Predicting the Costs of War

Phil Henrickson

March 7, 2018

Abstract

The expected cost of war 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. In this paper, I develop a proxy for the expected costs of war by focusing on one aspect of war costs - battle deaths. I train an ensemble 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 resulting model improves performance over that of a null model by 26% and a linear model by 9%. I apply the ensemble model to all interstate dyads in the Correlates of War dataverse from 1816-2007 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.

1 Introduction

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 is 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 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 about the costs of war. At the 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 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 is the simply 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 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

[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.

[2]For example, see Stiglitz and Bilmes (2008) for a thorough discussion of the task of estimating the economic cost of the 2003 Iraq War.

to produce a measure of this important theoretical concept in international relations.[3]

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 run an ensemble of models 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: my ensemble model improves out of sample predictions of conflict battle deaths over that of a null model by 26% and over a standard practice model (OLS 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. Having established 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 the ensemble model, 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.[4]

In predicting battle deaths I 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. Additionally, I find that all of these predictors are conditional based 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.

[3]All data is available for replication at https://github.com/phenrickson/DiCE, or by email at phenrickson@fsu.edu. Many thanks to Rob Caroll, Mark Souva, Sean Ehrlich, Matt Haunstein, and Sydney Gann, as well as those present at the conference for Forecasting in the Social Sciences for National Security.

[4]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.

2 The Expected Costs of Conflict

2.1 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.[5] 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.[6]

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).[7] 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 observe how other states behave in similar situations and use this observations as a precedent, or prior, for the currentsituation"(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 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.[8]

[5]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 a war.

[6]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, 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.

[7]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).

[8]Here it is important to explicitly note that the task at hand is developing the best possible prediction for a

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.

2.2 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, 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" (Blainey, 1988). 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 (BDM 919)". Smith and Stam (2004) develop a model of war in which state's 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

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, 32-46) for a discussion of the Johnson administrations decision to begin the Vietnam War and Kaufmann (2004) for a discussion of the Bush administration in the 2003 Iraq War 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.

2.3 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 train 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.

3 Learning Cost Expectations

3.1 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 fit models of the battle deaths from these wars, and use cross validation to assess the out of sample predictive performance of these models. I then use this model to make predictions about the costs of hypothetical disputes between all dyads in the international system. The most useful measure for researchers would be dyadic, reflecting the ex ante costs of fighting for two states should they choose to fight. This approach is better theoretically motivated, as it reflects that the nature of international relations consists of interactions which produce outcomes for both states.

3.2 Average, Aggregate, or Strongest Opponent?

The problem with the dyadic approach is that the majority of interstate conflicts are multilateral. This presents an issue for modeling battle deaths using directed dyads. 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 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. In order to potentially gain from each potential way of setting up the data, I combine predictions from models run on each of these approaches in the ensemble model.[9]

3.3 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 amongst all available datasets, which not only offers the more observations for training the

[9] I briefly discuss the effects of alliances on war costs more directly in the appendix, though I aim to more directly incorporate alliances in a future paper.

[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.

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 does 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 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]

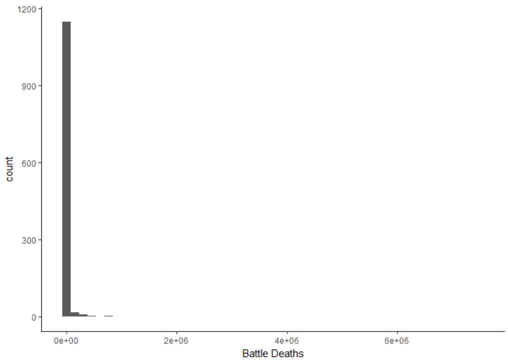
[11] Including conflicts as far back as 1816 might provide reason for caution, as Jenke and Gelpi (2017) find there to be significant temporal variation in international conflict: their results suggest that causes of international conflict are substantially different in the Cold War era than in all other 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

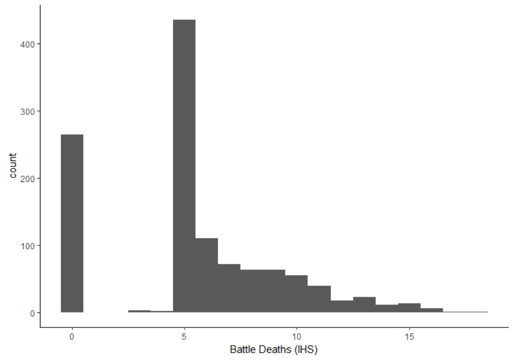
[13] Another possibility comes from the UCDP/PRIO dataset, which has compiled a dataset of all battle deaths in interstate war from 1945-2001. This dataset is notable because it is aggregated at the country-year level and distinguishes between the types of fatalities. This dataset has a total number 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 datset 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

[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] 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.



(a) Untransformed



(b) Inverse Hyperbolic Sine

Figure 1: Battle deaths for all fatal MIDs and interstate wars, before and after inverse hyerbolic sine transformation

3.4 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 COW, 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). I additionally can include each dyad's contiguity from COW. To capture time effects, I run the models with year included as a predictor.

3.5 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).

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

[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.

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]

3.6 The Candidate Models

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 algorithms 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 select the following models as candidates for predicting battle deaths:

1. An intercept-only model to serve as the baseline for predictive performance.
2. An ordinary least squares model a subset of predictors (CINC scores and Year)
3. An ordinary least squares model with all predictors
4. Partial least squares (PLS) Wold (1985), a supervised dimension reduction method.
5. An elastic net model Zou and Hastie (2005).
6. Multivariate adaptive regressive splines (MARS) (Friedman, 1991)
7. A K-nearest neighbors model (Cover and Hart, 1967)
8. Classification and regression trees (CART) Breiman et al. (1984).
9. Random forests (Breiman, 2001). Random forests have seen recent applications 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.
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 (Quinlan et al., 1992) M5 decision tree algorithm for the regression setting.
12. Support vector machines with a radial kernel (Scholkopf et al., 1997).
13. Neural networks. These results can be sensitive to the starting values selected in fitting the model , so I use averaged neural nets as described by Ripley (1996) and Hastie, Tibshirani and Friedman (2009).

For each of these models, I use cross-validation to estimate how well the models perform in predicting out of sample. For the models 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.

[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.

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 the model's 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 mode'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 candidate models using the same set of predictors, using each of the different dyadic data approaches in turn.

4 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 predictive power. Random forests in particular consistently produce the lowest test error in cross validation. Table 2 displays all of this information numerically, highlighting the three methods which show the best performance.[19]

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

Table 1: Out of sample performance estimated using nested 5-fold cross validation, tuned to minimize the root mean squared error. Results averaged across 10 imputations.

[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)

[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.

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

Table 2: Proportional reduction in loss (PRL) from each of the candidate models compared to the null model. Results averaged across 10 imputations.

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 deaths (5.53) for all conflicts. Table 3 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 26% reduction in test error. 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 an OLS model which has CINC scores and 'year' as predictors, and a 7-11% improvement over an OLS 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. Random forests and boosted trees in particular seem to be well suited to the task at hand, and that they improve over models in the linear family is likely because of their ability to easily detect nonlinearities and interactions present in the data generating process which would not be modeled unless specifically assumed by the applied researcher. Figure 2 illustrates this graphically.

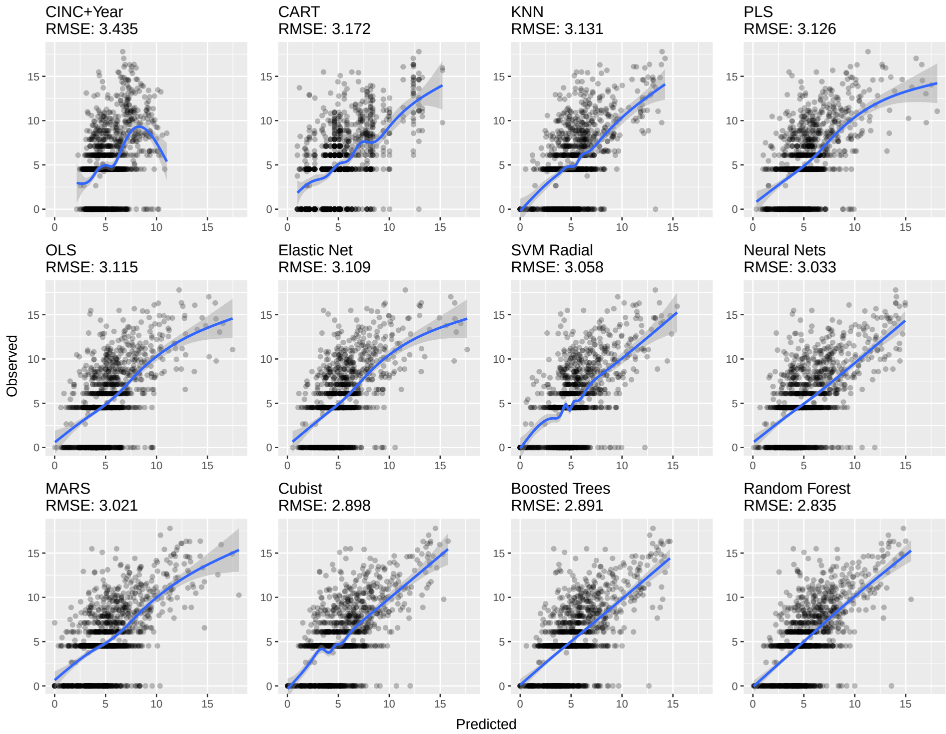


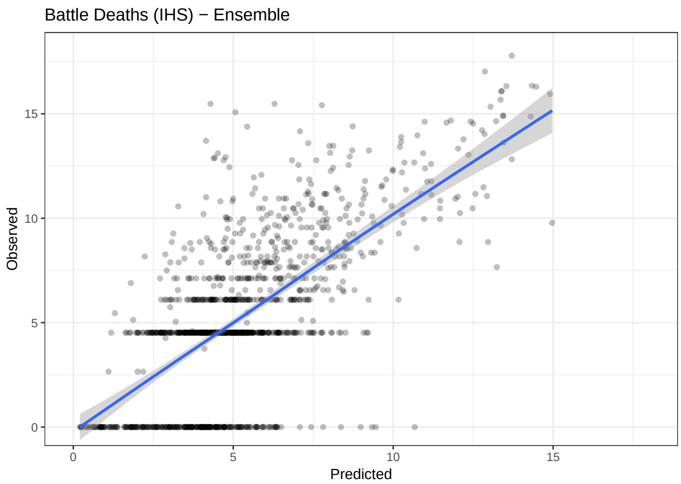
Figure 2: Scatter plots of model performance across each of the candidate models. A model which predicts perfectly would see all observations fitting along a 45 degree line from the bottom left to top right. As evidenced by the scatter plots, models from the linear family - OLS, PLS, and elastic net - 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.

4.1 The Ensemble Model

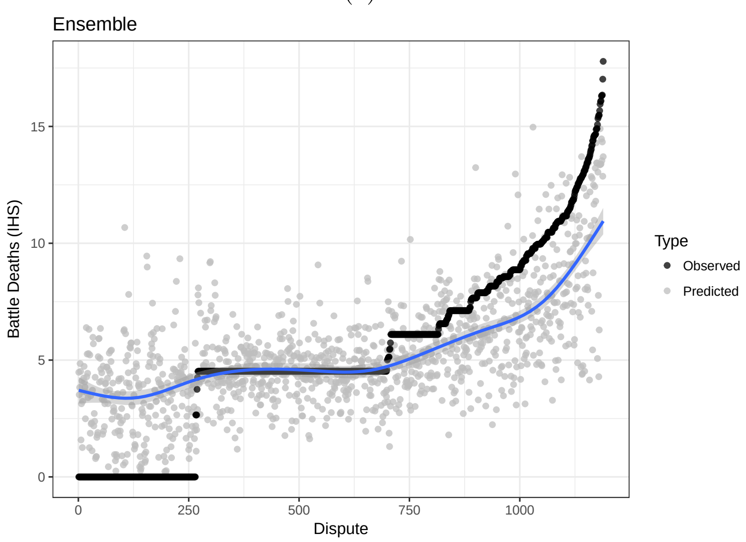
At this point, one option would be to select the model which performs best in cross-validation - either random forests, boosted trees, or Cubist - and proceed with its predictions for creating the measure of expected war costs. Instead, I assess whether I can achieve even better predictive performance 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.

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

Table 3: 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.



(a)



(b)

Figure 3: 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. Plot (b) shows the observed value 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.

5 Explaining Battle Deaths

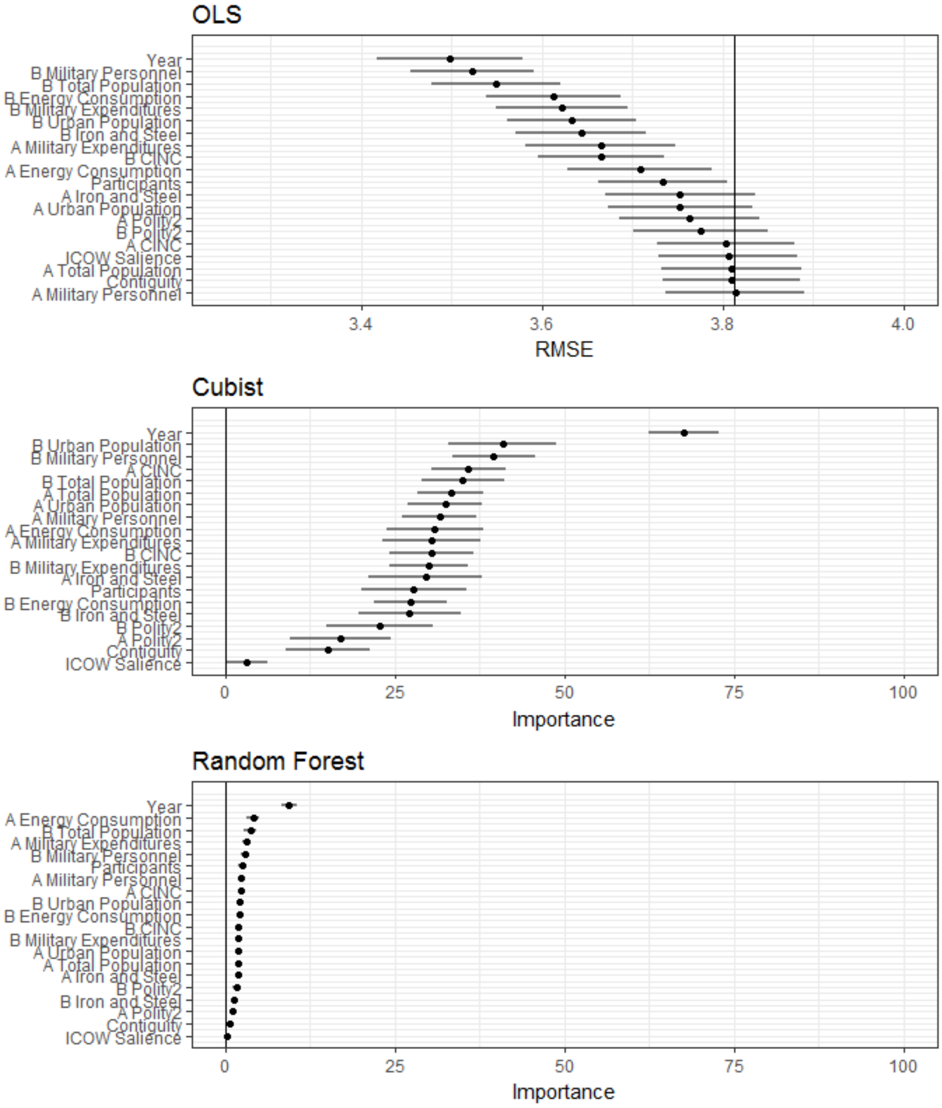


Figure 4: Variable importance plots from OLS models, Random Forests, and Cubist, fit to the aggregate data. Bootstrapped 1000 times to produce the resulting confidence interval.

It would be understandable to be hesitant to embrace '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 the ensemble model and some of the algorithms I use are opaque - neural nets in particular - this is largely not the case for the models which have performed well with this dataset. I am able to examine the relationship between our country characteristics and battle deaths using random forests while also improving our predictions (Jones and Linder, 2015). Namely, what matters for predicting battle deaths, and what is the relationship between the predictors and the outcome?

The first way I assess 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 OLS, 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 the test predictions from permuting a variable across all of the trees in the forest. The resulting scores can be used to assess the relative importance of each predictor in the model.

Figure 4 displays these variable importance scores. ‘Year‘ emerges as the most important predictor, which is in line with the decision tree shown earlier which found an important split in the outcome variable around the year 1950. 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 battel 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 are the most important thing for predicting a war outcome given that the 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 each of the other variables. 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.

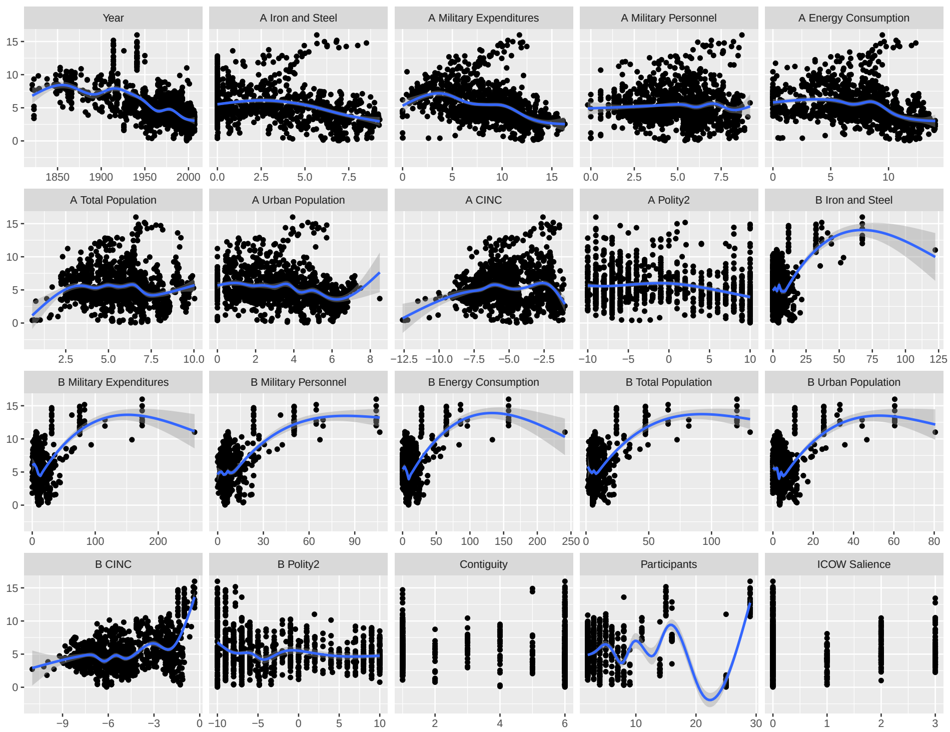


Figure 5: Partial dependence plots from each of the variables in a random forest fit to the entirety of the aggregated data, with and without year included as a predictor (with 7 randomly selected predictors and 500 trees).

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 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 decision tree and 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 signficant 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.

	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	

Table 4: 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.

6 The New Measure: Dispute Casualty Expectations (DiCE)

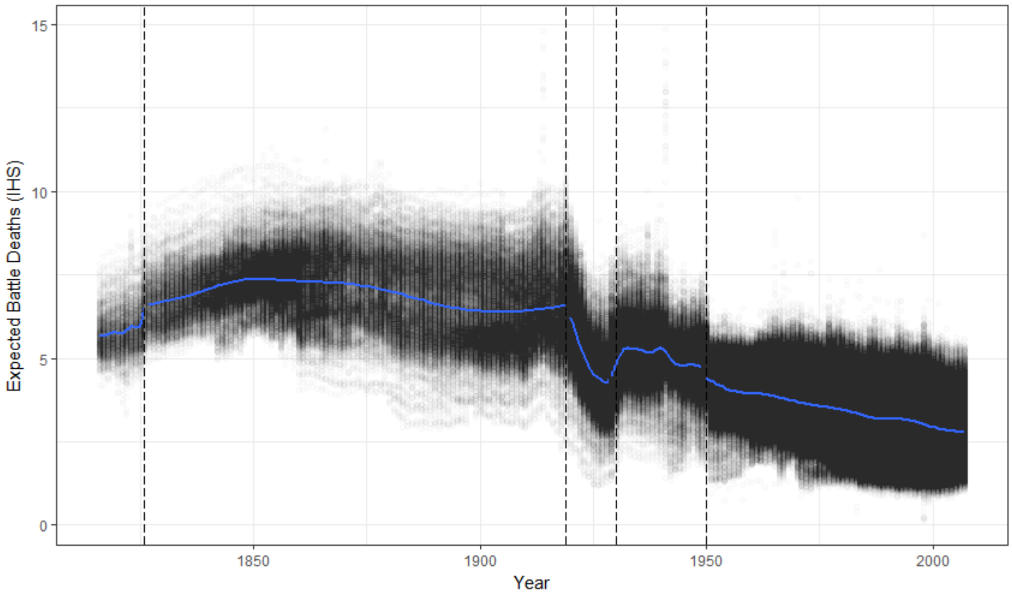


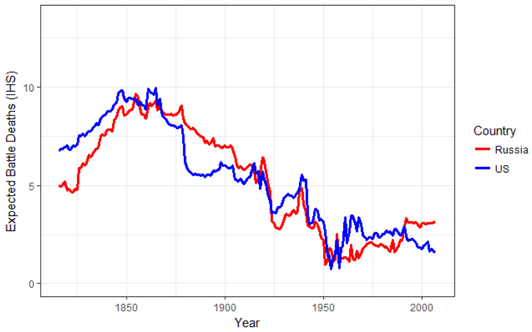
Figure 6: Expected battle deaths from the ensemble model, with LOESS lines and breaks at 1826, 1919, 1930, and 1945

I now turn to the final task of this paper, which is to produce the best estimates for hypothetical disputes in the international system. With the ensemble model in hand, I apply it to 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), and Figure 6 shows a visualized form of this variable. 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, 1930, and 1950. 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]

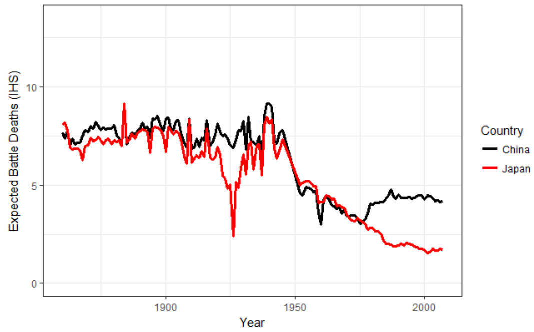
At this stage in the paper, I originally sought to replicate existing models in international conflict and show how a measure of war costs can be used to improve model fit and inform existing theory in the literature on conflict. But I have generally found that there is very little existing empirical 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. That is, 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.

[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.

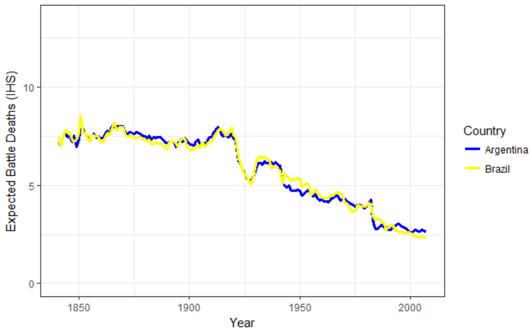
6.1 Substantive Examples



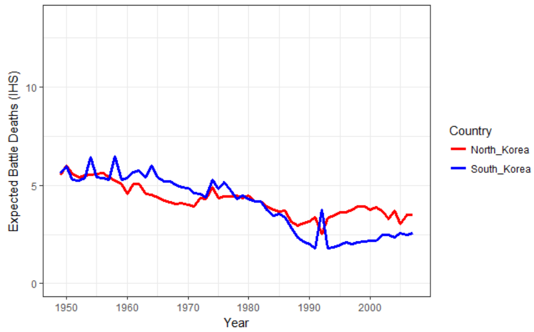
(a) US-Russia



(b) China-Japan



(c) Brazil-Argentina



(d) North Korea - South Korea

Figure 7: Examples of expected dyadic battle deaths over time.

To gain some sense of what the measure looks like, Figure 7 shows 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.

Though there is some variation after 1950, the expected costs for hypothetical conflicts between these countries has 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 8 shows the expected battle deaths for conflicts which occured after 1980 against the models bootstrapped predictions. Here we can see that the model 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.

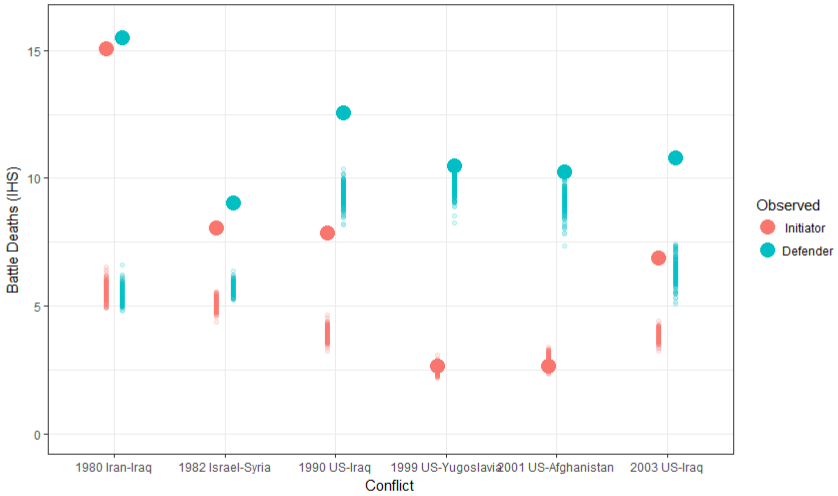


Figure 8: Expected battle deaths for conflicts post 1990 compared to the observed battle deaths, bootstrapped 100 times.

7 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. This new measure costs can be used to develop and test new theory in addition to revisiting old hypotheses.

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 while using the same methodology.

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.

References

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

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.".

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

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

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.

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

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.

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

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

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.

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

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.

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

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.

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.

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

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.

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* .

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

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.

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.

Jones, Zachary M and Fridolin Linder. 2015. "edarf: Exploratory data analysis using random forests." *R package version 0.1* .

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.

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

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

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.

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

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

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.

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

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

Quinlan, John R et al. 1992. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence.* Vol. 92 Singapore pp. 343–348.

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

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

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

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.

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.

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

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.

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.

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

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.

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.