Prediction, Proxies, and Power

Robert J. Carroll[*] Brenton Kenkel[†]

March 1, 2016

Abstract

Many enduring questions in international relations theory focus on power relations between states, so it is important that scholars have a good measure of relative power. But the standard measure of relative military power, the capability ratio, is barely better than random guessing at predicting military dispute outcomes. We use machine learning tools to build a superior proxy, the Dispute Outcome Expectations score, from the same underlying data. Our measure is an order of magnitude better than the capability ratio at predicting dispute outcomes. In replications of 18 recent empirical studies in international relations, we find that replacing the standard measure with DOE scores usually improves both in-sample and out-of-sample goodness of fit. More broadly, we argue that scholars should focus on out-of-sample predictive power when constructing proxies for important concepts in political science. Our approach illustrates how machine learning tools can automate this process.

[*] Florida State University. Email: `rjcarroll@fsu.edu`

[†] Vanderbilt University. Email: `brenton.kenkel@vanderbilt.edu`

For all its progress—more nuanced arguments, more useful theories, bigger data and more systematic ways to analyze them—international relations remains, in many ways, a study of power. This is best reflected in the questions that have endured. Is the world safer when power is concentrated in a few states or broadly distributed (Waltz 1979)? How does the balance of power between states, or shifts thereof, affect the likelihood of war (Organski and Kugler 1980; Powell 1999, 2006)? Do international organizations allow states to gain benefits they would not receive from power politics alone (Keohane and Nye 2001)? But without good measures of power, we cannot provide good empirical answers to these fundamental questions. Consequently, the importance of measuring power to the study of international politics cannot be overstated.

Like many other important concepts in political science—say, ideology or democracy—power cannot be measured directly. Indeed, measurement problems in political science often entail the construction of proxies. Recent advances in computing and modeling have allowed political scientists to build sophisticated, data-driven proxies for variables as diverse as legislator ideology (Clinton, Jackman and Rivers 2004), judicial independence (Linzer and Staton 2014), and country regime types (Jackman and Treier 2008). But despite the centrality of power to many important hypotheses in international relations, its measurement has seen far less innovation.[1] In this article, we devise a new approach to measuring power—specifically, the balance of material power between a pair of countries. Our focus on the contributions of material capabilities follows the example set by most existing efforts to measure power in the international sphere, starting with the work by Singer, Bremer and Stuckey (1972). In contrast with previous approaches, ours is data-driven: we aim to learn what combination of observable material capability variables best predicts international dispute outcomes. We show that the standard measure, the ratio of Composite Index of National Capability scores (Singer, Bremer and Stuckey 1972), predicts militarized dispute outcomes terribly—only 1 percent better than a null model with random guessing. Our new proxy, the Dispute Outcome Expectations score, is much better, providing a 20 percent predictive improvement.

Before constructing a new proxy for relative material power, we first consider what makes a good proxy more generally. Despite the innovations in measurement in various fields, political scientists have not reached a consen-

[1] A recent exception is Arena (2012).

sus on what makes for a good proxy, nor is there a common evaluatory metric. We argue for a predictive criterion: if the concept of interest is supposed to be associated with some observable outcome, then its proxy should predict the outcome well. By prediction, we mean out-of-sample prediction, with data not used to construct the proxy itself. Happily, contemporary machine learning tools make it easy to construct proxy variables according to this criterion so long as data on the relevant outcomes are available. In the case we consider, a proxy for relative military power, the natural outcome of interest is who wins in military disputes. A good measure of relative military power ought to predict dispute outcomes well. It is surprising that the standard measure does so poorly by this criterion, given its ubiquitousness: as we document below, dozens of recent publications in international relations use CINC-derived measures as proxies for power. Our predictive approach, combined with the use of modern machine learning tools, allows us to yield a far superior measure from the same data underlying the usual measure.

Like Ulysses or Goldilocks, the proxy maker must strike a delicate balance. She must learn from the data to construct the measure, else it will fail to capture important dimensions of the concept under study. *A priori* measures like summed rating scales suffer from this *underfitting* problem: they fail to take advantage of the wealth of data scholars now possess. But the analyst who employs a data model for proxy construction faces pitfalls, too. She may misidentify chance features of her data as systematic, a problem called *overfitting*. A good proxy should fit the data well, but not so well that it fails to generalize. The criterion we advocate, out-of-sample prediction, balances these two considerations. An underfit proxy will, of course, be a poor predictor, but so too will a data-driven proxy that maximizes in-sample fit at the expense of generalizability.

Supervised learning techniques, having been designed to navigate the straits between underfitting and overfitting, are ideal for data-driven proxy construction. Machine learning models are flexible enough to model relationships far more complex than possible in ordinary regression or measurement models, but they also guard against connecting the dots too aggressively or misinterpreting noise in the data as a complex relationship. Virtually every supervised learning method has a set of tuning parameters that govern how much flexibility to allow for—in effect, to what extent to treat variation in the data as signal rather than noise. To develop an optimal model for out-of-sample prediction, an analyst simply chooses appropriate tuning parameters, usually by a method like cross-validation that estimates prediction error (Efron and Gong 1983).

Our approach mirrors that of Hill and Jones (2014), who use cross-validation to assess the relative predictive power of many variables all thought to affect the same outcome. Our focus, however, is on constructing variables rather than comparing them.

Our output is the Dispute Outcome Expectations (DOE) score. We use the data on the outcomes of militarized interstate disputes (Palmer et al. 2015) to model the relationship between material capability holdings and dispute outcomes, all while optimizing for predictive power. For every dyad-year from 1816 to 2007, we use this model to estimate the probability that each state would win a hypothetical dispute (and the probability it would end in a stalemate). DOE scores therefore have the same temporal and spatial coverage as the current state of the art, the capability ratio, but have two additional advantages. First, in the cases where disputes did occur, the DOE scores are much better predictors of the outcome than the capability ratio. Second, the DOE score is directly interpretable as the probability of victory, an important concept in the literature on bargaining and war (Fearon 1995; Powell 1996).

When we construct a new proxy by optimizing over predictive power for a given outcome, it is almost tautological that it will predict that outcome better than the extant alternatives. For a fairer comparison, we can take a sample of typical applications of the old proxy and see whether the new one accomplishes them better. For example, international relations scholars include the capability ratio, the standard proxy for relative military power, in models of dependent variables other than dispute outcomes (most commonly, whether a dispute takes place at all). We reanalyze 18 such empirical models to see whether they fit better when we replace the standard proxy with DOE scores. Since these studies examine outcomes besides the one we use to construct our proxy—namely, victory or loss in international disputes—there is no guarantee that our new proxy will do better. Nonetheless, we yield an improvement in fit in at least 14 of the 18 cases. We encourage the creators of future proxies, both in this domain and others, to conduct similarly systematic and comparative studies of their variables' performance in typical applications.

Although our main goal is to develop a better proxy for military power rather than to test hypotheses about its determinants, we do gain some broad substantive insights from the model-building process. Most simply, material capabilities indeed matter for military power, as we can explain a substantial amount of the variation in militarized dispute outcomes just with variables on material capabilities. This finding runs in contrast to the classic study by Maoz (1983), who finds no relationship between matériel and militarized dispute outcomes. Our results suggest that this finding is the artifact of relying on ratios of capability holdings, which are a poor proxy for relative material power. We also find that the martial effectiveness of the various material capability components varies over time, which the standard measure does not allow for.

The paper proceeds in five sections. In the first, we lay out our general argument about proxy construction and its application to the case of military power. Section 2 describes the data and methods we use to construct a new proxy for expected dispute outcomes. In Section 3, we discuss the advantages and disadvantages of our measure. Section 4 contains the results of our replications and advice for using the DOE score. The final section addresses next steps and concludes.

1 Proxies and Power

A proxy is a function of observable variables that aims to measure an unobservable quantity. By definition, we can never know for sure how well a particular proxy captures the concept of interest. We can still try to gauge the quality of a proxy by testing its association with some observable outcome (or outcomes) that we would expect the underlying concept to be related to; in the context of summed rating scales, Spector (1992, 46–47) notes that "validation can only occur within a system of hypothesized relations between the construct of interest and other constructs," and thus that such validation "demonstrates the potential utility of the construct." We can take this logic further, extending it to how we build proxies in the first place. Instead of constructing a proxy from observable indicators according to an *a priori* formula and then testing whether it is associated with some observable outcome, we can select as our proxy the function of observables most strongly associated with the given outcome. Measurement models automate this process. For example, ideal point estimates of legislator ideology are selected to maximize the likelihood or posterior probability of the observed roll-call matrix (Poole and Rosenthal 1985; Clinton, Jackman and Rivers 2004).

But without some kind of regularization or correction, these data-driven approaches to proxy construction run the risk of overfitting—amplifying the noise inherent in data and mistakenly treating it as a signal. The risk of overfitting is particularly high when there are too many degrees of freedom relative to the amount of data available. If the outcome of interest is only rarely observed, it might be hard to separate signal from noise. Similarly, overfitting is a concern if we are modeling the proxy as a function of many observable indicators, or we do not have the domain knowledge we would need to impose a specific functional form for the relationship between these indicators and the outcome of interest. Situations like these are common in political science, including the current context. To prevent overfitting, we can use informative priors in Bayesian contexts (Clinton, Jackman and Rivers 2004) or cross-validation in frequentist contexts (Efron and Gong 1983).

We want to measure relative military power, or the balance of power between two states at a particular point in time. Singer (1963, 420) argues that "power is to [political scientists] what money is to the economist." But power, unlike money, is not directly observable. We follow Singer, and virtually all of the international relations literature of the previous half-century, in developing a proxy for relative power that is a function of each country's observable material capabilities. Though power may in truth be a function of many variables, including non-material factors, we restrict our attention to the set of variables used in the Composite Index of National Capabilities (Singer, Bremer and Stuckey 1972). By holding the set of variables fixed, we ensure that any observed improvements are due to our modeling approach and not to additional information. Still, to reflect the limited scope of the variables we consider, we refer to both CINC-based measures and our own as proxies for relative *material* power.

If we want to use an observable outcome to validate or construct a proxy for relative military power, the obvious choice is war outcomes. Or, since full-scale wars are (thankfully) rare, we may broaden our scope—as conflict scholars often do—to consider the outcomes of all militarized disputes (Palmer et al. 2015). A good proxy for relative power should be a good predictor of which side prevails in militarized disputes. Indeed, the probability of victory by each side is itself an important concept in formal theories of bargaining and war (Fearon 1995; Powell 1996) and empirical examinations thereof (e.g. Reed et al. 2008).

The standard proxy for relative material power, the capability ratio, is an *a priori* creation. It is based on the CINC score, which is the average of a state's shares of the global totals of six raw material holdings in a given year.[2] In a dyadic analysis, a state's capability ratio is the ratio of its own CINC score

[2] The six components are iron and steel production, primary energy consumption, military expenditures, military personnel, total population, and urban population (Singer, Bremer and Stuckey 1972).

to the total CINC score in the dyad. The capability ratio was popularized by Bueno de Mesquita (1981, 108), who treated each side's capability ratio as a proxy for its probability of victory in a potential dispute. Since then, the capability ratio and its cousins have become by far the most common proxies for relative material power. Examining publications from 2005 to 2014 in five top journals for empirical international relations research,[3] we found at least 94 articles that control for the capability ratio or other proxies based on CINC scores.

The key question for international relations scholars is whether the capability ratio is a good proxy for relative material power. Is it a good predictor of actual dispute outcomes? Writing over three decades ago, Maoz (1983) found no evidence that ratios of military expenditures and military personnel are associated with dispute outcomes; this important finding has helped shape debates since. Our findings on ratios of CINC scores, reported in the next section, echo Maoz's: the capability ratio is only 1 percent better than random guessing at predicting dispute outcomes. The next question is whether the methods we advocate would make better use of the data. We find that by using machine learning to develop a predictive model of dispute outcomes as a function of material capabilities, we yield a superior proxy for relative material power.

Measuring expected dispute outcomes is in many ways a hard case for out-of-sample prediction. There are relatively few interstate disputes, and even fewer that involve just a single pair of states. Even if we restrict ourselves just to the National Material Capabilities data, there is an abundance of variables: six capability components for each side of the dispute, along with the six annual shares associated with each raw component, for a total of 24. There is no consensus (and little developed theory in the first place) on how these components ought to map into power, so our models must be flexible. Yet amid all these potential sources of noise, we are able to extract a decent signal: our measure is 20 percent better than a null model at predicting dispute outcomes—acceptable performance in absolute terms, and a major improvement over current practice.

[3] *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, *International Organization*, and *International Studies Quarterly*.

2 Building a Better Proxy for Relative Military Power

Our goal now is to squeeze as much predictive power as we can from data on states' material capabilities. When prediction is the goal, "black box" algorithmic techniques usually outpace standard regression models (Breiman 2001). So, to build our new measure, we augment traditional approaches with methods from machine learning.

2.1 Data

To evaluate the predictive performance of the capability ratio and then to build an alternative measure, we use data on the outcomes of international disputes. We combine the National Material Capabilities data (Singer, Bremer and Stuckey 1972) with information on the outcomes and participants of Militarized International Disputes between 1816 and 2007 (Palmer et al. 2015). Our data consist of $N = 1{,}740$ disputes, each between an "initiator," or Country A, and a "target," or Country B.[4] Every dispute outcome is either A Wins, B Wins, or Stalemate, denoted $Y_i \in \{A, B, \emptyset\}$. Most disputes end in a stalemate, and victory by the initiator is over twice as likely as victory by the target, as shown in Table 1.

	Count	Proportion
A Wins	201	0.12
Stalemate	1460	0.84
B Wins	79	0.05

Table 1. Distribution of the three dispute outcomes.

We model dispute outcomes as a function of the participants' military capabilities. Our data source, the National Material Capabilities dataset, records annual observations of six characteristics of a country's military capability: military expenditures, military personnel, iron and steel production, primary energy consumption, total population, and urban population.[5] We also calculate

[4] See the Appendix for the data construction and coding specifics.

[5] There are missing observations in the National Material Capabilities data. Consequently, about 17 percent of the disputes we observe contain at least one missing cell. We use multiple imputation to deal with missingness (Honaker and King 2010); see the Appendix for details.

each country's share of the global total of each component, giving us 12 variables per dispute participant. The matrix of predictors has 26 columns: the 24 individual capability characteristics of the initiator and target, the standard capability ratio, and the year the dispute began. Collect these predictors for the i'th dispute into the vector X_i.

2.2 A Metric for Predictive Power

We face two challenges in evaluating a model's predictive power. The first is to define a metric—one that is appropriate to the task at hand and reasonably interpretable. The second is to measure each model's ability to predict *out of sample*. Our main purpose, which is to measure the chances of victory for each side in a hypothetical interstate dispute, is inherently an out-of-sample prediction task.

As fortune plays a role in every military engagement, it is impossible to perfectly predict the outcome of every dispute. We therefore want a measure of predictive power that respects the probabilistic nature of militarized disputes. Classification metrics like the accuracy statistic, also known as the percentage correctly predicted, do not fit the bill.[6] Instead, we employ the log loss, which is the negative of the average log-likelihood, as our metric for predictive power (Hastie, Tibshirani and Friedman 2009, 221). Let a *model* be a function $\hat{f}$ that maps from the dispute-level predictors X_i into the probability of each potential dispute outcome, $\hat{f}(X_i) = (\hat{f}_A(X_i), \hat{f}_B(X_i), \hat{f}_\emptyset(X_i))$. The "hat" on $\hat{f}$ emphasizes that the form of the function has been learned from the data, whether by estimating regression coefficients or by a more flexible predictive algorithm. The log loss of model $\hat{f}$ on the data (X, Y) is[7]

$$\ell(\hat{f}, X, Y) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t \in \{A,B,\emptyset\}} \mathbf{1}\{Y_i = t\} \log \hat{f}_t(X_i). \tag{1}$$

Smaller values of the log loss represent better predictive power, with the lower bound of 0 indicating perfect prediction.

We care mainly about the generalization error of our models—the expected quality of their predictions for new data that was not used to fit the models. Our small sample size of $N = 1{,}740$ makes this tricky. Had we a surplus of

[6] Such classification metrics also discriminate poorly with imbalanced classes like ours (Kuhn and Johnson 2013, 420–423).

[7] To avoid numerical problems, very low probabilities are trimmed at $\epsilon = 10^{-14}$.

observations, we could use some suitably large number to fit our models and hold out the remainder to assess the models' predictive power. But with as little data as we have, splitting the sample is ill-advised: we cannot hold out enough observations to estimate the generalization error precisely without harming the precision of the model itself. So, to measure out-of-sample predictive power without losing data, we turn to K-fold cross-validation (Hastie, Tibshirani and Friedman 2009, 241–249). We randomly assign each dispute observation to a "fold" $k \in \{1, \ldots, K\}$, where we follow standard practice by setting $K = 10$.[8] For each k, we split the data into a "test" sample containing fold k and a "training" sample containing the remainder of the data. We fit a model only on the training sample and then calculate its predicted probabilities for the data in the test sample.[9] After repeating this K times, we have an out-of-sample prediction for each observation in our data calculated from a model that did not see that observation. We compare these predicted probabilities to the observed outcomes to estimate our models' generalization error. Formally, the cross-validation loss of the model $\hat{f}$ is the average out-of-fold log loss,

$$\text{CVL}(\hat{f}) = \frac{1}{K} \sum_{k=1}^{K} \ell \left(\hat{f}^{(-k)}, X^{(k)}, Y^{(k)} \right), \tag{2}$$

where $(X^{(k)}, Y^{(k)})$ is the data in the k'th fold and $\hat{f}^{(-k)}$ is the model $\hat{f}$ fit to the data excluding the k'th fold.

Because it is measured on the log-likelihood scale, the log loss metric is hard to interpret. To ease the interpretation, we compare models' log loss to that of a null model, whose predicted probabilities always equal the sample proportions of each outcome. The proportional reduction in cross-validation loss of the model $\hat{f}$ is

$$\text{PRL}(\hat{f}) = \frac{\text{CVL}(\hat{f}_{\text{null}}) - \text{CVL}(\hat{f})}{\text{CVL}(\hat{f}_{\text{null}})}. \tag{3}$$

[8] Standard practice here stands on firm ground; Molinaro, Simon and Pfeiffer (2005) find that 10-fold cross-validation performs quite similarly to leave-one-out cross validation without having to take on massive computational costs. 10-fold cross-validation also performs better than other techniques, particularly in smaller samples like ours.

[9] When dealing with models with tuning parameters that are themselves selected by cross-validation, we choose tuning parameters separately within each of the K iterations via another cross-validation loop. This nested cross-validation is necessary to keep our estimates of generalization error from being too optimistic (Varma and Simon 2006).

	Estimate	SE	Z	p
Capability Ratio (logged)	0.26	0.06	4.16	<0.01
Cutpoint: B Wins to Stalemate	−3.31	0.14		
Cutpoint: Stalemate to A Wins	1.84	0.09		

Table 2. Results of an ordered logistic regression of dispute outcomes on the capability ratio using the training data. Because there are no missing values in the CINC scores, these estimates are identical across imputed datasets.

The theoretical maximum, for a model that predicts perfectly, is 1. If a model predicts even worse than the null model—meaning it is worse than random guessing—its proportional reduction in loss is negative.

2.3 Modeling Dispute Outcomes

Our task now is twofold: to assess the predictive power of the capability ratio and, should we find it lacking (as we do), to build a better alternative.

We model dispute outcomes as a function of the capability ratio via ordered logistic regression (McKelvey and Zavoina 1975). To reduce skewness, we take the natural logarithm of the capability ratio. The parameter estimates from the capability ratio model on the full sample appear in Table 2. Although these results do not speak directly to the capability ratio's out-of-sample performance, they foreshadow why its predictive power is so limited. The coefficient on the capability ratio is statistically significant but small relative to the cutpoints, indicating a substantively weak relationship. Dividing the cutpoints by the coefficient, we see that we would need a logged capability ratio below −13 or above +7 to predict any outcome other than a stalemate. These bounds lie well outside the observed range of capability ratios in the dispute data, which are bounded below by −9.1 (Palau–Philippines 2000) and above by −0.0004 (Germany–Panama 1940). In other words, the capability ratio always predicts a stalemate within the sample. This does not bode well for its out-of-sample performance.

We want a better model than what the capability ratio gives us, but we do not have a strong *a priori* sense of what the true relationship between material capabilities and dispute outcomes looks like. So, we use tools from machine learning that are designed to predict well without imposing much structure on the data. Ideally, we would select the predictive model that is best for our data, but there are too many algorithms to try them all. To narrow it down, we defer to the machine learning experts on which algorithms are best. We draw our set of candidate models from the top-ten list by Wu et al. (2007) and from the best performers in the tests by Fernández-Delgado et al. (2014). After excluding those unsuited to our data,[10] we end up with six predictive algorithms: C5.0, support vector machines, k-nearest neighbors, classification and regression trees, random forests, and ensembles of neural nets.[11] Each algorithm is widely used for prediction and can predict dispute outcome probabilities as a complex, potentially nonlinear function of the material capability components. As a compromise between these flexible "black box" models and the rigid capability ratio model, we also test ordered logistic regression models on the capability components.

In the spirit of flexibility, we try each model with different sets of predictors from the capability data. We examine four sets of variables: the raw capability components and the annual component shares, each with and without the year the dispute began. All of our models allow for interactive relationships, so including the year of the dispute lets the effect of each capability component vary over time. With two sides per dispute and six capability variables per side, each model has 12 or 13 variables, depending on whether the year is included. All told, we have 30 candidate models: four sets of variables for each of our seven algorithms, plus the capability ratio model and a null model used as a baseline.

We use cross-validation to estimate how well each of our candidate models predicts out of sample. The final problem, given those estimates, is to choose a model to construct an alternative to the capability ratio as a measure of expected dispute outcomes. It is tempting to simply pick the model with the lowest cross-validation loss. We can do even better at prediction, however, by taking a weighted average of all the models. We use the super learner algorithm (van der Laan, Polley and Hubbard 2007) to select the optimal model weights via cross-validation. Given a set of M candidate models $\hat{f}_1, \ldots, \hat{f}_M$, we

[10] Four of the algorithms named in Wu et al. (2007)—k-means, Apriori, expectation maximization, and PageRank—are not suited for the prediction task at hand. We also excluded AdaBoost due to long computation time and naive Bayes due to poor performance in initial tests.

[11] See the Appendix for full details of each method.

select weights $\hat{w}_1, \ldots, \hat{w}_M$ to solve the constrained optimization problem

$$\begin{aligned} \min_{w_1,\ldots,w_M} \quad & \text{CVL}\left(\sum_{m=1}^{M} w_m \hat{f}_m\right) \\ \text{s.t.} \quad & w_1, \ldots, w_m \geq 0, \\ & w_1 + \ldots + w_m = 1, \end{aligned} \tag{4}$$

Our final model is the super learner, $\hat{f} = \sum_m \hat{w}_m \hat{f}_m$. Each individual model is a special case of the super learner, with full weight $\hat{w}_m = 1$ placed on a single $\hat{f}_m$. Hence, by the cross-validation criterion, we should prefer the super learner over any individual model.[12] That said, the super learner does provide the capability ratio with an opportunity to defend itself; should it earn a high weight, then our costly enterprise may not be worth the effort.

To summarize, we fit and cross-validate $M = 30$ candidate models, then combine them into a super learner that we will use to construct a better proxy for expected dispute outcomes. The biggest downside of our approach is that the results are not easily interpretable. Because the super learner entails averaging a large set of models—some of which, like random forests, are themselves difficult to interpret—it gives us no simple summary of how each predictor affects dispute outcomes. This is not a problem, given our aims. Certainly, we would not recommend the super learner as a means of testing hypotheses about the determinants of dispute outcomes. However, our goal is not to test a hypothesis—it is to construct the best proxy possible for how a dispute between two countries is likely to end. In this context, it is worth sacrificing interpretability for the sake of predictive power.

2.4 Cross-Validation Results

We now turn to the cross-validation results, which are summarized along with the super learner weights in Table 3. As the in-sample analysis hinted, the capability ratio is indeed a poor predictor of dispute outcomes. Its proportional reduction in loss is 0.01, which means its predicted probabilities are just 1 percent more accurate than the null model. This number is not encouraging,

[12] As usual when selecting tuning parameters via cross-validation, the value of equation (4) is not an unbiased estimate of the generalization error of the super learner. Nested cross-validation is computationally infeasible for the super learner, so we calculate the bias correction recommended by Tibshirani and Tibshirani (2009) to estimate its generalization error.

but what matters even more is whether we can do better. A glance at Table 3 confirms that we can: all but one of our 28 alternative models have greater predictive power than the capability ratio, many of them considerably better. With these results in hand, we feel comfortable dismissing the capability ratio as a suboptimal proxy for expected dispute outcomes.

As we expected, the super learner ensemble performs better than any of the candidate models from which it is constructed. The ensemble's proportional reduction in loss is about 23 percent, or four percentage points better than the best candidate model. Even after we apply a bias correction (see footnotes 9 and 12), the super learner's predictive power is still the best among our models. Looking at the weights, what stands out is how few models are substantial components of the super learner: just five models have a weight of at least 5 percent. More generally, while models with lower generalization error tend to receive more weight, the relationship is by no means one-to-one. We see this because the ensemble prefers not only predictive power, but also diversity. Different classes of models have different blind spots; the more diverse the ensemble is, the more these blind spots are minimized. A model that looks bad on its own might still merit non-negligible weight in the optimal ensemble if it captures a slice of the data missed by the models that are best on their own.

The super learner predicts dispute outcomes much better than the capability ratio does. As we have just shown, the capability ratio only improves by 1 percent on a null model, whereas the super learner gives a 20 percent improvement. For another illustration of the difference in predictive power, see the plots of out-of-fold predicted probabilities—the ones we use in cross-validation—in Figure 1. Under the capability ratio model, all but a handful of disputes are predicted to have an 80–90 percent chance of ending in stalemate. Seeing how narrow the capability ratio's predictive range is, it is little surprise that it barely does better than a null model at prediction. Conversely, the super learner makes much better use of the material capability data. Its predictive range is greater, which in turn allows it to achieve a stronger, though hardly perfect, relationship between predicted and observed outcomes.

2.5 Implications for International Relations

Our main focus is on developing a proxy for relative power that predicts the outcomes of militarized disputes, and predictive approaches like ours are not optimal for testing specific hypotheses (Shmueli 2010). Nonetheless, we can

Method	Data	Year	CV Loss	P.R.L.	Weight
Null Model	Intercept Only		0.54		<0.01
Ordered Logit	Capability Ratio		0.53	0.01	<0.01
Ordered Logit	Components		0.49	0.09	<0.01
Ordered Logit	Components	✓	0.48	0.10	<0.01
Ordered Logit	Proportions		0.51	0.04	<0.01
Ordered Logit	Proportions	✓	0.49	0.08	<0.01
C5.0	Components		0.53	0.02	0.01
C5.0	Components	✓	0.51	0.04	0.04
C5.0	Proportions		0.52	0.03	0.02
C5.0	Proportions	✓	0.51	0.05	0.01
Support Vector Machine	Components		0.46	0.14	<0.01
Support Vector Machine	Components	✓	0.46	0.14	<0.01
Support Vector Machine	Proportions		0.49	0.09	<0.01
Support Vector Machine	Proportions	✓	0.48	0.10	<0.01
k-Nearest Neighbors	Components		0.47	0.12	<0.01
k-Nearest Neighbors	Components	✓	0.45	0.16	0.02
k-Nearest Neighbors	Proportions		0.51	0.05	<0.01
k-Nearest Neighbors	Proportions	✓	0.48	0.11	<0.01
CART	Components		0.52	0.02	<0.01
CART	Components	✓	0.44	0.19	0.28
CART	Proportions		0.55	−0.03	<0.01
CART	Proportions	✓	0.50	0.06	<0.01
Random Forests	Components		0.49	0.08	0.04
Random Forests	Components	✓	0.48	0.11	0.19
Random Forests	Proportions		0.47	0.12	<0.01
Random Forests	Proportions	✓	0.48	0.11	0.01
Averaged Neural Nets	Components		0.44	0.19	0.08
Averaged Neural Nets	Components	✓	0.43	0.19	0.13
Averaged Neural Nets	Proportions		0.48	0.11	<0.01
Averaged Neural Nets	Proportions	✓	0.44	0.19	0.16
Super Learner			0.41	0.23	
(bias-corrected)			0.43	0.20	

Table 3. Summary of cross-validation results and super learner weights. All quantities represent the average across imputed datasets.

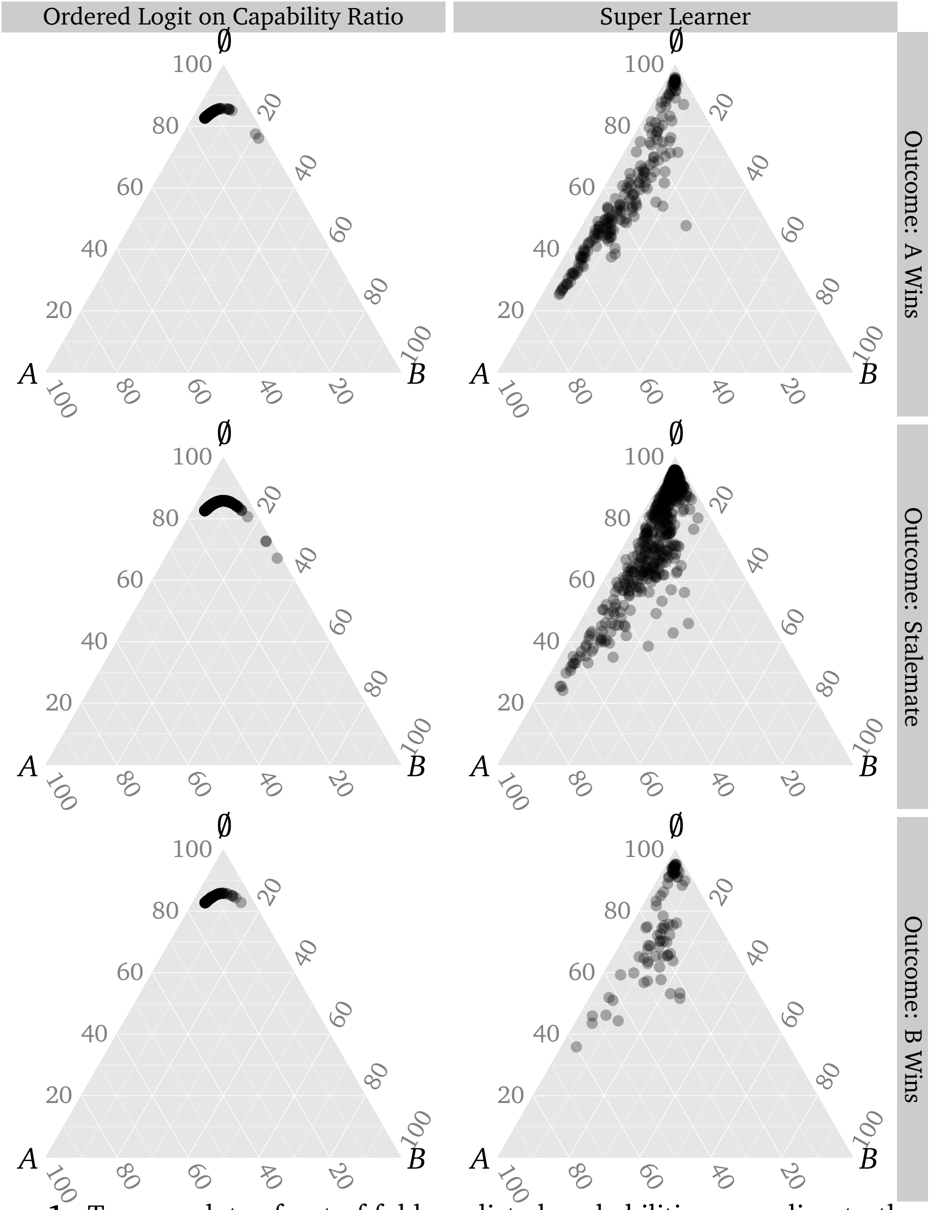

Figure 1. Ternary plots of out-of-fold predicted probabilities according to the capability ratio model and the super learner. Each predicted probability is calculated by fitting the model to 9/10 of the data, not including the observation in question—an approach that simulates true out-of-sample prediction.

glean from our results a few important insights about the nature of the relationship between capabilities and power. The first is that there *is* a relationship—that variation in dispute outcomes is associated with variation in the disputants' raw capabilities. This finding contrasts with previous studies concluding that military capabilities do not affect dispute outcomes (Maoz 1983). The problem is the use of the capability ratio to measure material power, which is only weakly related to dispute outcomes. Consequently, studies that rely on it will conclude that material power is unimportant. We find, however, that material power is related to dispute outcomes, albeit in ways that transcend the capability ratio's usefulness.

But material power is not all that matters. Even after an intense, diverse predictive effort, we explain only 20 percent of the variation in dispute outcomes with material capability variables. To some extent this reflects the inherent unpredictability of military affairs; we would never expect to predict outcomes perfectly. We suspect, however, that we could better predict dispute outcomes even better by conditioning on even more observable indicators. That is a task for future work, as the purpose of this paper is only to develop a proxy for the material components of relative power.

A second important finding is that the determinants of material power change over time. This conclusion may sound obvious, but it raises the question of why international relations scholars continue to use a proxy for power that assumes the relationship is unchanging. The simplest way to observe that time matters is to compare the predictive power of the models with and without the year variable: in 13 out of 14 cases, the model that includes time predicts better than its closest time-less counterpart.[13] The improvement in performance does not just reflect the fact that the distribution of dispute outcomes has changed over time (though, to be clear, it has). All of our models that include the year of dispute allow for the effects of each individual capability component to vary with time. For example, in our random forests on the components and year, every single tree contains at least one node that splits on a capability component whose parent node splits on the year of dispute, indicating a capability effect that is time-dependent. In fact, in about 27 percent of trees, the initial split is on the year of dispute.

Finally, our results show that a country's material power cannot be summarized in a single index. Our ordered logit models construct a flexible index that allows for different weights on the components, variation over time in those

[13] The difference in log loss is statistically significant (paired $t = -3.25$, $p = 0.006$).

weights (in the model with year included), and different weights for initiators and targets. Nonetheless, even the best ordered logit model has less than half the predictive power of our flexible ensemble. What this means is that the relationship between each capability component and overall material power is conditional, not absolute. The usefulness of a particular component may depend in part on the composition of the opponent's capability holdings. The super learner results are too complex to allow us to dig into the specifics of these interactions, but they alert us to their existence—and to the limitations of monadic indices as a measure of power. This ties well to broader theoretical conceptions of power; Dowding (1991, 48) observes that power necessarily involves a social relation among multiple actors.

3 The New Measure: Dispute Outcome Expectations

We use the super learner results to construct a new proxy for expected dispute outcomes—one that predicts actual dispute outcomes much more accurately than the capability ratio does. For any pair of countries at a particular point in time, whether or not they actually had a dispute with each other, we can use the super learner to ask, "Based on what we know about their material capabilities, how would a dispute between these countries be likely to end?" To construct the new proxy, we use the super learner to make predictions for every directed dyad–year in the international system between 1816 and 2007, the range of years covered by the National Material Capabilities data. We call the resulting dataset the Dispute Outcome Expectations data, or DOE. The DOE data contains predictions for more than 1.5 million directed dyad–years.[14]

The DOE scores are naturally directed, since each dispute in our training data contains an initiating side and a target side. However, many analyses in the international conflict literature (e.g., of dispute occurrence) use undirected data. We calculate undirected DOE scores through a simple average of the directed values. For example, to calculate the probability that the United States would win a dispute against the United Kingdom in 1816, we average its estimated chances of victory as an initiator (36 percent) and as a target (11 percent) to yield 23.5 percent. If an analyst using the DOE data believed that the likely identity of an initiator in a hypothetical dispute were not a coin

[14] About 19 percent of directed dyad–years contain missing values of the capability components for at least one country. We average across imputations of the capabilities data to calculate the DOE scores for these cases. See the Appendix for details.

flip, she could take a different average of the directed scores to produce a more representative undirected score.

The DOE measures have two advantages over the capability ratio as a proxy for expected dispute outcomes. First, they are direct measures of the quantity of primary interest to scholars of conflict: the probability that each state would win in a hypothetical dispute. Although the capability ratio is a proportion, it cannot be interpreted as the probability of victory. The ease of interpretation is particularly important for scholars who wish to control for expected dispute outcomes in a regression model. The coefficient on a DOE score can be interpreted directly as the marginal effect of a state's chance of victory; the coefficient on the capability ratio cannot. Second, as we have already seen, within the set of state pairs where disputes occur, the DOE measures are much better predictors of the outcome than the capability ratio is. In short, they are superior proxies, and therefore are the appropriate choice for scholars who need an accurate measure of expected dispute outcomes. The canonical correlation between the DOE scores and the capability ratio is 0.44 (for both the directed and undirected DOE scores), so the measures are related but distinct.

The DOE scores have one drawback worth mentioning: they should not be included as controls in regressions whose dependent variable is the outcome of a dispute or war. This may seem contradictory, given how much effort we have just spent showing that DOE scores are superior predictors of dispute outcomes. The reason they are superior is that, unlike the capability ratio, they are calibrated using real dispute data. But this in turn means that DOE scores would be endogenous in a regression whose dependent variable is dispute outcomes—i.e., the same data we used to construct the DOE scores. Another way to think about it is that the DOE score measures expectations of dispute outcomes, and there is no reason to think these expectations themselves have an independent effect on the outcomes. So when we test causal hypotheses about dispute outcomes, we should control for raw capabilities, not expectations. But when we are modeling dependent variables that might be affected by expectations, such as the onset of a crisis or a state's decision to join an ongoing conflict, we should use the best available proxy for those expectations—namely, the DOE scores.

Another potential concern about the DOE scores is selection bias. We calculate the DOE scores by learning from actual dispute outcomes, but there may be systematic differences between dyad-years with disputes and those without. We do not think this concern should discourage conflict scholars from including DOE scores in their empirical analyses. The relevant consideration

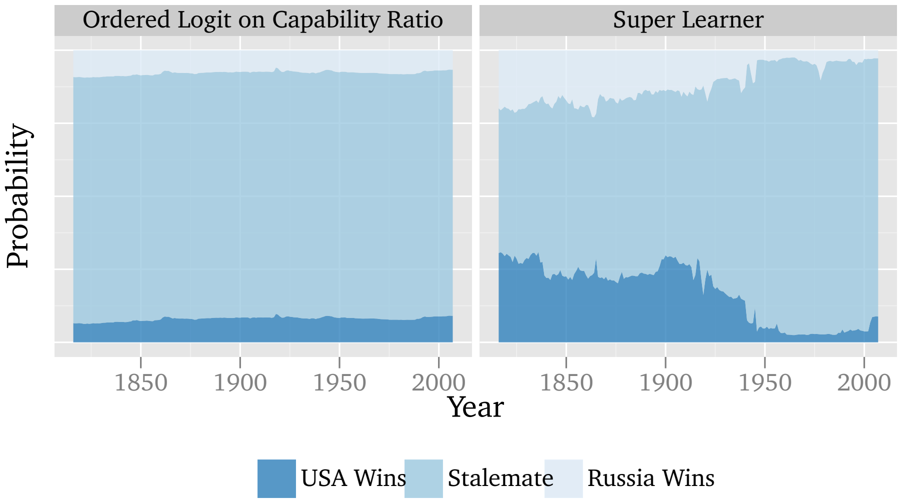

Figure 2. Comparison of predicted outcome probabilities over time from the capability ratio and the super learner (DOE scores) for the United States versus Russia. These plots use the undirected scores.

is not whether DOE scores are perfect—obviously, they are not—but rather whether they are better than the existing alternatives. We cannot evaluate the performance of any proxy at predicting the outcomes of disputes that did not happen in the first place. Seeing as the capability ratio barely predicts outcomes of actual disputes, we see no reason to assume it does well predicting counterfactual outcomes. Scholars who want the best available proxy for relative material power should rely on DOE scores, while keeping in mind its limitations.

To illustrate the contrast between the capability ratio and DOE scores for forecasting hypothetical dispute outcomes, Figure 2 plots the two models' predictions over time for the United States versus Russia. According to the capability ratio model, there was essentially no change between 1816 and 2007 in the likelihood of either side winning a dispute. We need not dwell on the implausibility of this prediction. Conversely, the DOE scores stack up with our intuitions relatively well: the predicted chance of a stalemate balloons during the Cold War, but the chance of victory by the United States picks up afterward.

In light of the DOE scores' superior predictive performance in the Militarized Interstate Disputes data, we are inclined to believe they dominate the capability ratio as a proxy for expected dispute outcomes. Next, we test this conjecture by seeing if replacing the capability ratio with DOE scores in empirical models of international conflict improves their in-sample fit and out-of-sample predictive power.

4 Using the New Measure

We have shown that our measure, the DOE score, predicts international dispute outcomes better than the capability ratio. But most recent studies of conflict do not aim to predict how disputes will end. They focus on other dependent variables, and they usually only treat the capability ratio (or other functions of raw capabilities) as a control. In this section, we investigate how well the DOE score performs as a replacement for the capability ratio in this more typical setting. First, we replicate 18 recent empirical models of various international phenomena, finding that these models usually fit and predict better when we replace CINC-based measures with DOE scores. After that, we provide some advice to practitioners on how to decide which measure—or measures—to include.

4.1 Replication Analysis

International relations scholars often control for the capability ratio as a proxy for expected outcomes when modeling dependent variables besides who wins, such as the onset or escalation of a dispute. We have shown that the DOE score is a better proxy than the capability ratio, but it does not follow immediately that it is a better control variable in studies of conflict more generally. Can we improve on these analyses—i.e., do our regressions fit the data better—if we replace the capability ratio with our new measure? To address this question, we replicate 18 recent analyses of conflict using DOE scores in place of the capability ratio or other functions of CINC scores. On the whole, we see that the models with DOE scores tend to have better in- and out-of-sample fit, though not always.

We constructed the set of replications by looking for empirical analyses of dyad-years (directed or undirected) that included the capability ratio or another function of CINC scores as a covariate. Each study was published re-

		AIC			P.R.L.	
Replication	*N*	CINC	DOE	Vuong	CINC	DOE
Bennett (2006)	1,065,755†	29712	30969	−13.89	0.245	0.213
Weeks (2012)	766,272†	15816	15568	4.70	0.310	0.321
Jung (2014)	742,414†	10659	10588	2.44	0.350	0.354
Park and Colaresi (2014)	379,821†	10632	10587	2.68	0.315	0.318
Sobek, Abouharb and Ingram (2006)	183,451†	5344	5199	4.33	0.326	0.344
Gartzke (2007)	171,509†	4284	4167	4.04	0.442	0.457
Salehyan (2008*b*)	86,497	8864	8821	0.58	0.279	0.282
Fuhrmann and Sechser (2014)	85,306	2614	2580	1.36	0.203	0.208
Arena and Palmer (2009)	54,403†	1152	1061	2.77	0.071	0.137
Owsiak (2012)	15,806	5805	5750	2.36	0.117	0.125
Zawahri and Mitchell (2011)	12,186	814	809	0.66	0.062	0.066
Salehyan (2008*a*)	10,197	3003	2981	1.43	0.101	0.107
Fordham (2008)	7,788	537	604	−2.27	0.275	0.188
Dreyer (2010)	5,316	3676	3635	2.48	0.239	0.248
Huth, Croco and Appel (2012)	3,826	5938	5935	−0.64	0.053	0.052
Uzonyi, Souva and Golder (2012)	1,667	2008	1986	1.54	0.128	0.137
Weeks (2008)	1,276	1574	1568	0.03	0.101	0.105
Morrow (2007)	864	1488	1504	−2.81	0.260	0.251

Table 4. Summary of results from the replication analysis. In-sample goodness of fit is measured by the AIC and the Vuong (1989) test. Positive values of the Vuong test statistic indicate that the model with DOE terms fits better than the model with CINC terms, and vice versa for negative values. The Vuong test statistic has a standard normal distribution under the null hypothesis of no difference between the models, so values with a magnitude of 1.96 or greater would lead us to reject the null hypothesis at the 0.05 significance level. Out-of-sample fit is measured by proportional reduction in log loss relative to the null model, as reported in the last two columns. We use repeated 10-fold cross-validation to estimate each model's out-of-sample log loss, with 10 repetitions for models indicated by a dagger (†) and 100 repetitions for all others. The null model's log loss is estimated via leave-one-out cross-validation.

cently in a prominent political science or international relations journal.[15] We examined only studies with publicly available replication data. If we could not reproduce a study's main result or were unable to merge the DOE scores into the replication data (e.g., because of missing dyad-year identifiers), we excluded it from the analysis. We also excluded studies that employed duration models or selection models, due to conceptual and technical problems with assessing their out-of-sample performance. Lastly, we excluded studies in which our measure of expected dispute outcomes would be endogenous, namely those whose dependent variable was MID outcomes—the same data we used to construct the DOE scores—or a closely related quantity. In the end, we were left with the 18 studies listed in Table 4.

For each analysis in our sample, we begin by identifying the main statistical model reported in the paper, or at least a representative one.[16] We then estimate two models: the original model, and a replicated model where we replace any functions of CINC scores with their natural equivalents in DOE scores. For example, if the capability ratio is logged in the original model, we log the DOE scores in the replicated model. As a basic measure of each model's in-sample goodness of fit, we compute the Akaike (1974) Information Criterion,[17]

$$\text{AIC} = 2(\text{number of coefficients}) - 2(\text{log-likelihood}).$$

The AIC is commonly used in model selection, with lower values representing better fit. In addition, we compute the Vuong (1989) test of the null hypothesis that the original and replicated models fit equally well.[18] To estimate each model's out-of-sample fit, we perform repeated 10-fold cross-validation. Because each study has a discrete dependent variable, we again employ the log loss (equation (1)) to measure out-of-sample fit.

Table 4 summarizes the results of the replication analysis. In general, the models that include DOE scores do better than those with CINC scores by both in- and out-of-sample criteria. Starting with in-sample fit, we see that the DOE

[15] For details, see footnote 3.

[16] When no main model is apparent, our heuristic is to pick one on the log-likelihood–sample size frontier. Details of the model chosen from each paper and the functions of CINC and DOE scores used are in the Appendix.

[17] Because DOE scores are ternary, the replicated models typically have more parameters than their original counterparts. Hence we measure in-sample fit with the AIC, which penalizes over-parameterization, rather than the log-likelihood.

[18] We employ the standard Bayesian Information Criterion (Schwarz 1978) correction to the Vuong test statistic.

model has a lower AIC than the CINC model in 15 of 18 cases. Moreover, in more than half of those cases (8), under the Vuong test we would reject at the 0.05 significance level the null hypothesis that the models fit equally well. The difference in fit is also statistically significant in all three cases where the CINC model has a lower AIC. The results are similar for out-of-sample fit, with the DOE model having a greater proportional reduction in log loss in 14 cases. The improvement due to using DOE scores is typically modest—about a single percentage point increase in the proportional reduction in log loss. For context, keep in mind that these studies do not take capability measures as a key theoretical variable of interest. In each regression, the capability measure is just one of a battery of control variables. With so much else going on in these models, even a substantial improvement in the quality of the capability measure may lead to just a small increase in overall model fit or predictive power.

With such a small sample of replicated studies, we can only conjecture about why DOE performs better in some cases and worse in others. We see that the cases where DOE is significantly better according to the Vuong test tend to have large sample sizes—but, then again, the study where it does worst has the largest *N* in our sample. In two of the replications where DOE performs worst, namely Bennett (2006) and Fordham (2008), we see that both specifications include the raw CINC scores alongside or in lieu of the capability ratio. These terms may be capturing monadic effects that the purely dyadic DOE scores miss. On the other hand, in the other three analyses that include raw CINC scores (Arena and Palmer 2009; Zawahri and Mitchell 2011; Weeks 2012), the replication with DOE scores performs better by both AIC and cross-validation loss.

4.2 Advice to Practitioners

Seeing as neither the capability ratio nor DOE scores are uniformly better in typical applications, how should empirical scholars choose which one to include in their analysis? Our main recommendation is a theory-driven approach. When theory provides no guidance, we recommend either a data-driven approach or dropping capability measures altogether.

If theory suggests that material capabilities only affect the outcome of interest insofar as they shape expectations about how a dispute would end, then DOE scores are the best measure to control for. Figure 3 contains a causal graph of this situation. The clearest examples of theories where only expecta-

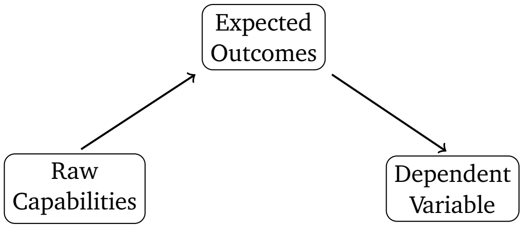

Figure 3. Raw capabilities only affect the outcome of interest through expectations.

tions matter come from the formal literature on crisis bargaining. Take Powell's (1999) theory of bargaining in the shadow of power. War is possible only if the status quo distribution of benefits is far enough from the expected outcome of conflict that at least one state is dissatisfied. An empirical model derived from this theory should control for DOE scores rather than the capability ratio or other poor proxies for expectations.

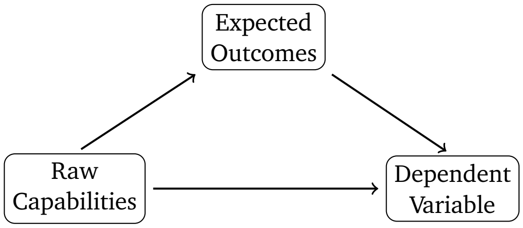

Figure 4. Raw capabilities affect the outcome of interest both directly and through expectations.

If material capabilities affect the outcome both directly and indirectly via expectations, then it would be appropriate to control for both raw capabilities and expected dispute outcomes. Figure 4 illustrates this scenario. For example, imagine an empirical study of "sinking costs" via military mobilization in international crises (Fearon 1997). The initial movement of peaceful relations into a crisis, as well as early behavior at the bargaining table, might be shaped solely by states' expectations about dispute outcomes. But if states build up their military as a way to signal resolve, independently of the effect on likely outcomes, then raw capabilities matter too. When empirically modeling a theory like this, scholars should include both DOE scores and raw capability measures. The ratio of CINC scores may or may not be the most appropriate way to capture raw capabilities—that, too, depends on the specifics of the theory.

The last possibility to consider is that expectations do not affect the out-

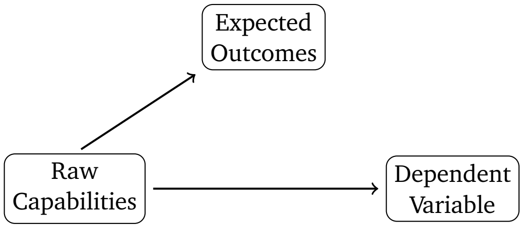

Figure 5. Raw capabilities directly affect the outcome of interest, but expectations do not.

come of interest. In this case, empirical models should only include raw capability measures, not DOE scores. The clearest example is when the dispute outcome itself is the dependent variable. First, absent some kind of self-fulfilling prophecy effect, we would expect actual capability holdings to drive outcomes more so than expectations. Second, because DOE scores are calculated using the dispute outcome data, the DOE scores themselves are endogenous to observed outcomes, and thus should not be included as an independent variable when outcome is the dependent variable.[19]

When there is no theory or the theory does not specify how material capabilities affect the outcome of interest, we recommend a data-driven approach. The steps are the same ones we took in the replication analysis: determine a metric for model fit (whether in- or out-of-sample), run the model separately for each potential measure, and choose the best-fitting model. Alternatively, if your theory says nothing about the relationship between capabilities and the outcome of interest, it may be best not to include capability measures at all. A confounding variable, by definition, must be related to both the treatment and outcome of interest. If raw capabilities are not supposed to affect the outcome either directly or through expectations, then they are not a confounder and there is no need to control for them.

5 Conclusion

In this paper, we have argued that proxies should be constructed using predictive power as the criterion of interest, provided a method for doing so, and demonstrated the usefulness of the method in an application to measuring rel-

[19] In principle, this latter problem could be solved, albeit at great computational expense, by only using data for years up to $t-1$ to calculate the DOE score for year t.

ative military power. We hope that our efforts will be of use both for the DOE scores we provide and for the theoretical merits of our general argument.

In our application, the DOE scores outperform the extant proxy—the CINC-based capability ratio—in a number of important ways. In pure terms, the DOE score more closely relates to what international relations scholars care about: the expected outcome of a dispute between two nations. It therefore has a more natural interpretation than the capability ratio. It also lacks the *ad hoc* assumptions imposed by both the CINC score and the ratio-based transformation used in most studies. On the practical side, our replications suggest that the DOE score is a better contributor to the usual battery of variables included in the ever-expanding universe of international relations regressions. We hope, then, that it will find use as scholars advance and test new claims.

Though it represents a massive improvement over the *status quo*, the DOE score could still be improved. We have only included those variables that could be extracted from the data used to construct the capability ratio—namely, the six Correlates of War National Material Capabilities variables. We did so consciously, as we wanted to demonstrate that our method could improve measures holding the covariates fixed. Having made our point, we look forward to seeing what the future holds for coming versions of DOE when new data is brought to bear on the problem. At the risk of belaboring: we created DOE using open-source software and have made our replication code available, and so anybody with a computer—and some patience!—could create a new version with new covariates.

These improvements and opportunities for further refinement are especially exciting given the broad substantive points that we have managed to glean from our rather general analysis. Most obviously, our results demonstrate an important effect of material holdings on dispute outcomes; this is especially interesting since we use all militarized disputes, and not only full-fledged wars, as our training class. This result alone makes the flexible measurement of powers a fruitful avenue for further substantive research, and one would not have arrived at the same conclusion armed only with the capability ratio. Those interested in taking on the challenge may proceed our two more nuanced substantive findings. First, power (and its relationship with capability) should be thought of as dynamic, not static. That which makes a state strong today might be worthless tomorrow, while unforseen advances may render previously unappreciated resources vital. Second, power is inherently relational. Paraphrasing Dowding (1996, 4), it makes little sense to say "State A has power;" instead, we should focus on what State A has the power to do, which necessarily involves her social power over some State B. Our results suggest that such thinking is far more relevant than a strictly monadic approach.

On the methodological side, we believe that our data-driven approach to measurement will prove useful for those wishing to proxy for other quantities. All one needs is a set of predictor variables X and some outcome of interest Y—the procedure we provide to produce a mapping f from X to Y will work. Just as with introducing new covariates in any given application, future scholars can improve their proxies by including new models for evaluation in the super learner—the general approach remains unchanged. Our application tasked us to create a proxy of a probabilistic expectation like those seen in formal models of choice under uncertainty, and similar applications provide a natural starting point for our method. Doing so, however, requires good theory for just what it is that we hope to predict with our abstractions. As such theories continue to develop, we hope political scientists across subfields will turn their attention to examples like these as they construct new measures and improve existing ones.

We would like to conclude with a still broader point. Breiman (2001) argues that statistical modelers fall into one of two cultures: data modelers, who interpret models' estimates after assessing overall quality via in-sample goodness of fit; and algorithmic modelers, who seek algorithms that predict responses as well as possible given some set of covariates.[20] The method we advance is certainly algorithmic. Our decision to adopt algorithmic modeling based on prediction, however, was not culture-driven—it was purpose-driven (Clarke and Primo 2012). Most simply, prediction matters for measurement, so algorithmic tools should play a larger role. But as we show in the replication analysis, an algorithmically constructed proxy can be useful to include in traditional models. As new problems emerge and new solutions arise to solve them, we believe methodological pragmatism will be an important virtue. We neither expect nor encourage empirical political science to turn its focus from causal hypothesis testing to prediction. But good hypothesis testing depends on good measures, and sometimes the best way to build a measure is to assume the persona of the algorithmic modeler. By doing just that, this paper has developed one measure that improves on the previous state of the art along a

[20] In case it is not obvious from our previous citations, Breiman self-identifies as an algorithmic modeler. He claims that 98% of statisticians fall into the data modeling camp, or at least did as of 2001. We are comfortable positing that the percentage is similar, if not greater, for empirical political scientists in 2015.

number of dimensions.

References

Akaike, Hirotugu. 1974. “A New Look at the Statistical Model Identification.” *IEEE Transactions on Automatic Control* 19(6):716–723.

Arena, Phil. 2012. “Measuring Military Capabilities.” Blog post.
URL: http://fparena.blogspot.com/2012/11/once-more-on-military-capabilities.html

Arena, Philip and Glenn Palmer. 2009. “Politics or the Economy? Domestic Correlates of Dispute Involvement in Developed Democracies.” *International Studies Quarterly* 53(4):955–975.

Bennett, D. Scott. 2006. “Toward a Continuous Specification of the Democracy–Autocracy Connection.” *International Studies Quarterly* 50(2):313–338.

Breiman, Leo. 2001. “Statistical Modeling: The Two Cultures.” *Statistical Science* 16(3):199–231.

Bueno de Mesquita, Bruce. 1981. *The War Trap*. Yale University Press.

Clarke, Kevin A. and David M. Primo. 2012. *A Model Discipline: Political Science and the Logic of Representations*. Oxford, UK: Oxford University Press.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. “The Statistical Analysis of Roll Call Data.” *American Political Science Review* 98(2):355–370.

Dowding, Keith. 1991. *Rational Choice and Political Power*. Aldershot, UK: Edward Elgar.

Dowding, Keith. 1996. *Power*. Minneapolis, MN: University of Minnesota Press.

Dreyer, David R. 2010. “Issue Conflict Accumulation and the Dynamics of Strategic Rivalry.” *International Studies Quarterly* 54(3):779–795.

Efron, Bradley and Gail Gong. 1983. “A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation.” *The American Statistician* 37(1):36–48.

Fearon, James D. 1995. “Rationalist Explanations for War.” *International Organization* 49(3):379–414.

Fearon, James D. 1997. “Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs.” *Journal of Conflict Resolution* 41(1):68–90.

Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro and Dinani Amorim. 2014. “Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?” *The Journal of Machine Learning Research* 15(1):3133–3181.

Fordham, Benjamin O. 2008. “Power or Plenty? Economic Interests, Security Concerns, and American Intervention.” *International Studies Quarterly* 52(4):737–758.

Fuhrmann, Matthew and Todd S. Sechser. 2014. “Signaling Alliance Commitments: Hand-Tying and Sunk Costs in Extended Nuclear Deterrence.” *American Journal of Political Science* 58(4):919–935.

Gartzke, Erik. 2007. “The Capitalist Peace.” *American Journal of Political Science* 51(1):166–191.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Second ed. New York: Springer.

Hill, Daniel W. and Zachary M. Jones. 2014. “An Empirical Evaluation of Explanations for State Repression.” *American Political Science Review* 108(3):661–687.

Honaker, James and Gary King. 2010. “What to Do about Missing Values in Time-Series Cross-Section Data.” *American Journal of Political Science* 54(2):561–581.

Huth, Paul, Sarah Croco and Benjamin Appel. 2012. “Law and the Use of Force in World Politics: The Varied Effects of Law on the Exercise of Military Power in Territorial Disputes.” *International Studies Quarterly* 56(1):17–31.

Jackman, Simon and Shawn Treier. 2008. “Democracy as a Latent Variable.” *American Journal of Political Science* 52(1):201–217.

Jung, Sung Chul. 2014. “Foreign Targets and Diversionary Conflict.” *International Studies Quarterly* 58(3):566–578.

Keohane, Robert O. and Joseph S. Nye. 2001. *Power and Interdependence*. New York: Longman.

Kuhn, Max and Kjell Johnson. 2013. *Applied Predictive Modeling*. New York: Springer.

Linzer, Drew and Jeffrey K. Staton. 2014. "A Measurement Model for Synthesizing Multiple Comparative Indicators: The Case of Judicial Independence." Working paper.
URL: http://polisci.emory.edu/faculty/jkstato/resources/WorkingPapers/LS-scaling-140430.pdf

Maoz, Zeev. 1983. "Resolve, Capabilities, and the Outcomes of Interstate Disputes, 1816-1976." *Journal of Conflict Resolution* 27(2):195–229.

McKelvey, Richard D. and William Zavoina. 1975. "A Statistical Model for the Analysis of Ordinal Level Dependent Variables." *Journal of Mathematical Sociology* 4(1):103–120.

Molinaro, Annette M., Richard Simon and Ruth M. Pfeiffer. 2005. "Prediction Error Estimation: A Comparison of Resampling Methods." *Bioinformatics* 21(15):3301–3307.

Morrow, James D. 2007. "When Do States Follow the Laws of War?" *American Political Science Review* 101(3):559–572.

Organski, A.F.K. and Jacek Kugler. 1980. *The War Ledger*. Chicago: University of Chicago Press.

Owsiak, Andrew P. 2012. "Signing Up for Peace: International Boundary Agreements, Democracy, and Militarized Interstate Conflict." *International Studies Quarterly* 56(1):51–66.

Palmer, Glenn, Vito D'Orazio, Michael Kenwick and Matthew Lane. 2015. "The MID4 dataset, 2002–2010: Procedures, Coding Rules and Description." *Conflict Management and Peace Science* 32(2):222–242.

Park, Johann and Michael Colaresi. 2014. "Safe Across the Border: The Continued Significance of the Democratic Peace When Controlling for Stable Borders." *International Studies Quarterly* 58(1):118–125.

Poole, Keith T. and Howard Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29(2):357–384.

Powell, Robert. 1996. "Stability and the Distribution of Power." *World Politics* 48(2):239–267.

Powell, Robert. 1999. *In the Shadow of Power: States and Strategies in International Politics*. Princeton, NJ: Princeton University Press.

Powell, Robert. 2006. "War as a Commitment Problem." *International Organization* 60(1):169–203.

Reed, William, David H. Clark, Timothy Nordstrom and Wonjae Hwang. 2008. "War, Power, and Bargaining." *Journal of Politics* 70(4):1203–1216.

Salehyan, Idean. 2008*a*. "No Shelter Here: Rebel Sanctuaries and International Conflict." *Journal of Politics* 70(1):54–66.

Salehyan, Idean. 2008*b*. "The Externalities of Civil Strife: Refugees as a Source of International Conflict." *American Journal of Political Science* 52(4):787–801.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6(2):461–464.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25(3):289–310.

Singer, J. David. 1963. "Inter-nation Influence: A Formal Model." *American Political Science Review* 57(2):420–430.

Singer, J. David, Stuart Bremer and John Stuckey. 1972. Capability Distribution, Uncertainty, and Major Power War, 1820–1965. In *Peace, War, and Numbers*, ed. Bruce Russett. Beverley Hills, CA: Sage.

Sobek, David, M. Rodwan Abouharb and Christopher G. Ingram. 2006. "The Human Rights Peace: How the Respect for Human Rights at Home Leads to Peace Abroad." *Journal of Politics* 68(3):519–529.

Spector, Paul E. 1992. *Summated Rating Scale Construction: An Introduction*. Quantitative Applications in the Social Sciences Newbury Park, CA: Sage.

Tibshirani, Ryan J. and Robert Tibshirani. 2009. "A Bias Correction for the Minimum Error Rate in Cross-Validation." *The Annals of Applied Statistics* 3(2):822–829.

Uzonyi, Gary, Mark Souva and Sona N Golder. 2012. "Domestic Institutions and Credible Signals." *International Studies Quarterly* 56(4):765–776.

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):91.

Vuong, Quang H. 1989. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* 57(2):307–333.

Waltz, Kenneth N. 1979. *Theory of International Politics*. Boston, MA: McGraw Hill.

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.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand and Dan Steinberg. 2007. "Top 10 Algorithms in Data Mining." *Knowledge and Information Systems* 14(1):1–37.

Zawahri, Neda A. and Sara McLaughlin Mitchell. 2011. "Fragmented Governance of International Rivers: Negotiating Bilateral versus Multilateral Treaties." *International Studies Quarterly* 55(3):835–858.

A Appendix

A.1 National Material Capabilities Data

Our predictors are taken from the National Material Capabilities (v4.0) dataset from the Correlates of War project (Singer, Bremer and Stuckey 1972).[21] The dataset contains observations on six variables for 14,199 country-years from 1816 to 2007. For details on the variables and their measurement, see the NMC Codebook.[22] Table 5 lists the proportions of zeroes and missing values among each variable.

Component	Pr(Zero)	Pr(Missing)	θ
Iron and Steel Production	0.558	0.006	2^{-5}
Military Expenditures	0.034	0.139	2^{-7}
Military Personnel	0.066	0.027	2^{-1}
Primary Energy Consumption	0.097	0.030	2^{-3}
Total Population	0.000	0.002	2^{-7}
Urban Population	0.210	0.007	2^{-8}

Table 5. Proportions of zeroes and missing values in each National Military Capability component variable.

All six variables are strongly right-skewed. Since five of the six variables are sometimes zero-valued (though all are non-negative), a logarithmic transformation is not appropriate. Instead, to correct for skewness, we apply an inverse hyperbolic sine transformation (Burbidge, Magee and Robb 1988) to each component:

$$h(x,\theta) = \sinh^{-1}(\theta x) = \log\left(\theta x + \sqrt{(\theta x)^2 + 1}\right). \tag{5}$$

We set the scale θ separately for each component variable with the aim of making the transformed variable approximately normally distributed. For each variable, we choose the value of $\theta \in \{2^d\}_{d=-10}^{10}$ that minimizes the Kolmogorov–Smirnov test statistic (Massey Jr 1951) against a normal distribution with the

[21] Downloaded from http://correlatesofwar.org/data-sets/national-material-capabilities/nmc-v4-data/at_download/file.

[22] Available at http://correlatesofwar.org/data-sets/national-material-capabilities/nmc-codebook/at_download/file.

same mean and variance. Table 5 gives the scale selected for each component. We use the transformed components in both the multiple imputation (see below) and the super learner training.

A.2 Militarized Interstate Dispute Data

Our sample and outcome variable are taken from the Militarized Interstate Disputes (v4.1) dataset from the Correlates of War project (Palmer et al. 2015).[23] The dataset records the participants and outcomes of interstate disputes from 1816 to 2010. To avoid the problem of aggregating capabilities across multiple states, we exclude disputes with more than one state on either side. We drop disputes that end in an outcome other than one side winning, one side yielding, or a stalemate;[24] we then collapse "A Wins" and "B Yields" into a single coding, and similarly for "B Wins" and "A Yields." Finally, since the capabilities data only run through 2007, we exclude disputes that end after 2007. In the end, we have $N = 1{,}740$ cases.

For each dispute in our dataset, we code the participating countries' capabilities using the values in the year the dispute began. About 17 percent of disputes have at least one missing capability component for at least one participant.

A.3 Multiple Imputation

As noted above, all of the National Material Capabilities variables contain some missing values. Following standard practice, we multiply impute the missing observations. We perform the imputations via the `Amelia` software package (Honaker, King and Blackwell 2011).

Rather than just impute the missing values in the final dataset of disputes, we impute the entire National Material Capabilities dataset. This allows us to fully exploit the dataset's time-series cross-sectional structure in the imputation process (Honaker and King 2010). We include in the imputation model a cubic polynomial for time, interacted with country dummy variables. As this results in an explosion in the number of parameters in the imputation model, we then impose a ridge prior equal to 0.1 percent of the observations in the dataset

[23] Downloaded from http://correlatesofwar.org/data-sets/MIDs/mid-level/at_download/file.

[24] For details on other kinds of outcomes, see the MID Codebook.

(see Section 4.7.1 of the `Amelia` package vignette). We enforce the constraint that every imputed value be non-negative. Finally, we impose an observation-level prior with mean zero and variance equal to that of the observed values of the corresponding component variable for every missing cell that meets the following criteria:

- There are no non-zero observed values in the time series preceding the cell
- The first observed value that comes after the cell is zero

So, for example, if a country's urban population is zero from 1816 to 1840, missing from 1841 to 1849, and zero in 1850, we would impose this form of prior on the 1841–1849 values. Diagnostic time series plots of observed versus imputed values within each data series, generated by the `tscsPlot()` function in `Amelia`, will be made available in the project's Dataverse.

The presence of missing data also complicates the calculations of country-by-country proportions of the total amount of each component by year. One option is to recompute the annual totals in each imputed dataset, so that the resulting data will be logically consistent—in particular, all proportions will sum to one. The drawback of this approach is that virtually every observation of the proportions will differ across the imputed datasets, even for countries with no missing data, since the annual totals will differ across imputations. An alternative approach is to compute the annual totals using only the observed values. The advantage is that non-missing observations will not vary across imputed datasets; the downside is that the proportions within each imputation will generally sum to more than one. For our purposes in this paper, we think it is preferable to reduce variation across imputations, even at the expense of some internal consistency in the imputed datasets, so we take the latter approach: annual totals are the sums of only the observed values.

We impute $I = 10$ datasets of national capabilities according to the procedure laid out above, and we merge each with the training subset of our dispute data to yield I training data imputations. We run the super learner separately on each imputation, and our final model is an (unweighted) average of the I super learners.

After training is complete, we run into missing data problems once again when calculating DOE scores. To calculate predicted probabilities for dyads with missing values, we calculate a *new* set of $I = 10$ imputations of the capabilities data and take an (unweighted) average of our model's predictions across the imputations.

A.4 Super Learner Candidate Models

We use the R statistical environment (R Core Team 2015) for all data analysis. We fit, cross-validate, and calculate predictions from each candidate model through the `caret` package (Kuhn 2008). We then construct the super learner by solving (4) via base R's `constrOptim()` function for optimization with linear constraints. Further details about each candidate model are summarized below.

- Ordered Logit (McKelvey and Zavoina 1975)

  Package `MASS` (Venables and Ripley 2002)

  Tuning Parameters None

  Notes In the "Year" models, the year of the dispute is included directly and interacted with each capability variable

- C5.0 (Quinlan 2015)

  Package `C50` (Kuhn et al. 2015)

  Tuning Parameters

  - Number of boosting iterations (`trials`): selected via cross-validation from $\{1, 10, 20, 30, 40, 50\}$
  - Whether to decompose the tree into a rule-based classifier (`model`): selected via cross-validation
  - Whether to perform feature selection (`winnow`): selected via cross-validation

- Support Vector Machine (Cortes and Vapnik 1995)

  Package `kernlab` (Karatzoglou et al. 2004)

  Tuning Parameters

  - Kernel width (`sigma`): selected via cross-validation from $\{0.2, 0.4, 0.6, 0.8, 1\}$

  - Constraint violation cost (`C`): selected via cross-validation from $\{\frac{1}{4}, \frac{1}{2}, 1, 2, 4\}$

  Notes Radial basis kernel

- k-Nearest Neighbors (Cover and Hart 1967)

  Package `caret` (Kuhn 2008)

  Tuning Parameters

  - Number of nearest neighbors to average (`k`): selected via cross-validation from $\{25, 50, \ldots, 250\}$

  Notes All predictors centered and scaled to have zero mean and unit variance

- CART (Breiman et al. 1984)

  Package `rpart` (Therneau, Atkinson and Ripley 2015)

  Tuning Parameters

  - Maximum tree depth (`maxdepth`): selected via cross-validation from $\{2, 3, \ldots, 9, 10\}$ (only up to 9 for models without year included)

- Random Forest (Breiman 2001)

  Package `randomForest` (Liaw and Wiener 2002)

  Tuning Parameters

  - Number of predictors randomly sampled at each split (`mtry`): selected via cross-validation from $\{2, 4, \ldots, 12\}$

  Notes 1,000 trees per fit

- Averaged Neural Nets (Ripley 1996)

  Package `nnet` (Venables and Ripley 2002), `caret` (Kuhn 2008)

  Tuning Parameters

  - Number of hidden layer units (`size`): selected via cross-validation from $\{1, 3, 5, 7, 9\}$
  - Weight decay parameter (`decay`): selected via cross-validation from $\{10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$

  Notes Creates an ensemble of 10 neural nets, each initialized with different random number seeds

A.5 Replications

The following list contains basic information about each model in the replication study. We carry out logistic and probit regressions via `glm()` in base R (R Core Team 2015), multinomial logit via `multinom()` in the `nnet` package (Venables and Ripley 2002), ordered probit via `polr()` in the `MASS` package (Venables and Ripley 2002), and heteroskedastic probit via `hetglm()` in the `glmx` package (Zeileis, Koenker and Doebler 2013).

- Arena and Palmer (2009)

  Model Replicated Table 3

  Unit of Analysis Directed Dyads

  Estimator Heteroskedastic Probit

  CINC Terms CINC_A

  DOE Terms p_A, p_B

  Notes CINC and DOE terms are included in both the mean and dispersion equations.

- Bennett (2006)

  Model Replicated Table 1, Column 1

  Unit of Analysis Directed Dyads

  Estimator Logistic Regression

  CINC Terms CINC_A, CINC_B, $\text{CINC}_{\text{min}} / \text{CINC}_{\text{max}}$

  DOE Terms p_A, p_B, $|p_A - p_B|$

- Dreyer (2010)

  Model Replicated Table 2, Model 2

  Unit of Analysis Undirected Dyads

  Estimator Logistic Regression

  CINC Terms $\log(\text{CINC}_{\text{min}} / \text{CINC}_{\text{max}})$

  DOE Terms $\log p_{\text{min}}$, $\log p_{\text{max}}$

- Fordham (2008)

Model Replicated Table 2, third column (alliance onset with full set of controls)

Unit of Analysis Undirected Dyads

Estimator Probit Regression

CINC Terms $\log \mathrm{CINC}_{\mathrm{US}}$, $\log \mathrm{CINC}_2$

DOE Terms $\log p_{\mathrm{US}}$, $\log p_2$

- Fuhrmann and Sechser (2014)

  Model Replicated Table 2, Model 1

  Unit of Analysis Directed Dyads

  Estimator Probit Regression

  CINC Terms $\mathrm{CINC}_A / (\mathrm{CINC}_A + \mathrm{CINC}_B)$

  DOE Terms p_A, p_B

- Gartzke (2007)

  Model Replicated Table 1, Model 4

  Unit of Analysis Undirected Dyads

  Estimator Logistic Regression

  CINC Terms $\log(\mathrm{CINC}_{\max} / \mathrm{CINC}_{\min})$

  DOE Terms $\log p_{\min}$, $\log p_{\max}$

- Huth, Croco and Appel (2012)

  Model Replicated Table 2

  Unit of Analysis Directed Dyads

  Estimator Multinomial Logistic Regression

  CINC Terms Average of *A*'s respective shares of total dyadic military personnel, military expenditures, and military expenditures per soldier

  DOE Terms p_A, p_B

- Jung (2014)

  Model Replicated Table 1, Model 1

Unit of Analysis Directed Dyads

Estimator Logistic Regression

CINC Terms $\text{CINC}_A / (\text{CINC}_A + \text{CINC}_B)$

DOE Terms p_A, p_B

- Morrow (2007)

  Model Replicated Table 1, first column (no weighting for data quality)

  Unit of Analysis Directed Dyads

  Estimator Ordered Probit Regression

  CINC Terms $\text{CINC}_A / (\text{CINC}_A + \text{CINC}_B)$, interaction with joint ratification

  DOE Terms p_A, p_B, interactions of each with joint ratification

  Notes Capability ratio is "corrected for distance to the battlefield and aggregated across actors with a unified command." We drop the cases with coalitional actors in both models, hence the difference in sample size from the original article. No distance correction is applied to the DOE scores.

- Owsiak (2012)

  Model Replicated Table 3, Model 3

  Unit of Analysis Undirected Dyads

  Estimator Logistic Regression

  CINC Terms $\log(\text{CINC}_{\text{min}} / \text{CINC}_{\text{max}})$

  DOE Terms $\log p_{\text{min}}$, $\log p_{\text{max}}$

- Park and Colaresi (2014)

  Model Replicated Table 1, Model 2

  Unit of Analysis Undirected Dyads

  Estimator Logistic Regression

  CINC Terms $\text{CINC}_{\text{min}} / \text{CINC}_{\text{max}}$, interaction with contiguity

  DOE Terms $|p_A - p_B|$, interaction with contiguity

- Salehyan (2008*a*)

Model Replicated Table 1, Model 1

Unit of Analysis Undirected Dyads

Estimator Logistic Regression

CINC Terms $\log(\text{CINC}_{\max}/(\text{CINC}_{\max}+\text{CINC}_{\min}))$

DOE Terms $\log p_{\min}$, $\log p_{\max}$

- Salehyan (2008*b*)

  Model Replicated Table 1, Model 2

  Unit of Analysis Directed Dyads

  Estimator Probit Regression

  CINC Terms $\text{CINC}_A/(\text{CINC}_A+\text{CINC}_B)$, interaction with refugee stock in *A*, interaction with refugee stock from *A*

  DOE Terms p_A, p_B, interaction of each with refugee stock in *A*, interaction of each with refugee stock from *A*

- Sobek, Abouharb and Ingram (2006)

  Model Replicated Table 1, first row (political prisoners model)

  Unit of Analysis Undirected Dyads

  Estimator Logistic Regression

  CINC Terms $(\text{CINC}_{\max}-\text{CINC}_{\min})/(\text{CINC}_{\max}+\text{CINC}_{\min})$

  DOE Terms $p_{\min}$, $p_{\max}$

- Uzonyi, Souva and Golder (2012)

  Model Replicated Table 3, Model 3

  Unit of Analysis Directed Dyads

  Estimator Logistic Regression

  CINC Terms $\text{CINC}_A/(\text{CINC}_A+\text{CINC}_B)$

  DOE Terms p_A, p_B

- Weeks (2008)

  Model Replicated Table 4, Model 3

Unit of Analysis Directed Dyads

Estimator Logistic Regression

CINC Terms $\text{CINC}_A / (\text{CINC}_A + \text{CINC}_B)$

DOE Terms p_A, p_B

- Weeks (2012)

  Model Replicated Table 1, Model 2

  Unit of Analysis Directed Dyads

  Estimator Logistic Regression

  CINC Terms CINC_A, CINC_B, $\text{CINC}_A / (\text{CINC}_A + \text{CINC}_B)$

  DOE Terms p_A, p_B

- Zawahri and Mitchell (2011)

  Model Replicated Table 2, Model 1

  Unit of Analysis Directed Dyads

  Estimator Logistic Regression

  CINC Terms CINC_A, CINC_B

  DOE Terms p_A, p_B

  Notes Dyads are directed, but A is the upstream state in a river basin rather than the (prospective) initiator of conflict, so we use the undirected form of the DOE scores.

- Hafner-Burton and Montgomery (2006)

  Model Replicated Table 1, Model 2

  Unit of Analysis Undirected Dyads

  Estimator Logistic Regression

  CINC Terms $\log(\text{CINC}_{\max} / \text{CINC}_{\min})$

  DOE Terms $\log p_{\min}$, $\log p_{\max}$

Additional References

Arena, Philip and Glenn Palmer. 2009. "Politics or the Economy? Domestic Correlates of Dispute Involvement in Developed Democracies." *International Studies Quarterly* 53(4):955–975.

Bennett, D. Scott. 2006. "Toward a Continuous Specification of the Democracy–Autocracy Connection." *International Studies Quarterly* 50(2):313–338.

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

Breiman, Leo, Jerome Friedman, Charles J. Stone and R. A. Olshen. 1984. *Classification and Regression Trees*. Chapman and Hall/CRC.

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.

Cortes, Corinna and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20(3):273–297.

Cover, T. M. and P. E. Hart. 1967. "Nearest Neighbor Pattern Classification." *IEEE Transactions on Information Theory* 13(1):21–27.

Dreyer, David R. 2010. "Issue Conflict Accumulation and the Dynamics of Strategic Rivalry." *International Studies Quarterly* 54(3):779–795.

Fordham, Benjamin O. 2008. "Power or Plenty? Economic Interests, Security Concerns, and American Intervention." *International Studies Quarterly* 52(4):737–758.

Fuhrmann, Matthew and Todd S. Sechser. 2014. "Signaling Alliance Commitments: Hand-Tying and Sunk Costs in Extended Nuclear Deterrence." *American Journal of Political Science* 58(4):919–935.

Gartzke, Erik. 2007. "The Capitalist Peace." *American Journal of Political Science* 51(1):166–191.

Hafner-Burton, Emilie M and Alexander H Montgomery. 2006. "Power Positions International Organizations, Social Networks, and Conflict." *Journal of Conflict Resolution* 50(1):3–27.

Honaker, James and Gary King. 2010. "What to Do about Missing Values in Time-Series Cross-Section Data." *American Journal of Political Science* 54(2):561–581.

Honaker, James, Gary King and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45(7):1–47.
URL: `http://www.jstatsoft.org/v45/i07/`

Huth, Paul, Sarah Croco and Benjamin Appel. 2012. "Law and the Use of Force in World Politics: The Varied Effects of Law on the Exercise of Military Power in Territorial Disputes." *International Studies Quarterly* 56(1):17–31.

Jung, Sung Chul. 2014. "Foreign Targets and Diversionary Conflict." *International Studies Quarterly* 58(3):566–578.

Karatzoglou, Alexandros, Alex Smola, Kurt Hornik and Achim Zeileis. 2004. "kernlab – An S4 Package for Kernel Methods in R." *Journal of Statistical Software* 11(9):1–20.
URL: `http://www.jstatsoft.org/v11/i09/`

Kuhn, Max. 2008. "Building Predictive Models in R Using the `caret` Package." *Journal of Statistical Software* 28(5):1–26.
URL: `http://www.jstatsoft.org/v28/i05`

Kuhn, Max, Steve Weston, Nathan Coulter, Mark Culp and Ross Quinlan. 2015. *C50: C5.0 Decision Trees and Rule-Based Models*. R package version 0.1.0-24.
URL: `http://CRAN.R-project.org/package=C50`

Liaw, Andy and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2(3):18–22.
URL: `http://CRAN.R-project.org/doc/Rnews/`

Massey Jr, Frank J. 1951. "The Kolmogorov-Smirnov Test for Goodness of Fit." *Journal of the American Statistical Association* 46(253):68–78.

McKelvey, Richard D. and William Zavoina. 1975. "A Statistical Model for the Analysis of Ordinal Level Dependent Variables." *Journal of Mathematical Sociology* 4(1):103–120.

Morrow, James D. 2007. "When Do States Follow the Laws of War?" *American Political Science Review* 101(3):559–572.

Owsiak, Andrew P. 2012. "Signing Up for Peace: International Boundary Agreements, Democracy, and Militarized Interstate Conflict." *International Studies Quarterly* 56(1):51–66.

Palmer, Glenn, Vito D'Orazio, Michael Kenwick and Matthew Lane. 2015. "The MID4 dataset, 2002–2010: Procedures, Coding Rules and Description." *Conflict Management and Peace Science* 32(2):222–242.

Park, Johann and Michael Colaresi. 2014. "Safe Across the Border: The Continued Significance of the Democratic Peace When Controlling for Stable Borders." *International Studies Quarterly* 58(1):118–125.

Quinlan, Ross. 2015. "Data Mining Tools See5 and C5.0." RuleQuest website.
URL: https://www.rulequest.com/see5-info.html

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: http://www.R-project.org/

Ripley, Brian D. 1996. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Salehyan, Idean. 2008*a*. "No Shelter Here: Rebel Sanctuaries and International Conflict." *Journal of Politics* 70(1):54–66.

Salehyan, Idean. 2008*b*. "The Externalities of Civil Strife: Refugees as a Source of International Conflict." *American Journal of Political Science* 52(4):787–801.

Singer, J. David, Stuart Bremer and John Stuckey. 1972. Capability Distribution, Uncertainty, and Major Power War, 1820–1965. In *Peace, War, and Numbers*, ed. Bruce Russett. Beverley Hills, CA: Sage.

Sobek, David, M. Rodwan Abouharb and Christopher G. Ingram. 2006. "The Human Rights Peace: How the Respect for Human Rights at Home Leads to Peace Abroad." *Journal of Politics* 68(3):519–529.

Therneau, Terry, Beth Atkinson and Brian Ripley. 2015. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-10.
URL: http://CRAN.R-project.org/package=rpart

Uzonyi, Gary, Mark Souva and Sona N Golder. 2012. "Domestic Institutions and Credible Signals." *International Studies Quarterly* 56(4):765–776.

Venables, W. N. and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth ed. New York: Springer. ISBN 0-387-95457-0.
URL: http://www.stats.ox.ac.uk/pub/MASS4

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.

Zawahri, Neda A. and Sara McLaughlin Mitchell. 2011. "Fragmented Governance of International Rivers: Negotiating Bilateral versus Multilateral Treaties." *International Studies Quarterly* 55(3):835–858.

Zeileis, Achim, Roger Koenker and Philipp Doebler. 2013. *glmx: Generalized Linear Models Extended*. R package version 0.1-0.
URL: http://CRAN.R-project.org/package=glmx