Making Useful Conflict Predictions:
Methods for Addressing Skewed Classes and Implementing Cost-Sensitive Learning in the Study of State Failure

Ryan Kennedy
University of Houston/Northeastern University/Harvard University
rkennedy@uh.edu

Word Count: 11,106
Page Count: 41

Keywords: state failure, case-control studies, class skew, rare events, cost-sensitive learning, political forecasting.

Abstract: One of the major issues in predicting state failure is the relatively rare occurrence of event onset. This class skew problem can cause difficulties in both estimating a model and selecting a decision boundary. Since the publication of King and Zeng's (2001) study, scholars have utilized case-control methods to address this issue. This paper re-analyzes the landmark research of the Political Instability Task Force (Goldstone et al. 2010), comparing the case-control approach to several other methods from the machine learning field and some original to this study. Case-control methods are outperformed by almost all of the alternatives. A multilevel model on the raw data performs best. The article also introduces cost sensitive methods for determining a decision boundary. This explication reveals problems in the Task Force's formulation of a decision boundary and suggests methods for making useful predictions for policy.

Jacqueline Stevens (2012) caused a firestorm of controversy when she labeled political scientists "lousy forecasters" in the *New York Times*, and included an illustration suggesting that political scientists were comparable to monkeys throwing darts. In response, scholars argued that she had misinterpreted several of the key studies cited in the article, and that the preeminent goal of political science is understanding, and not prediction per se (e.g. Farrell 2012; Johnson 2012).

Both Stevens and her critics missed several remarkable, if rare, successful quantitative forecasting enterprises in political science. In American politics, Martin et al. (2004) were able to construct a simple model of Supreme Court decisions that predicted case outcomes better than legal experts. Models to predict US elections are a consistent subject of inquiry, often perform as well as or better than pre-election polls, and are frequently used by news and financial organizations (e.g. Campbell 2012; Lewis-Beck and Tien 2012; Klarner 2012).

In international relations, the US government has invested heavily in early detection of political crises. Andriole and Young's (1977) crisis detection system was reportedly part of Ronald Reagan's daily brief. Bueno de Mesquita's (1981) expected utility model (called Policon and later updated to Senturion) has been utilized by the CIA and private companies to predict numerous events.[1] More recently, the Political Instability Task Force (PITF, Goldstone et al. 2010) and Integrated Crisis Early Warning System (ICEWS, O'Brien 2010) have received generous support from the CIA and Department of Defense, reporting impressive results. PITF, which was initiated in 1994 by then Vice President Al Gore, recently reported successful prediction in nearly 86% of out-of-sample state failures (Goldstone et al. 2010: 201). A more

[1] Feder (2002) reports that the CIA found the Policon model was correct in 90% of the real-world applications for which the CIA used it, but, as Bueno de Mesquita acknowledges, the process for arriving at these accuracy estimates are unclear.

recent attempt at predicting incidences of genocide reports success in predicting 90.9% of genocide onsets correctly (Goldsmith et al. 2013).

Moreover, some influential scholars have made the case that prediction should play a more important role in political science than either Stevens or her critics allow. Ward et al. (2010), in their re-analysis of two influential studies on civil conflict (Fearon and Laitin 2003; Collier and Hoeffler 2004), persuasively argue that the traditional focus on statistical significance may emphasize variables that have little substantive significance. In addition, they point out that substantive interpretations of probabilistic models are, in fact, predictions. Schrodt (Forthcoming: 4) uses stronger words: "[E]xplanation in the absence of prediction is not scientifically superior to predictive analysis, it isn't scientific at all! It is, instead, 'pre-scientific'." In response, several ambitious efforts at predicting have begun to appear in major international relations journals (e.g. Hegre et al. 2013; Goldsmith et al. 2013; Gleditsch and Ward 2013). Yet, prediction remains under-emphasized in political science, and many of the basic lessons from more prediction-oriented fields, like machine learning, are rarely addressed.

This study looks at one of the major issues in political science prediction, the relatively rare occurrence of certain important phenomena, within the context of a re-analysis of PITF's global model for forecasting state failure (Esty et al. 1995; Esty et al. 1998; Esty et al. 1999; Goldstone et al. 2000; King and Zeng 2001b; Bates et al. 2003; Goldstone et al. 2010). In the machine learning field, this issue of rare events is referred to as the skewed class problem. The social sciences face this problem in studies of everything from incidence of warfare to changes in electoral systems. Skewed classes can produce two inter-related maladies. First, models can have low overall error rates while still not producing useful or interesting results. For example, a naïve model that always predicts state failure will not take place will be correct about 95 percent of the

time. Yet, a policy-maker would likely prefer a model that has more overall error if it could correctly predict the onset of state failure in some cases. We demonstrate that the formulation of decision boundaries – the probability above which we predict the event will occur – is inseparable from expected utility analysis (Elkan 2001; Abu-Mostafa et al. 2012: 28-30). Cost-sensitive learning – using the cost of outcomes to formulate an optimum decision boundary – is explained and demonstrated in PITF's out-of-sample estimates. We find that PITF's decision to use quartiles to form a decision boundary would only be preferred to a higher threshold, which predicts fewer state failure cases correctly, if false positives were only about 2.5 percent the cost of true positives. Moreover, this threshold would only be preferred to a completely naïve threshold, which always predicts state failure will not occur, if false positives were only about 12 percent of the cost of true positives. This exhibits how the class skew problem can produce very impressive prediction accuracy, while still falling short of policy-maker needs and masking the need for further work. The results are not just relevant for PITF's results, but demonstrate that traditional methods for drawing decision boundaries in political science, which are often arbitrary, are heavily affected by class skew and should be informed by their policy implications.

Second, when certain events are particularly rare and there is substantial mixing between outcomes, formulating a generalizable decision boundary can be difficult. Case-control methods, a specialized version of random under-sampling, can be useful and are the default in some applications, but, as King and Zeng (2001: 707) note, the decision of how to deal with class skew "depends on the goals of the particular application." Which of these methods are best for generating out-of-sample predictions is an open question, and depends on the characteristics of the data being analyzed (Kubat and Matwin 1997; Batista et al. 2004; Japkovicz 2000; Weiss and Provost 2001). Our analysis of a reconstructed version of PITF's data demonstrates that case-

control methods are sub-optimal in this case. The case-control method is outperformed by almost all the alternative machine learning algorithms and is substantially worse than using the full dataset for out-of-sample prediction. In fact, the best model uses a simple multilevel logit on the full dataset. This study introduces new methods for handling skewed classes in political science data and demonstrates that scholars should not default to any one particular method. Indeed, in this application, the leverage from using the full data trumps the more methodologically sophisticated and difficult to implement methods of data cleaning, producing better decision boundaries. Our application also impacts the machine learning literature, since the performance of these methods in large-scale machine learning problems like spam detection and character recognition is likely quite different from performance in the relatively small-N and time-series cross-section (TSCS) data common in political science.

What is Our Goal?

Political scientists usually assign low priority to prediction, preferring to focus on the statistical significance of particular variables. Yet, to analyze the substantive importance of variables, we often utilize the language of prediction. Whenever a study states that a certain increase/decrease in one variable leads to an increase/decrease in another variable within particular bounds, it makes a prediction about the likely effect on an out-of-sample subject (i.e. the universe of cases).

When model predictions are analyzed, scholars usually utilize the same sample to formulate and test the model's predictions (via R^2 statistics). This leads to several maladies, including over-fitting the model to the particular data or using time-lagged dependent variables, producing well-fitted but substantively uninteresting results. As O'Brien (2010: 98) argues,

"Though such a naïve model may retrospectively achieve acceptable levels of overall performance, it is useless for real world applications."

There are, of course, subtle distinctions in the role that prediction plays in different fields. This study draws significant insight from the field of machine learning – broadly defined as the study of how machines can be programmed to draw conclusions about unobserved data from observed data (Abu-Mustafa et al. 2012: 14-15). This concentration on application to out-of-sample data in machine learning stands in contrast to econometrics, which generally views prediction as separate from and subsequent to model estimation (Gujarati and Porter 2009: 3); statistics, which makes more restrictive assumptions and deals with less general models; and data mining, which focuses on finding patterns in large relational databases (Abu-Mustafa et al. 2012: 14-15). This does not mean that machine learning tools cannot be used for studies which do not explicitly involve out-of-sample prediction. Indeed, some specific methods from machine learning have even found their way into the canonical experimental literature in political science (see e.g. Gerber and Green 2012: 310; Green and Kern 2012). Political science, however, has only recently begun to appreciate the importance of techniques specifically designed for prediction on unobserved data.

For prediction to be useful, it must be done on a different sample from the one used to formulate the model (Alpaydin 2010: 38-39; Abu-Mostafa et al. 2012: 59-60). There must be a division between the "training" data – used to build the model – and the "testing" data, used to evaluate the model. Intuitively, the practice of using the same data to both formulate and test a model is somewhat akin to giving the answer key to a student along with the test.

Out-of-sample prediction also addresses a more fundamental issue. Political scientists usually distinguish between internal and external validity based on Campbell's (1957)

dichotomy. Internal validity deals with how accurate the results are in the target population, while external validity refers to how generalizable the results are for observations outside the target population. Often, however, external validity is conflated with what Cook and Campbell (1979) later described as statistical validity, a part of internal validity dealing with whether the results are sizable and statistically significant in the population of interest. External validity, in contrast, deals with whether the results developed in one set of data can be applied to new data drawn from other circumstances, times, or places. While internal validity is necessary for generalization (see e.g. Morton and Williams 2010), conflating it with statistical validity leads analysts to assume, when their samples are drawn randomly from or include the full population of cases, their results are also externally valid. Ward et al. (2010) demonstrate this is incorrect – statistically significant variables in census datasets may add little to in-sample or out-of-sample prediction, and, in some cases, make prediction worse. These problems are likely to be especially severe in areas where the sample includes data from a set of non-randomly selected cases or where we are technically analyzing "census" data. This situation is common in the fields of comparative politics and international relations, where "samples" are actually census data of, for example, all interstate wars (Sarkees and Wayman 2010), all competitive elections (Hyde and Marinov 2012), all independent country-years (Ross 2006), or a subset of all cases for which there is data (King et al. 2001). In these cases, there is no true "sample" from a larger "population," rendering our traditional explanations of statistical significance quite awkward (Jackman 2009: xxxii). A p-value of 0.01 indicates that we would get non-zero results in 99 percent of new samples from the population, but there is no larger population from which to draw more samples. This leads to discussions of hypothetical alternative realities, which have little philosophical or statistical foundation (see e.g. Fair 2011). From this perspective, out-of-

sample prediction is not a luxury for those primarily interested in forecasting, it is a fundamental aspect of external validity in large, non-sample based comparative politics and international relations data sets.

Once we have out-of-sample predictions, how do we evaluate the accuracy of the model? For prediction on a binary dependent variable, the four possible outcomes are displayed in a "confusion matrix." Table 1 shows the confusion matrix with the different types of errors and accuracy measures labeled.

Table 1: Confusion Matrix

	Actual Positive	**Actual Negative**	
Predict Positive	True Positive	False Positive	Precision (Positive Predictive Value) = $\frac{\sum True\ Positive}{\sum Predict\ Positive}$
Predict Negative	False Negative	True Negative	Negative Predictive Value = $\frac{\sum True\ Negative}{\sum Predict\ Negative}$
	Sensitivity (Recall) = $\frac{\sum True\ Positive}{\sum Actual\ Positive}$	Specificity = $\frac{\sum True\ Negative}{\sum Actual\ Negative}$	

The accuracy of predictions can be evaluated by the ratios in the margins. Political scientists, however, are rarely trained in evaluating them. Goldstone et al. (2010), for example, only evaluate the sensitivity and specificity of their out-of-sample predictions – implicitly ignoring the relative cost of false positives to true positives. As demonstrated below, highly skewed classes can give us very high values of sensitivity and specificity, but still make dramatic assumptions about the relative costs of true and false positives.

In sum, we argue that the choice of modeling strategy is inseparable from the evaluation of predictive success. This study deals with both ends of this equation, evaluating data cleaning techniques used in estimation and the methods by which scholars draw the boundaries for their predictions to evaluate their success.

The Problem of Skewed Classes, Part 1: Evaluating Success

In many fields, we encounter data where the event of interest is relatively rare. Credit card fraud, loan defaults, exotic illness, onset of civil war, and incidence of warfare are but a few examples. These rare events are often important and their prediction valuable. The machine learning literature calls this the problem of "skewed classes" – when one category has many more examples of occurrence than another in the training and test data.

Skewed classes pose several problems. Most basically, researchers are forced to modify their definition of successful prediction. If we have bivariate data where 98 percent of the data have a value of 0 and 2 percent a value of 1, a model which predicts $P(Y=0|X) = 1$ will be correct 98 percent of the time. Even worse, changing the decision boundary to correctly predict some of the true positives will often perform worse than the naïve model.

Still more problematic, occurrences of the rarer class are often especially costly. Cases of securities fraud or particular illnesses may be costly enough that we are willing to accept a lot of false positives to correctly predict their onset. Similarly, to predict the occurrence of war or state failure, policy-makers will likely tolerate some false positives to allow preventative action in risky cases. How tolerant the consumer of a predictive finding will be depends on the utility associated with true positives and false positives. As demonstrated further below, when classes are severely skewed, we can have very high sensitivity and specificity, while still having a very

large proportion of false positives. Getting a high percentage of cases correctly classified is not a sufficient condition for evaluating model success in skewed classes. Indeed, it is a relatively weak criterion, and models can be highly accurate but still not be useful for policy-making.

Generally speaking, there are two ways to address the issue of rare but important outcomes. The first is related to the data pre-processing methods we explore below. Essentially, a researcher can under-sample from the majority case or use synthetic cases to over-sample the minority case, making the two classes more equal or reflecting their relative costs (Elkan 2001; Leetaru 2012; Witten et al. 2011: 167-168). While this method is often utilized, we demonstrate below that such pre-processing can significantly reduce the performance of the model. It is also heuristically less useful than the alternative. We leave it aside for this article and point the interested reader to the above-cited sources.

The other option is to change the level of probability at which we predict the minority case event will occur. Several prominent studies have bemoaned the fact that no empirical model has produced a predicted probability of about 0.5 for certain types of conflict (e.g. Beck et al. 2000; O'Brien 2010). Others have argued that requiring a 0.5 decision boundary is arbitrary and the low predicted probabilities are an accurate reflection of events that are quite rare (e.g. De Marchi et al. 2004; Ward et al. 2010). This study sides with the latter group, arguing, for example, that a 30 percent chance of state failure onset would likely be enough that a policy-maker would take some type of action. A common illustration of such decision-making is as follows: if you found out that you had a 1 percent chance of dying today, it would likely be enough to change your behavior, given the loss of utility associated with death and the degree to which 1 percent exceeds the standard probability of death on any particular date.

These examples illustrate that the choice of where to draw the decision boundary should not be determined by arbitrary standards, but should be determined by the utility of that decision boundary. Table 2 changes the confusion matrix into a cost matrix by inserting the costs for each outcome in the associated cell.

Table 2: Prediction Cost Matrix

	Actual Positive	Actual Negative
Predict Positive	c_{11}	c_{10}
Predict Negative	c_{01}	c_{00}

From the cost matrix in Table 2, a decision boundary for predicting a positive outcome, will be optimal if and only if

$$p(c_{11}) + (1-p)(c_{10}) \leq p(c_{01}) + (1-p)(c_{00}) \qquad (1)$$

where *p* is the probability of an actual positive, given the set of predictor variables, $P(Y = 1|X)$. The right-hand side of this equation is the expected cost of predicting a positive outcome and the left-hand side is the expected cost of predicting a negative outcome. From this, the optimal decision boundary, p^*, is calculated as

$$p^* = \frac{c_{10}-c_{00}}{c_{10}-c_{00}+c_{01}-c_{11}}. \qquad (2)$$

Traditionally, bivariate models in political science usually draw a boundary at $p^* = 0.5$, which predicts a positive outcome when it is more likely than a negative outcome. This is the result we get if all incorrect predictions are equal in cost ($c_{10} = c_{01}$) and all correct predictions are equal in cost ($c_{11} = c_{00}$). The reader can verify that, in these circumstances, the denominator will always be twice the numerator.

This is rarely the case. If, for example, the cost of a false positive is especially high, we would want to choose a higher value of p^*, as reflected in equation 2. For example, in many legal systems, where a guilty verdict in a criminal case can deprive a person of their liberty, the standard for guilt is "beyond a reasonable doubt" (i.e. p^* approaches 1). In contrast, the civil system, where imprisonment is not a potential outcome, the standard for guilt is "the preponderance of the evidence" (i.e. $p^* > 0.5$).

Of course, there are some instances where p^* may not be directly calculable, as when we doubt the estimates of p produced by our estimator, $\hat{p}$, are accurate reflections of the true probabilities.[2] As long as the predicted values of $\hat{p}$ are consistent (ie. where $\hat{p}$ is still generally higher for positive outcomes than for negative outcomes), p^* can be computed by following the algorithm below. A proposed threshold, $\bar{p}$, will distribute cases into the cells of the cost matrix at a particular rate: n_{11}, n_{10}, n_{01}, and n_{00}. A value of $\bar{p}$ will equal p^* if and only if it satisfies the following criterion

$$p^* = \underset{\bar{p} \to \{n_{11}, n_{10}, n_{01}, n_{00}\}}{\operatorname{argmin}} f(n, c) = n_{11}(c_{11}) + n_{10}(c_{10}) + n_{01}(c_{01}) + n_{00}(c_{00}). \quad (3)$$

The link between decision boundaries and the utility calculations of the intended policy-makers should now be clear. There are, however, additional factors that might be important in the study of conflict that are not well developed in the machine learning literature. The cost of credit defaults, fraud, and spam messages tend to be relatively homogenous or are a direct function of the size of the outcome (e.g. the size of a loan). The cost of state failures, however, is likely to vary considerably depending on the state that is failing. For example, the impact on the international system of state failure in Fiji may be different from the impact of state failure in

[2] This is a common problem when using certain types of estimators. See, for example, Domingos and Pazzani (1996) on naïve Bayes estimators and Mease and Wyner (2008) on adaptive boosting. Both methods, however, produce very good, and often better, classification results than alternatives.

Russia. Additionally, the cost of intervention depends on how difficult intervention would be. If a state has a probability of failure much higher than our decision boundary, intervention may be futile. Policy-makers may prefer to intervene in cases close to the decision boundary, where a small change in the probability of the adverse event would be more effective.

To some extent, these problems can be addresses by recalibrating equation (2) and calculating a separate *p** for each of the *i* involved cases

$$p^{*i} = \frac{c_{10}^{i} - c_{00}^{i}}{c_{10}^{i} - c_{00}^{i} + c_{01}^{i} - c_{11}^{i}}.$$

This obviously becomes more complex if, as is the case with some estimators, $\hat{p}$ does not reflect actual probabilities. However, the evidence from these estimators suggests that their tendency is to inflate $\hat{p}$ for cases where the event occurs and deflate $\hat{p}$ in cases where the event does not occur (Mease and Wyner 2008), meaning that the threshold will still be valid. There have also been some proposals for post-estimation correction of probability estimates that obviate these problems (Witten et. al. 2011: 343-346).

The main point is that none of the major projects on the prediction of state failure and conflict, insofar as it is reflected in the public record, have dealt with the issue of drawing useful decision boundaries. Policy-maker input has been limited to setting pre-estimation goals, when it should also be incorporated in post-estimation evaluation. Further, as we demonstrate in the PITF data below, this can result in overly optimistic evaluations of model success.

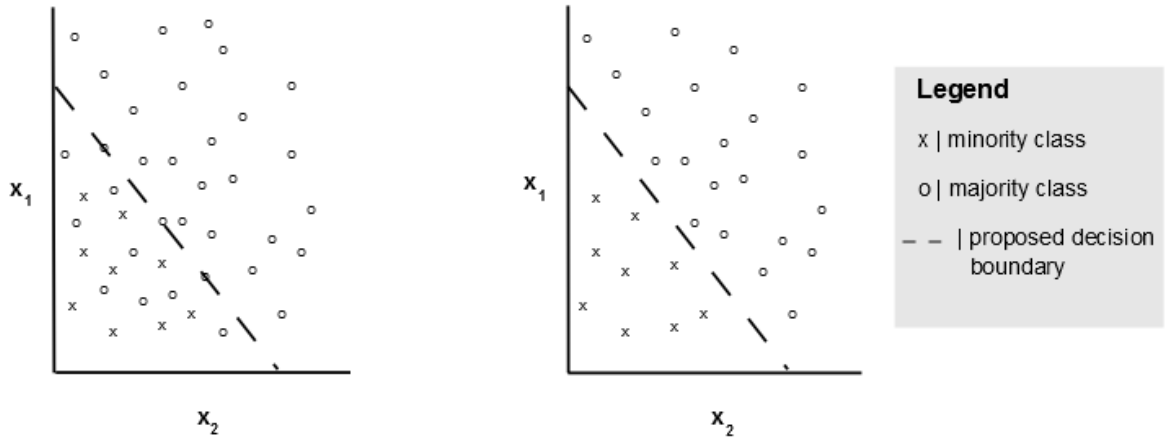
The Problem of Skewed Classes, Part 2: Generalizability and Noise

Before we can evaluate the success of our models, we must have models producing the results. Skewed classes can also hinder the development of effective classification models. Figure 1 shows a classic illustration from the machine learning literature (see e.g. Kubat and

Matwin 1997; Batista et al. 2004), where there are two independent variables, x_1 and x_2, and a bivariate dependent variable with the two classes represented by x and o. Where one class, the majority class (in this case o), has many more examples than the other, the minority class (in this case x), and where there is significant mixing between the classes (left-hand graph), it becomes difficult to draw a decision boundary. As the reader can see in the left-hand graph, the majority class has a number of examples that fall in the space primarily associated with the minority class. If the mixing is due to systematic issues, our results will be misleading. If it is due to noisy measurement, the decision boundary may be less efficient. Similarly, focusing on in-sample results in heavily skewed data increases the temptation to over-fit the training data in a way that is not generalizable. For most political science studies to date, the issue of over-fitting has been mitigated by the assumptions of linearity and additivity in most of our estimation techniques and the relatively rare usage of interaction terms or higher-order polynomials in our specifications. As scholars begin adding more characteristics to their regression models and adapt more flexible estimation techniques, these problems are likely to become severe.[3] With a flexible enough model, a curved decision boundary that correctly classifies nearly all the cases on the left-hand graph could be formulated, but it would likely not be generalizable to new data. John von Neumann famously warned of this danger: "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk" (Dyson 2004: 297).

[3] Several top scholars have emphasized the advantages to more flexible estimation techniques and adapted them to the needs of political scientists – for example: kernel regularized least squares (Hainmueller and Hazlet 2013), LASSO (Kenkel and Signorino 2013), and spline techniques (Keele 2008). This has led some to speculate that political science has a distinctly "non-linear future" (Box-Steffensmeier et al. Forthcoming). It will also be a future, however, in which the danger of over-fitting will loom much larger.

Figure 1: Decision Boundaries in Skewed Classes



There are a number of ways to deal with this problem. The first is simply to ignore the issue and use the full data with our traditional models. While some may be suspicious of doing this, experimental results indicate this is the best option in some cases (e.g. Weiss and Provost 2001).

We can also pre-process the data to make the classes more equal. Pre-processing alternatives fall into two categories:

- Under-sampling – where cases in the majority class are eliminated under a particular criterion; and
- Over-sampling – where cases in the minority class are duplicated or new examples of the minority class are produced through simulation.

The goal of both is making our data more like the right-hand side of Figure 1, where the decision boundary is clear and noise is reduced. As mentioned above, under- and over-sampling can also be used to increase the weight of the minority class, better reflecting the cost of events in this class (Elkan 2001; Witten et al. 2011). Both approaches carry costs. Under-sampling eliminates

potentially useful data. Over-sampling can lead to over-fitting when cases are randomly duplicated or produce inconsistent new cases when simulation is used.

Several techniques are available for under-sampling. The first is random or block-random under-sampling, also known as a 1:*k* case-control technique (see King and Zeng 2001; Rothman et al. 2008). For each example of the minority class, *k* examples of the majority class are selected completely at random or separated into matching blocks on critical characteristics and randomly selected from within these blocks. PITF uses block-random under-sampling, where the *k* majority class cases are selected to match the minority class example on region and year, and they must not have experienced a state failure onset two years prior and four years after the particular year (Goldstone et al. 2010: 193).

This technique has advantages where the goal is to reduce a large dataset to a more manageable size for coding difficult variables (King and Zeng 2001), but it tends to be the least popular technique in machine learning, since it does not explicitly address the class mixing problem and can eliminate a large portion of the data (84 percent in our replication of PITF's data). We found it also the most time-consuming and difficult to reproduce. More popular are techniques using the distance between minority and majority cases to determine which cases are likely to be problematic or due to noise. Given that PITF's dataset does not suffer from issues of difficult new data collection, which would provide a-priori justification for reducing the size of the data set with matched cases, these techniques seem promising.

Wilson's Edited Nearest Neighborhood Rule (ENN) and the Neighborhood Cleaning Rule (NCR) utilize k-nearest neighbor (KNN) methods to achieve better balance and remove cases that are problematic for the decision boundary (Wilson 1972; Japkowicz and Stephen 2002). In a two-class problem, for each example, E_i, in the training set, the three nearest

neighbors are found. In ENN, if E_i is from the majority class and at least two of its nearest neighbors are from the minority class, E_i is removed from the dataset. ENN can be pictured as something like a majority rules voting system for removing majority class examples. If two of a majority class example's three closes neighbors are from the minority class, we might suspect that this majority class case is an outlier and/or is a result of noisy measurement and it is removed. NCR applies the ENN rule and also adds an additional rule. If E_i is from the minority class and more than two of its nearest neighbors are from the majority class, the examples from the majority class are also removed.[4] NCR is more aggressive than ENN, essentially giving minority class examples the ability to unilaterally remove majority class cases that are its nearest neighbors.

Political science data, however, often exhibits structures that are not present in common machine learning problems. Namely, machine learning experimentation is usually conducted on the data available from UCI machine learning repository, which has relatively independent observations in one-time interactions (iris classification, English pronunciation, spam detection, etc.). This has given rise to concerns that experimental results on well-studied UCI datasets may not be generalizable (see e.g. Salzberg 1997). In political science, our datasets tend to be much smaller and have a time-series cross-sectional (TSCS) structure. This entails that our data will tend to cluster by units (e.g. countries), meaning that the *k*-nearest neighbors for any unit *i* at time *t* will likely be the same unit at time *t-1*, *t+1*, etc.

[4] Another popular algorithm, Tomek links, was also tried, but, because of the TSCS structure of the data, did not produce any links in this data. We utilize the "Fast Nearest Neighbor Search Algorithms and Applications (FNN)" package in R (Beygelzimer et al. 2013) to calculate nearest neighbors. No substantive difference in outcomes was observed using various search algorithms. Replication code for all author-composed R functions available on author's website.

We experiment with two methods for addressing this. The first method we label the aggressive edited nearest neighbor rule (AENN), where the number of *k* nearest neighbors found by the algorithm is expanded and the number of minority examples necessary for removal, *l*, is decreased. We thus make both *k* and *l* into tuning parameters for the algorithm.[5] Increasing *k* means that more cases will be allowed to vote for the removal of a majority class case, while decreasing *l* reduces the number of minority class examples needed to remove a majority class case – essentially increasing the voting power of minority cases. The second method is to model the TSCS structure as a multilevel problem, using the full dataset with the unit effects modeled as a random variable from the normal distribution (Gelman and Hill 2006).

Over-sampling methods are also explored in this study. Random over-sampling, in which random minority cases are simply duplicated, is frowned upon because it tends to encourage over-fitted models. Thus, we utilize the synthetic minority over-sampling technique (SMOTE), which forms new minority class examples by interpolating between several minority class examples that lie in proximity (e.g. Chawla et al. 2002).[6] For example, say that we have two minority class examples in one-dimensional (one variable) space. SMOTE essentially assumes that if one minority class example has a value of 6 and another a value of 8, a likely location for a third example would be at a value of 7. This is extended to higher dimensions using KNN to identify neighboring minority class members. This method avoids the over-fitting problem while allowing the minority class to spread further into the majority class space.

[5] Tuning parameters are those that can be manipulated by the user, as opposed to being directly estimated from the data, to alter the characteristics of a model. In this case, the tuning parameters allow us to experiment with how aggressive our algorithm is in removing majority case examples.

[6] We modify the SMOTE function from the "Data Mining with R (DMwR)" package by Torgo (2013) for the purposes of this paper. Replication code for this function can be found on the author's website.

We also experimented with several combinations of the above methods (e.g. Batista et al. 2004), but find no significant benefits to their combination. In what follows, we evaluate all of these methods to find which works best in both in-sample and out-of-sample forecasting.

Data

The first step in this analysis is to, as best as possible, re-create PITF's original data. Unfortunately, the replication data posted by PITF only includes the under-sampled data for their conditional logit models – data for which out-of-sample testing is impossible.[7] The full data is unavailable due to licenses and other agreements with the producers of some constituent data sets.[8] We thus had to reconstruct the dataset. This means that there will be some differences between our results and theirs. In particular, our region classifications do not completely overlap.[9] Similarly, Goldstone et al. (2010: 197) state that in 2003, there were 77 countries with more than four state failure events ongoing in neighboring countries. Using the least restrictive definition of neighbor available in the Correlates of War (COW), we were only able to find ten cases with three or more neighbors experiencing said failures in that year. In reproducing their data, however, we are able to make some important additions, including using infant mortality data with more comprehensive coverage.[10] Since the focus of this study is on the methods for

[7] These replication data sets can be found at http://globalpolicy.gmu.edu/political-instability-task-force-home/pitf-reports-and-replicant-data-sets/.
[8] Personal correspondence with PITF author, August 30, 2012. Comment from *Journal of Peace Research* reviewer, November 2, 2013.
[9] PITF uses a very basic, continent based, delineation of region. Many studies have found these classifications unclear and/or uninformative (e.g. Hadenius and Teorell 2005). Without knowledge of where they draw these boundaries, we default to Hadenius and Teorell's (2005) more detailed classification, which defines region more by inter-country relevance.
[10] For infant mortality, we extend the Abouharb and Kinball (2007), which provides the most comprehensive historical data within the COW framework and extended it using more current World Bank (2008) World Development Indicators and the Institute for Health Metrics and Evaluation data on neonatal mortality (Rajaratnam et al. 2010).

evaluation and not on producing an alternative model of state failure, these issues are not severe. For brevity, we focus on their "full problem set," which counts all of the different types of state failures (civil war onset, adverse regime change, and genocide), rather than on each individual type of event.

With all the above caveats, Table 3 shows that we are able to approximate their results using three case-control samples from the data drawn from 1955 to 2004 (as in Goldstone et al. 2005). The PITF model uses seven variables: dummy variables indicating full democracy, partial democracy with factionalism, partial democracy without factionalism, partial autocracy (full autocracy is the baseline); infant mortality; armed conflict in 4 or more border states (we move the threshold to three or more for the reasons discussed above); and the existence of state-led discrimination (Minorities at Risk 2009). All of the variables have the same directional effect and similar levels of statistical significance. The in-sample, in this case 1955 to 2004, sensitivity (onsets correctly classified) and specificity (controls correctly classified) are not as high as PITF's reported results. This is likely due to the reconstruction problems mentioned above and the lower level of missing data in our reconstructed dataset. The average number of cases in our samples is 664, while PITF's is 468. Despite these differences, since our goal is to test the relative performance of different methods for dealing with class skew, these results are close enough to proceed.

Table 3: Replication of PITF Conditional Logit Model

	Sample 1	Sample 2	Sample 3
Partial Autocracy	0.945***	0.744***	1.349***
	(0.281)	(0.277)	(0.331)
Partial Democracy Without Factionalism	1.296***	1.125***	1.366***
	(0.364)	(0.337)	(0.389)
Partial Democracy With Factionalism	2.852***	2.644***	2.972***
	(0.362)	(0.342)	(0.423)
Full Democracy	0.460	0.549	0.774
	(0.541)	(0.503)	(0.541)
Infant Mortality	0.927***	0.958***	0.907***
	(0.214)	(0.215)	(0.240)
Armed Conflict in 3+ Border States	1.012***	1.161***	1.014***
	(0.299)	(0.312)	(0.357)
State Led Discrimination	0.959***	0.981***	1.373***
	(0.206)	(0.213)	(0.251)
N	710	703	578
Onsets Correctly Predicted (Sensitivity)	76.5%	76.1%	74.3%
Controls Correctly Predicted (Specificity)	79.5%	79.2%	79.6%

Note: Reported values are logistic regression coefficients with standard errors in parentheses.
* p < .1, ** p < .05, *** p < .01

From this point on, we split the data between a training sample that runs from 1955 to 1995, and a test sample that runs from 1996 to 2004. The decision to split the sample by time has its roots in the TSCS structure of the data. In machine learning, we can often assume time invariance of the results (Alpaydin 2010), but this is not the case for political science data, where

we might think that the causes of state failure will change over time.[11] The choice of 1995 as the cutoff is arbitrary, simply reflecting the choice made by the PITF studies.[12]

Methods

Both in their original report (Goldstone et al. 2005) and in the later published version (Goldstone et al. 2010), the PITF concentrate on the results of conditional logit models on their case-controlled training data. While these methods are common in epidemiology (e.g. Rothman et al. 2008), they are problematic for forecasting purposes. This can be seen in the equation for prediction from conditional logit

$$\Pr(y_{it} = 1) = \frac{\exp(x_{it}\beta)}{1+\exp(x_{it}\beta)} \qquad (4)$$

where *t* denotes the 1:3 matched group to which the observation belongs. In other words, the predicted probability of the event is only relevant within the group to which it is being compared. Since there is no method for developing groups in out-of-sample data, the probabilities are likely to be wildly inaccurate (see also Goldsmith et al. 2013). When we attempt to calculate out-of-sample probabilities using this method, the probabilities are unrealistically high, ranging from 0.84 to 0.99. PITF recognize this and use unconditional logit for their out-of-sample predictions (Goldstone et al. 2010: 198). In our results, we also use unconditional logit models with different forms of under- and over-sampling prior to estimation.

When using an under-sampling or over-sampling technique, the slope estimates for the variables should be consistent. The intercept terms, however, will not translate from the training

[11] We find empirical evidence for this intuition in our extended data. The longer we extend the out-of-sample data, the worse the PITF model, even when using our best version, performs worse.

[12] If we were using more sophisticated estimation methods with flexible tuning parameters, we would need to also divide our training data into k-fold cross-validation sets so we could tune the models prior to using them on the out-of-sample data. Since we are limiting ourselves to linear models and are not trying to develop an alternative to their model, this step is omitted for space and clarity.

data to the test data without adjustment. We use the following formula to make this adjustment (see e.g. King and Zeng 2001):

$$\hat{\beta}_0 - ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right] \qquad (5)$$

where τ is the proportion of 1s in the population and $\bar{y}$ is the proportion of 1s in the sample.[13]

Finally, PITF does not provide a clear threshold for the decision boundary. In the conditional logit models, they use a threshold of $p > .24$ to draw the difference between 0 and 1. In the out-of-sample data, they use all of the cases in the top quintile, but do not give a definite value for the threshold. The final section will deal with this issue and demonstrate how to draw a consistent decision threshold in studies like this one.

Since there is no clear decision threshold, we use ROC curves to evaluate the effectiveness of the models (Fawcett 2005). ROC curves look at the necessary tradeoff between classifying positive cases correctly (sensitivity) and classifying negative cases correctly (specificity). If we draw the decision boundary at $p > 0$, we will correctly classify all positive cases, but incorrectly classify all negative cases. Similarly, if we draw the boundary at $p > 1$, we will incorrectly classify all the positive cases, but will also correctly classify all the negative cases. ROC curves plot the ratio of sensitivity to specificity across all possible decision boundaries. Curves further from 45 degrees indicate a better trade-off across all decision boundaries. The area under the curve (AUC) provides a strong one-number summary of how well a model performs in correctly discriminating positive cases from negative ones and is the standard benchmark in machine learning (Hanley and McNeil 1982). It also has an intuitive

[13] Changing the intercept does not change the prediction accuracy, it simply corrects the size of the probability estimates in under-sampled data.

interpretation as the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett 2005: 868).

Results

Figures 2 and 3 show our main findings. The Figures include the ROC curves, with the AUC reported in the lower-right-hand corner (along with the 95 percent confidence intervals), on the in-sample training set and the out-of-sample test set for each of the above methods of handling skewed classes.[14] The results suggest that the case-control methods used by PITF do not produce any significant improvement, and are usually quite worse, than most of the alternatives in our replication data.

In the in-sample data, the three case-control samples yield AUC values between .760 and .771. This compares unfavorably with simply using standard logit with the full data set, which produces an AUC of .798. While not an overwhelming difference, the spread is statistically significant ($p = .001$, $p = .010$ and $p = .011$ for the three case-control samples). When we explicitly model the multi-level structure of the data, we receive even more favorable results. Using country random effects, AUC dramatically increases to approximately .900, while the three-level model, with both country and region, produces a slightly lower AUC of .887. While not statistically distinguishable from each other, both are significantly better than the case-control models ($p < .001$ in all cases for both models). The three under-sampling techniques gleaned from the machine learning literature also out-perform the case-control method in the in-sample data, with AUC values ranging between .797 and .798. These results are fairly close to those for the logit model on the full dataset, suggesting little improvement from any data

[14] Statistical significance and 95% confidence intervals calculated using the methods introduced by DeLong et al. (1988).

cleaning method. Finally, the SMOTE over-sampling procedure is the one alternative that does not out-perform the case-control method. Its AUC of .763 is lower than all but one of the case-control samples and none of the differences are statistically significant.

We receive similar results when evaluating the different techniques on the out-of-sample data. The three case-control samples yield AUC values between .793 and .801. Results for the same simple logit model using the full dataset yield an AUC of about .818. This difference fails to achieve conventional levels of statistical significance ($p = .270$, $p = .449$ and $p = .398$ for the three case-control samples). This is not completely surprising, as our out-of-sample data is about one-fourth the size of our in-sample data.[15] Nevertheless, given that the case-control method was, by far, the most difficult of the various algorithms to implement, its failure to make any improvement on out-of-sample prediction over simply using all the available data, strongly undermines its oft-assumed status as the default method for data with rare events. Instead of clarifying the decision boundary for out-of-sample testing, the loss of information from block-random under-sampling method has produced less generalizable results.

None of the other methods for under- or over-sampling produce better results than straightforward logit estimation with the full dataset. The ENN procedure only identifies 37 problematic cases, and their removal does not produce a substantial improvement. The more aggressive AENN procedure finds more problematic cases, 91, but their removal from the training set worsens the out-of-sample prediction accuracy. Results inversely correlate with the aggressiveness of the AENN tuning parameters, indicating that under-sampling itself is ill-advised in this data. Finally, the NCR procedure is less aggressive than AENN, but more

[15] If we split the data at 1990, for example, these differences all surpass standard levels of statistical significance ($p < 0.05$).

aggressive than ENN, removing 49 cases. While NCR produces better results than ENN or AENN, it is still worse than the model utilizing the full training data.

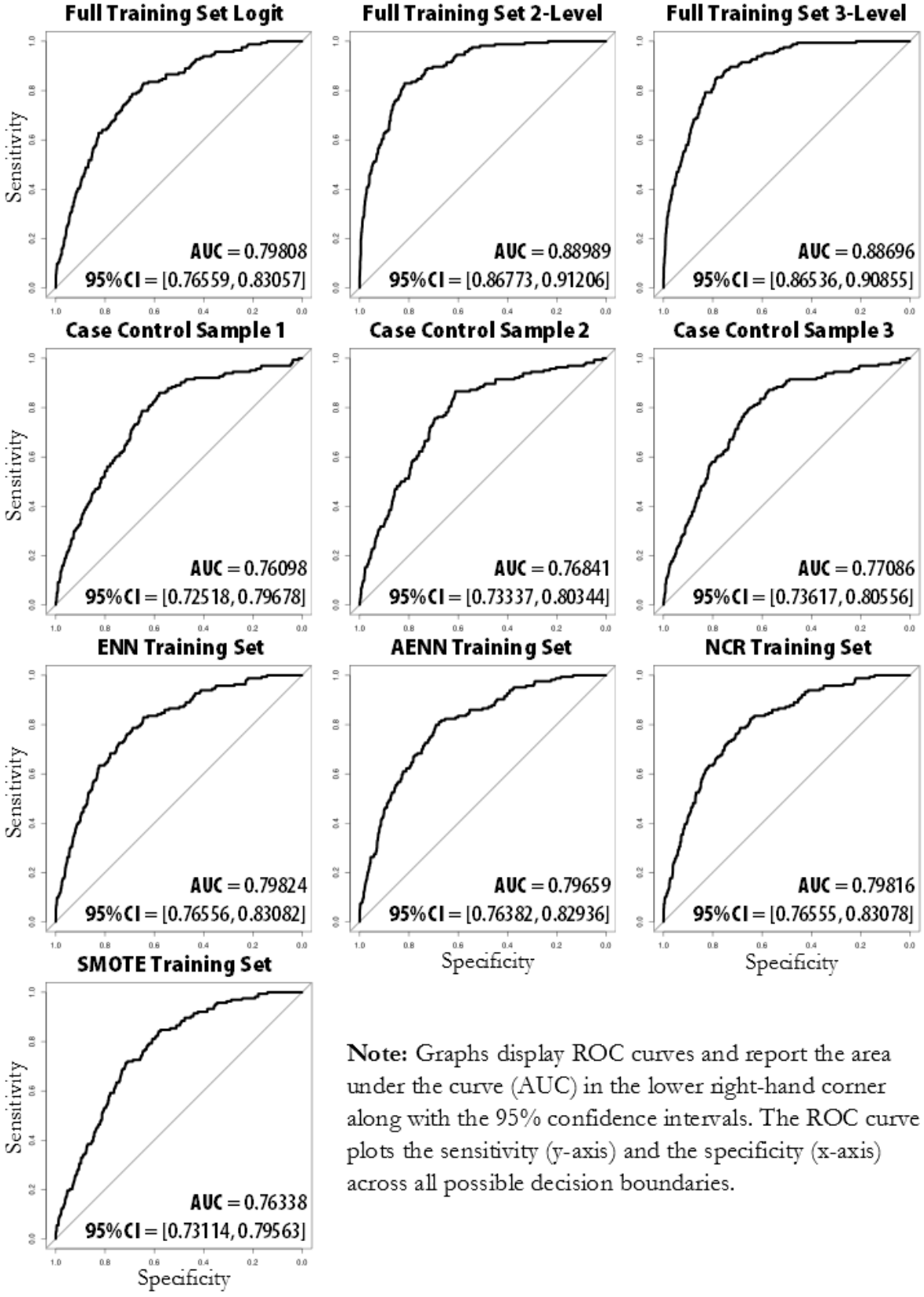
Over-sampling performs no better. The SMOTE procedure doubles the number of positive cases in the dataset, but performs worse in out-of-sample prediction, producing an AUC of .780. This performance can be improved by tuning the number of synthetic cases, but the results suggest better performance the fewer synthetic cases are imputed, arguing against its use in this data. Contrary to findings in some previous studies (Batista et al. 2004), we find no better results from combining SMOTE with any of our under-sampling techniques.

Conversely, both of the multilevel models produce noticeably better results. The three-level model, where both the country and region are given random intercepts, yields a significantly higher AUC of .845. The simpler two-level model, with a random effect for country and a fixed effect for region produces similar results, with an AUC of about .841. In comparison with the three case-control samples, the two-level ($p = .088$, $p = .146$ and $p = .129$) and three-level models ($p = .053$, $p = .087$ and $p = .075$) produce results that are only marginally statistically significant. Again, this is not too surprising, given the smaller number of cases. Given the superiority of these models in both the in-sample and out-of-sample data, these results are substantial enough to recommend their continued usage since this data is not so large as to introduce dramatically higher computation times on a standard office computer. We should note, however, that we make no claim that multilevel models will always perform better.[16] In this case,

[16] Indeed, the preferred model will vary, not only by the data analyzed, but also by the evaluation method. While ROC curves and AUC are the standard measures for most applications, there are a range of alternatives – e.g., F-measures, F1-scores, Rand Accuracy, Jaccard, Matthews correlation coefficient, and Cohen's kappa. For all of these, the best performance for Goldstone et al.'s (2010) threshold is produced by the multilevel models. If, on the other hand, we used the highest F1-score for decision boundary selection and model evaluation, the relative performance of the machine learning techniques is much better. The AENN procedure produces the best results under this criterion (results available from authors). The case control method is still among the worst performing methods, regardless of evaluation method. Some of these alternative model evaluation methods have substantial biases and assumptions of which users should be aware before using (see e.g. Powers 2011).

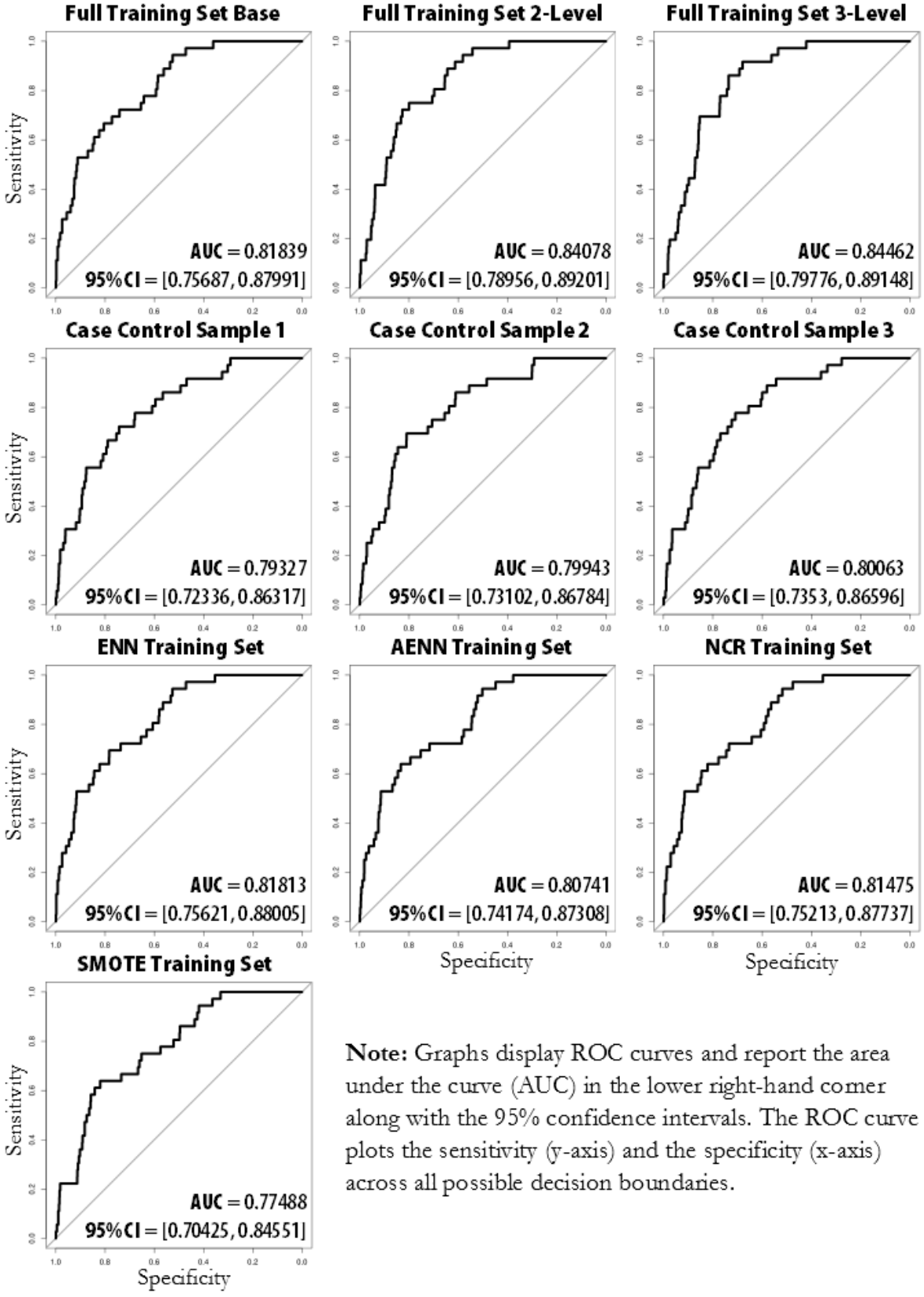
they are picking up on unmodeled heterogeneity between states that is useful for prediction, but this might be eliminated by more detailed information on the cases, such as better measures of contentious issues that might lead to conflict in addition to the structural characteristics on which previous studies have focused (e.g. in international conflict, see Gleditsch and Ward 2013).

Figure 2: In-Sample ROC Curves for Methods of Balancing Data



Note: Graphs display ROC curves and report the area under the curve (AUC) in the lower right-hand corner along with the 95% confidence intervals. The ROC curve plots the sensitivity (y-axis) and the specificity (x-axis) across all possible decision boundaries.

Figure 3: Out-Of-Sample ROC Curves for Methods of Balancing Data



Note: Graphs display ROC curves and report the area under the curve (AUC) in the lower right-hand corner along with the 95% confidence intervals. The ROC curve plots the sensitivity (y-axis) and the specificity (x-axis) across all possible decision boundaries.

The punch line of these findings is that social scientists should not use under- or over-sampling as a first resort when forecasting events. Indeed, some studies on more traditional machine learning data have suggested that scholars should always be cautious about under- or over-sapling their data in response to skewed classes (e.g. Weiss and Provost 2001). Previous scholars have put much work into developing comprehensive datasets of state failure and conflict, and researchers should be cautious about removing data. Where under- or over-sampling are tried, it makes sense to test a variety of tools, since the performance of these tools is likely to vary from problem to problem and since the other tools introduced above are significantly easier to implement.

Formulating a Cost Sensitive Decision Boundary for State Failure

Once we have chosen a model, we must formulate the decision boundary – the probability over which we will predict a state failure will take place. As noted above, we need to base our decision of where to draw the decision boundary on the expected utility of that decision boundary. To date, however, we know of no study in political science that justifies its decision boundary according to any utility calculation. Goldstone et al. (2010) simply draw their decision boundary by taking those cases in the top quartile and predicting these cases will experience state failure. In their data, this produces 18 correct positive predictions and 2 incorrect positive predictions, for a sensitivity score of .857. It also, however, produces 233 false positives, for a precision of .033. This suggests that they associate extremely low cost with false positives, and, by implication, with the commitment of resources towards preventing conflict.

In this section, we expound on cost sensitive decision boundaries. Using the two-level unconditional logit model on the full dataset, we reconstruct the confusion matrix. By adding a

few logical assumptions to our cost matrix from above, we are able to gain an approximate calculation of the relative utility implicitly assumed by Goldstone et al. (2010) for state failure outcomes, at least for the results from our reconstructed data. We should note that we are not suggesting that Goldstone et al. (2010) overtly support these utility cost decisions – this is an "as if" exercise to demonstrate the importance of making these decisions explicit. The results confirm that they implicitly assume a very low cost of intervention to prevent state failure.

We start by introducing two "reasonableness conditions" to the cost matrix in Table 2 (Elkan 2001):

$$c_{10} > c_{00} \qquad (6)$$

and

$$c_{01} > c_{11}. \qquad (7)$$

The first condition essentially states that there must be some cost to false positives, making them more costly than correctly labeling negative outcomes. If not, then the policy-maker should always predict a positive outcome. The second condition states that there is a cost to false negatives. If not, then the policy-maker should always predict a negative outcome and there is no reason for estimation.

We add two additional conditions:

$$c_{10} > c_{11} \qquad (8)$$

and

$$c_{01} > c_{00}. \qquad (9)$$

These conditions state that there is some benefit to correct prediction over incorrect prediction. While not necessary to prevent row dominance, they are both intuitive and will allow interesting later calculations.

Using the first two reasonableness conditions (equations (6) and (7)), we can standardize costs to simplify the Table 2. This is illustrated on the left-hand side of Table 4, where

$$c'_{11} = (c_{11} - c_{00})/(c_{10} - c_{00})$$

and

$$c'_{01} = (c_{01} - c_{00})/(c_{10} - c_{00}).$$

Using the latter two conditions (equations (8) and (9)), we can similarly standardize the cost matrix as in the right-hand side of Table 4, where

$$c'_{11} = (c_{11} - c_{00})/(c_{01} - c_{00})$$

and

$$c'_{10} = (c_{10} - c_{00})/(c_{01} - c_{00}).$$

Table 4: Standardizations of Cost Matrix

Standardization Using Reasonableness Conditions

	Actual Positive	Actual Negative
Predict Positive	c'_{11}	*1*
Predict Negative	c'_{01}	*0*

Standardization Using Additional Conditions

	Actual Positive	Actual Negative
Predict Positive	c'_{11}	c'_{10}
Predict Negative	*1*	*0*

Unfortunately, the unavailability of PITF's original data makes exact computation of the implicit cost structure impossible. We therefore use the 2-level model from above and replicate the quartile decision boundary ($p > .044$), along with a boundary that produces marginally lower ($p > .053$) and higher ($p > .039$) true positives in out-of-sample data from 1995 to 2004. While not exact, they replicate the decision structure given by Goldstone et al. (2010) and serve as a

useful illustration of why their decision not to evaluate the precision of their results is important. The results are shown in Table 5.

Table 5: Confusion Matrix at Three Values of p^* Using PITF Model on Out-of-Sample, 1995-2004, Data.

Quartile Threshold, $p^* > .044$

	Actual Positive	Actual Negative
Predict Positive	27	305
Predict Negative	9	970

Lower Threshold, $p^* > .039$

	Actual Positive	Actual Negative
Predict Positive	29	367
Predict Negative	7	908

Upper Threshold, $p^* > .053$

	Actual Positive	Actual Negative
Predict Positive	26	266
Predict Negative	10	1009

Combined with the right-hand standardization from Table 4, this yields two inequalities and allows us to estimate the assumed cost of intervention to prevent state failure.

The two inequalities are:

$$27c_{11} + 305c_{10} + 9 < 26c_{11} + 266c_{10} + 10$$

and

$$27c_{11} + 305c_{10} + 9 < 29c_{11} + 367c_{10} + 7.$$

When we solve the two inequalities, the lowest estimate for c_{11} is $c_{11} < -39c_{10} + 1$. Assuming that a successful forecast allows successful intervention to prevent state failure, the only cost associated with correct prediction is the cost of intervention. Since the cost of stability and non-intervention is $c_{00} = 0$, the cost of a true positive, which involves intervention, must be greater than 0. This means we would only prefer the PITF threshold to the higher threshold if $c_{10} < .026$. In other words, a false positive, which would recommend intervention when there was no danger of state failure, would have to be less than 2.5% of the cost of allowing state failure to

take place for us to prefer this threshold over the higher threshold. Taking this a step further, if we compare the PITF threshold with a completely naïve model, where we always predict no state failure ($p^* > 1$), we get the inequality:

$$27c_{11} + 305c_{10} + 9 < 0c_{11} + 0c_{10} + 36.$$

Carrying out the same calculations, we would only prefer PITF's threshold to a completely naïve model if the cost of intervention were less than 11.8% the cost of allowing state failure to take place.

Collier (2004) puts the average cost of civil conflict at about $50 billion. Whether PITF's implicit cost structure holds will depend, of course, on the intervention proposed. Some proposals, like reorienting foreign aid, entail mostly opportunity costs related to domestic and foreign relations (Bueno de Mesquita and Smith 2009). Military intervention entails much more direct costs. The interventions in Bosnia, Cambodia, El Salvador, Haiti, Rwanda and Somalia cost an average of $14.1 billion – and had a mixed record in averting subsequent costs (Collier et al. 2003: 174). These are rough numbers – as noted above, there is likely to be substantial heterogeneity in cost estimates, depending on where the conflict takes place and the proposed intervention.

This result leads to two conclusions. First, there is a lot more room for improving the prediction model than the headline number of 86% correct prediction would indicate. There is a clear false positive problem. This should not be surprising, given that PITF attribute the success of their model to the ability to clearly distinguish cases where state failure is unlikely. Future research may wish to focus on the factors that distinguish cases that are likely to become unstable.

Second, political scientists in this area of study have rarely bothered to estimate the cost structure of state failure and intervention. The work that has been done in this area suggests a wide range of costs and benefits, depending on the intervention involved (Collier et al. 2003: 173-186). From the perspective of cost sensitive learning, the criteria for successful prediction is not whether researchers can produce a decision boundary above 0.5 or produce an arbitrary level of correct positive predictions, but whether they can produce a useful decision boundary for policy-makers.

Before concluding, we must be clear that these results should not detract from the achievements of the PITF. Their work will likely stand the test of time as a landmark achievement, both for the success of their models and for helping to bring ambitious prediction goals into the mainstream of political science. In our interactions with scholars close to the group, we have been informed that they recognize the false positive issue and are continuing to work on the issue. Our main purpose is not to detract from their achievement, but to make clear to the rest of the field the issues involved in pre-processing data and evaluating success. Based on what we have learned in soliciting comments for this project, we strongly suspect that most political scientists have a limited understanding of how to evaluate prediction models, automatically assuming that they must use case-control methods in the presence of skewed classes. Such issues must be addressed as machine learning methods, new data sets, and prediction-based projects become more common (see e.g. Hegre et al. 2013; Leetaru and Schrodt 2013; Goldsmith et al. 2013; Gleditsch and Ward 2013).

Conclusions

We draw several conclusions from the above results. First, consistent with Leamer (2010), we doubt that every asymptotically unbiased estimation strategy will be equally successful in every situation. Out-of-sample prediction is a powerful mechanism for evaluating the relative merits of estimation strategies. Moreover, our results recommend trying several strategies for dealing with the class skew problem. In this particular case, this study finds using the full dataset is most successful. We do not believe this will always be the result, but we should not default to a class balancing strategy without testing.

Second, once we develop an estimator, we must decide on a decision boundary above which we predict the event will happen. Political scientists have generally ignored the theoretical implications of this choice, either setting the threshold at $p^* > 0.5$ or choosing another arbitrary value that produces a *prima-facie* "reasonable" distribution of outcomes. This study shows that the choice of this threshold is tied inherently to the relative costs of potential outcomes, and ignoring this can lead to underestimating the cost of false positives produced by a model of a rare event.

Finally, this study opens several new lines of research. It reveals how much further we have to go in regards to predicting state failure. The PITF model, at least in our best effort at replication, produces outstanding results in terms of sensitivity and specificity, but weaker results in terms of precision. This is an accurate reflection of their approach to explaining conflict by looking at the institutions that make state failure especially unlikely. From a utility perspective, improving the precision is important and would likely benefit from exploring both the factors preventing state failure and those which contribute to its onset. Research also needs to be done into the relative costs associated with state failure and intervention. We often hear that political

scientists should do more to make their work relevant to policy-makers, this, however, also involves developing a clearer understanding of policy makers' preferences than we have currently.

Works Cited

Abouharb, M. Rodwan and Anessa A. Kimball. 2007. "A New Dataset of Infant Mortality Rates, 1816-2002." *Journal of Peace Research* 44(6): 743-754.

Abu-Mostafa, Yasar S., Malik Magdon-Ismail, and Hsuan-Tien Lin. 2012. *Learning from Data*. New York: AMLbook.

Alpaydin, Ethem. 2010. *Introduction to Machine Learning*, 2nd ed. Cambridge, MA: MIT Press.

Andriole, Stephen. J., and Robert A. Young. 1977. "Toward the Development of an Integrated Crisis Warning System." *International Studies Quarterly* 21: 107–150.

Bates, Robert H., David L. Epstein, Jack A. Goldstone, Ted Robert Gurr, Barbara Harff, Colin H. Kahl, Kristen Knight, Marc A. Levy, Michael Lustik, Monty G. Marshall, Thomas M. Parris, Jay Ulfelder, and Mark R. "Political Instability Task Force Report: Phase IV Findings." Working Paper. McLean, VA: Science Applications International Corporation.

Batista, Gustavo E.A.P.A., Ronaldo C. Prati and Maria Carolina Monard. 2004. "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data." *Sigkdd Explorations* 6(1): 20-29.

Beck, Nathaniel, Gary King and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review* 94(1): 21-35.

Beygelzimer, Alina, Sham Kakadet, John Langford, Sunil Arya, David Mount and Shengqiao Li. 2013. "Fast Nearest Neighbor Search Algorithms and Applications (FNN)." Package manual available online: http://cran.r-project.org/web/packages/FNN/FNN.pdf.

Box-Steffensmeier, Janet, John Freeman, and Jon Pevehouse. Forthcoming. *Time Series for Social Scientists*. Cambridge, UK: Cambridge University.

Bueno de Mesquita, Bruce. 1981. *The War Trap*. New Haven, CT: Yale University Press.

Bueno de Mesquita, Bruce and Alastair Smith. 2009. "A Political Economy of Aid." *International Organization* 63: 309-340.

Campbell, James E. 2012. “Forecasting the 2012 US National Election – Editor’s Introduction.” *P.S. Political Science and Politics* 45(4): 610-613.

Campbell, Donald T. 1957. “Factors Relevant to the Validity of Experiments in Social Settings.” *Psychological Bulletin* 54: 297-312.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer. 2002. “SMOTE: Synthetic Minority Over-sampling Technique.” *Journal of Artificial Intelligence Research* 16: 321-357.

Cook, T.D. and D.T. Campbell. 1979.*Quasi-Experimentation*. Chicago: Rand McNally.

Collier, Paul, V.L. Elliott, Håvard Hegre, Anke Hoeffler, Marta Rynal-Querol, and Nicholas Sambanis. 2003. *Breaking the Conflict Trap: Civil War and Development Policy*. Washington, DC: World Bank.

Collier, Paul. 2004. “Development and Conflict.” Working Paper: Centre for the Study of African Economies, Oxford University. Available online: http://www.un.org/esa/documents/Development.and.Conflict2.pdf (last accessed 11/20/2013).

Collier, Paul and Anke Hoeffler. 2004. “Greed and grievance in civil war.” *Oxford Economic Papers* 56(4): 563–595.

DeLong, Elisabeth R., David M. DeLong and Daniel L. Clarke-Pearson. 1988. “Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach.” *Biometrics* 44: 837–845.

De Marchi, Scott, Christopher Gelphi, and Jeffrey D. Grynaviski. 2004. “Untangling Neural Nets.” *American Political Science Review* 98(2): 371-378.

Domingos, Pedro and Michael Pazzani. 1996. “Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier.” In *Proceedings of the Thirteenth International Conference on Machine Learning*. Burlington, MA: Morgan Kaufmann, p. 105-112.

Dyson, Freeman. 2004. “A Meeting with Enrico Fermi.” *Nature* 427: 297.

Elkan, Charles. 2001. “The Foundations of Cost-Sensitive Learning.” In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*. Burlington, MA: Morgan Kaufmann, 973-978.

Esty, Daniel C., Jack A. Goldstone, Ted Robert Gurr, Pamela T. Surko, and Alan N. Unger. 1995. “State Failure Task Force Report.” Working Paper. McLean, VA: Science Applications International Corporation.

Esty, Daniel C., Jack Goldstone, Ted Robert Gurr, Barbara Harff, Pamela T. Surko, Alan N. Unger, and Robert Chen. 1998. "The State Failure Project: Early Warning Research for US Foreign Policy Planning." In John L. Davies and Ted Robert Gurr (eds.), *Preventive Measures: Building Risk Assessment and Crisis Early Warning Systems*. Boulder, CO: Rowman and Littlefield, chp. 3.

Esty, Daniel C., Jack A. Goldstone, Ted Robert Gurr, Barbara Harff, Marc Levy, Geoffrey D. Dabelko, Pamela T. Surko, and Alan N. Unger. 1999. "State Failure Task Force Report: Phase II Findings." In *Environmental Change and Security Project Report*. Washington, DC: The Woodrow Wilson Center.

Fair, Ray. 2011. *Predicting Presidential Elections and Other Things*, 2nd ed. Palo Alto, CA: Stanford.

Farrell, Henry. 24 June 2012. "Why the Stevens Op-Ed is Wrong." *The Monkey Cage*. Available at: http://themonkeycage.org/blog/2012/06/24/why-the-stevens-op-ed-is-wrong/. (Accessed 3 October 2012.)

Fawcett, Tom. 2006. "An Introduction to ROC Analysis." *Pattern Recognition Letters* 27: 861-874.

Fearon, James D and David D Laitin. 2003. "Ethnicity, insurgency, and civil war." *American Political Science Review* 97(1): 75–90.

Feder, Stanley. (2002) Forecasting for Policy Making in the Post-Cold War Period. Annual Review of Political Science 5: 111–125.

Gelman, Andres and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press.

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiments*. New York: W.W. Norton & Company.

Gleditsch, Kristian Skrede and Michal D. Ward. 2013. "Forecasting is Difficult, Especially About the Future: Using Contentious Issues to Forecast Interstate Disputes." *Journal of Peace Research* 50(1): 17-31.

Goldsmith, Benjamin E., Charles R. Butcher, Dimitri Semenovich, and Arcot Sowmya. 2013. "Forecasting the Onset of Genocide and Politicide: Annual Out-of-Sample Forecasts on a Global Dataset, 1988-2003." *Journal of Political Research* 50(4): 437-452.

Goldstone, Jack A., Ted Robert Gurr, Barbara Harff, Marc A. Levy, Monty G. Marshall, Robert H. Bates, David L. Epstein, Colin H. Kahl, Pamela T. Surko, John C. Ulfelder, and Alan N. Unger. 2000. "State Failure Task Force Report: Phase III Findings." Working Paper. McLean, VA: Science Applications International Corporation.

Goldstone, Jack A., Robert H. Bates, David L. Epstein, Ted Robert Gurr, Michael B. Lustik, Monty G. Marshall, Jay Ulfelder, and Mark Woodward. 2005. "A Global Model for Forecasting Political Instability." Paper Presented at the Annual Meeting of the American Political Science Association, Washington, DC.

Goldstone, Jack A., Robert H. Bates, David L. Epstein, Ted Robert Gurr, Michael B. Lustik, Monty G. Marshall, Jay Ulfelder, and Mark Woodward. 2010. "A Global Model for Forecasting Political Instability." *American Journal of Political Science* 50(1): 190-208.

Green, Donald P. and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3): 491-511.

Gujarati, Damodar N. and Dawn C. Porter. 2009. *Basic Econometrics*, 5th ed. New York: McGraw-Hill.

Hainmueller, Jens and Chad Hazlet. 2013. "Kernel Regularized Least Squares: Moving Beyond Linearity and Additivity Without Sacrificing Interpretability." MIT Political Science Department Research Paper No. 2012-8.

Hegre, Håvard, Joakim Karlsen, Håvard Mokleiv Nygård, Håvard Strand, and Henrik Urdal. 2013. "Predicting Armed Conflict, 2010-2050." *International Studies Quarterly* 57(2): 250-270.

Hyde, Susan D. and Nikolay Mainov. 2012. "Which Elections Can be Lost?" *Political Analysis* 20(2): 191-210.

Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. Hoboken, NJ: Wiley.

Japkowicz, Nathalie. 2000. "Learning from Imbalanced Data Sets: A Comparison of Various Strategies." Working Paper: DalTech/Dalhousie University.

Japkowicz, N. and S. Stephen. 2002. "The Class Imbalance Problem: A Systematic Study." *IDA Journal* 6(5): 429-449.

Johnson, Jim. 24 June 2012. "Incoherent Criticisms of Political Science." *(Notes On) Politics, Theory & Photography*. Available at: http://politicstheoryphotography.blogspot.com/2012/06/incoherent-criticisms-of-political.html. (Accessed 3 October 2012.)

Kenkel, Brenton and Curtis Signorino. 2013. "Bootstrapped Based Regression with Variable Selection: A New Method for Flexible Form Estimation." Working Paper: University of Rochester.

Keele, Luke. 2008. "Splines." In *Semi-parametric Regression for the Social Sciences*. Chichester, UK: Wiley.

King, Gary and Langche Zeng. 2001a. “Explaining Rare Events in International Relations.” *International Organization* 55(3): 693-715.

King, Gary and Langche Zeng. 2001b. “Improving Forecasts of State Failure.” *World Politics* 53(4): 623-658.

King, Gary, James Honaker, Anne Joseph, and Kenneth Schreve. 2001. “Analyzing Incomplete Political Science Data.” *American Political Science Review* 95(1): 49-69.

Klarner, Carl E. 2012. “State-level Forecasts of the 2012 Federal and Gubernatorial Elections.” *P.S. Political Science and Politics* 45(4): 655-662.

Kubat, Miroslav and Stan Matwin. 1997. “Addressing the Curse of Imbalanced Training Sets: One-Sided Selection.” In *Proceedings of the Fourteenth International Conference on Machine Learning*. Burlington, MA: Morgan Kaufmann, 179-186.

Leamer, Edward E. 2010. “Tantalus on the Road to Asymptopia.” *Journal of Economic Perspectives* 24(2): 31-46.

Leetaru, Kalev and Philip A. Schrodt. 2013. “GDELT: Global Data on Events, Location and Tone, 1979-2012.” Paper presented at the International Studies Association Annual Meeting, San Francisco, CA.

Lewis-Beck, Michael S. and Charles Tien. 2012. “Election Forecasting for Turbulent Times.” *P.S. Political Science and Politics* 45(4): 625-629.

Martin, Andrew D., Kevin M. Quinn, Theodore W. Ruger, and Pauline T. Kim. 2004. “Competing Approaches to Predicting Supreme Court Decision Making.” *Perspectives on Politics* 2(4): 761-767.

Mease, David and Abraham Wyner. 2008. “Evidence Contrary to the Statistical View of Boosting.” *Journal of Machine Learning Research* 9: 131-156.

Morton, Rebecca B. and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality*. Cambridge, UK: Cambridge University Press.

O’Brien, Sean P. 2010. “Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research.” *International Studies Review* 12: 87-104.

Powers, D.M.W. 2011. “Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation.” *Journal of Machine Learning Technologies* 2(1): 37-63.

Rajaratnam, J.K. et al. 2010. “Neonatal, postneonatal, childhood, and under-5 mortality for 187 countries, 1970–2010: a systematic analysis of progress towards Millennium Development Goal 4.” *Lancet*, 375: 1988-2008.

Salzberg, Steven L. 1997. “On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach.” *Data Mining and Knowledge Discovery* 1(3): 317-328.

Sarkees, Meredeth Reid and Frank Wayman. 2010. *Resort to War: 1816-2007*. Washington, DC: CQ Press.

Schrodt, Philip A. Forthcoming. “Seven Deadly Sins of Contemporary Quantitative Political Analysis.” *Journal of Peace Research*. DOI: 10.1177/0022343313499597.

Stevens, Jacqueline. 23 June 2012. “Political Scientists are Lousy Forecasters.” *The New York Times*. Available at: http://www.nytimes.com/2012/06/24/opinion/sunday/political-scientists-are-lousy-forecasters.html?pagewanted=all. (Accessed 3 October 2012.)

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

Weiss, Gary M. and Foster Provost. 2001. “The Effect of Class Distribution on Classifier Learning: An Empirical Study.” Technical Report ML-TR-44. Department of Computer Science, Rutgers University.

Wilson, D.L. 1972. “Asymptotic Properties of Nearest Neighbor Rules for Using Edited Data.” *IEEE Transactions on Systems, Man, and Communications* 2(3): 408-421.

Witten, Ian H., Eibe Frank, and Mark A. Hall. 2011. *Data Mining*, 3rd ed. Burlington, MA: Morgan Kaufman.

World Bank 2008. World Development Indicators. Available at: http://data.worldbank.org/data-catalog.