repressive as they will face less of a threat from domestic actors. Motivated by the idea that leaders use repression in response to challenges to the status quo, we should expect leaders to become more repressive as the military becomes more loyal and the leader, by extension, more secure in power. This is in essence the argument of Arbatli and Arbatli (2016), who examine the effect of military disputes on coup initiation. They argue that an external threats - as proxied by the presence of a military dispute - can induce loyalty to the regime in a rally-around-the-flag manner. As popular support shifts in favor of the leader, fewer challenges to the status quo take place and less repression is needed in order to maintain power.[5]

Alternatively, other arguments in the literature imply that foreign threats will *increase* state repression. First, by increasing military loyalty as outlined above, leaders may become more able to invest in their military and use it for the task of repression. There is a large literature on how military spending and relates to the task of state repression, with the general expectation that a powerful military lowers the cost of repression for the state. The literature has not reached a consensus on the empirical effects of the military on state repression, as some authors find that states with stronger militaries engage in more repression (Poe, Tate and Keith, 1999) while others

[5] Arbatli and Arbatli (2016) also argue that the same effect may be achieved because militaries will have to split between two tasks while facing an international crisis, being forced to engage in both international conflict and domestic repression. As a result of this mixed role, militaries will become less effective as a repressive tool of the state.

find no relationship between military spending and various types of political instability (Goldstone et al., 2005). Colaresi and Carey (2008) can somewhat account for this inconsistency, as they argue that the relationship between military spending and mass killings is conditioned by the political institutions present within the state. Leaders constrained by domestic institutions (in their study, an independent judiciary) will be more likely to use a large military for the public good while unconstrained leaders will be more likely to engage in repressive behavior. While they focus on genocides and politicides, this argument can easily extend itself to the broader study of repression, as we would expect democracies and autocracies to make different use of their militaries emboldened by the presence of a foreign threat.

Another argument that foreign threats should increase repression is indirect, focusing on the effect international security has on political institutions themselves. German essayist Otto Hintze 1975 suggested that the internal political environment of the state is a function of its external factors. His emphasis was on the formation of the state's political institutions, arguing that states which are secure from the threat of invasion are able to organize around a minimalist state political structure. Lacking any imminent external threat, the state does not need to develop a powerful governmental apparatus in order to mobilize resources. Pointing to the United States and Britain as examples, Hintze argues that states which do not face threats from the international system will be able to democratize. In contrast, states which experience a highly threatening external environment will choose to remain authoritarian, as national security depends on the ability to quickly mobilize the state's resources for war. This sentiment is also expressed in the state-making literature from sociology, echoed by both Tilly et al. (1992) and Herbst (2000). footnoteUltimately, this argument suggests that international environment creates a selection process where certain

states are able to democratize while others are not. The greater the presence of external threat, the less likely is a state to democratize. Thompson (1996) revisits Hintze's core argument:

> In essence, most of the states that became (and remained) democratic in the nineteenth and early twentieth centuries had created or found themselves in relatively cooperative niches that insulated them from extremely competitive, regional international politic. In contrast, the presence of foreign threats leads states to tend towards authoritarianism as they must consolidate resources for the purpose of defense. (142)

From this perspective, foreign threats might influence a state's human rights practice simply via the formation of its political institutions. However, the logic also implies that states may become more repressive while in a heightened state of conflict. Facing an external threat, a leader may seek to curtail freedoms and suppress dissidence in order to appear unified to its an external foes. Pierskalla (2010) focuses on this aspect directly, as he argues repression might increase with the presence of foreign threats because of the strategic interaction between the state and its foreign adversaries. In modeling the interaction between the state and a dissenting group, he also considers the role of a third party. While he is principally concerned with the idea of a military or hard-liner faction within the government, he also allows for the possibility of a neighboring country willing to claim territory from the state. In this case, the leader must make a decision about the level of repression considering its effect not only on the domestic opposition but on the third party. Pierskalla argues, among other things, that when the third party is able to take a costly action against the leader, the leader has an incentive to repress in order to appear strong with regards to the third party actor. From this we would expect that, in the presence of a foreign adversary which may engage in international conflict with the state, the leader may become more repressive.

In sum, there are two existing strands in the literature which yield conflicting expectations for how foreign threats will shape a leader's decision to repress. From this, we can view international security as a structural factor to include in a simple decision-theoretic approach to repression and for the most part expect it to a) matter in predicting measures of repression and human rights practices used by the literature and b) increase levels of observed repression.[6] The goal of this paper is to examine the broader influence of the international system on internal political violence within a state. This is in line with the work of Moore (1995), who urged scholars of intrastate conflict to consider features of the international system in their examination of domestic phenomena.

3.2 Methodology

The first step in assessing whether foreign threats matter for state repression is to evaluate the predictive importance of the measure. Political science research has privileged the role of statistical significance and in-sample goodness of fit in assessing individual variables, but both of these quantities fail to capture whether a variable is important for predicting new data (Ward, Greenhill and Bakke, 2010). Assessing predictive importance amounts to testing the external validity of a model - how well a model generalizes to data that was not used in fitting the model - and thereby offers a different lens through which to evaluate both models and individual variables. As Breiman et al. (2001) puts it, prediction offers a clean method of assessing the ability of a model to accurately capture nature: "the extent to which the model box emulates nature's box is a measure of how well our model can reproduce the natural phenomenon producing the data." (204). From this, we can increase our confidence in the external validity of a model if it is able to

[6]However, if we view repression as a strategic choice by a leader, further work should consider how international security alters the incentives of the actors involved - both dissidents and leaders. As (Ritter, 2014) demonstrates, dissent and repression are strategic choices selected by an opposition group and a leader because of disagreements over policy. For the purpose of this paper, I focus on simply on repression, but another question of interest may be to consider how a state's international security environment influences dissent in all of its forms.

accurately predict out of sample, and we can extend this same line of thinking to assess individual variables. If a model performs worse in the absence of a variable, then that variable is important for prediction. The question of which inputs matter in modeling an output is not a trivial question, and identifying the optimal set of inputs via feature selection and feature engineering are vital for applied predictive modeling problems.[7]

For the purpose of this paper, I use a number of techniques to evaluate variables that are commonly used in models of state repression. First, I mimic the work of Hill and Jones (2014) to assess the importance of individual variables in predicting measures of observed state repression out of sample. I first use cross validations with linear models, adding variables one at a time to standard linear models and assessing their out of sample performance. I then examine variable permutation scores from conditional random forests, and use the Boruta variable importance algorithm (Kursa, Rudnicki et al., 2010) with Ranger to provide a final assessment of the relative importance of individual variables. Having assessed the predictive importance of the variable, I then turn to the question of inference: what is the effect of foreign threats on measures of observed repression? Here I follow the literature and make use of linear regression to estimate the partial effect of the measure on the same outcomes as before.

3.3 Data

The key concept of interest is the presence of foreign adversaries facing a state. To capture this concept, I predicted the probability of fatal MID onset between all dyads in the international system over the years 1870-2001. I trained a stacked ensemble using extreme gradient boosted trees as a meta learner on an ensemble of candidate learners (neural networks, random forests,

[7] See chapters 3, 18, and 19 from Kuhn and Johnson.

logistic regression) to maximize out of sample performance in predicting fatal MID onset during this time period. I then applied the results of this ensemble model to all interstate dyads and summed these probabilities for each country in each year to create a state-level measure of the international security environment for each state. This, I argue, better proxies for foreign threats than current approaches, offering a new measure of an important concept for researchers in the study of international relations.

In order to measure repression, I use three different scales commonly employed in the literature. The Political Terror Scale (PTS) and Cingranelli and Richards Human Rights Data (CIRI) are both ordinal measures of state repressive behavior coded from annual Amnesty International and United States State Department reports which seek to identify instances of torture, disappearances, political imprisonment, and state executions. These datasets cover the time periods 1976-2010 and 1981-2010, respectively. A third measure comes from Fariss (2014), who uses IRT models to incorporate existing measures of repression into one aggregated scale. In modeling these measures of state repression, I use predictors employed in the literature which are identified by Hill and Jones (2014). These are too numerous to discuss detail, but they include domestic features of states such as demographics, economics, and political institutions, as well as international factors such as INGOs, international law, and effects of globalization.[8]

Table 3.1: List of outcome variables for state repression and human rights. Listed with description of the variable and its source.

Outcomes	Description	Source
Dynamic Latent Score	Scores indicate a country's respect for human rights, with positive scores indicating less repression.	Fariss (2014)
Political Terror Scale (PTS)	Ordinal indicator (scale 1-5) of a state's level of political terror from AI and US state department reports.	Poe and Tate (1994); Poe, Tate and Keith (1999); Wood and Gibney (2010)
Physical Integrity Index	Ordinal indicator (scale 1-5) of a state's respect for human rights from AI and US state department reports.	Cingranelli and Richards (2010)

Table 3.2: Predictor variables used in models of state repression and human rights and source of the data.

Predictors	Source
Population	Gleditsch et al. (2002)
GDP per capita	Gleditsch et al. (2002)
Polity	Polity IV
Executive Constraints	Polity IV
Executive Open	Polity IV
Participation (Political) Competition	Polity IV
Military Regime	Database of Political Institutions
Left/Right Executive	Database of Political Institutions
Judicial Independence	CIRI
Common Law System	Mitchell, Ring and Spellman (2013)
Constitutional Provisions	Keith, Tate and Poe (2009)
Trade Openness	World Bank
FDI	World Bank
Oil Revenue	Ross (2006)
WB/IMF Structural Adjustment	Abouharb and Cingranelli (2007)
Preferential Trade Agreement w/ HR Clause	Spilker and Böhmelt (2013)
Western Media Shaming	Ron, Ramos and Rodgers (2005)
HRO Shaming	Murdie and Davis (2012)
Civil War	UCDP/PRIO
International War	UCDP/PRIO
Contract Intensive Money	Clague et al. (1999)
Foreign Threats (Raw)	Henrickson, working
Foreign Threats (DOE)	citecarroll2016prediction

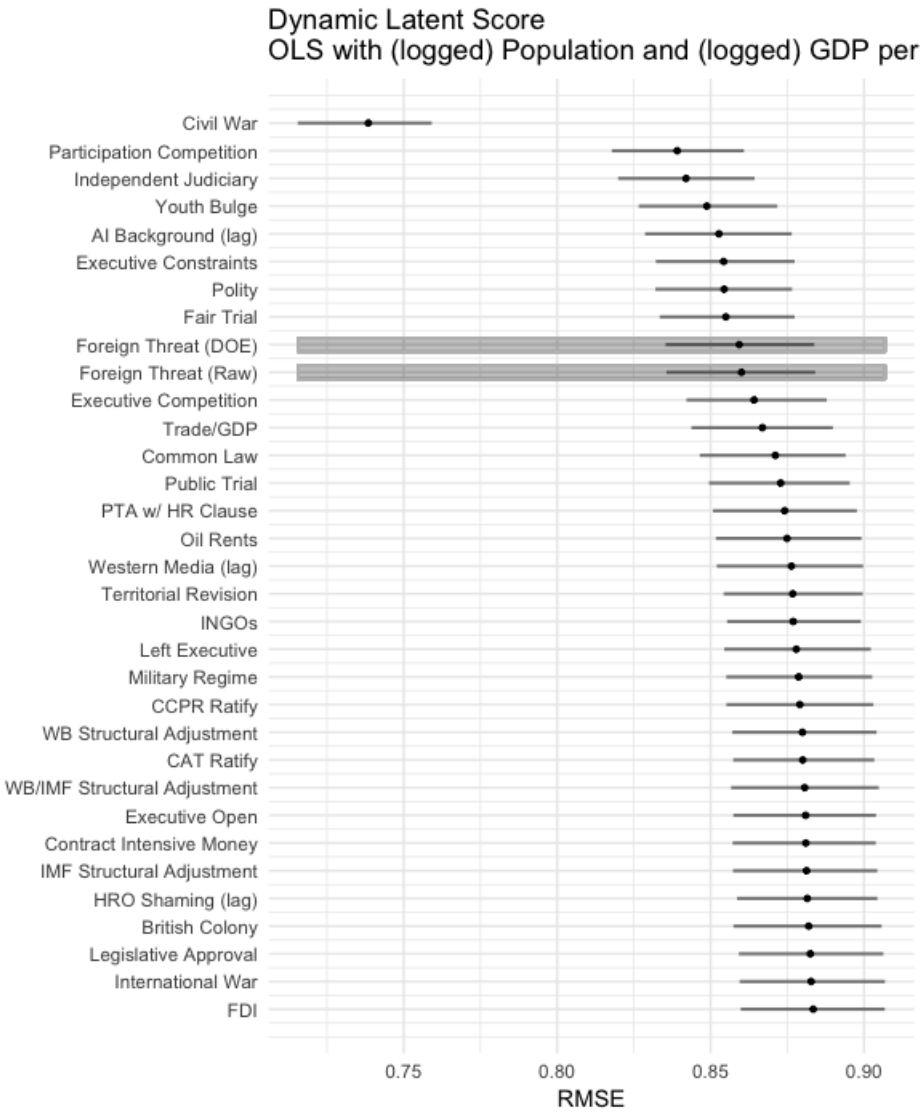

Figure 3.1: Cross validated variable importance scores for the Dynamic Latent Score from Fariss (2014). Variable importance estimated using linear regression models with 10 fold cross validation, iteratively adding each variable to a baseline specification including logged GDP per capita and logged population. Results bootstrapped 1000 times.

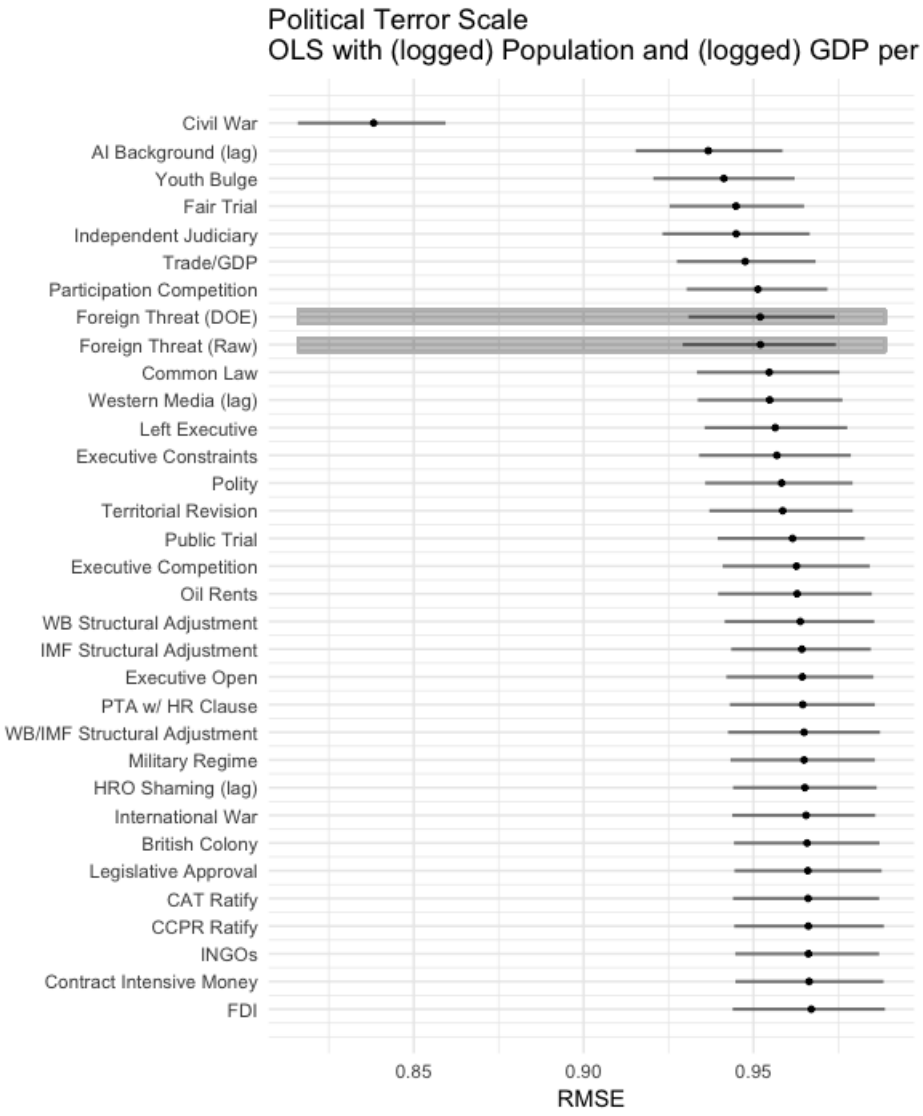

Figure 3.2: Cross validated variable importance scores for the Political Terror Scale Variable importance estimated using linear regression models with 10 fold cross validation, iteratively adding each variable to a baseline specification including logged GDP per capita and logged population. Results bootstrapped 1000 times.

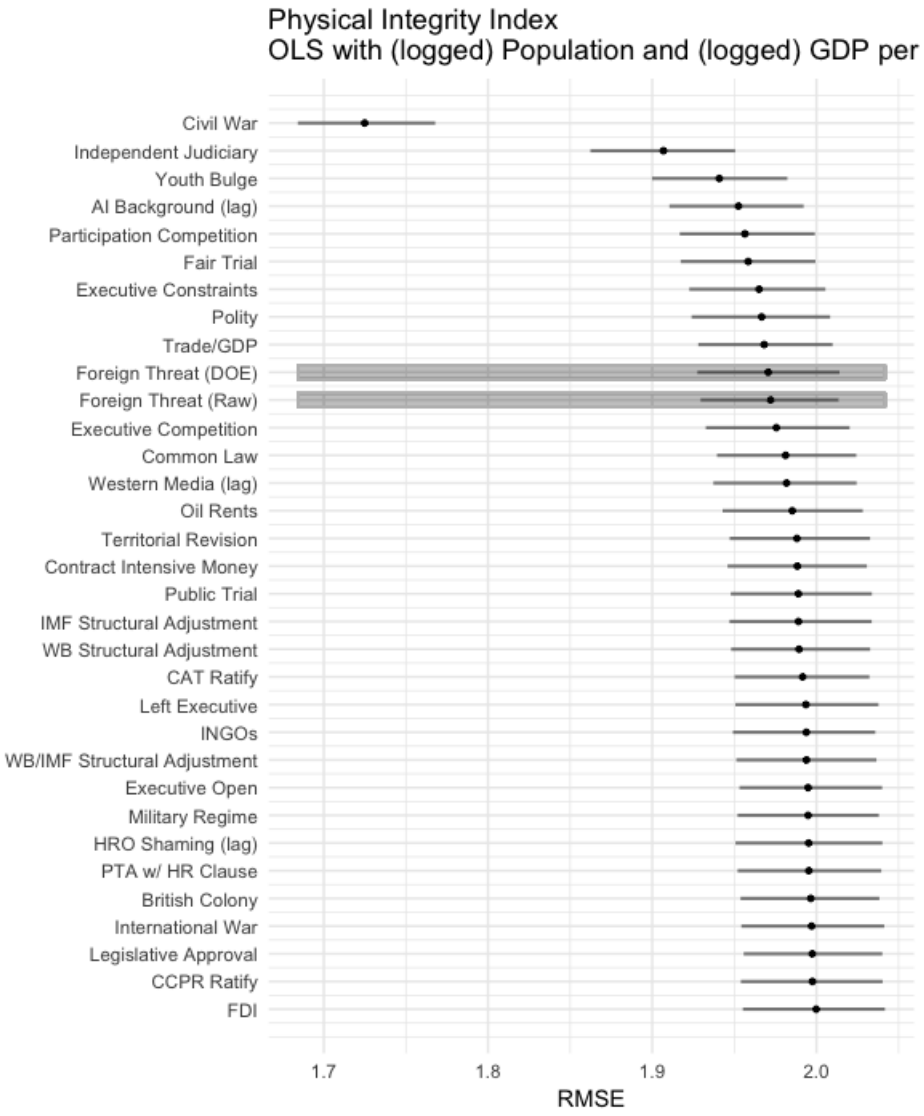

Figure 3.3: Cross validated variable importance scores for CIRI's Physical Integrity Index. Variable importance estimated using linear regression models with 10 fold cross validation, iteratively adding each variable to a baseline specification including logged GDP per capita and logged population. Results bootstrapped 1000 times.

3.4 Results

3.4.1 OLS Variable Permutations

First, I use cross validation with linear models to assess the importance of individual variables for each of the repression outcomes. To do this, I split the dataset into k folds (in this case, 10). I then select a model specification, adding a variable to a baseline model of logged population and logged GDP per capita. Using these three variables, I train a model on $k-1$ folds, then assess the model's out of sample error on the remaining fold. I repeat this process for all k folds, for all variables, for each of the dependent variables, bootstrapped 1000 times.[9]

These results should only be seen as suggestive, indicating the relative decrease in test error between common variables used in models of state repression and human rights practices. While in principle it is possible to define a threshold indicating whether a variable is important or not, I defer to looking at the entirety of the evidence in evaluating whether certain variables are said to 'matter'. In looking only at cross validation using OLS permutations, we can see, for instance, that the mean reduction in test error is greatest among a common set of variables across each of the three measures, and only marginal among a different set. Civil war consistently emerges as the most important feature, typically followed by some combination of participation competition, judicial independence, AI background reports, and youth bulges. Of these, in the cross validation setting, only civil war immediately stands out as an important predictor for repression.

The evidence for foreign threats here is mixed; while there are indications of a slight reduction in test error when using this variable, the wide bootstrapped confidence interval makes it difficult to determine its importance relative to other variables. Despite being towards the top of the list, it

[8]Data missingness is a problem for a number of these measures, so I make used of multiple chained imputation using random forests to create a dataset with coverage from 1984 to 1999.

[9]I additionally repeated this process with a base model specification including a lagged DV, as well as the separate components of CIRI. See the appendix for these results.

would be incorrect to draw a definitive conclusion from these estimates. For instance, in ranking the variables by their mean reduction in test error, we can see that foreign threats outranks international war, but we cannot conclude that this difference is meaningful due to their overlapping confidence intervals. It is somewhat encouraging that the variable does not rank at the bottom of the list, for instance, but further inspection needs to be done in assessing the predictive importance of the variable.

To that end, I additionally stayed in the linear family and examined how regularization affected each variable in the model. I used a lasso Tibshirani (1996) to conduct feature selection and a ridge regression to shrink the coefficient estimates for linear models with all of the predictors for each of the three dependent variables. Using the one standard error rule advocated by Hastie, Tibshirani and Friedman (2009), I examined the minimal set of variables needed to achieve similar out of sample performance to the full set of variables. I found that foreign threats remained in the model in both settings for each of the three dependent variables; see the appendix for the full results. This by itself is not definitive evidence that a variable is important for the purpose of modeling an outcome, as Mullainathan and Spiess (2017) demonstrate that penalized regression methods will often select different variables across different samples of the same dataset. This is similar to the Rashomon effect as described by Breiman (2001), in that different models with similar performance may be built with different subsets of variables. Identifying which variables 'matter' in this way is difficult, so it is important to use a wide variety of methods. I therefore proceed with variable importance scores from nonparametric methods in the following section.

3.4.2 Random Forest Variable Importance

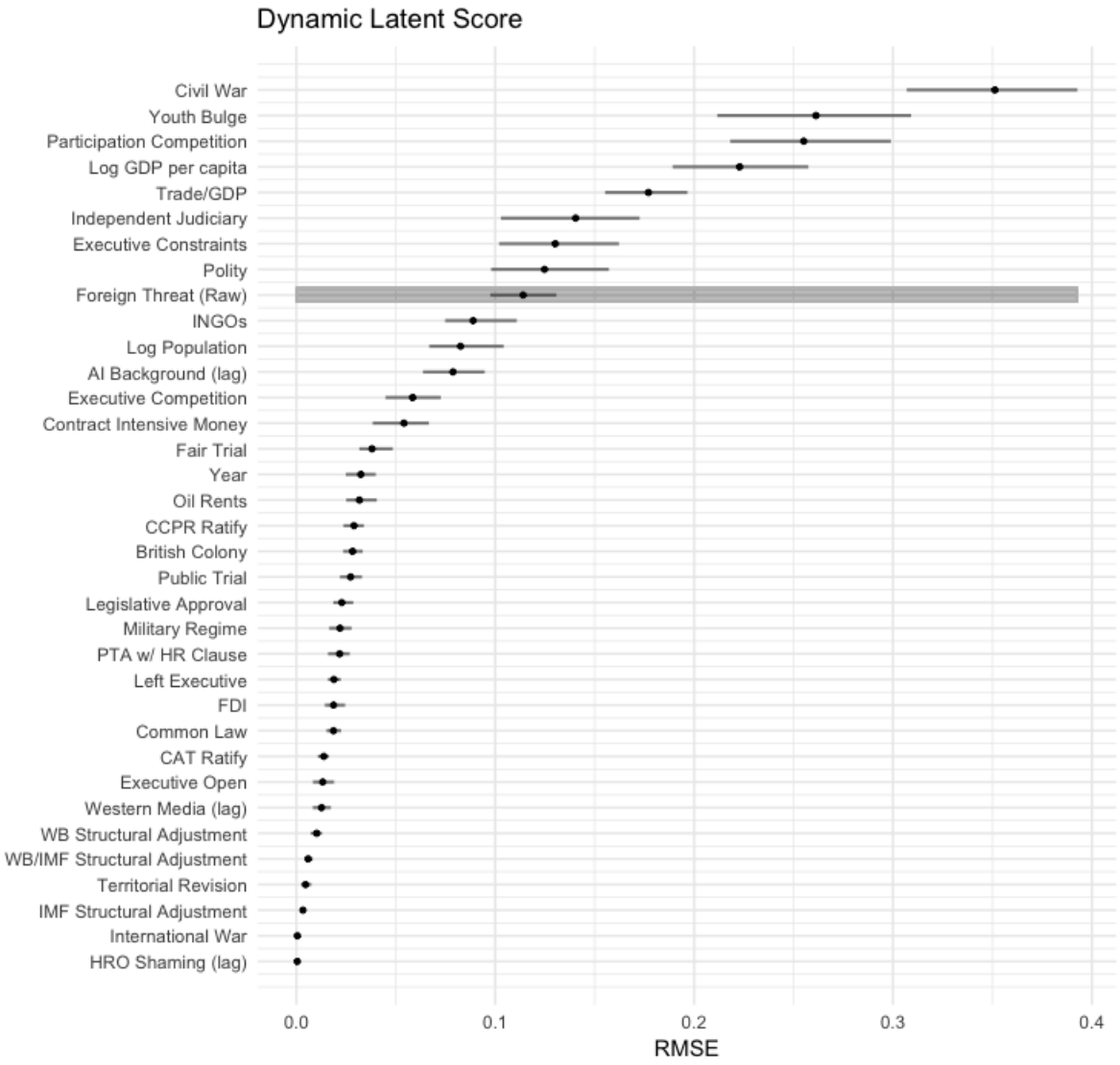

Figure 3.4: Variable permutation scores from a conditional random forest for the Dynamic Latent Score from Fariss (2014). Forest grown with 500 trees and 10 randomly selected predictors.

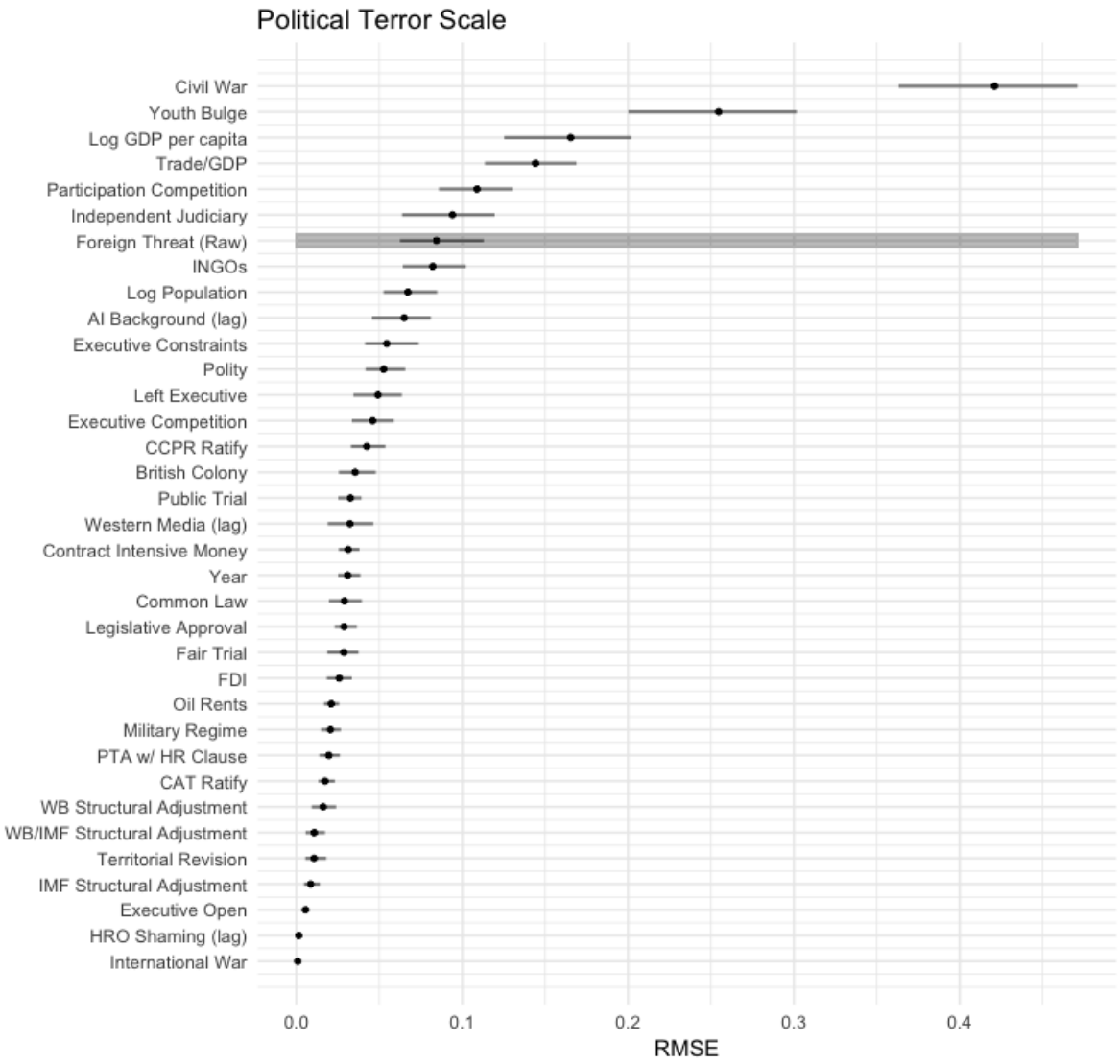

Figure 3.5: Variable permutation scores from a conditional random forest for the Political Terror Scale. Forest grown with 500 trees and 10 randomly selected predictors.

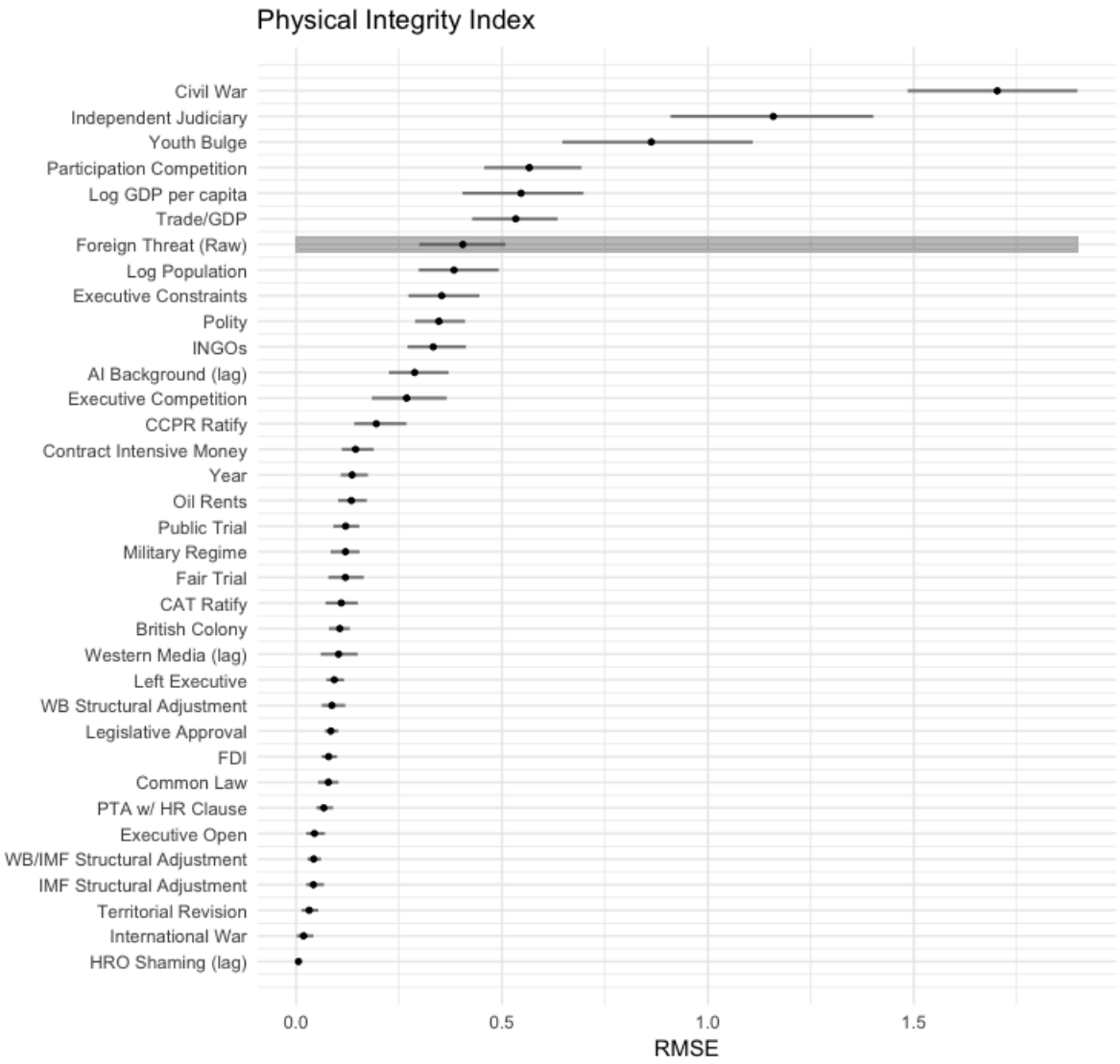

Figure 3.6: Variable permutation scores from a conditional random forest for CIRI's Physical Integrity Index. Forest grown with 500 trees and 10 randomly selected predictors.

I next use a different means of assessing variable importance, examining permutation scores from random forests. To recap, a random forest is an ensemble method consisting of bootstrapped nonparametric decision trees which are grown using a subset of randomly selected predictors. Random forests are a useful tool because they are able to handle complex relationships within the

data and make accurate predictions without having any a priori sense of the relationship between the predictors and the outcome. Variable permutation scores reflect how much worse the random forest model performs when a specific variable is removed from the model: in this case, important variables will be associated with a higher levels of RMSE, indicating that the model performs worse in their absence.

The traditional variable importance scores from decision trees and random forests in the regression setting relied on the average impurity reduction associated with the variable. But, because of this metric's tendency to gravitate towards variables with more splits, the resulting variable importance scores were biased towards these variables (Strobl et al., 2007). Instead, researchers have shifted towards using marginal permutation scores – the average increase in test error when a variable is permuted from trees in the forest. While improved, these can tend to be unstable across different runs of the algorithm, especially in the presence of correlated predictors. Instead, I run conditional random forests, an implementation of random forests grown using the unbiased decision tree algorithm.[10] While still recursively splitting across all variables in the dataset, a conditional inference tree first assesses the global hypothesis that there is information about the outcome variable covered by any of the available predictors. If this global hypothesis is rejected, the algorithm will then measure the association between the outcome and each predictor. It will then select the predictor with the strongest association with the outcome, and then proceed from the beginning until the global hypothesis cannot be rejected.

First, it is important to note that these scores indicate the marginal importance of each variable. That is, the scores do not reflect the importance of each variable conditional on all of the other variables. For highly correlated variables this can be a problem, as results can be spurious and

[10] I additionally ran Ranger, a faster implementation of the normal random forest algorithm; see the appendix for a direct comparison between the two different forests for variable importance.

unimportant predictors can emerge as important if they are highly correlated with meaningful variables. Polity's component scores, for instance, are highly correlated with each other, and this may result in an issue in assessing their importance.[11] In order to address both the instability of permutation scores and the possible inflation of correlated variables, I followed Hill and Jones and selected 10 variables to be randomly selected for each tree and upped the number of trees to 500. An initial check of the results indicated that the results were stable across different levels of the tuning parameters, though this could merit further inspection.[12]

With these caveats aside, the results generate some similar findings to that of using cross-validation with OLS. As observed by Hill and Jones, domestic level outcomes continue to be most strongly related to repression. Civil war is still associated with the largest marginal reduction in test error across each of the three separate outcome measures in both forests. Also as before, judicial independence, executive constraints, youth bulges, GDP per capita, and participation competition are generally top tier variables in reducing error in both settings. There are some differences, however, as certain international factors such as trade and INGO presence shift upwards in importance compared to the variable permutations from linear models.

Most interestingly, the measure for foreign threats is associated with a noticeable decrease in prediction error. It is important to recall that this measure is intended to proxy for international threats facing a state – how likely it is to experience international conflict in a given year. That there is a reduction in test error using this variable indicates there is a meaningful - yet still possibly spurious - relationship between international conflict expectations and domestic behavior within a state. It is, for instance, possible that foreign threats are correlated with GDP per capita, or

[11] See the appendix for a correlation heat map for all variables used in this analysis; most importantly, Foreign Threat (Raw) and Foreign Threat (DOE) are highly correlated, so I only included one in each conditional random forest at a time. I display the results using the raw scores here.

[12] Grömping (2009) for a full discussion of how variable importance can differ with the number of variables selected in random forests.

one of the other variables, and that it is a spurious relationship between threats and repression. Fortunately, the measure of threats is not highly correlated with any other independent variable in this dataset: see the appendix. This is the strongest evidence so far to tip the scale in favor of concluding that foreign threats matter for repression.

Critically, international war offers little explanatory power for predicting repression, while the measure of conflict expectations is consistently among the top 10 predictors across each of the three outcomes. The comparison between these two variables is of some theoretical interest. An applied researcher might have reasonably inferred that international conflict has no meaningful relationship with human rights behavior based on existing research. But, once we measure conflict expectations rather than observed conflict, we see that there is indeed evidence of a relationship. In this case whether international conflict matters depends on how we measure the concept, rather than with the theoretical relationship.

This speaks to the larger importance of being careful with measurement in applied research within political science, as conclusions for the theoretical concept are based on the measures we use. While we can theoretically justify the construct validity of a measure, it is important where possible to use prediction as a criterion to judge the empirical validity of a measure. For this reason, data-driven methods which focus on out of sample prediction can be an effective tool not only for exploratory analysis but also for assessing and improving measurement.

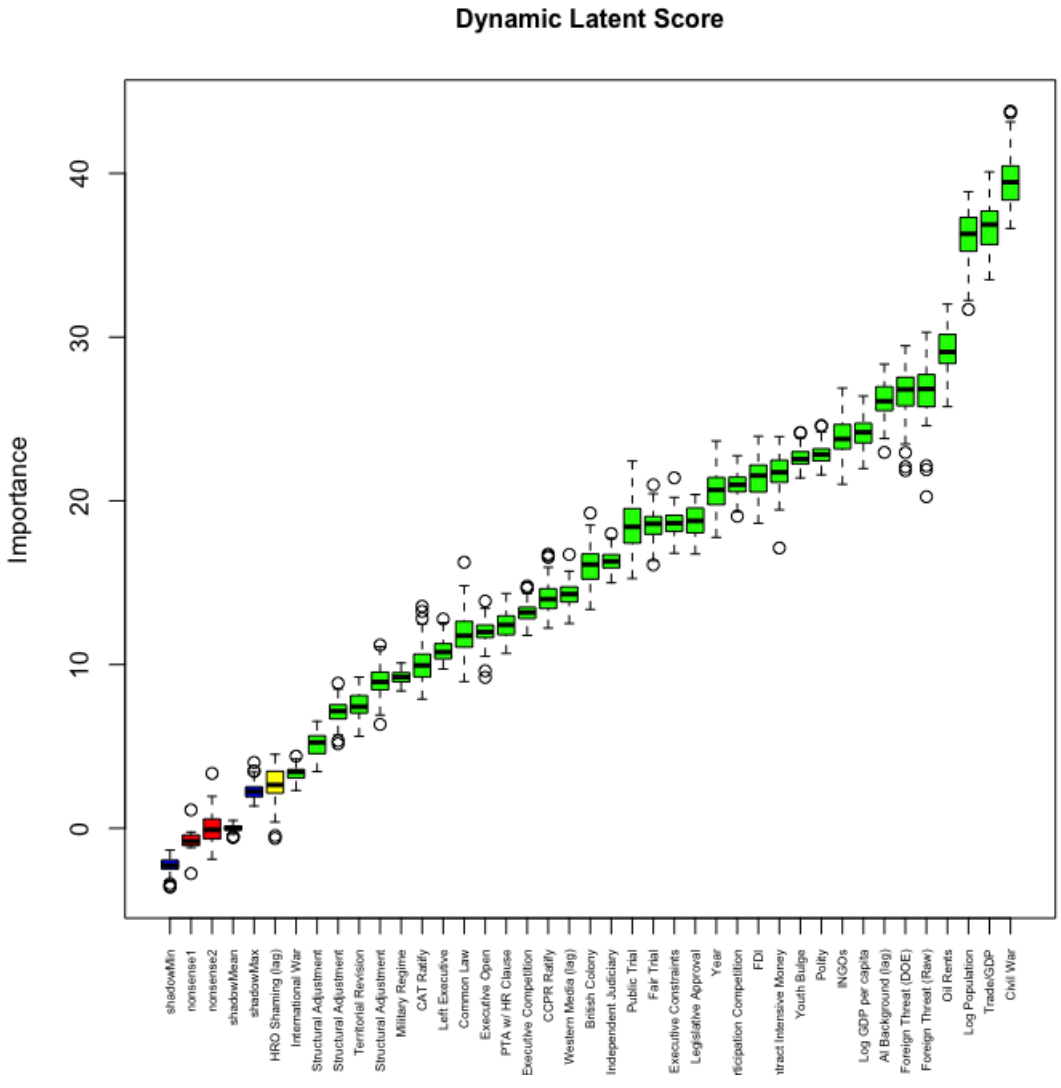

Figure 3.7: Boruta variable importance for Fariss' Dynamic Latent Score

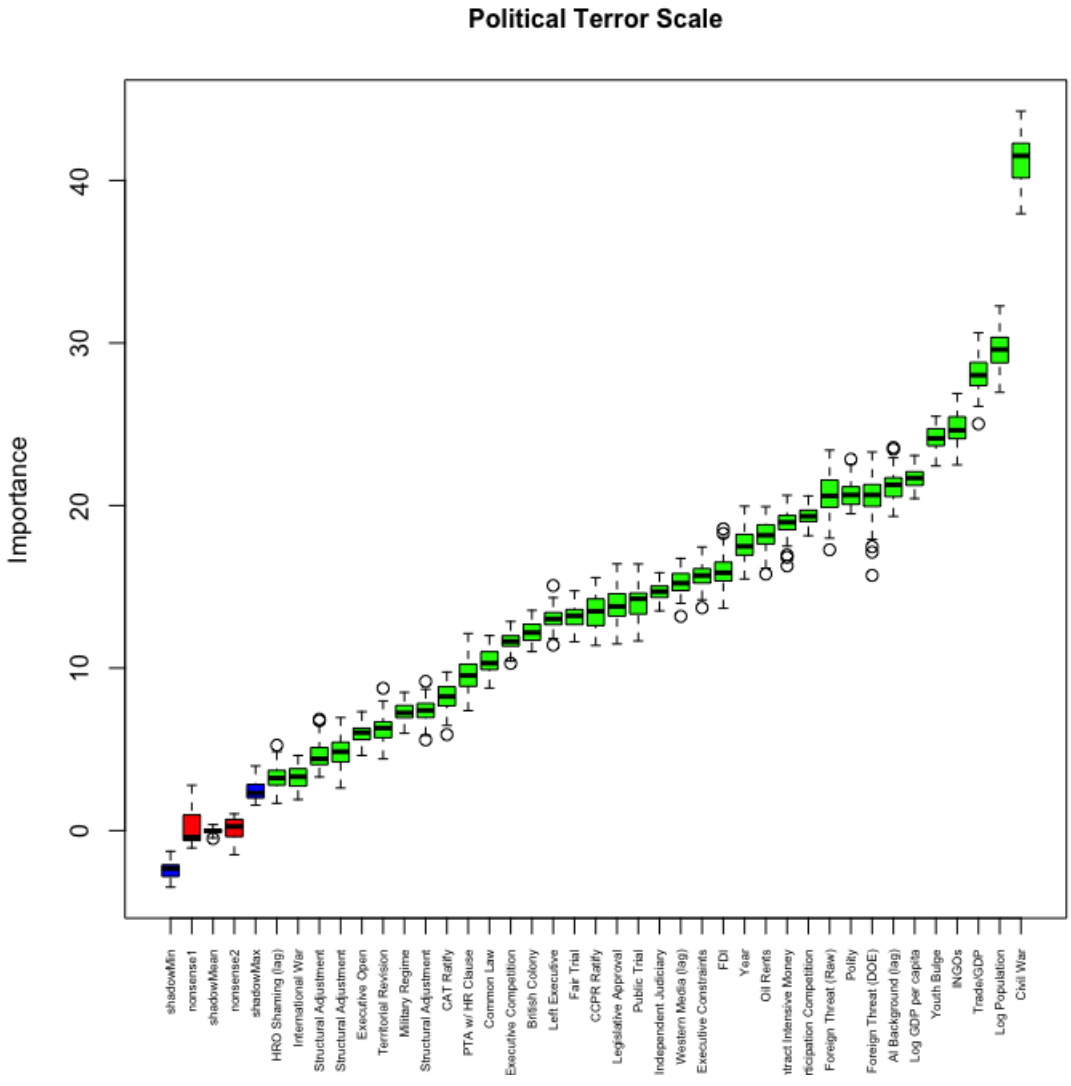

Figure 3.8: Boruta variable importance plots for the Political Terror Scale

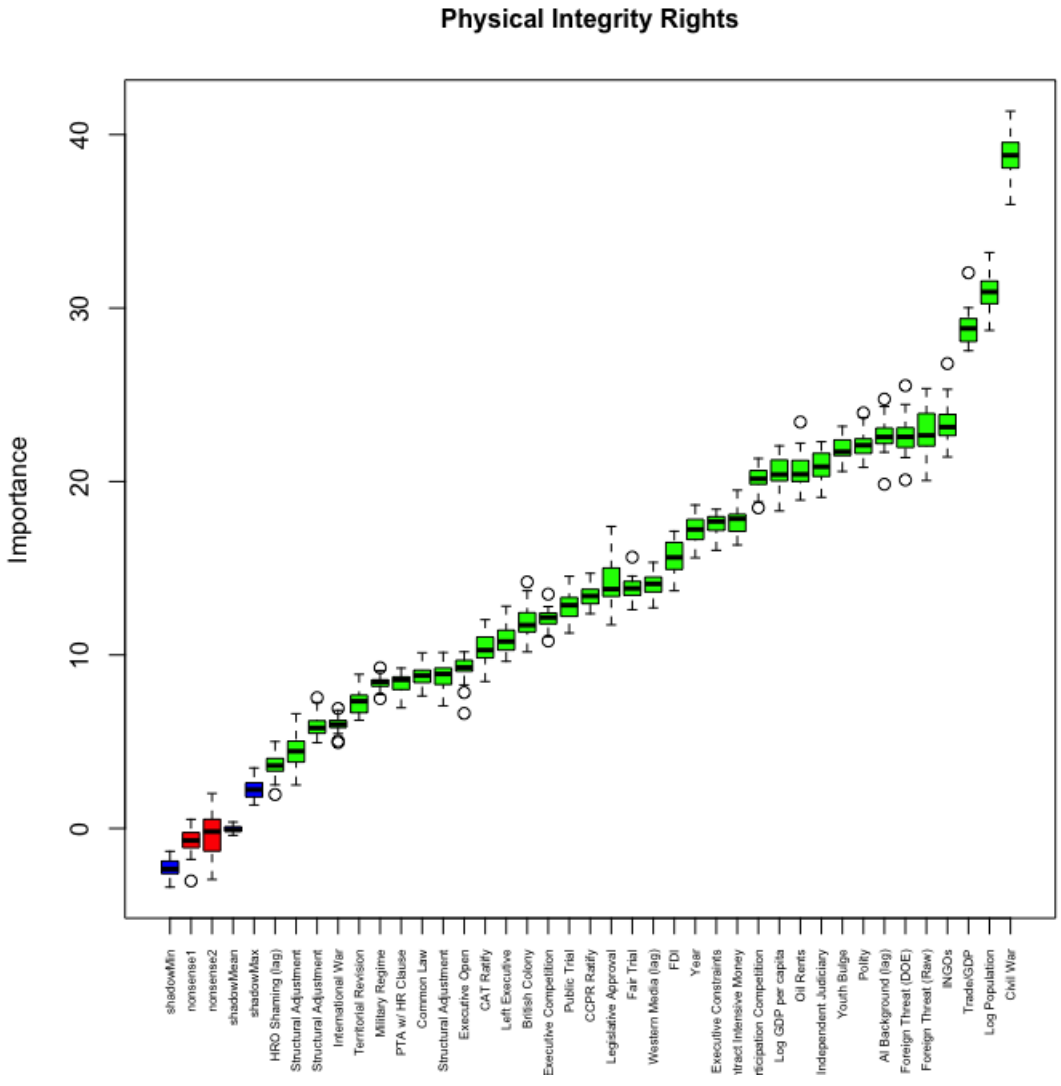

Figure 3.9: Boruta variable importance plots for CIRI's Physical Integrity Index

As a final inspection of variable importance, I use the Boruta algorithm (Kursa, Rudnicki et al., 2010) as a wrapper around Ranger, a faster implementation of standard random forests. In this method, each predictor is shuffled to remove its correlation with the outcome variable, generating a random 'shadow' variable with a similar distribution for comparison with the actual variable. The

shadow variable represents a baseline for predictive performance, as it is by definition randomly associated with the response. By repeatedly comparing the true variable with randomly generated shadow variables, it is possible to directly assess the importance of an individual variable. The main purpose of this test is to identify predictors which are needed for modeling the outcome with the aim of trimming irrelevant predictors. But in iteratively growing random forests, the algorithm also offers a means of assessing the stability of the importance of individual variables. Variables which emerge as the highest on this scale can be thought of as the most important in predicting the outcome. I assessed each variable in this way, growing up to a maximum of 100 random forests with each run.[13]

Here the results continue to align with what the literature has found so far, though international trade shifts upwards in importance relative to previous methods. Civil war, population, and trade emerge as the top three variables for the Dynamic Latent Score, the Political Terror Scale, and the Physical Integrity Index. While these three variables represent the top tier for predictive importance, there is generally a second tier dominated by the other variables studied in the literature of repression: judicial independence, youth bulges, AI background reports, and oil rents. The measure of foreign threats is in this second tier of predictive importance. This is consistent with the findings from the conditional random forests, indicating that foreign threats continue to be meaningful for the task of prediction.

The evidence to this point suggests that the measure of is a meaningful variable for predicting observed repression within states. This stands in contrast to a key takeaway from Hill and Jones' evaluation of the literature for predicting repression :

[13]I additionally included two randomly generated variables which were completely unassociated with the outcome variables as an additional check as to whether the algorithm could identify these variables as unimportant. Both were identified as irrelevant.

> For the most part features of domestic politics, rather than international politics, are adding the most explanatory power to these models... The contrast between results for domestic/international factors suggests that the institutional (political and legal) constraints that exist at the domestic level are more important for the decision to repress than are any international constraints arising from treaties, NGO activity, or a state's situation in the global economy.... But this analysis suggests that international political factors are, in general, not as useful for predicting the level of repression as domestic factors. (674)

The results here do certainly reaffirm that a number of domestic factors are highly important to the study of human rights. Judicial independence and youth bulges in particular continue to receive attention in the study of political violence (Davenport and Armstrong, 2004), along with the effects of international trade (Hafner-Burton, 2005*b*). But, international security and conflict have been relatively unexamined in the literature which is devoted to studying how states wield their coercive power. In the words of Moore 1995, there has been "a paradox between the robust theoretical support for the presence of some kind of a nexus and the dearth of robust evidence for any kind of nexus" (130). To this end, the original findings from Hill and Jones 2014 might steer researchers away from searching for renewed evidence of a relationship between international and intrastate conflict. The results shown here suggest that international factors should not be discounted and merit further investigation.

To that end, future work needs to continue to explore the nexus between international and domestic conflict, especially pertaining to both state repression and citizen dissent. Ritter (2014) develops a breakthrough model which speaks to the importance of a leader's political security in the decision to use coercive means to stay in power. Yet states do not exist in a vacuum, and political security is not solely determined by domestic politics. It is vital, then, that we continue to

explore how international factors relate to domestic outcomes, learning from the data rather than imposing a theoretical bound a priori.

3.4.3 Inference

While the evidence so far suggests the proxy for foreign threats is meaningful in predicting state repression, we might next want to understand the direction of this effect. Do foreign threats reduce or increase state repression? For that I turn to the task of inference, evaluating directly the relationship between the measure of foreign threats and the same outcomes used by human rights and political violence scholars. It is important to stress here that this analysis is not a causal test; there are surely some expectations of endogeneity and selection effects at work (ie, repressive countries may be targeted by other countries; highly repressive countries might select into international conflict) as the predictor of interest is not randomly assigned. Such concerns permeate most studies of these outcomes: for instance, civil war is associated a large reduction in out of sample error and routinely is associated with a degradation in human rights protections and an increase in protection. Young (2012) argues that repression is a cause in the path towards civil war, but it might simply be the other way around as civil wars might lead states to engage in repression.

As this paper is largely an exploratory analysis, I am not aiming to evaluate the evidence through the perspective of testing a causal hypothesis. Instead, in the language of King, Keohane and Verba (1994), the goal here is to make a descriptive, rather than, causal inference. By first establishing a pattern between conflict expectations and human rights behavior, future work can be better enabled to formulate hypotheses about the intersection of international conflict and state repression. While it would be possible to try to overcome the identification problem via the use of an instrument or matching with the data that is available, a more fruitful path forward, in my

estimation, would be to explore disaggregated data to better inspect the chain of causality from international security to human rights behavior. For now, limited to country-year data, I simply aim to unpack the relationship between my measure of foreign threats and the common outcomes studied by this subfield.

As a first inspection of the relationship between foreign threats and each of the outcome variables, I display bivariate scatter plots with LOESS lines (with 95% confidence intervals). At an initial glance, we can see a negative relationship between for human rights and foreign threats. That is, as the measure for threats increases, we see a decrease in respect for human rights (Dynamic Latent Score, Physical Integrity Index) and an increase in observed repression (Political Terror Scale). There is a slight nonlinearity present in each of these cases, but this seems to be driven by a few data points rather than by a clear trend in the data and modeling this in a nonlinear fashion may lead to overfitting.

To estimate the partial effect of foreign threats and control for spuriousness, I next estimate linear models for each dependent variable. Each table displays the results of two different linear models: a reduced model and a full model.[14] The reduced model specification mimics that of (Poe, Tate and Keith, 1999), using the most common set of predictors in the literature along with a lagged DV and a polynomial for year to capture time dependence. The full model specification makes use of the full set of variables along with a lagged DV and a year polynomial. In both model specifications I use bootstrapped standard errors.

In terms of evaluating the evidence, the results are mixed. Using the reduced model specification with standard predictors from the literature, I find that, holding all else constant, foreign threats are significantly associated (at the 10%) level with a reduction in a state's respect for human rights

[14]Note: I did not include international war in these model specifications, replacing it instead with my measure of conflict expectations. The results are essentially unchanged with and without this variable.

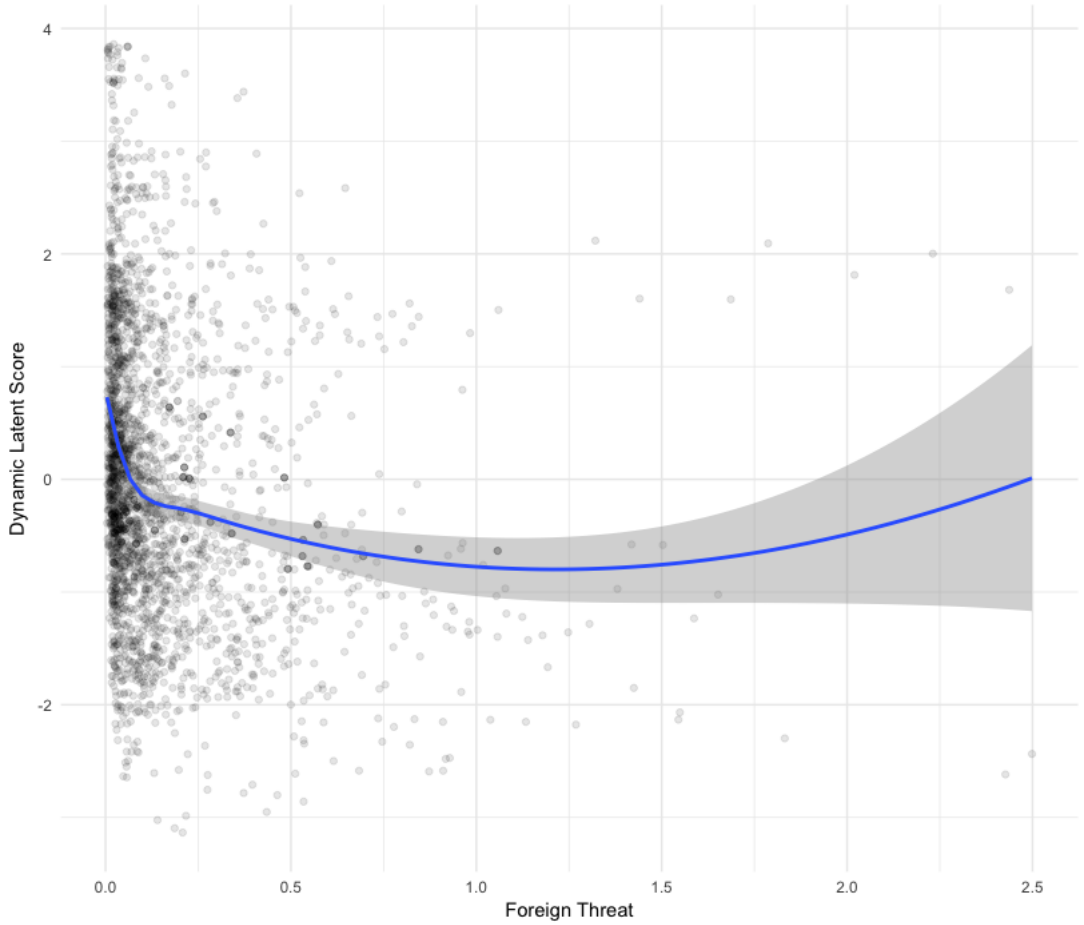

Figure 3.10: Scatter plot of Fariss' dynamic latent score and the measure of foreign threats. LOESS line fit with 95% confidence interval

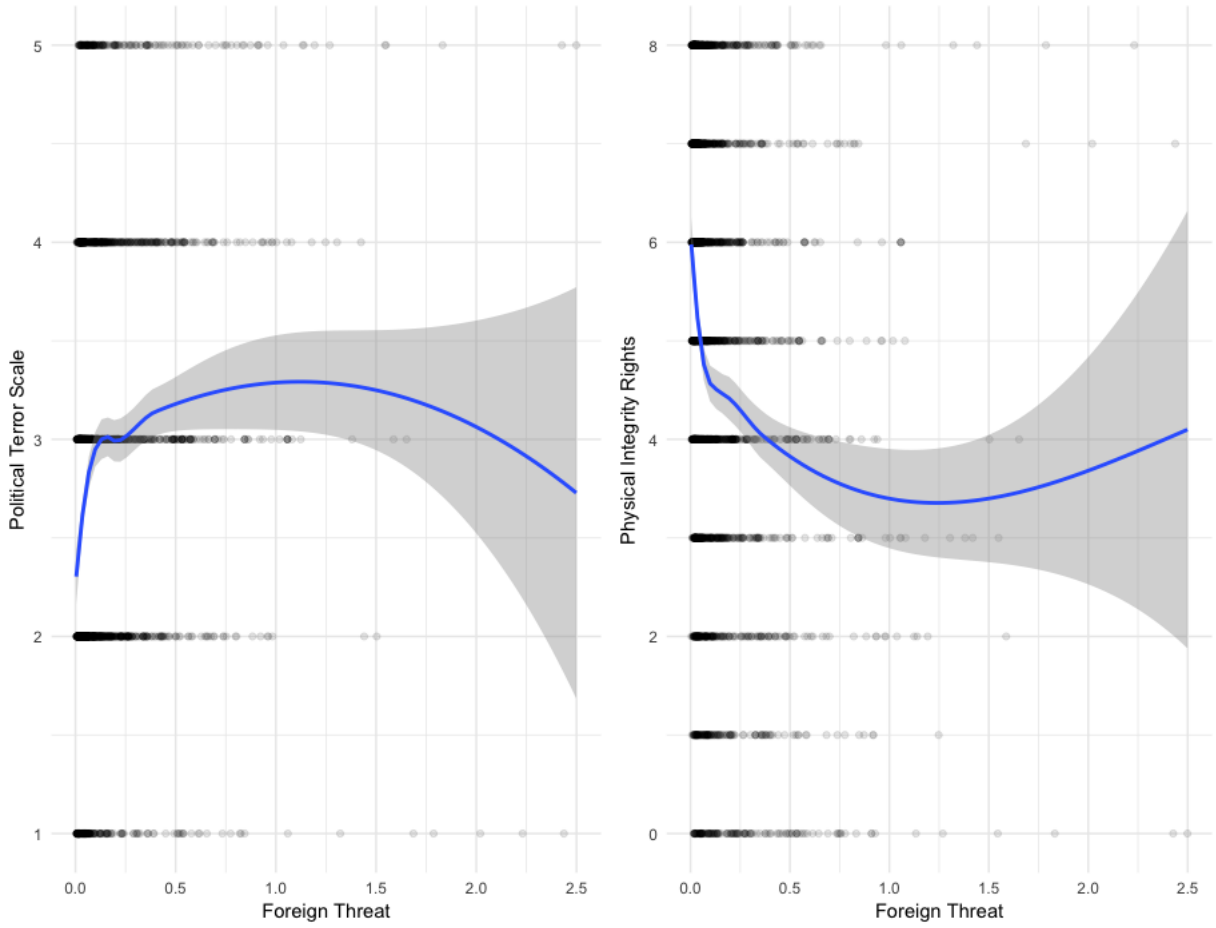

Figure 3.11: Scatter plots of the Political Terror Scale and the measure of foreign threats (left) and CIRI's Physical Integrity Index and the measure of foreign threats (right). LOESS line fit with 95% confidence interval

(Dynamic Latent Score) and an increase in political violence (Political Terror Scale). I find results in the same direction for the Physical Integrity Index (an increase in threat decreases respect for human rights), but it is not significant at conventional levels.[15]

But, in the full model specification, I find no evidence of a linear effect of foreign threats on human rights behavior after controlling for all of the most common variables found in the literature. Specifically, in modeling the Dynamic Latent Score, the coefficient for 'Foreign Threat' is only slightly reduced and is still negative, and the difference between the two coefficients may not necessarily be meaningful. But in the full model of the Political Terror Scale, the coefficient for the measure flips in direction, while in the full model of the Physical Integrity Index the coefficient is reduced in magnitude.

There are a couple of takeways for the question of inference. First, the results from the reduced models the dynamic latent score and political terror scale suggest that expectations of international conflict are negatively associated with state repression. But, these results are highly sensitive to both model specification as well as the variable used to measure repression. The literature generates the implication that external threats may either increase or reduce repression; it is difficult to reach a definitive conclusion supporting either way.

Second, why might a variable be important for prediction without evidence of a statistically significant effect in a linear model? The answer likely lies in interactions and nonlinearities. Foreign threats in particular may play a mediating role in affecting other outcomes within a state. A flexible model which is suited to handling interactions and nonlinearities (random forests) will be able to recover these relationships within the data. A standard linear model will not be able to pick up

[15]It's worth noting here that as both PTS and CIRI are ordinal, it might be more appropriate to use ordered logit models. While I did do so (see the appendix; the results are largely unchanged), I display linear models here in order to examine the models which were assessed for predictive validity earlier. The main reason is to mimic the modeling approach commonly used in this particular literature. While some efficiency might be lost and nonsensical predictions may be possible, inferences are safe after fixing the standard errors.

these relationships unless they are imposed a priori by the applied researcher. Given that this paper is devoted to learning from the data rather than testing hypoptheses with data, I do not investigate interactive hypotheses here, as I did not begin the enterprise with a theoretical model generating these hypotheses. But, future work can begin to shed some light in this area by first developing theories which posit meaningful interaction and then testing them (with additional care taken to see that the model is externally valid).

3.5 Conclusion

Under what circumstances are domestic decisions driven by international factors? I find in this paper that a measure proxying for foreign threats - the probability of being attacked by other states - is an important predictor for measures of state repression and human rights behavior. The evidence as to *how* foreign threats are related to these outcomes is unclear: I find some indications that foreign threats increase repression, but the evidence is far from certain. Moving beyond the case of repression, the field must continue to examine under what circumstances domestic politics are produced because of international considerations. What if a leader's winning coalition is not domestic but international, and their political survival depends on foreign actors? We simply do not know. I hope to draw attention once again to exploring the intersection between international and domestic conflict in the aim of understanding how the international arena affects domestic politics.

Table 3.3: Linear models of Fariss' dynamic latent score, using a reduced and full set of predictors, with bootstrapped standard errors.

Dynamic Latent Score

	Reduced Model			Full Model		
	Coef	SE	P.Value	Coef	SE	P.Value
(Intercept)	0.1787	0.0615	0.0037	0.1640	0.0716	0.0220
Dynamic Latent Score (lag)	0.9552	0.0044	0.0000	0.9582	0.0047	0.0000
Polity	0.0025	0.0006	0.0001	0.0019	0.0007	0.0051
Independent Judiciary	−0.0001	0.0056	0.9888	0.0029	0.0060	0.6309
Log Population	−0.0105	0.0025	0.0000	−0.0133	0.0036	0.0002
Log GDP per capita	0.0074	0.0045	0.1035	0.0088	0.0055	0.1086
CAT Ratify	−0.0011	0.0084	0.8953	−0.0047	0.0086	0.5868
CCPR Ratify	−0.0017	0.0076	0.8259	0.0065	0.0078	0.4066
Youth Bulge	−0.0036	0.0008	0.0000	−0.0029	0.0008	0.0005
Civil War	−0.0589	0.0110	0.0000	−0.0551	0.0111	0.0000
Foreign Threat (Raw)	−0.0245	0.0130	0.0598	−0.0182	0.0142	0.1982
Oil Rents				−0.0012	0.0013	0.3887
Military Regime				0.0025	0.0089	0.7738
Left Executive				0.0001	0.0027	0.9566
Trade/GDP				0.0000	0.0001	0.9747
FDI				−0.0005	0.0006	0.3663
Public Trial				−0.0052	0.0056	0.3508
Fair Trial				−0.0025	0.0064	0.6941
Legislative Approval				0.0086	0.0037	0.0182
WB/IMF Structural Adjustment				−0.0211	0.0189	0.2642
IMF Structural Adjustment				−0.0093	0.0179	0.6023
WB Structural Adjustment				0.0278	0.0215	0.1949
British Colony				−0.0140	0.0095	0.1394
Common Law				0.0138	0.0104	0.1869
PTA w/ HR Clause				−0.0209	0.0082	0.0105
Territorial Revision				−0.0021	0.0115	0.8543
AI Background (lag)				0.0002	0.0005	0.6976
Western Media (lag)				0.0095	0.0053	0.0713
HRO Shaming (lag)				−0.0030	0.0041	0.4671
INGOs				0.0000	0.0000	0.3336
N		2671			2671	
Adjusted R^2		0.983			0.984	
Year Polynomial?		Yes			Yes	

Table 3.4: Linear models of the Political Terror Scale, using a reduced and full set of predictors, with bootstrapped standard errors.

Political Terror Scale

	Reduced Model			Full Model		
	Coef	SE	P.Value	Coef	SE	P.Value
(Intercept)	−0.0568	0.2523	0.8218	0.5609	0.2982	0.0601
Political Terror Scale (lag)	0.6103	0.0176	0.0000	0.5732	0.0184	0.0000
Polity	−0.0053	0.0027	0.0506	−0.0012	0.0029	0.6815
Independent Judiciary	−0.0873	0.0233	0.0002	−0.0642	0.0252	0.0109
Log Population	0.0580	0.0111	0.0000	0.0542	0.0143	0.0002
Log GDP per capita	−0.0028	0.0194	0.8857	−0.0465	0.0250	0.0634
CAT Ratify	0.0190	0.0361	0.6000	0.0170	0.0367	0.6432
CCPR Ratify	0.0104	0.0331	0.7523	0.0203	0.0346	0.5574
Youth Bulge	0.0206	0.0031	0.0000	0.0192	0.0035	0.0000
Civil War	0.4650	0.0444	0.0000	0.4546	0.0446	0.0000
Foreign Threat (Raw)	0.0905	0.0544	0.0961	−0.0132	0.0563	0.8147
Oil Rents				0.0096	0.0061	0.1182
Military Regime				−0.0424	0.0356	0.2336
Left Executive				−0.0354	0.0118	0.0028
Trade/GDP				−0.0005	0.0004	0.2380
FDI				0.0036	0.0024	0.1274
Public Trial				0.0010	0.0245	0.9681
Fair Trial				−0.0643	0.0277	0.0205
Legislative Approval				0.0075	0.0153	0.6225
WB/IMF Structural Adjustment				−0.0787	0.0716	0.2715
IMF Structural Adjustment				−0.0561	0.0687	0.4144
WB Structural Adjustment				0.0380	0.0846	0.6533
British Colony				0.0696	0.0389	0.0735
Common Law				−0.1479	0.0454	0.0012
PTA w/ HR Clause				−0.0201	0.0354	0.5715
Territorial Revision				0.0355	0.0522	0.4960
AI Background (lag)				0.0062	0.0022	0.0038
Western Media (lag)				0.0188	0.0177	0.2889
HRO Shaming (lag)				0.0012	0.0177	0.9461
INGOs				0.0000	0.0000	0.2525
N		2150			2150	
Adjusted R^2		0.673			0.679	
Year Polynomial?		Yes			Yes	

Table 3.5: Linear models of CIRI's Physical Integrity Index, using a reduced and full set of predictors, with bootstrapped standard errors.

Physical Integrity Index

	Reduced Model			Full Model		
	Coef	SE	P.Value	Coef	SE	P.Value
(Intercept)	4.5359	0.4881	0.0000	4.3765	0.5602	0.0000
Physical Integrity Index (lag)	0.5978	0.0163	0.0000	0.5708	0.0170	0.0000
Polity	0.0211	0.0048	0.0000	0.0185	0.0054	0.0006
Independent Judiciary	0.2822	0.0472	0.0000	0.2437	0.0502	0.0000
Log Population	−0.1604	0.0197	0.0000	−0.1864	0.0267	0.0000
Log GDP per capita	−0.0023	0.0357	0.9478	0.0525	0.0460	0.2539
CAT Ratify	−0.0815	0.0648	0.2085	−0.1157	0.0655	0.0778
CCPR Ratify	−0.1406	0.0585	0.0163	−0.1605	0.0616	0.0092
Youth Bulge	−0.0404	0.0055	0.0000	−0.0379	0.0061	0.0000
Civil War	−0.9823	0.0915	0.0000	−0.9802	0.0910	0.0000
Foreign Threat (Raw)	−0.1556	0.1033	0.1320	−0.0355	0.1067	0.7392
Oil Rents				−0.0103	0.0107	0.3349
Military Regime				0.0461	0.0724	0.5246
Left Executive				−0.0387	0.0213	0.0690
Trade/GDP				−0.0002	0.0007	0.7314
FDI				−0.0009	0.0038	0.8062
Public Trial				−0.0317	0.0443	0.4748
Fair Trial				0.1285	0.0482	0.0077
Legislative Approval				−0.0015	0.0276	0.9580
WB/IMF Structural Adjustment				−0.1342	0.1342	0.3176
IMF Structural Adjustment				−0.1009	0.1342	0.4521
WB Structural Adjustment				0.3241	0.1615	0.0449
British Colony				−0.1479	0.0682	0.0302
Common Law				0.1437	0.0744	0.0535
PTA w/ HR Clause				0.1145	0.0633	0.0706
Territorial Revision				−0.1199	0.0978	0.2203
AI Background (lag)				−0.0085	0.0041	0.0394
Western Media (lag)				0.0012	0.0284	0.9669
HRO Shaming (lag)				0.0190	0.0338	0.5749
INGOs				0.0001	0.0001	0.0246
N		2500			2500	
Adjusted R^2		0.734			0.739	
Year Polynomial?		Yes			Yes	

APPENDIX A

PREDICTING THE COSTS OF WAR

A.1 The Ensemble Model

I assess whether I can achieve even better predictive performance by using an ensemble of all of the candidate models (Van der Laan, Polley and Hubbard, 2007). The intuition is that a weighted average of all of the models will outperform the results of one model alone. I take the out of sample predictions from every candidate model at each value of its tuning parameters (excluding the null models) and bind them into a matrix. I use Y.Ye's general nonlinear augmented Lagrange multiplier method solver (Ye, 1987; Ghalanos and Theussl, 2015) to select the optimal model weights for minimizing the loss function, in this case the square root of the difference between $\hat{Y}$ and Y. Table 3 displays the weights given to each of the candidate models with a weight greater than 0.0001 in the ensemble. The weights are primarily assigned to the predictions of the random forest, followed by Cubist, boosted trees, MARs, and neural nets. The random forest using the strongest opponent criterion receives the largest weight, though each of the three data approaches does have models which are ultimately used in the ensemble. I take these weights and the predictions from each candidate model to produce the ensembled predictions, which achieve a final test performance of 2.795 RMSE, achieving a final improvement over the null of 26%. This represents an incremental improvement over the performance of the random forest in the strong setting (RMSE = 2.833). If computational time was a key constraint, as it often is with predictive tasks, I would proceed with the simpler model. But given that the goal here is to simply find the best predictions for Y using X, I proceed with the ensemble model.

Table A.1: Results for each candidate model with a weight greater than 0.001 in the ensemble model. The performance of the ensemble in terms of RMSE and proportional reduction in loss from the null model for comparison to each of the candidate models.

Data	Model	Weight	RMSE	PRL
Average	Cubist: committees=20, neighbors=1	0.017	3.567	0.066
Strong	Neural Nets: size=11, decay=0.06	0.025	3.055	0.200
Aggregate	Neural Nets: size7, decay=0.04	0.034	3.040	0.203
Strong	MARS: nprune=26, degree=1	0.016	3.040	0.204
Aggregate	MARS: nprune18, degree=1	0.124	3.021	0.208
Aggregate	MARS: nprune22, degree=1	0.031	3.021	0.209
Aggregate	Cubist: committees1 .neighbors=9	0.103	2.996	0.215
Average	Boosted Trees: ntrees=200, depth=9, shrinkage=0.1	0.107	2.975	0.221
Aggregate	Random Forest: mtry=5	0.134	2.835	0.257
Average	Random Forest: mtry=13	0.086	2.837	0.257
Strong	Random Forest: mtry=13	0.324	2.833	0.258
All	Ensemble	-	2.795	0.268

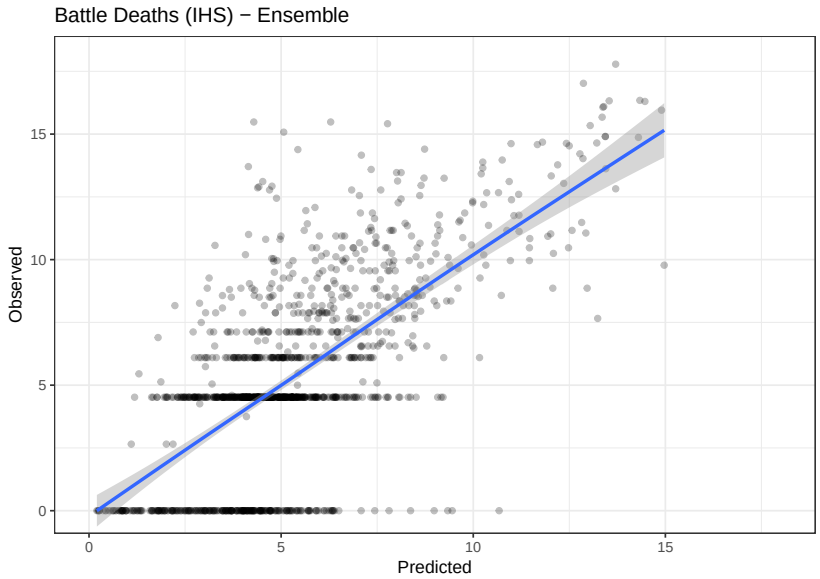

Figure A.1: Out of sample performance of the ensemble model. The scatter plot shows the ensemble model's out of sample predictions regressed against the observed values with a LOESS line with a 95% confidence interval.

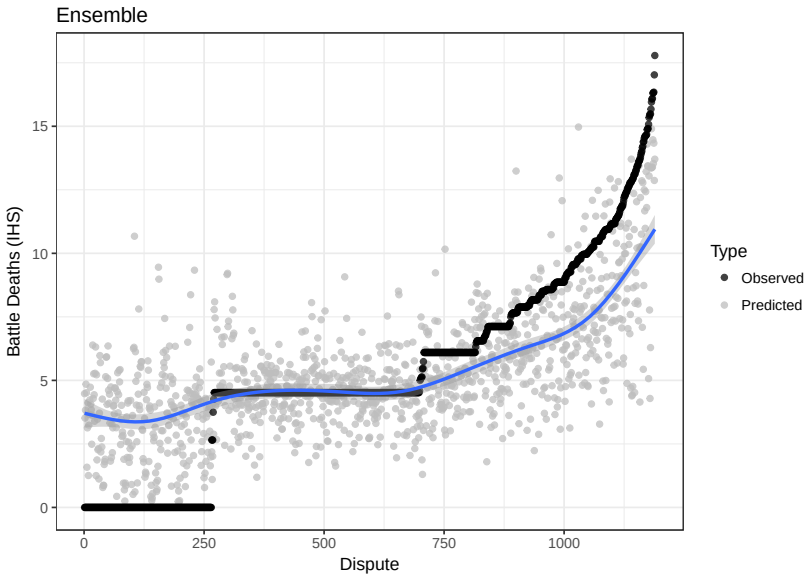

Figure A.2: Observed values of battle deaths sorted from least to greatest against the model's predictions for each of these observations, LOESS line with a 95% confidence interval added.

A.2 Alliances and War Costs

The DiCE results shown in this paper have been predicted battle deaths for bilateral conflicts. One immediate counterpoint would be that wars are not typically fought between two states. That is, if North Korea and South Korea were to fight a war, we would not expect that conflict to be confined to those two states. We would reasonably expect the United States and China to intervene, which would affect the war costs for both sides. We should therefore expect allies to be an important factor in predicting the costs of war and the effort to predict war costs should be able to account for alliances. Fortunately, the modeling approach I have used is flexible enough to incorporate the features of allies.

Return quickly to the set up of state features described earlier. Consider a scenario where states A and B are in an alliance and fight an alliance of states C and D. I can model each state's battle

deaths in the following way:

Modeling Ally Features:
A Battle Deaths = f(A Features, Ally (B) Features, (C+D Features), Dyad Features, Year)
B Battle Deaths = f(A Features, Ally (A) Features, (C+D Features), Dyad Features, Year)
C Battle Deaths = f(A Features, Ally (C) Features, (A+B Features), Dyad Features, Year)
D Battle Deaths = f(C Features, Ally (D) Features, (A+B Features), Dyad Features, Year)

This again poses the question of how to incorporate features from additional states. As evidenced by the prior results, aggregating capabilities provided the best out of sample performance, so I aggregate ally features and incorporate their features directly. Here it is important to note that the original approach I adopt may be failing to account for alliance features. Presumably, if A and B fight alongside each other, their costs of conflict will be affected by each other's presence. That is, a state fighting alongside an ally should reasonably face different costs than if they fought alone. The original set up does not explicitly allow for this by not directly modeling ally features. With this in mind, we might expect incorporating ally features to improve our predictions. This is ultimately an empirical question: does modeling ally features lead to better out of sample predictions?

To answer this, I re ran the same methodology presented so far while including aggregated ally capabilities. I then ensembled these models to produce an another ensemble out of sample prediction for battle deaths. The end result produces very similar to that of the original ensemble (RMSE = 2.818 compared to RMSE 2.795). Some models perform better with the addition of ally features - the neural networks markedly improve over their performance in the original aggregated modeling approach. The tree based models continue to be the best performers, with similar results to that of the original results. At a glance, the results here do not demonstrate the utility of adding more features to accomodate allies. However, for the purpose of predicting multilateral conflicts, this approach may still be preferred for scholars seeking an estimate of costs. As there are too many

Table A.2: Test performance of the candidate models having incorporated ally features more directly.

	RMSE	SD	Weight
Null	3.816	0.116	0.000
CINC+Year	3.434	0.077	0.000
OLS	3.027	0.134	0.000
PLS	3.016	0.147	0.074
Elastic Net	3.014	0.140	0.000
KNN	3.102	0.090	0.000
Cart	3.189	0.057	0.000
Random Forests	2.843	0.075	0.556
Boosted Trees	2.895	0.081	0.047
Cubist	2.889	0.087	0.198
SVM - Radial	3.000	0.086	0.000
Neural Nets	2.965	0.085	0.125
Ensemble RMSE: 2.818			

hpothetical multilateral conflicts to predict them all, for the applied researcher I have written a function which permits the user to estimate the expected casualties for hypothetical pairings using the methodology here, and the function is available with instructions on GitHub. One extension in this area would be to predict alliance fulfillment and add a dimension of uncertainty to whether alliances will come to aid.

A.3 Political Institutions and War Costs

The enterprise here is to provide a measure of war costs with which applied researchers can examine outcomes. Though I have primarily sought to develop the measure in this paper, one immediate topic to examine is the interaction of political institutions and expected war costs. One of the justifications for this enterprise is that of the democratic peace theory. One argument in this realm is that because democracies must win the wars they fight, democracies will fight harder, leading to a much more costly war for their opponent.(Bueno de Mesquita et al., 1999) Because a

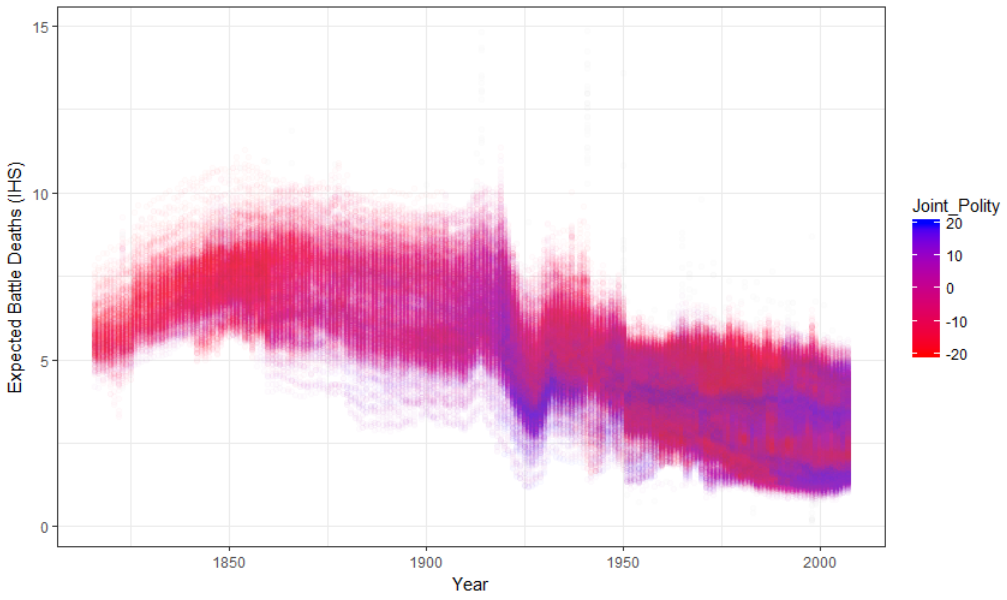

Figure A.3: Expected battle deaths by dyad Polity scores

war between two democracies would be prohibitively costly for both parties, we rarely if ever observe wars between democratic states. With this intuition we would expect to see very high expected battle deaths for democratic pairings relative to that of mixed or autocratic pairings. Figure 9 shows the DiCE estimates scaled by the combined polity score of the dyad. The main observation here is that democracies do not exhibit different patterns of war costs than do other country pairings. Indeed, if anything there is a small pattern in the opposite direction, that pairings of democracies have lower expected war costs. A simple t-test comparing the average costs of democratic dyads vs all others lends evidence to this proposition: democratic dyads have lower costs than others. This is in line with the evidence shown prior that institutional variables offered little value to the task of predicting battle deaths. But this result should be interpreted with some care. The costs of conflict produced here are conditional on states having entered into a dispute. That democracies experience lower expected costs might simply indicate that democracies are better at signaling information

(cite audience costs literature) and thereby better able to avert costly conflicts.

A.4 Lasso and Ridge Regressions

Though I show the results of variable importance plots following (Hill and Jones, 2014) in the paper, I show here the results of a lasso and ridge regression to further identify the subset of predictors which offer improvements in out of sample performance. Using the one standard error rule advocated by (Hastie, Tibshirani and Friedman, 2009), I tuned a lasso and ridge using cross validation to minimize the RMSE. Table 6 shows the coefficients from these models, while FIgure 10 shows the coefficient paths with the a dotted line to indiciate the value of lambda which was selected via cross validation. The lasso shrinks the coefficients of the NMC components for each state, which should be expected as these predictors are highly correlated. Interestingly, the lasso retains different components for states A and B, with iron and steel, military personnel, and Polity 2 remaining for state A and only CINC remaining for side B. Taken as a whole, the relatively low number of predictors in the model does not fully illustrate the utility of penalized methods, but this offers another cut at determining the relevant subset of features used in modeling battle deaths. I place more emphasis on variable importance scores from the random forest and Cubist models largely because these models perform much better in cross validation. The elastic net, which combines the regularization of the ridge and the feature selection of the lasso, achieves largely the same performance as a standard linear model during cross validation, which the tree based models consistently outperform.

A.5 Tuning Parameters

I relied on the caret package (Kuhn, 2008) in R for estimation and cross validation. I present here the tuning parameters for each of the candidate models used in this paper. Though I separately

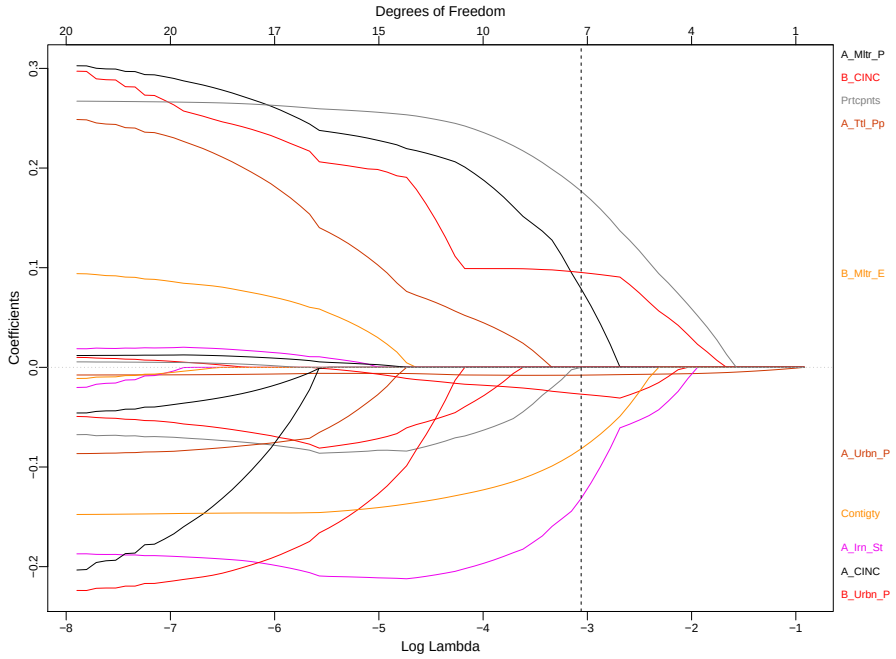

Figure A.4: Lasso variable trace plots for all predictors used in modeling battle deaths

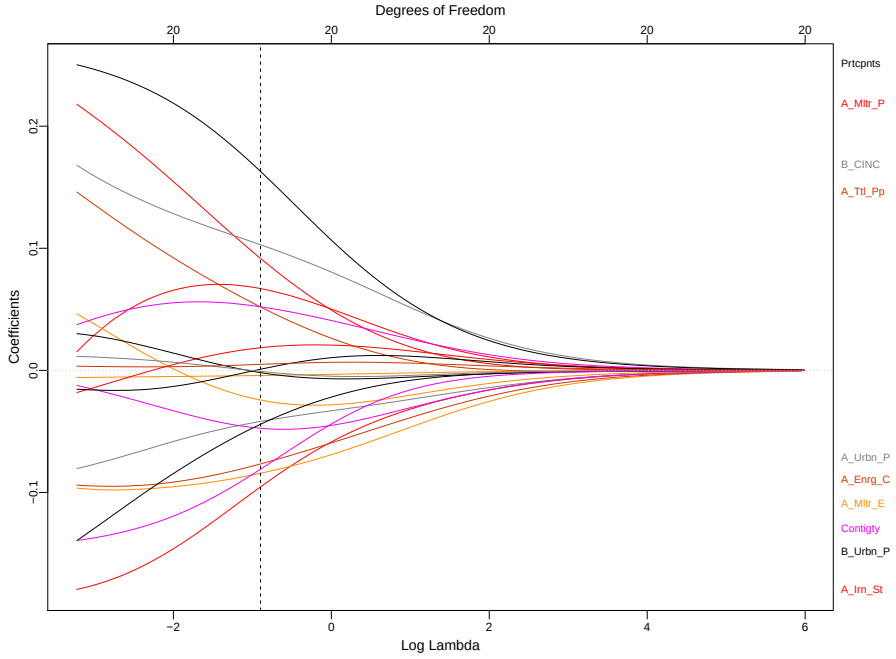

Figure A.5: Ridge regression variable trace plots for all predictors used in modeling battle deaths

Table A.3: Coefficient estimates from an OLS compared to shrunken coefficients from the lasso and ridge. Predictors centered and scaled.

	OLS	Lasso	Ridge
(Intercept)	15.207	15.303	7.700
Year	−0.008	−0.008	−0.004
A Iron Steel	−0.185	−0.132	−0.083
A Military Expenditures	−0.042	–	−0.080
A Military Personnel	0.314	0.079	0.077
A Energy Consumption	−0.064	–	−0.072
A Total Population	0.266	–	0.043
A Urban Population	−0.086	–	−0.039
A CINC	−0.240	–	0.063
A Polity2	0.013	−0.027	−0.048
B Iron Steel	−0.054	–	0.020
B Military Expenditures	0.099	–	−0.027
B Military Personnel	−0.038	–	0.049
B Energy Consumption	0.015	–	−0.004
B Total Population	−0.020	–	0.005
B Urban Population	−0.231	–	−0.036
B CINC	0.336	0.095	0.096
B Polity2	0.011	–	−0.003
Contiguity	−0.148	−0.082	−0.069
Participants	0.268	0.177	0.146
ICOW Salience	0.006	–	0.006

tuned models for the strongest opponent, average, and aggregate models, I list only the aggregate tuning parameters here in order to save space.

1. Partial Least Squares (Wold, 1985)
    - Packages: pls
    - Tuning Parameters: components = 10
2. Elastic Net (Zou and Hastie, 2005)
    - Packages: elasticnet
    - Tuning Parameters: fraction = 0.55, lambda = 0
3. k-Nearest Neighbors (Cover and Hart, 1967)
    - Packages: knn
    - Tuning Parameters: k = 11
4. Classification and regression trees (CART) (Breiman et al., 1984)
    - Packages: rpart
    - Tuning Parameters: maxdepth = 7
5. Random forests (Breiman, 2001)
    - Packages: randomForest
    - Tuning Parameters: mtry = 5
6. MARs (Friedman, 1991)
    - Packages: earth
    - Tuning Parameters: nprune = 22, degree =1
7. Stochasic Gradient Boosted Trees (Friedman, Hastie and Tibshirani, 2001) (Elith, Leathwick and Hastie, 2008)

- Packages: gbm
- Tuning Parameters: mstop = 150, maxdepth = 3, nu = 0.1

8. Cubist (Kuhn et al., 2012)
    - Packages: Cubist
    - Tuning Parameters: committees = 20, neighbors = 9
9. Support vector machines with a radial kernel (Scholkopf et al., 1997)
    - Packages: kernlab
    - Tuning Parameters: sigma=0.045, C=1
10. Averaged Neural Networks (Scholkopf et al., 1997)
    - Packages: nnet
    - Tuning Parameters: size = 5, decay = 0.09, bag = T

APPENDIX B

A NEW MEASURE OF FOREIGN THREAT

B.1 Logistic Regression Coefficient Plot

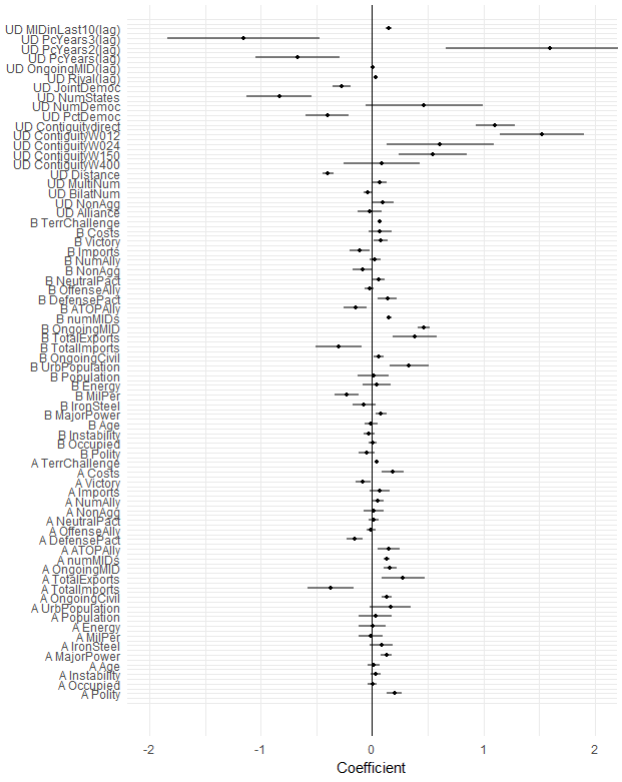

Figure B.1: Logistic regression using all predictors for the full dataset. Standardized coefficients to facilitate comparison; 95% robust standard errors reported around each point estimate.

B.2 Modeling with Lagged Inputs

Since the data involved in the paper is country-year, I wondered whether it was possible that the strong performance of the models was due to the inputs being affected by the presence of militarized disputes. Ie, military personnel that is reported for Russia in 1915 might be lower than

Table B.1: Results from cross validation on training set using lagged inputs

Model	LL	AUC
Random Forest	0.024	0.984
Neural Nets	0.029	0.952
Elastic Net	0.0317	0.959
Logit	0.0318	0.9608
SVM Radial	0.033	0.930
MARS	0.03464	0.929
Boosted Logit	0.0381	0.916
CART	0.043	0.751
Null	0.055	0.499

previous years because a conflict is underway, and it would then be easier to pick out when conflicts are occuring based on drops in the inputs. One way to check for this was to re run the models using lagged versions of the inputs, as these would not be affected in anyway by military disputes occuring in the following year. If there was a big drop in performance as a result of this change in the data structure, that would indicate the approach of using contemporaneous years for prediction is misleading. Fortunately, there was no substantive change in model performance using lagged inputs, indicating that this was not the explanation for the performance of the models.

B.3 Modeling with Incremental Years

Starting in 1875, I trained a logistic regression on all years prior to t and classified observations in years *t+1,2*. I proceeded in this way until 2001, classifying observations using only data in the years preceding to train the models. In order to reduce the computational strain of this task, I first downsampled the dataset to roughly 100 observations without conflict (B Attacked=0) for every observation with conflict (B Attacked = 1), having first estimated that this proportion best balanced computational and predictive performance. Though it takes a few decades for the model to stabilize (as there are few observations prior to 1900), the result is generally the same. The world

wars create a degradation in test performance while the model continues to otherwise steadily improve over time. The figure displays performance using first the log-likelihood, then the area under the receiver operating curve.

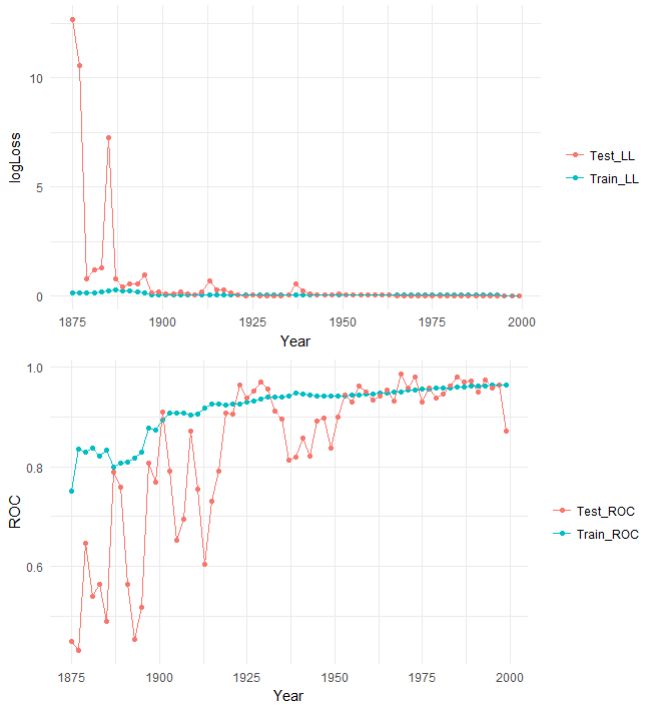

Figure B.2: Training and test results using log likelihood and area under the receiver operating curve for incremental runs of a logistic regression

B.4 Examining Additional Year Comparisons

In the paper I show a comparison between the ensemble and logit during the late 1930s. For a comparison to other times, I present a few other year segments here. In the 1970s, for instance, the models perform somewhat comparably, though the ensemble typically is better at picking out true positives even when conflicts are few. The biggest gains for the ensemble are years in which there are many conflicts, as evidenced by looking at the years 1991 through 1993. Though this comes at the cost of more false positives, the ensemble is able to successfully identify many conflicts in these

years. Logistic regression, by comparison, fails to pick up the uptick in conflict which happened in these years.

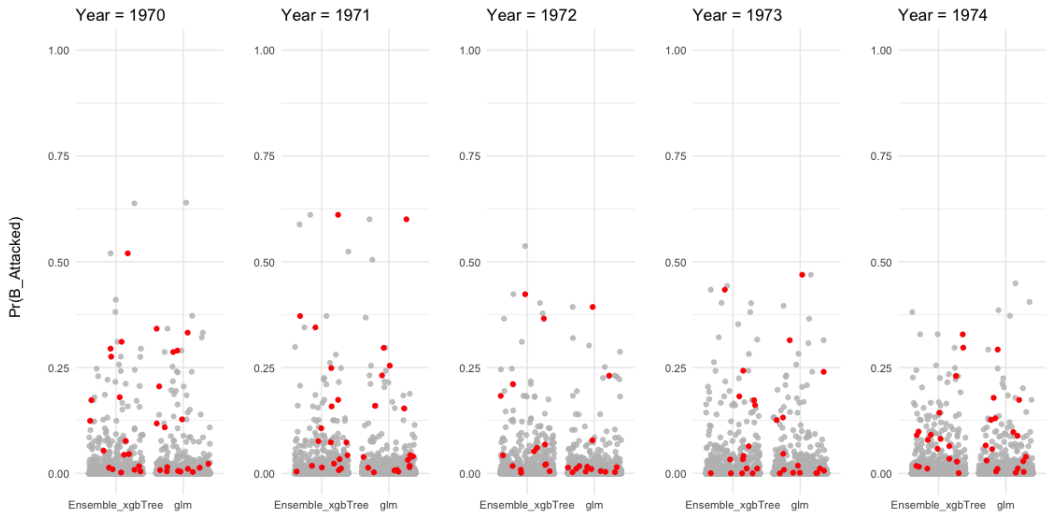

Figure B.3: Comparing predictions for the best performing ensemble and logit for the years 1970-1974. Observations in which B attacked are highlighted in red; jitter used to help distinguish between observations.

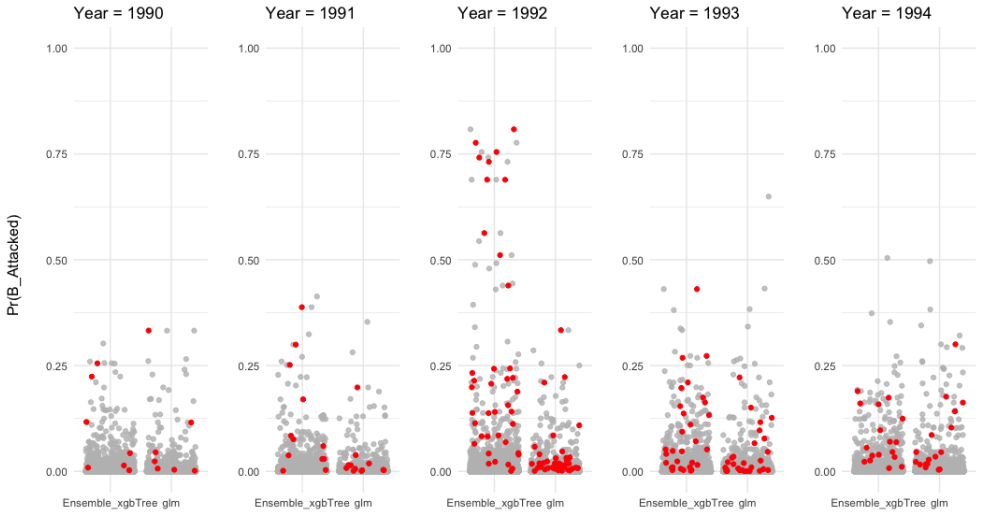

Figure B.4: Comparing predictions for the best performing ensemble and logit for the years 1990-1994. Observations in which B attacked are highlighted in red; jitter used to help distinguish between observations.

B.5 Tuning Parameters

Tuning parameters for models discussed in the paper. I relied on the caret package (Kuhn, 2008) in R for estimation and cross validation. I present here the tuning parameters for each of the methods.

1. Averaged Neural Networks (Scholkopf et al., 1997)
    - Packages: avNNet
    - Tuning Parameters: size = 5, decay = 0.1
2. Logistic regression with elastic net regularization (Zou and Hastie, 2005)
    - Packages: glmnet
    - Tuning Parameters: alpha = 0.2, lambda = 0
3. Boosted logistic regression (Friedman et al., 2000)
    - Packages: LogitBoost
    - Tuning Parameters: niter = 41
4. Classification and regression trees (CART) (Breiman et al., 1984)
    - Packages: rpart
    - Tuning Parameters: cp=0.0012
5. MARs (Friedman, 1991)
    - Packages: earth
    - Tuning Parameters: nprune = 11, degree = 1
6. Ranger (Breiman, 2001) (Wright and Ziegler, 2015)
    - Packages: ranger
    - Tuning Parameters: mtry = 2, splitrule = gini, min.node.size = 1

7. Support vector machines with a radial kernel (Scholkopf et al., 1997)

    - Packages: kernlab
    - Tuning Parameters: sigma=0.011, C=4, weight = 1

APPENDIX C

FOREIGN THREATS AND REPRESSION

C.1 Correlation Heatplot

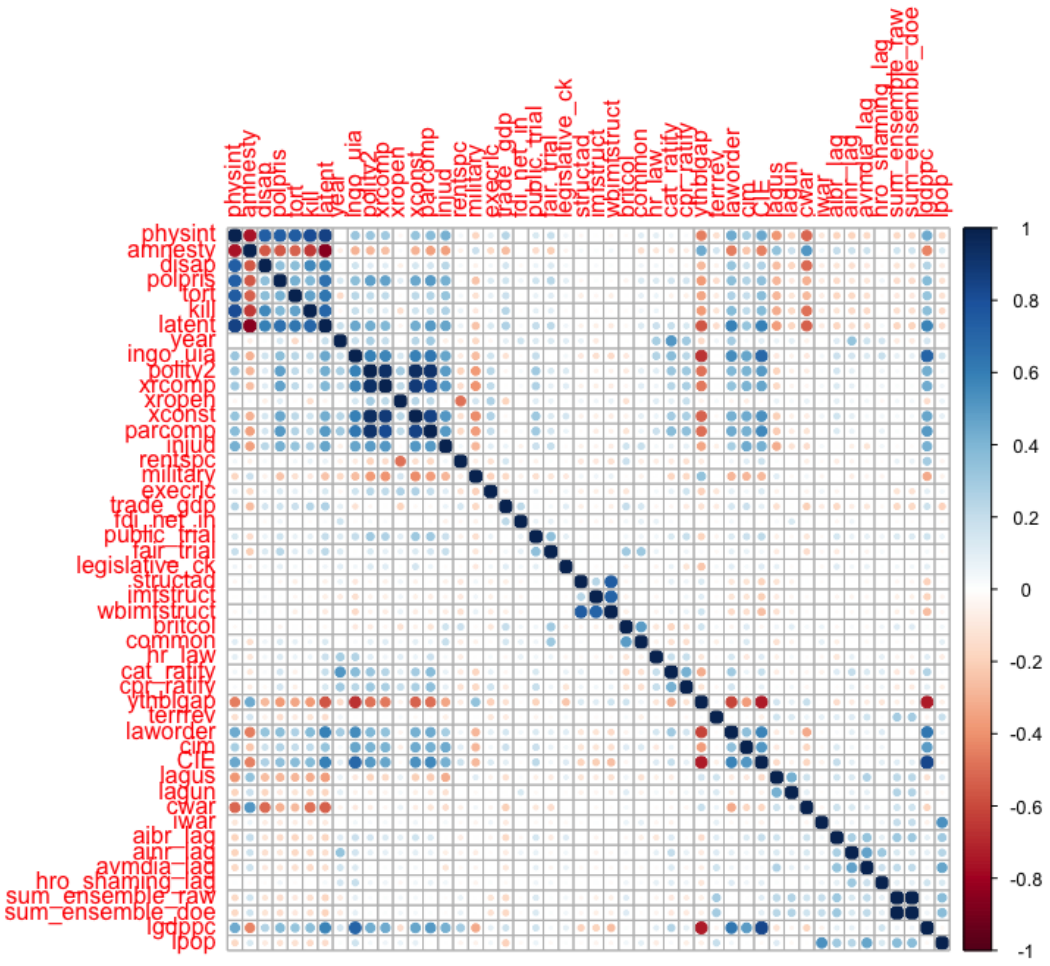

Figure C.1: Correlation plot for all predictors and outcomes.

C.2 Additional OLS Variable Permutations

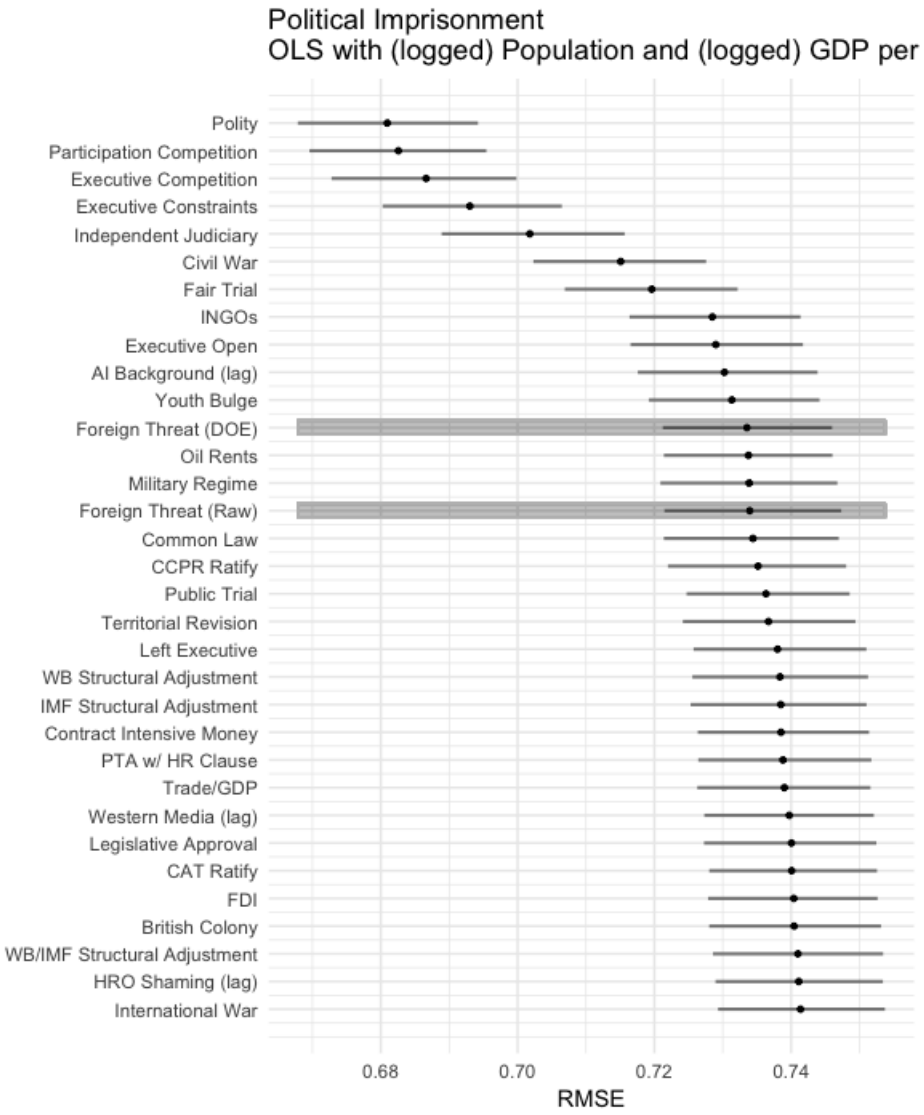

Figure C.2: Cross validated variable importance scores for CIRI's political prisoners measure. Variable importance estimated using linear regression models with 10 fold cross validation, iteratively adding each variable to a baseline specification including logged GDP per capita and logged population. Results bootstrapped 1000 times.

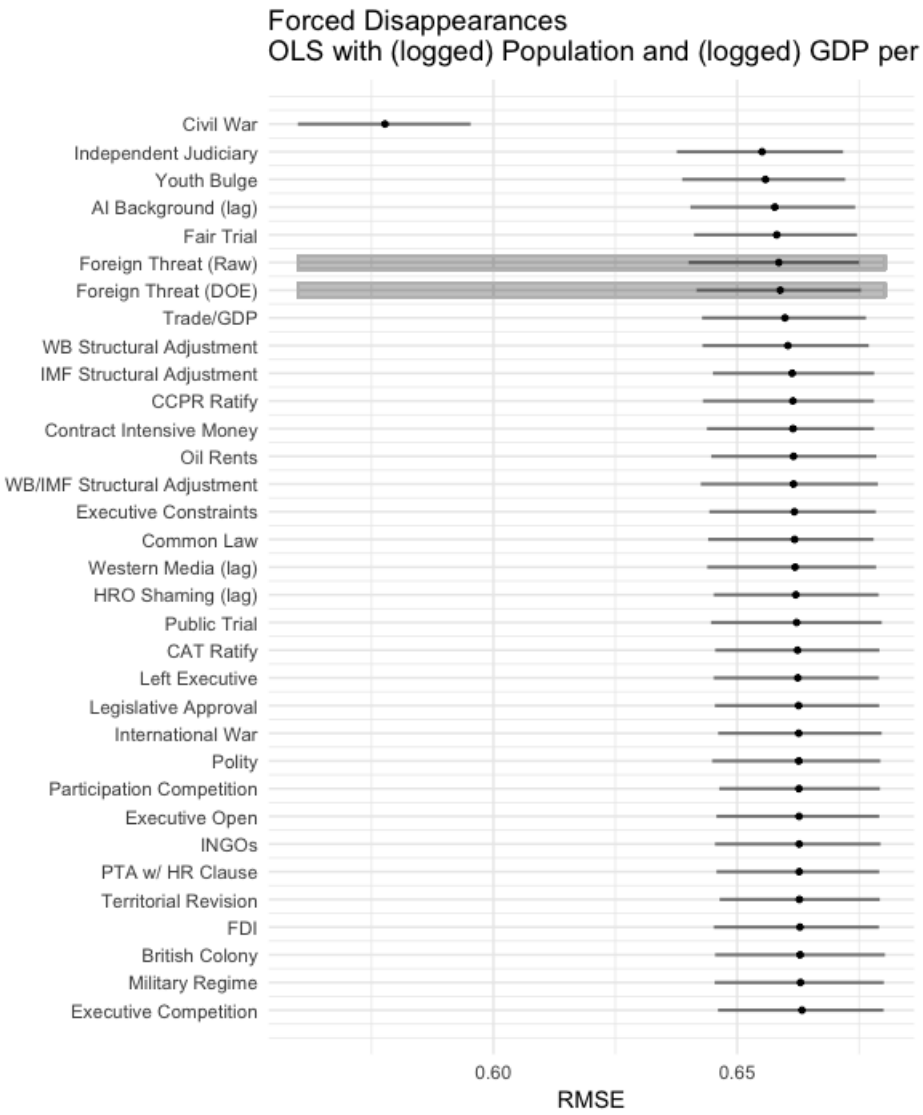

Figure C.3: Cross validated variable importance scores for CIRI's political disappearances measure. Variable importance estimated using linear regression models with 10 fold cross validation, iteratively adding each variable to a baseline specification including logged GDP per capita and logged population. Results bootstrapped 1000 times.

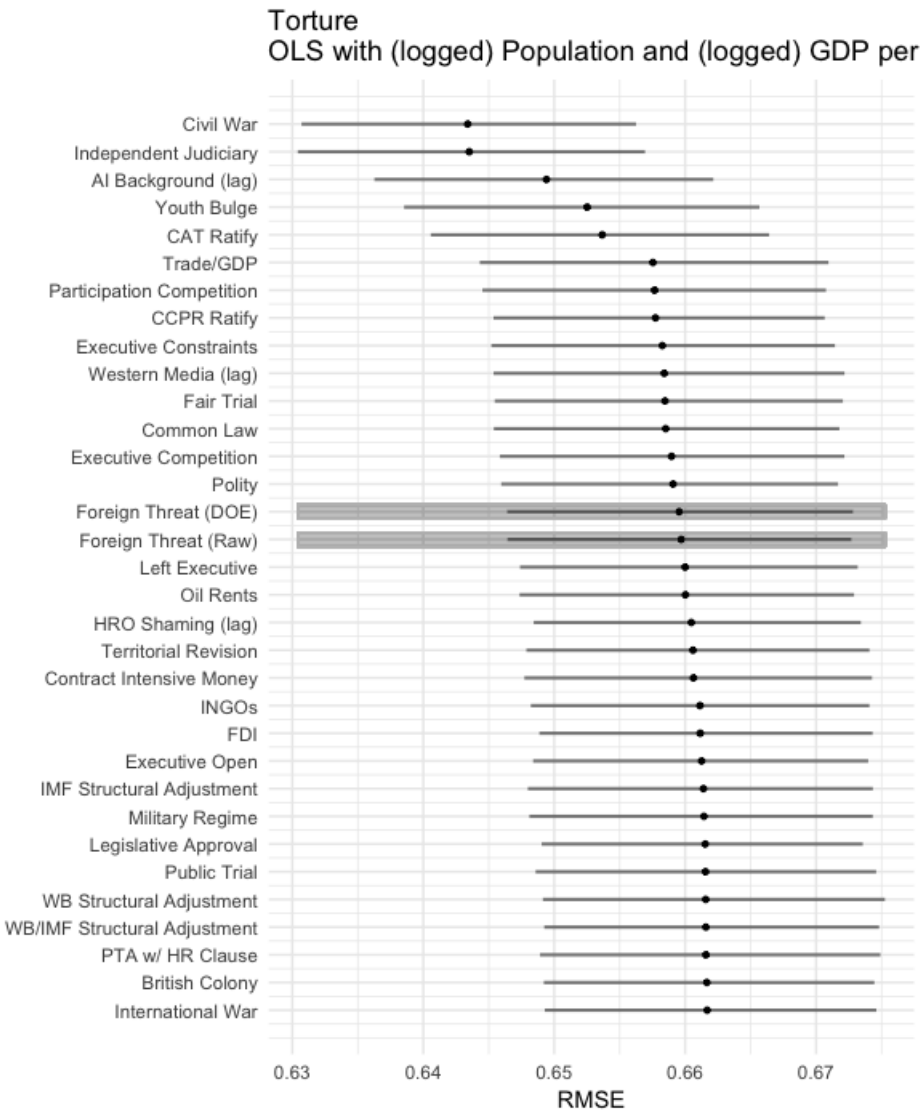

Figure C.4: Cross validated variable importance scores for CIRI's torture measure. Variable importance estimated using linear regression models with 10 fold cross validation, iteratively adding each variable to a baseline specification including logged GDP per capita and logged population. Results bootstrapped 1000 times.

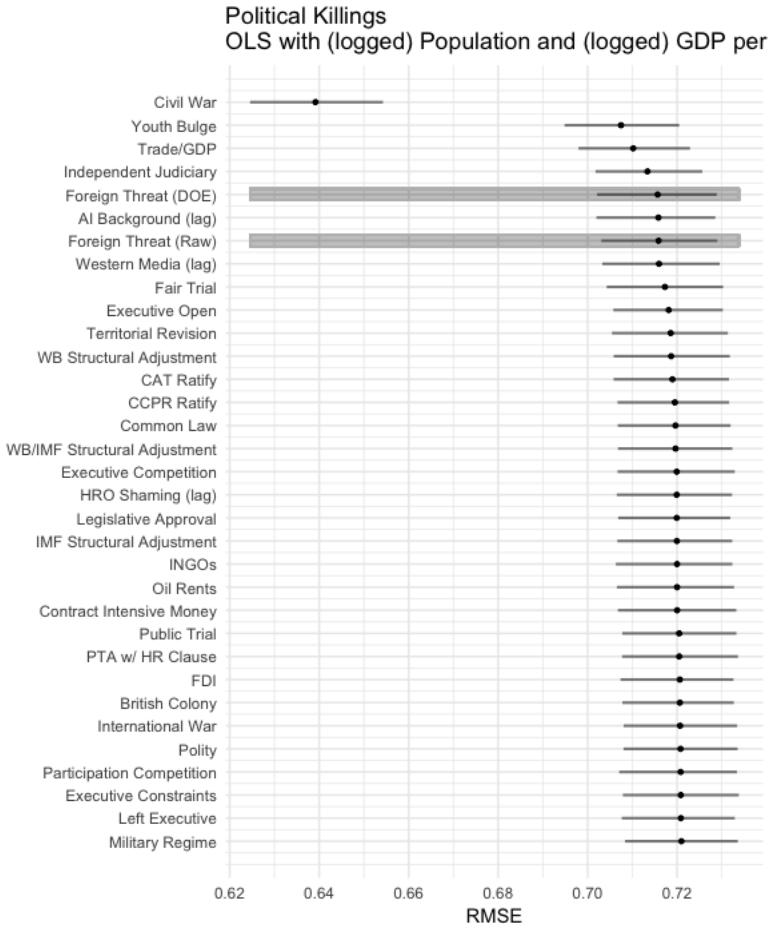

Figure C.5: Cross validated variable importance scores for CIRI's political killings measure. Variable importance estimated using linear regression models with 10 fold cross validation, iteratively adding each variable to a baseline specification including logged GDP per capita and logged population. Results bootstrapped 1000 times.

C.3 Additional Conditional Random Forest and Ranger Variable Permutations

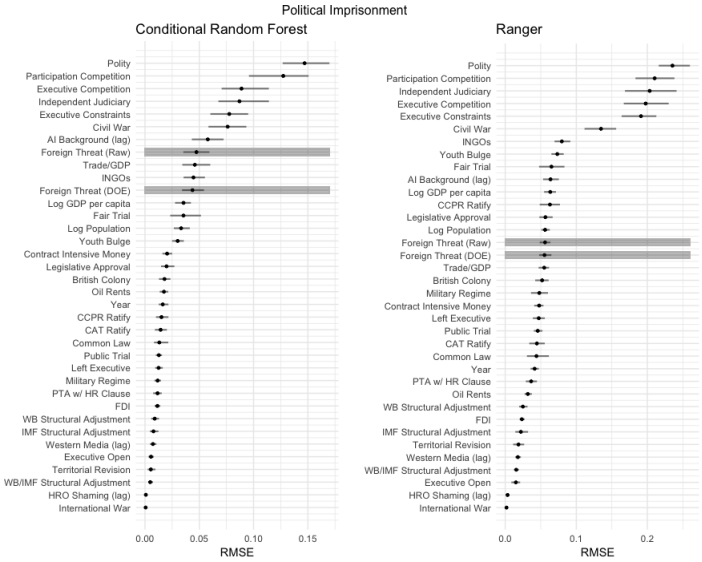

Figure C.6: Variable permutation scores from a conditional random forest for CIRI's political prisoners measure. Forest grown with 500 trees and 10 randomly selected predictors.

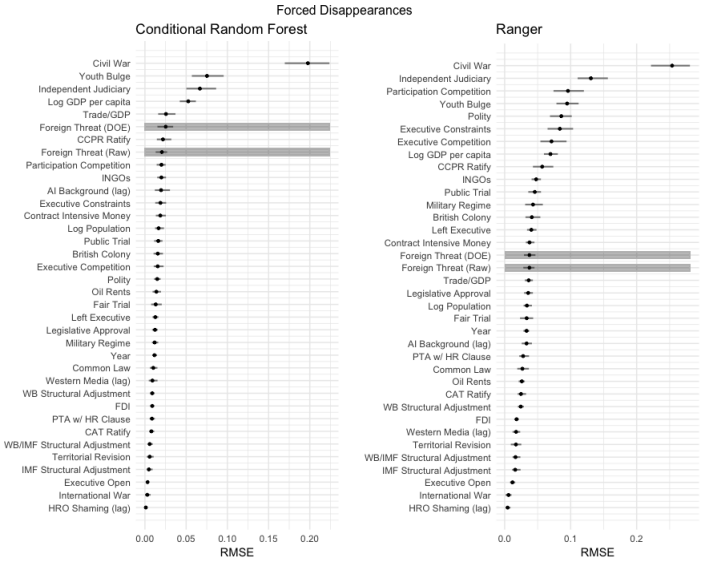

Figure C.7: Variable permutation scores from a conditional random forest for CIRI's political disappearances measure. Forest grown with 500 trees and 10 randomly selected predictors.

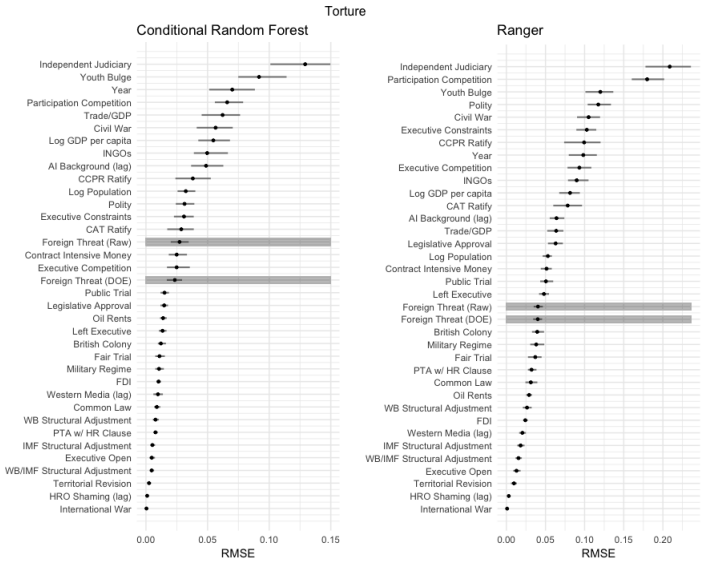

Figure C.8: Variable permutation scores from a conditional random forest for CIRI's torture measure. Forest grown with 500 trees and 10 randomly selected predictors.

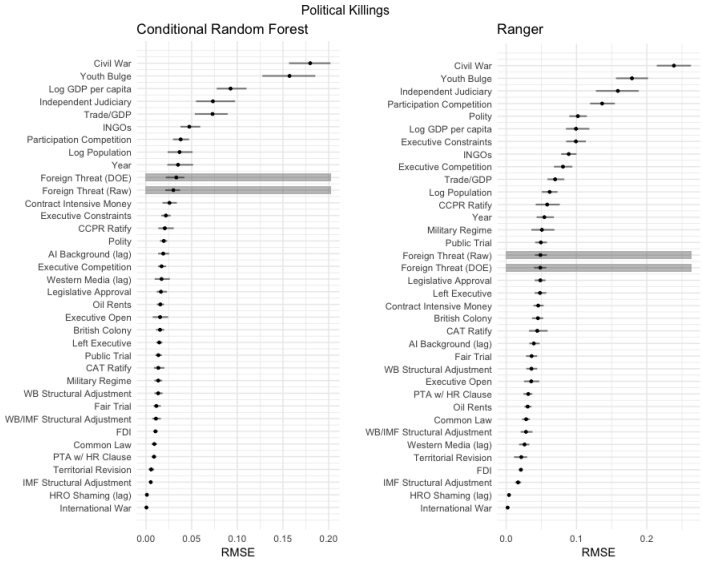

Figure C.9: Variable permutation scores from a conditional random forest for CIRI's Physical Integrity Index. Forest grown with 500 trees and 10 randomly selected predictors.

C.4 Lasso, Ridge, and OLS Coefficients

Table C.1: Coefficients from lasso, ridge, and ordinary least squares regression models of Fariss' dynamic latent score using all predictors without a lagged dependent variable.

Dynamic Latent Score			
	OLS	Lasso	Ridge
(Intercept)	0.001	0.001	0.001
Year	−0.048	−0.028	−0.021
INGOs	−0.134	−0.033	0.035
Polity	−0.391	−−	0.033
Executive Competition	0.111	−−	0.012
Executive Open	−0.036	−0.026	−0.052
Executive Constraints	0.246	0.054	0.080
Participation Competition	0.368	0.227	0.156
Independent Judiciary	0.144	0.142	0.136
Oil Rents	−0.081	−0.038	−0.017
Military Regime	0.035	0.004	0.010
Left Executive	0.007	−−	0.015
Trade/GDP	0.068	0.065	0.091
FDI	−0.010	−0.004	−0.017
Public Trial	0.031	0.025	0.033
Fair Trial	0.079	0.067	0.076
Legislative Approval	−0.012	−−	−0.005
WB/IMF Structural Adjustment	0.004	−−	0.005
IMF Structural Adjustment	−0.009	−−	0.001
WB Structural Adjustment	0.036	0.013	0.008
British Colony	−0.076	−0.045	−0.050
Common Law	0.095	0.073	0.071
PTA w/ HR Clause	0.106	0.080	0.080
CAT Ratify	−0.044	−0.037	−0.034
CCPR Ratify	0.010	−−	0.005
Youth Bulge	−0.220	−0.221	−0.219
Territorial Revision	−0.035	−0.038	−0.043
Contract Intensive Money	−0.036	−0.011	0.002
Civil War	−0.387	−0.386	−0.336
International War	−0.038	−0.031	−0.037
AI Background (lag)	−0.179	−0.182	−0.161
Western Media (lag)	−0.003	−0.005	−0.024
HRO Shaming (lag)	−0.022	−0.016	−0.018
Foreign Threat (Raw)	−0.095	−0.092	−0.085
Log GDP per capita	0.467	0.379	0.274
Log Population	−0.030	−0.038	−0.044

Table C.2: Coefficients from lasso, ridge, and ordinary least squares regression models of the Political Terror Scale using all predictors without a lagged dependent variable.

Political Terror Scale			
	OLS	Lasso	Ridge
(Intercept)	2.837	2.837	2.837
Year	0.032	0.009	0.020
INGOs	0.151	−−	−0.004
Polity	0.385	−−	0.001
Executive Competition	−0.086	−−	0.018
Executive Open	−0.008	−−	0.011
Executive Constraints	−0.131	−−	−0.021
Participation Competition	−0.302	−0.080	−0.094
Independent Judiciary	−0.100	−0.079	−0.086
Oil Rents	0.066	0.005	0.014
Military Regime	−0.046	−0.006	−0.019
Left Executive	−0.095	−0.070	−0.079
Trade/GDP	−0.088	−0.082	−0.103
FDI	0.032	0.011	0.030
Public Trial	−0.001	−−	−0.009
Fair Trial	−0.068	−0.050	−0.070
Legislative Approval	0.010	−−	0.004
WB/IMF Structural Adjustment	−0.022	−−	−0.020
IMF Structural Adjustment	−0.015	−−	−0.018
WB Structural Adjustment	−0.041	−0.033	−0.022
British Colony	0.053	0.012	0.034
Common Law	−0.123	−0.086	−0.081
PTA w/ HR Clause	−0.039	−0.002	−0.024
CAT Ratify	0.022	0.004	0.019
CCPR Ratify	0.026	0.010	0.026
Youth Bulge	0.222	0.207	0.185
Territorial Revision	0.032	0.034	0.039
Contract Intensive Money	0.064	0.028	0.026
Civil War	0.404	0.404	0.325
International War	0.056	0.041	0.039
AI Background (lag)	0.132	0.127	0.114
Western Media (lag)	0.065	0.065	0.063
HRO Shaming (lag)	0.009	−−	0.006
Foreign Threat (Raw)	0.026	0.012	0.034
Log GDP per capita	−0.318	−0.178	−0.178
Log Population	−0.006	0.016	0.021

Table C.3: Coefficients from lasso, ridge, and ordinary least squares regression models of CIRI's Physical Integrity Index using all predictors without a lagged dependent variable.

Physical Integrity Index			
	OLS	Lasso	Ridge
(Intercept)	4.786	4.786	4.786
Year	−0.297	−0.229	−0.228
INGOs	−0.361	−0.087	−0.064
Polity	−0.683	−−	0.050
Executive Competition	0.219	−−	0.026
Executive Open	−0.091	−0.056	−0.106
Executive Constraints	0.481	0.094	0.159
Participation Competition	0.563	0.304	0.242
Independent Judiciary	0.402	0.400	0.357
Oil Rents	−0.170	−0.045	−0.066
Military Regime	0.090	0.011	0.041
Left Executive	−0.059	−0.023	−0.034
Trade/GDP	0.140	0.134	0.177
FDI	−0.029	−0.010	−0.045
Public Trial	−0.059	−0.003	−0.024
Fair Trial	0.217	0.164	0.198
Legislative Approval	−0.022	−−	−0.005
WB/IMF Structural Adjustment	−0.035	−−	0.002
IMF Structural Adjustment	−0.026	−−	0.027
WB Structural Adjustment	0.177	0.083	0.077
British Colony	−0.156	−0.082	−0.122
Common Law	0.125	0.079	0.098
PTA w/ HR Clause	0.185	0.111	0.128
CAT Ratify	−0.120	−0.113	−0.122
CCPR Ratify	−0.138	−0.123	−0.121
Youth Bulge	−0.491	−0.469	−0.423
Territorial Revision	−0.079	−0.075	−0.090
Contract Intensive Money	−0.079	−0.006	−0.023
Civil War	−0.803	−0.797	−0.707
International War	−0.116	−0.092	−0.113
AI Background (lag)	−0.255	−0.257	−0.247
Western Media (lag)	−0.011	−0.013	−0.050
HRO Shaming (lag)	0.013	−−	0.001
Foreign Threat (Raw)	−0.104	−0.087	−0.102
Log GDP per capita	0.587	0.348	0.363
Log Population	−0.130	−0.161	−0.131

BIBLIOGRAPHY

Abouharb, M Rodwan and David Cingranelli. 2007. *Human rights and structural adjustment*. Cambridge University Press.

Adcock, Robert. 2001. "Measurement validity: A shared standard for qualitative and quantitative research." *American political science review* 95(3):529–546.

Agüero, Felipe. 1995. *Soldiers, civilians, and democracy: Post-Franco Spain in comparative perspective*. Johns Hopkins University Press.

Allison, Graham. 1999. *Essence of decision: explaining the Cuban missile crisis*. Boston: Little, Brown and Company.

Apodaca, Clair. 2001. "Global economic patterns and personal integrity rights after the Cold War." *International Studies Quarterly* pp. 587–602.

Arbatli, Cemal Eren and Ekim Arbatli. 2016. "External threats and political survival: Can dispute involvement deter coup attempts?" *Conflict Management and Peace Science* 33(2):115–152.

Barrilleaux, Charles and Carlisle Rainey. 2014. "The politics of need: Examining governors' decisions to oppose the "Obamacare" Medicaid expansion." *State Politics & Policy Quarterly* 14(4):437–460.

Beger, Andreas. working. "Explaining and Predicting Interstate Battle Deaths.".

Beger, Andreas, Cassy L Dorff and Michael D Ward. 2016. "Irregular leadership changes in 2014: Forecasts using ensemble, split-population duration models." *International Journal of Forecasting* 32(1):98–111.

Bennett, D Scott. 1997. "Measuring rivalry termination, 1816-1992." *Journal of Conflict Resolution* 41(2):227–254.

Blainey, Geoffrey. 1988. *Causes of War*. Simon and Schuster.

Braumoeller, Bear. 2013. "Is War Disappearing?".

Breiman, Leo. 1996. "Bagging predictors." *Machine learning* 24(2):123–140.

Breiman, Leo. 2001. "Random forests." *Machine learning* 45(1):5–32.

Breiman, Leo, Jerome Friedman, Charles J Stone and Richard A Olshen. 1984. *Classification and regression trees*. CRC press.

Breiman, Leo et al. 2001. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." *Statistical Science* 16(3):199–231.

Bueno de Mesquita, Bruce. 1980. "An expected utility theory of international conflict." *American Political Science Review* 74(04):917–931.

Bueno de Mesquita, Bruce. 1983. "The costs of war: a rational expectations approach." *American Political Science Review* 77(02):347–357.

Bueno de Mesquita, Bruce, Feryal Marie Cherif, George W Downs and Alastair Smith. 2005. "Thinking inside the box: A closer look at democracy and human rights." *International Studies Quarterly* 49(3):439–458.

Bueno de Mesquita, Bruce, James D Morrow, Randolph M Siverson and Alastair Smith. 1999. "An institutional explanation of the democratic peace." *American Political Science Review* pp. 791–807.

Bueno de Mesquita, Bruce and Randolph M Siverson. 1995. "War and the survival of political leaders: A comparative study of regime types and political accountability." *American Political Science Review* 89(04):841–855.

Burbidge, John B, Lonnie Magee and A Leslie Robb. 1988. "Alternative transformations to handle extreme values of the dependent variable." *Journal of the American Statistical Association* 83(401):123–127.

Carey, Sabine C, Michael P Colaresi and Neil J Mitchell. 2015. "Governments, informal links to militias, and accountability." *Journal of Conflict Resolution* 59(5):850–876.

Carroll, Robert J and Brenton Kenkel. 2016. "Prediction, Proxies, and Power.".

Caruana, Rich, Art Munson and Alexandru Niculescu-Mizil. 2006. Getting the most out of ensemble selection. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE pp. 828–833.

Cingranelli, David L and David L Richards. 2010. "The Cingranelli and Richards (CIRI) human rights data project." *Human Rights Quarterly* 32(2):401–424.

Clague, Christopher, Philip Keefer, Stephen Knack and Mancur Olson. 1999. "Contract-intensive money: contract enforcement, property rights, and economic performance." *Journal of economic growth* 4(2):185–211.

Clausewitz, Carl von. 1976. "On War, trans. Michael Howard and Peter Paret.".

Colaresi, Michael and Sabine C Carey. 2008. "To Kill or to Protect Security Forces, Domestic Institutions, and Genocide." *Journal of Conflict Resolution* 52(1):39–67.

Colaresi, Michael and William R Thompson. 2002. "Strategic rivalries, protracted conflict, and crisis escalation." *Journal of Peace Research* 39(3):263–287.

Conrad, Courtenay R and Emily Hencken Ritter. 2013. "Treaties, tenure, and torture: the conflicting domestic effects of international law." *The Journal of Politics* 75(02):397–409.

Copeland, Dale C. 2000. *The origins of major war*. Cornell University Press.

Cortes, Corinna and Vladimir Vapnik. 1995. "Support-vector networks." *Machine learning* 20(3):273–297.

Cover, Thomas and Peter Hart. 1967. "Nearest neighbor pattern classification." *IEEE transactions on information theory* 13(1):21–27.

Crescenzi, Mark JC. 2007. "Reputation and interstate conflict." *American Journal of Political Science* 51(2):382–396.

Davenport, Christian. 1995. "Multi-dimensional threat perception and state repression: An inquiry into why states apply negative sanctions." *American Journal of Political Science* pp. 683–713.

Davenport, Christian. 2007. "State repression and political order." *Annu. Rev. Polit. Sci.* 10:1–23.

Davenport, Christian A. 1996. ""Constitutional Promises" and Repressive Reality: A Cross–National Time–Series Investigation of Why Political and Civil Liberties are Suppressed." *The Journal of Politics* 58(03):627–654.

Davenport, Christian and David A Armstrong. 2004. "Democracy and the violation of human rights: A statistical analysis from 1976 to 1996." *American Journal of Political Science* 48(3):538–554.

Davies, Graeme AM. 2016. "Policy selection in the face of political instability: Do states divert, repress, or make concessions?" *Journal of Conflict Resolution* 60(1):118–142.

De Marchi, Scott, Christopher Gelpi and Jeffrey D Grynaviski. 2004. "Untangling neural nets." *American Political Science Review* 98(2):371–378.

Desch, Michael C. 1996. "War and strong states, peace and weak states?" *International Organization* 50(2):237–268.

Desch, Michael C. 1998. "Soldiers, states, and structures: The end of the Cold War and weakening US civilian control." *Armed Forces & Society* 24(3):389–405.

Downes, Alexander B. 2009. "How smart and tough are democracies? Reassessing theories of democratic victory in war." *International Security* 33(4):9–51.

Efron, Bradley and Robert Tibshirani. 1997. "Improvements on cross-validation: the 632+ bootstrap method." *Journal of the American Statistical Association* 92(438):548–560.

Elith, Jane, John R Leathwick and Trevor Hastie. 2008. "A working guide to boosted regression trees." *Journal of Animal Ecology* 77(4):802–813.

Enterline, Andrew J and Kristian S Gleditsch. 2000. "Threats, opportunity, and force: Repression and diversion of domestic pressure, 1948–1982." *International Interactions* 26(1):21–53.

Fariss, Christopher J. 2014. "Respect for human rights has improved over time: Modeling the changing standard of accountability." *American Political Science Review* 108(02):297–318.

Fearon, James D. 1994. "Domestic political audiences and the escalation of international disputes." *American political science review* 88(3):577–592.

Fearon, James D. 1995. "Rationalist explanations for war." *International organization* 49(03):379–414.

Fearon, James D. 1997. "Signaling foreign policy interests: Tying hands versus sinking costs." *Journal of Conflict Resolution* 41(1):68–90.

Feaver, Peter D. 1999. "Civil-military relations." *Annual Review of Political Science* 2(1):211–241.

Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro and Dinani Amorim. 2014. "Do we need hundreds of classifiers to solve real world classification problems." *J. Mach. Learn. Res* 15(1):3133–3181.

Fey, Mark and Kristopher W Ramsay. 2007. "Mutual optimism and war." *American Journal of Political Science* 51(4):738–754.

Filson, Darren and Suzanne Werner. 2004. "Bargaining and fighting: The impact of regime type on war onset, duration, and outcomes." *American Journal of Political Science* 48(2):296–313.

Filson, Darren and Suzanne Werner. 2007. "Sensitivity to costs of fighting versus sensitivity to losing the conflict: Implications for war onset, duration, and outcomes." *Journal of Conflict Resolution* 51(5):691–714.

Frankopan, Peter. 2015. *The silk roads: A new history of the world*. Bloomsbury Publishing.

Friedman, Jerome H. 1991. "Multivariate adaptive regression splines." *The annals of statistics* pp. 1–67.

Friedman, Jerome, Trevor Hastie and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1 Springer series in statistics Springer, Berlin.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani et al. 2000. "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)." *The annals of statistics* 28(2):337–407.

Gartzke, Erik and Quan Li. 2003. "War, peace, and the invisible hand: Positive political externalities of economic globalization." *International Studies Quarterly* 47(4):561–586.

Ghalanos, Alexios and Stefan Theussl. 2015. *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*. R package version 1.16.

Ghatak, Sambuddha, Aaron Gold and Brandon C Prins. 2017. "External threat and the limits of democratic pacifism." *Conflict management and peace science* 34(2):141–159.

Gibler, Douglas M. 2007. "Bordering on peace: Democracy, territorial issues, and conflict." *International Studies Quarterly* 51(3):509–532.

Gibler, Douglas M and Jamil A Sewell. 2006. "External threat and democracy: the role of NATO revisited." *Journal of Peace Research* 43(4):413–431.

Gibler, Douglas M and Jaroslav Tir. 2010. "Settled borders and regime type: Democratic transitions as consequences of peaceful territorial transfers." *American Journal of Political Science* 54(4):951–968.

Gibler, Douglas M and Scott Wolford. 2006. "Alliances, then democracy: An examination of the relationship between regime type and alliance formation." *Journal of Conflict Resolution* 50(1):129–153.

Gibler, Douglas M, Steven V Miller and Erin K Little. 2016. "An analysis of the militarized interstate dispute (MID) dataset, 1816–2001." *International Studies Quarterly* 60(4):719–730.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg and Håvard Strand. 2002. "Armed conflict 1946-2001: A new dataset." *Journal of peace research* 39(5):615–637.

Goertz, Gary and Paul F Diehl. 1995. "The initiation and termination of enduring rivalries: The impact of political shocks." *American Journal of Political Science* pp. 30–52.

Goldstone, Jack A, Robert H Bates, Ted Robert Gurr, Michael Lustik, Monty G Marshall, Jay Ulfelder and Mark Woodward. 2005. A global forecasting model of political instability. In *Annual Meeting of the American Political Science*.

Grömping, Ulrike. 2009. "Variable importance assessment in regression: linear regression versus random forest." *The American Statistician* 63(4):308–319.

Guyon, Isabelle and André Elisseeff. 2003. "An introduction to variable and feature selection." *Journal of machine learning research* 3(Mar):1157–1182.

Hafner-Burton, Emilie M. 2005*a*. "Right or robust? The sensitive nature of repression to globalization." *Journal of Peace Research* 42(6):679–698.

Hafner-Burton, Emilie M. 2005*b*. "Trading human rights: How preferential trade agreements influence government repression." *International Organization* 59(3):593–629.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. Unsupervised learning. In *The elements of statistical learning*. Springer pp. 485–585.

Hensel, Paul R. 2009. "ICOW colonial history data set, version 0.4." *University of North Texas. http://www. paulhensel. org/icowcol. html* .

Herbst, Jeffrey. 2000. "Economic incentives, natural resources and conflict in Africa." *Journal of African Economies* 9(3):270–294.

Hill, Daniel W and Zachary M Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108(03):661–687.

Hindman, Matthew. 2015. "Building better models: Prediction, replication, and machine learning in the social sciences." *The ANNALS of the American Academy of Political and Social Science* 659(1):48–62.

Hintze, Otto, Felix Gilbert and Robert M Berdahl. 1975. "The historical essays of Otto Hintze.".

Hutchison, Marc L. 2011*a*. "Territorial threat and the decline of political trust in Africa: A multilevel analysis." *Polity* 43(4):432–461.

Hutchison, Marc L. 2011*b*. "Territorial threat, mobilization, and political participation in Africa." *Conflict Management and Peace Science* 28(3):183–208.

Jackson, Matthew O and Massimo Morelli. 2011. "The reasons for wars: an updated survey." *The handbook on the political economy of war* 34.

Jenke, Libby and Christopher Gelpi. 2017. "Theme and variations: Historical contingencies in the causal model of interstate conflict." *Journal of Conflict Resolution* 61(10):2262–2284.

Jervis, Robert. 1978. "Cooperation under the security dilemma." *World politics* 30(2):167–214.

Johnson, Dominic DP and Dominic Tierney. 2011. "The Rubicon theory of war: how the path to conflict reaches the point of no return." *International Security* 36(1):7–40.

Kant, Immanuel and Hans Siegert Reiss. 1970. *Kant's Political Writings: Transl. by HB Nisbet.* Cambridge University Press.

Kaufmann, Chaim. 2004. "Threat inflation and the failure of the marketplace of ideas: The selling of the Iraq war." *International Security* 29(1):5–48.

Keith, Linda Camp, C Neal Tate and Steven C Poe. 2009. "Is the law a mere parchment barrier to human rights abuse?" *The Journal of Politics* 71(02):644–660.

King, Gary, Robert O Keohane and Sidney Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research.* Princeton university press.

Kuhn, Max. 2008. "Caret package." *Journal of Statistical Software* 28(5):1–26.

Kuhn, Max and Kjell Johnson. 2013. *Applied predictive modeling.* Vol. 26 Springer.

Kuhn, Max, Steve Weston, Chris Keefer and Nathan Coulter. 2012. "Cubist Models For Regression." *R package Vignette R package version 0.0* 18.

Kursa, Miron B, Witold R Rudnicki et al. 2010. "Feature selection with the Boruta package." *J Stat Softw* 36(11):1–13.

Lacina, Bethany and Nils Petter Gleditsch. 2005. "Monitoring trends in global combat: A new dataset of battle deaths." *European Journal of Population/Revue européenne de Démographie* 21(2-3):145–166.

Lamborn, Alan C. 1985. "Risk and foreign policy choice." *International Studies Quarterly* 29(4):385–410.

Levy, Jack S. 1983. "Misperception and the causes of war: Theoretical linkages and analytical problems." *World Politics* 36(01):76–99.

Lu, Lingyu and Cameron G Thies. 2013. "War, rivalry, and state building in the Middle East." *Political Research Quarterly* 66(2):239–253.

Maoz, Zeev. 1983. "Resolve, capabilities, and the outcomes of interstate disputes, 1816-1976." *Journal of Conflict Resolution* 27(2):195–229.

McMahon, R Blake and Branislav L Slantchev. 2015. "The guardianship dilemma: Regime security through and from the armed forces." *American Political Science Review* 109(02):297–313.

Mearsheimer, John J. 1990. "Back to the future: Instability in Europe after the Cold War." *International security* 15(1):5–56.

Milner, Helen V and Bumba Mukherjee. 2009. "Democratization and economic globalization." *Annual Review of Political Science* 12:163–181.

Mitchell, Sara McLaughlin, Jonathan J Ring and Mary K Spellman. 2013. "Domestic legal traditions and states' human rights practices." *Journal of Peace Research* 50(2):189–202.

Montgomery, Jacob M, Florian M Hollenbach and Michael D Ward. 2015. "Calibrating ensemble forecasting models with sparse data in the social sciences." *International Journal of Forecasting* 31(3):930–942.

Moore, Will H. 1995. "Action-Reaction or Rational Expectations? Reciprocity and the Domestic-International Conflict Nexus during the "Rhodesia Problem"." *Journal of Conflict Resolution* 39(1):129–167.

Muchlinski, David, David Siroky, Jingrui He and Matthew Kocher. 2016. "Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data." *Political Analysis* 24(1):87–103.

Mullainathan, Sendhil and Jann Spiess. 2017. "Machine learning: an applied econometric approach." *Journal of Economic Perspectives* 31(2):87–106.

Murdie, Amanda M and David R Davis. 2012. "Shaming and blaming: Using events data to assess the impact of human rights INGOs." *International Studies Quarterly* 56(1):1–16.

Nordhaus, William, John R O'Neal and Bruce Russett. 2012. "The effects of the international security environment on national military expenditures: A multicountry study." *International Organization* pp. 491–513.

Organski, Abramo FK. 1958. *World politics*. Knopf.

Pierskalla, Jan Henryk. 2010. "Protest, deterrence, and escalation: The strategic calculus of government repression." *Journal of Conflict Resolution* .

Pinker, Steven. 2011. *The better angels of our nature: Why violence has declined*. Vol. 75 Viking New York.

Poast, Paul. 2010. "(Mis) using dyadic data to analyze multilateral events." *Political Analysis* 18(4):403–425.

Poe, Steven C and C Neal Tate. 1994. "Repression of human rights to personal integrity in the 1980s: a global analysis." *American Political Science Review* 88(04):853–872.

Poe, Steven C, C Neal Tate and Linda Camp Keith. 1999. "Repression of the Human Right to Personal Integrity Revisited: A Global Cross-National Study Covering the Years 1976–1993." *International studies quarterly* 43(2):291–313.

Powell, Emilia Justyna and Jeffrey K Staton. 2009. "Domestic judicial institutions and human rights treaty violation." *International Studies Quarterly* 53(1):149–174.

Powell, Robert. 2006. "War as a commitment problem." *International organization* 60(01):169–203.

Putnam, Robert D. 1988. "Diplomacy and domestic politics: the logic of two-level games." *International organization* 42(03):427–460.

Reiter, Dan. 2009. *How wars end*. Princeton University Press.

Reiter, Dan and Allan C Stam. 2002. *Democracies at war*. Princeton University Press.

Ribeiro, Marco Tulio, Sameer Singh and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM pp. 1135–1144.

Ripley, Brian D. 1996. *Pattern recognition and neural networks*. Cambridge university press.

Ritter, Emily Hencken. 2014. "Policy Disputes, Political Survival, and the Onset and Severity of State Repression." *Journal of Conflict Resolution* 58(1):143–168.

Ron, James, Howard Ramos and Kathleen Rodgers. 2005. "Transnational information politics: NGO human rights reporting, 1986–2000." *International Studies Quarterly* 49(3):557–587.

Ross, Jeffrey Ian. 2006. *Political terrorism: an interdisciplinary approach*. Peter Lang.

Rudra, Nita. 2005. "Globalization and the Strengthening of Democracy in the Developing World." *American Journal of Political Science* 49(4):704–730.

Sarkees, Meredith Reid and Phil Schafer. 2000. "The correlates of war data on war: An update to 1997." *Conflict Management and Peace Science* 18(1):123–144.

Schapire, Robert E. 2003. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*. Springer pp. 149–171.

Schelling, Arms. 1966. "Influence." *George and Simons* pp. 2–3.

Scholkopf, Bernhard, Kah-Kay Sung, Christopher JC Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio and Vladimir Vapnik. 1997. "Comparing support vector machines with Gaussian kernels to radial basis function classifiers." *IEEE transactions on Signal Processing* 45(11):2758–2765.

Schultz, Kenneth A. 1999. "Do democratic institutions constrain or inform? Contrasting two institutional perspectives on democracy and war." *International Organization* 53(2):233–266.

Sechser, Todd S. 2010. "Goliath's Curse: Coercive Threats and Asymmetric Power." *International Organization* 64(4):627–660.

Smith, Alastair and Allan C Stam. 2004. "Bargaining and the Nature of War." *Journal of Conflict Resolution* 48(6):783–813.

Spilker, Gabriele and Tobias Böhmelt. 2013. "The impact of preferential trade agreements on governmental repression revisited." *The Review of International Organizations* 8(3):343–361.

Stiglitz, Joseph and Linda Bilmes. 2008. *The Three Trillion Dollar War: The True Cost of the Iraq War*. New York: WW Norton and Company Inc.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis and Torsten Hothorn. 2007. "Bias in random forest variable importance measures: Illustrations, sources and a solution." *BMC bioinformatics* 8(1):25.

Thompson, William R. 1995. "Principal rivalries." *Journal of Conflict Resolution* 39(2):195–223.

Thompson, William R. 1996. "Democracy and peace: putting the cart before the horse?" *International Organization* 50(1):141–174.

Thompson, William R. 2001. "Identifying rivals and rivalries in world politics." *International Studies Quarterly* 45(4):557–586.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Tilly, Charles and Gabriel Ardant. 1975. *The formation of national states in Western Europe*. Vol. 8 Princeton Univ Pr.

Tilly, Charles et al. 1992. *Coercion, capital, and European states, AD 990-1990*. Oxford Blackwell.

Van der Laan, Mark J, Eric C Polley and Alan E Hubbard. 2007. "Super learner." *Statistical applications in genetics and molecular biology* 6(1).

Varma, Sudhir and Richard Simon. 2006. "Bias in error estimation when using cross-validation for model selection." *BMC bioinformatics* 7(1):1.

Wagner, R Harrison. 2000. "Bargaining and war." *American Journal of Political Science* pp. 469–484.

Waltz, Kenneth. 1979. "Theory of international relations." *Reading: Addison-Wesley* pp. 635–650.

Ward, Michael D, Brian D Greenhill and Kristin M Bakke. 2010. "The Perils of Policy by P-value: Predicting Civil Conflicts." *Journal of Peace Research* 47(4):363–375.

Weeks, Jessica L. 2008. "Autocratic audience costs: Regime type and signaling resolve." *International Organization* 62(1):35–64.

Weeks, Jessica L. 2012. "Strongmen and straw men: Authoritarian regimes and the initiation of international conflict." *American Political Science Review* 106(2):326–347.

Wendt, Alexander. 1992. "Anarchy is what states make of it: the social construction of power politics." *International organization* 46(2):391–425.

Whalen, Sean and Gaurav Pandey. 2013. A comparative analysis of ensemble classifiers: case studies in genomics. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE pp. 807–816.

Wold, Herman. 1985. "Partial least squares." *Encyclopedia of statistical sciences* .

Wolpert, David H. 1992. "Stacked generalization." *Neural networks* 5(2):241–259.

Wood, Reed M and Mark Gibney. 2010. "The Political Terror Scale (PTS): A re-introduction and a comparison to CIRI." *Human Rights Quarterly* 32(2):367–400.

Wright, Marvin N and Andreas Ziegler. 2015. "Ranger: a fast implementation of random forests for high dimensional data in C++ and R." *arXiv preprint arXiv:1508.04409* .

Wu, Xindong, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip et al. 2007. "Top 10 algorithms in data mining." *Knowledge and information systems* 14(1):1–37.

Ye, Yinyu. 1987. Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming PhD thesis Department of ESS, Stanford University.

Young, Joseph K. 2012. "Repression, dissent, and the onset of civil war." *Political Research Quarterly* p. 1065912912452485.

Zou, Hui and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320.

BIOGRAPHICAL SKETCH

I received my Bachelor's degree in Political Science at Texas A&M University in May of 2013. I came to the FSU Department of Political Science in 2013 in pursuit of a Master's Degree and a Doctor of Philosophy in Political Science. I received my Master's Degree in 2015.

My substantive research interests focus on political violence, human rights, and international conflict. My methodological interests include but are not limited to applied predictive modeling, machine learning, and causal inference.