Predicting Civil Wars with Higher Order Interactions

Adeline Lo*

May 6, 2015

This is a preliminary draft. Please do not cite or circulate.

Abstract

If we wish to predict civil wars, we may require a new approach, placing more emphasis on variable selection for better prediction. As the number of highly influential variables and variable sets available has grown with the size of data, though, how can we decide what to use? The universe of variables to sift through should include any and all information available, but also the higher-order interactions amongst all variables. However, higher order interactions become more difficult to capture as the number of explanatory variables grow due to the well-known curse of dimensionality. Given that civil wars are highly complex processes, it is likely that using only marginal information (information from single variables) may result in discarding important information embedded in higher orders interactions (interactions between several variables). Recent advances in big data analysis catered towards uncovering higher order interactions lend themselves to application to political science questions. I suggest that important higher order interactions in existing political science data can be uncovered by the Partition Retention method and illustrate with an application to civil wars data. My findings show that the Partition Retention method identifies variables *and* variable sets, with some variable sets as large as 4 or 5 variables interacting to predict civil wars. Using these identified variables and variable sets to predict boosts correct prediction rates on out of sample testing sets from 86.2% to 97.6%. True positive rates are improved from 5.67% to 90%. The application demonstrates possible gains in correct prediction rates for political science phenomena like civil wars when including a research step for identifying very predictive sets of variables.

Introduction

Civil wars are highly consequential and complex processes that are important to social scientists and policy analysts everywhere. A quick Google scholar search yields over 2 million results as of

*PhD Candidate, University of California, San Diego, Department of Political Science. Contact email: aylo@ucsd.edu. http://loadeline.com. Many thanks go to Rachel Fan, Shaw-Hwa Lo, Margaret Roberts and Tian Zheng for their insights and important feedback. I also thank Chien-Hsun Huang for computing support. I am grateful to Héctor Pifarré i Arolas, Erin Giffin, Veena Jeevanandam, Michael Levy, and Shanthi Manian for helpful commentary and discussion. All errors are my own.

the writing of this paper.[1] Information collected on conflicts and violent processes has also grown in the last few decades. With so much data, how do we best find variables and variable sets that can predict and explain violence?

There is prolific work exploring the causes of conflict in the literature. Proposed theories emphasize ethnic politics (e.g. [21]), economic greed or grievances [9, 10], the effects of geography and natural endowments (inter alia, [13, 28]), as well as factors that enter into the calculus of launching successful insurgencies [12], are a few among many. Thus, a common method for finding variables that explain civil war is to test hypotheses derived from civil war theories using regression models on quantitative data. Traditional significance testing then establishes whether key variables identified by the theory are significantly associated with civil wars.

How to best find variables and variable sets that predict wars has also become increasingly important. Knowing when and where a civil war might occur can actively guide political and humanitarian preparation and amelioration of the negative impacts of war. Recent work has focused on comparing different models and their prediction error rates as methods of selecting good, predictive models for observing civil war in a given country and year. Regressions are a particularly popular tool (see [2, 15, 33] among others). Prediction requires a different method of selecting covariates than that of testing theories, since the main goal is minimizing error rates in predicting civil wars, not establishing significance of key independent variables. However, current predictive models often turn to theoretical work on the causes of civil war to retrieve variables that are then included in different models; the models are then compared for out-of-sample prediction rates (see [19, 32, 33], among others). As Ward et al. (2010) warn however, selecting significant variables for prediction does not automatically lead to better predictions [32]. Indeed, in related work I show that variables that are highly predictive can appear as entirely insignificant under traditional significance-based tests [23]. How then do we select explanatory variables for predicting civil wars? The method I present and illustrate in this paper is designed to flexibly account for not only the individual predictivity of single explanatory variables, but also the joint predictive effects of any and all groups of explanatory variables we may have information on. Thus, our universe of collected independent variables — *and all of their possible interactions* — compose the full set of possible variables to use.

[1] Date of search: January 27, 2015.

This full set of possible variables and variable interactions (which I will refer to as variable sets in this paper) is not small, given the types of data we have available today. Higher order interactions — here defined as any 2-way or higher interaction between variables — become more difficult to identify as the number of explanatory variables grow due to the well-known curse of dimensionality. While the individual effects contributed by single variables is important, when answering political science questions we may require information contributed by interactions of variables, however. This is of particular importance when we care about predicting an outcome, such as predicting whether a given country year will see civil war, outcomes of elections or voter turnout, etc.

To see how the curse of dimensionality might quickly affect the size of variable sets amongst which we must select a smaller set of highly predictive variable sets, consider the following. We have a single binary outcome variable of interest, Y, and a set of m explanatory variables $\mathbf{X} = \{X_1, ..., X_m\}$. When m is small relative to the sample size, for example, 5, with a reasonably sized sample we can utilize a logistic regression with all possible interactions between all variables, which is 31 parameters (the sum of 5 choose all variable set sizes). This is manageable. Simply double the size of m, however, and the number of parameters the researcher must estimate becomes 1,023[2]. In the real world of sample constrained data, uncovering influential higher order interactions quickly becomes exponentially difficult.

This presents a dilemma — given the full set of possible independent variables, and their interactions, and limited sample sizes, how might we identify the most predictive variables and variable sets?

Advances in big data prediction have begun to focus on dealing with large numbers of independent variables for prediction, as well as uncovering higher order interactions. The Partition Retention Method is designed to tackle big data with larger numbers of variables. I propose using this approach specifically for variable selection and demonstrate its applicability towards predicting civil wars. The method allows for identifying previously unknown influential variable sets, which may aid political theory-building when there exists contention over the composition of the set of influential variables amongst a larger set of political variable sets. That is not to say the presentation in this paper is theory-building in and of itself. Rather, it is possible that one

[2]Number of parameters given m variables is $n^m - 1$.

externality of excavating higher-order interactions may be in adding to the considered variables and their interacting behaviors in the pre-theory building process of the causes of civil war.

Note that this paper considers "influence" in reference to the ability of a variable(s) or variable set(s) to predict an outcome variable. In particular, I demonstrate the methodological benefits of using the Partition Retention method on a civil war data set to predict civil war. While extensive research has been conducted towards building theories on the factors that cause civil wars (inter alia, [9, 12, 16, 21]), I will focus on the prediction of observing a civil war in a given country and year. Indeed, the prediction of civil wars has become an important endeavor in and of itself (see [6, 32, 33] among others). Furthermore, it is becoming increasingly apparent that good prediction is not an automatic result of utilizing highly significant variables [32].[3]

The paper proceeds as follows. The first section provides a brief discussion on current approaches to predicting civil wars and the need for a method of finding highly predictive variables and variable sets, thus encompassing higher-order interactions. The second section provides an introduction to the Partition Retention (PR) method and its associated influence measure as a candidate approach. The third section describes the adaptation process for Fearon and Laitin's 2003 dataset, discusses uncovered interactions in the civil war data and presents prediction results. Future work will include the adaptation and results of predicting civil war onset, duration and termination.

Prediction of Civil Wars

More recently, interest in the prediction of violent events has increased. Goldsmith et al. (2013) use a two stage model to first predict instability and then predict the onset of genocide and politicide. Clayton and Gleditsch (2014) similarly use a two-stage approach towards first predicting civil war mediation and then likely success. A drawback from these types of analyses is that the variable selection process is theoretical in part; Goldsmith et al. use theory guided variable selection to predict instability, which in turn forms a predictor for genocide onset. Clayton and Gleditsch similarly use "*ex ante* knowable features highlighted in previous research" to select variables to predict for genocide onset and outcome. Hazlett (2011) explores prediction of genocide using a

[3]This is also found to be true in other fields. See [1].

forecasting model that takes into account the probability of genocide over the course of a political instability event. While this treatment of prediction is less theoretical in variable selection, no interactions between collected variables of interest are considered in the prediction model. In a working paper, Blair et al. (2014) favor a neural networks approach towards identifying important predictors for local violence in Liberia. Again, while the approach is more atheoretical, higher order interactions are not considered as potential predictors. Throughout these works, the motivation behind prediction of violent events is primarily in the policy realm; limitations and adequate responses to genocide or violent conflict can be sufficiently aided via accurate prediction to curtail the negative effects, or possibly remove the occurence, of such political events. Given the growth of these types of analyses in the last five years however, there seems to be a growing community of scholars interested in forecasting a violent political event.

When targeting prediction of civil wars, two main questions emerge: (1) What variables give us good prediction? (2) What is the appropriate class of models to consider? Currently, much of the variable selection is done through consideration of key covariates that explain causes of civil war in the civil war literature. Different models and approaches are then compared for out-of-sample performance rates. I argue for a fundamental shift in our approach towards prediction by first rethinking how we identify the influential variables we use in our models for prediction. This includes consideration of higher order interactions between explanatory variables when defining our influential variable sets. The Partition Retention (PR) method is one such potential methodological candidate; upon recovering influential variable sets, the researcher is then open to incorporating the predictive variable sets in a model of her own discretion. The PR approach can thus be considered a pre-model analytical tool for excavating highly influential variables for prediction purposes.

Currently most of the variables used as predictors for civil war stem from political theories seeking to explain the causes of war. Common variables include economic variables, such as state revenues from primary commodity exports, per capita GDP or costs to rebellion (in the greed or grievance framework suggested by [9, 10]). Others include measures of ethnic fractionalization, identity shifts and/or ethnic defection [21]. Geographic variables have gained popularity, as how a civil war is fought or whether an insurgency may be mounted are constrained by geographic limitations (described by measures such as elevation or ruggedness of the terrain) [5, 12]. Natural

resources have also been cited as related to civil wars, though the jury seems to still be out on precisely how that relationship may be manifested [28]. Scholars have also noted the relationship between natural phenomena such as climate change and civil conflict [6]. Finally, micro-level variables such as civilian attitudes are being investigated as the number of in-country field experiments and field surveys increase [20].

With these theories in hand, prediction has focused on identifying models that help raise correct prediction rates. For instance, Havard Hegre et al. have conducted extensive analyses of models aimed towards predicting armed conflicts; they propose the usage of a multinomial logit on a split-sample design (1 training group and 1 testing group). Selecting appropriate predictors, however, was done by identifying key variables in main theories of civil war, subjecting the variables to trial and error of model specification and comparing prediction outcomes. The authors inherently note the importance of possible information in interactions — they include some 2-way interactions in the models. Hegre et al. use a dynamic multinomial logit model that inputs explanatory variables that are well-known in the literature such as population, conflict history and development indicators such as mortality rates [19]. While transition probabilities of each of the variables are updated in the dynamic model, these are all weights on the marginal contributions of the independent variables. Appropriate predictors were selected through trial and error of model specification and comparing prediction outcomes. In addition, the authors only focus on marginal and 2-way interactions in all specifications.

It is clear from the wealth of theories on civil war outbreak that many key variables have been identified as significantly influential. However, it would not be difficult to believe that many of these same variables interact to influence civil wars. In particular, as a methodological advantage, when we wish to predict civil war it is highly desirable to be able to identify highly predictive variable sets that jointly influence outcome as well. In addition, while scholars like Ward et al. (2010) note that significant variables do not necessarily lead to good prediction, this is not necessarily true of their higher order interactions with other known variables [32]. Yet, extricating these highly predictive variables requires some screening mechanism.

The suggested approach in this paper encompasses higher order interactions and allows for ease of selection for candidate variable sets through backwards dropping.

Advantages of PR

The Partition Retention method is desirable as a statistical tool for social scientists for several reasons. First and foremost, it flexibly captures higher order interactions (beyond 3+) for a very large number of independent variables. For the sake of this exposition, top influential variable sets include up to 5-way interactions.[4] As of the writing of this paper, interactions of these orders have yet to be identified in the existing literature.

Second, the approach can be considered as an analytical tool meant to aid in the theory-building stage of the scientific process. While identified variable sets are meant to predict outcomes of interest, the higher order interactions found can then be considered as spring boards for building (though certainly not testing!) more interactive or jointly influencing causal theories. That is, preliminary data-gathering of this sort can perhaps, while not test a theory, help inform the theory building process.

Third, the variable sets identified by PR does prediction at worst similar to and at best better than common alternatives [7]. This is unsurprising; when the true relationship between a given outcome variable and a set of independent variables is based mostly on marginal effects, then methods that focus on marginal effects are likely to do just fine in prediction. However, should the relationship involve more complicated dependencies amongst the independent variables, then PR stands at an advantage [7]. This last appeal is actually of some importance; we may desire to know whether a particular political phenomenon is influenced by a complex network of independent variables or may very well be a simpler process composed of mostly marginal effects of individual variables.

Partition Retention: the $I-$score & Backwards Dropping

Here I briefly introduce the Partition Retention (PR) method [7]. Originally designed to tackle the overwhelming number of SNPs in the human genome that can jointly result in different disease phenotypes (and the resulting exponential boom in number of possible different interactions), the PR method has found success in predicting diseases such as breast cancer, irritable bowel disease (IBD), and prostate cancer [11, 25, 31]. As a brief description, the PR approach involves

[4] Up to 8-way interactions are identified, though are less influential in predicting civil war.

randomly selecting a smaller subset of a full set of explanatory variables and analyzing if any of the variables in the subset are associated with the outcome variable. An influence measure, the I-score, attributes an amount of influence to the subset of variables and measures the association with the outcome variable. Next, a step-wise elimination decreases the subset of variables to a returned, smaller set of potentially influential explanatory variables. These steps are repeated many times and returned subsets are considered potentially highly influencial towards the outcome variable.

Two main innovations form the core of the approach. First, the PR method offers an alternative measure of influence, the I-score, to the classical significance-based measures (chi-square, F-statistics etc.). The I-score is designed to retrieve variables and variable sets that exhibit maximum predictivity and can be seen as relatable to a multiple correlation coefficient. The second innovation is in the approach of *backwards-dropping* variables from variable sets in order to identify maximally predictive variable sets. This is in contrast to the common forward searching (e.g. recursive partition methods), which are more likely to miss higher order interactions amongst variables if their lower order signals are weak or mild. I present the I-score first:

Suppose we have a dataset of n observations with a binary observed variable Y of interest and a set S of explanatory variables measured as $\boldsymbol{X} = \{X_1, ..., X_S\}$. If all X take at most 3 discrete values (say 0, 1 or 2), then the idea is to partition the n observations into 3^S partition elements identified with the values of $\boldsymbol{X}$. There are n_i observations in each ith element. The I-score, or influence measure for how well the particular partition separates our observations into reasonably similar subsets (classifying between the two Y outcomes) is:

$$I = \frac{1}{n} \sum_i n_i^2 (\bar{Y}_i - \bar{Y})^2 \tag{1}$$

where $\bar{Y} = \sum_i \frac{n_i \bar{Y}_i}{n}$ is the average of Y overall and $\bar{Y}_i$ is the average in the ith element. Suppose we wish to measure the influence of a single variable X_1 on I. We would simply consider the coarser partition formed by the $\boldsymbol{X} = \{X_2, X_3, ..., X_S\}$ not including X_1, as if X_1 did not exist. The I-score of this new, coarse partition, differenced from the partition including the X_1 variable is a measure of the influence of X_1 on Y when appearing with the other X variables in $\boldsymbol{X}$. Should the new coarse partition produce a smaller I-score, we regard the X_1 variable as influential for Y. We repeat this process for all other X variables and consider the difference in I-scores, ultimately discarding the

X variable that produces the lowest I-score. This procedure is repeated until discarding another X variable from the remaining set of variables produces only increases in the I-score; at this point the set of variables remaining are all kept. This is the *backwards-dropping* element to the PR approach.

When S becomes too large to estimate the finest partition, we can select a subset or group of m variables from $\boldsymbol{X} = (X_1, X_2, ..., X_S)$ which defines a partition Π^* of the sample of n observations into $m_1 = 3^m$ subsets. We denote these partition elements $\{A_1, A_2, ..., A_{m_1}\}$. These are all possible values taken by our subset of m variables. For ease of presentation, let $\{X_1, X_2, ..., X_m\}$ denote the group of m variables selected from the original $\boldsymbol{X}$. Each A_j is a subset of $n_j Y$ values and $\sum n_j = n$. Every A_j (nonempty) has a mean value $\bar{Y}_j$. The overall mean is $\bar{Y} = \sum \frac{n_j \bar{Y}_j}{n}$. The I-score is then:

$$I_{\Pi^*} = \frac{1}{n} \sum n_j^2 (\bar{Y}_j - \bar{Y})^2 \tag{2}$$

Again, we can compare a coarser partition of $\{X_2, X_3, ..., X_m\}$ against the original partition $\{X_1, X_2, ..., X_m\}$ by comparing the difference in I-scores under the full set variables, including X_1, that form the original partition against the I-score retained under the coarser partition leaving out X_1. This difference is a measure of how much X_1 contributes in influence on Y in the presence of $\boldsymbol{X} = \{X_1, X_2, ..., X_m\}$. The equation for the difference between a coarse and finer partition when X_1 can take a finite set of values is:

$$D_I = -\frac{1}{n} \sum_i \sum_{j<k} n_{ij} n_{ik} (\bar{Y}_{ij} - \bar{Y})(\bar{Y}_{ik} - \bar{Y}) \tag{3}$$

where A_{ij} is the subset in A_i where $X_1 = j$ and has n_{ij} elements averaging to $\bar{Y}_{ij}$. Asymptotic properties of the I-score and D can be found in [7].

With this new influence score in hand, we turn to the second arm of the PR approach, the Backwards Dropping Algorithm (BDA). This step essentially involves repeated random sampling of subsets of variables from the total number of independent variables, finding joint I-scores, finding differences in I-scores and dropping variables from the subsets and rescoring until a maximum I-score is retrieved. The resulting subsets of variables are then ranked by I-score to find the most

influential variable subsets. Below are the steps to BDA (also illustrated in Figure 1).

Steps in the PR approach

1. Randomly draw a subgroup m of variables from the total number S variables. Calculate the I-score of this group of m variables. Call this $I(m)$.

2. From m, randomly draw a single variable from the group and calculate the I-score for the remaining $m-1$ variables. Do this for each variable in m. Compare I-scores in (2) with the I-score in (1), $I(m)$. If $I(m)$ is larger than all the I-scores in (2), keep all m variables as influential, and the process ends here. If not, find the variable that, when removed, results in the largest positive I-score difference. This variable is regarded as not influential towards the outcome and is discarded; there remain $m-1$ variables.

3. Repeat step 2 until removal of any remaining variables from a set of size m^* results in a lower I-score than the I-score of set m^*. m^* is the set of interacting and influential variables for the outcome of interest.

4. Repeat steps 1-3 many times for coverage of the full set of combinations of set m variables.

5. Repeat steps 1-4 for other random draws from S of size m.

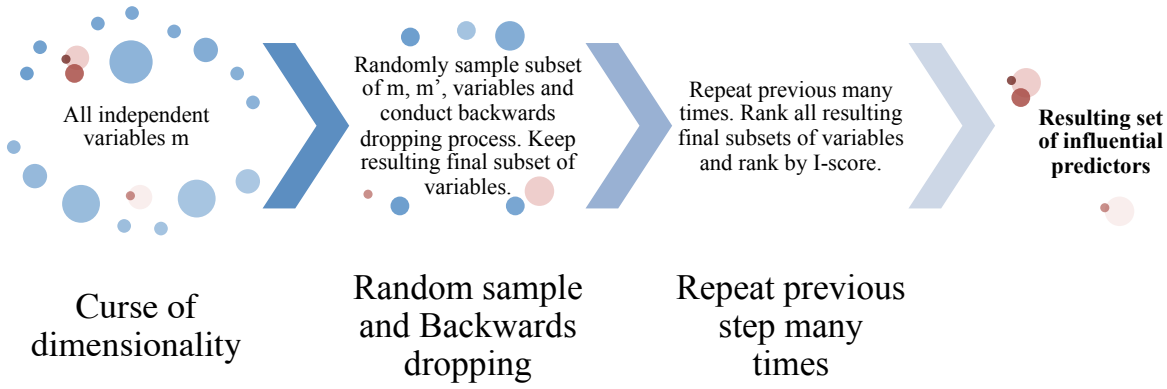

Figure 1: Backwards Dropping Process

A key characteristic of this process is its recursive search for influential variable sets. Forward seeking algorithms (inter alia, Random Forest) rely on starting with single variables and deciphering further partitions of the data according to measures of marginal effects; the resulting variable sets are thus partitioned along specific types of dependency amongst the variables chosen. The boon of backwards dropping is that removing variables (not adding them) leave the joint dependencies amongst the remaining variables in the variable set intact [7].

Although a returned influential variable set may be of higher order, lower-order influences in the same variable set may also exist, albeit masked by the higher I-score of the entire variable set. That is, suppose a variable set composed of 3 variables, $\{X_1, X_2, X_3\}$ is returned as highly influential. While the three-way interaction is clearly influential, this may be masking (slightly less) influential interactions at the lower order; any two-way combination of the three variables or their marginal effects may also be influential. Consequentially, when incorporating the returned variables and variable sets into our model of choice for prediction, we should include all combinations of lower-order interactions as well. While this will result in more parameters to estimate, a) this is already a heavy reduction of number of parameters, compared to the paramters for the original universe of possible variables and their interactions, and b) we can easily harness common models such as LASSO or other shrinkage and selection methods for regressions.

Applying PR to Civil Wars data

In 2010, Ward et al. considered the problem of statistically significant variables for civil conflicts performing poorly when predicting the onset of said conflicts. Ward and collaborators used data gathered by Fearon and Laitin as well as Collier and Hoeffler to demonstrate the poor predictive abilities of the models built from statistically significant variables. I argue that two issues are at play. First, high statistical significance does not necessarily lead to high predictivity. I discuss this phenomenon in further detail in other dissertation work [24]. Second, important information helpful towards predicting may be lost when only using marginal effects of independent variables. This paper focuses on this second issue.

The number of values each variable can take directly factors into the number of possible partitions of the data. Through simulations, the maximum number of values a variable can take before

the partitions grow too large and cumbersome to properly analyze seems to be 3 or 4. As such, I have recoded variables so they take a maximum of 3 possible values. While this process surely removes some information from each variable, the retained information boosts the number of variables we can analyze and the number of interactions we can consider. The recoded variables are only used in the variable selection process. When constructing final predictors for the prediction process we can return to the original and full versions of the variables.

Description of Analysis Process

Independent variables

The original Fearon and Laitin dataset included a total of 82 variables. After removing variables that contained too much missing data[5], as well as multiple measures of "war", 37 variables were included in the variables list (see 4) including the dependent variable for civil war, "war".[6] I use lagged versions of variables that change over time (for instance, country-years coded as 1 for "Asia" do not change over time).

As mentioned earlier, in the variable selection process I have recoded variables accordingly so they maximally reflect 3 possible values (recall that for the m variables selected this equates to 3^m partition elements). For instance, the variable *popl*, which is lagged population, is a variable that can take any positive integer value. For variables such as these, I have used k-means clustering to recode the variable, where k is equal to 3.

I conduct a 5-fold cross validation on the dataset, consisting of 600 cases and 3600 controls. Cross-validation is a common approach for tackling the problem of model overfitting [14, 22] and is desirable as a form of identifying out-of-sample prediction rates. Currently the literature is short on widely agreed upon cross-validation techniques for time-series or panel data. For the purposes of this illustration, I make the strong assumption of one-period dependency Markov chain in order to achieve exchangeability. Further work will include incorporation of a new cross validation technique designed for time series data in [27]. For each of the five training sets, I run the backwards dropping algorithm (PR steps 1 through 4) 10,000 times with a maximum starting

[5] More than 70% NAs.

[6] Imputation is a highly important and well-studied field in and of itself but I do not venture into this here. Rather than multiple imputation of missing values, which could change the makeup of the data depending on the assumptions and techniques used, and for the purposes of this paper, I dropped variables with too many NAs.

variable set size of 8. This is equivalent to allowing for capture of up to 8-way interactions, or 8-variable sets. Sample size and computational power constraints determine the maximum starting variable set size in the backwards dropping process. Given my training set sizes of 480 cases and 2880 controls, choosing from 8 to 10 as max variable set sizes are feasible. I choose 8 ultimately as this made my processing time faster (dropping from nearly 3 hours per training set to 30 minutes on a common PC desktop) and also because the top returned influential variable sets were very similar when using 8, 9 or 10 starting variable set sizes.

I retain variable sets with the top normalized I-scores to create predictors. These predictors are then used on the testing data in a logistic LASSO regression with war as the outcome variable.[7] Correct prediction (and error rates) rates are found for each fold, using 0-1 loss. Finally, I use the predictors constructed in the five folds on the independent testing set.

Results of Backwards Dropping

Table 1 shows the top returned variable sets from one of the five fold training sets. Recall that this is the result of the novel variable selection stage I propose as helpful in prediction efforts. Using the PR's I-score and backwards dropping from randomly drawn 8-variable groups, I identify variable sets with the highest I-scores and keep these as variable sets of interest for prediction. Most if not all of the variables are unsurprising; after all, the population of variables collected by Fearon and Laitin were specifically meant to either control for or be significantly associated with civil wars. However, short of variables that code for a lagged civil war, all other top returned variable sets are of higher order. The results also mirror Fearon and Laitin's in that ethnic and religious fractionalization are perhaps less influential than per capita income or factors that facilitate insurgent activities such as poverty, instability, rough terrain and large populations.

[7] I use logistic LASSO as my model. Logistic regression is a common binary choice model for the civil war literature. Because of the inclusion of all lower orders of each higher-order influential interaction from the backwards dropping process however, we are left with a slightly large number of parameters to estimate (though this is still orders of magnitude smaller than the number of parameters needed to estimate all variable interactions).The LASSO is a common shrinkage and selection approach. I also use the LASSO on a comparison model to demonstrate that it is not the LASSO itself that brings higher prediction rates. See [17, 30].

1 variable	2 variables	3 variables	4 variables
war lagged	Population, colonial war	# of languages, new state, instability	Log population, GDP type, Eastern European
	Population, previous war	Elevation difference, western, anocracy	Log population, GDP, Eastern European, former French colony
	# of languages, former French colony	GDP, elevation difference, Eastern Europe	Asia, noncontiguous states, instability, anocracy
	log, population, anocracy	Log population, Eastern Europe, instability	# of languages, Muslim, North Africa/Middle East, former French colony
	# of languages, anocracy		

Table 1: **Example of top returned variable sets**. Returned variable sets with the highest I-scores from one of the five-fold training sets are illustrated here. All variables are lagged unless otherwise indicated.

Preliminary results from models

After identifying our sets of influential variables under the PR approach, for each of the five fold training sets, I turn to incorporating returned influential variable sets into model form and test out-of-sample prediction error rates. Here each training set and its associated set of influential variables (found through PR) are subjected to the logistic LASSO with a 10-fold cross validation to determine the tuning parameter (λ) value that minimizes training error. Figure 2 shows the variable coefficient paths as the tuning parameter changes in one training set. Each curve on the graph corresponds to a variable or variable set. The curve traces the path of the coefficient against the L1-norm of the whole coefficient vector as λ (the tuning parameter in lasso regression) varies. The blue and red curves on the top half of the figure are coefficient paths drawn from marginal effects — these variables are retained for most values of λ. Thus, when simply looking at the training set, the data already demonstrates reasonably strong marginal effects from variables such as whether there was a previous year of war. This is logical given that a previous year of war can be indicative of a continuation of war while a year of peace more often leads to another war-free

year.

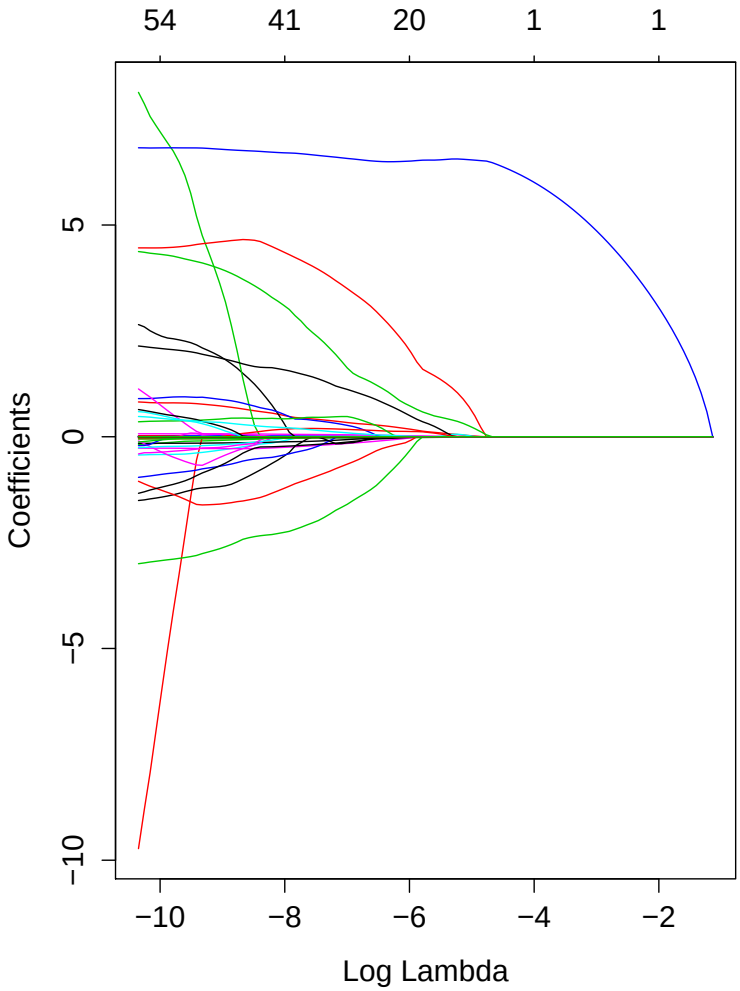

Figure 2: **Variable coefficient paths against L1-norm as λ tuning parameter varies**. The axis shows the number of nonzero coefficients at each λ (the degrees of freedom for the lasso). Results are from one training set only.

Again, the λ that minimizes the training error (λ^*) can be identified through simple further cross validation within each training set. As such, I use 10-fold cross-validation within each of the five training sets to locate the λ^* . This is not to be confused with the 5-fold cross validation design for training influential variable sets under the PR. This 10-fold cross validation is within each of the 5-fold training sets. No testing sets have been touched at this stage. Figure 3 shows the cross-validation curve (red dots) and the upper and lower standard deviation curves along the λ sequence. The λ^* is the first dotted horizontal line on the left, while the λ of the first standard

deviation is the horizontal dotted line on the right. These correspond to the lowest dips in the cross-validation curve, minimizing the binomial deviance. The top set of numbers are the number of parameters kept by the model at each value of λ. At λ^* this is 28 parameters.

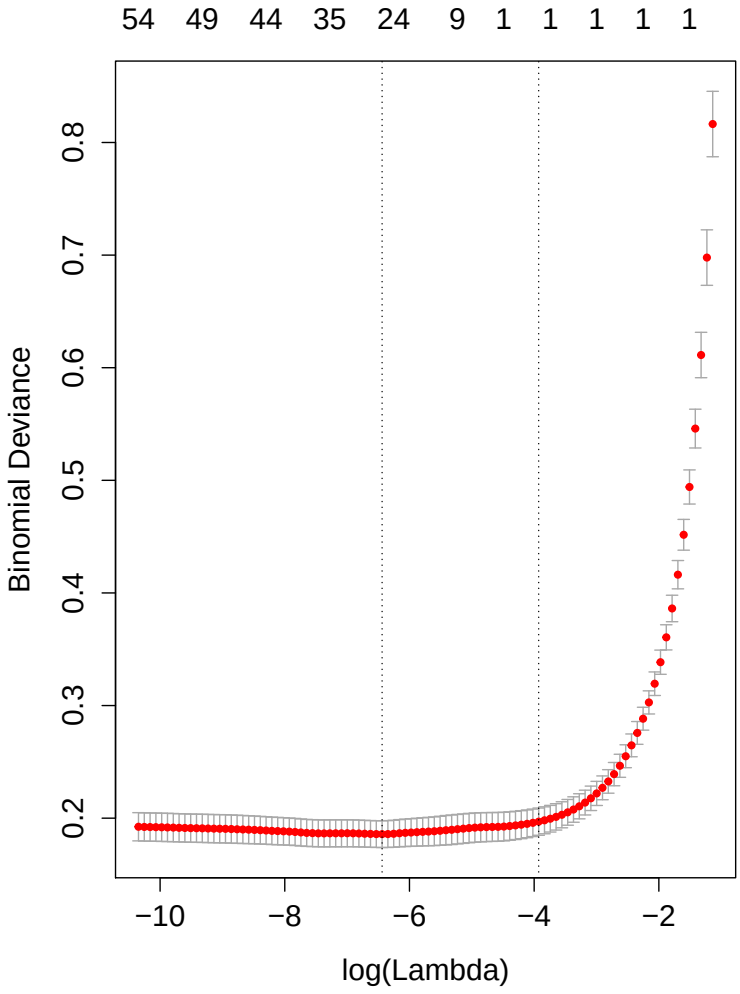

Figure 3: **Cross-validation curve fit**. Results are shown for training set 1 only.

The cross-validating PR logistic LASSO tunes certain variables to coefficients of 0. Figure 4 illustrates the retained variables and variable sets, as well as their coefficients for one example training set. Strong marginal effects result from lagged war (warl), whether a country is a new state (nwstate), a measure of instability (instabl), and whether a country is noncontiguous (ncontig). Most remaining coefficients result from higher order interactions.

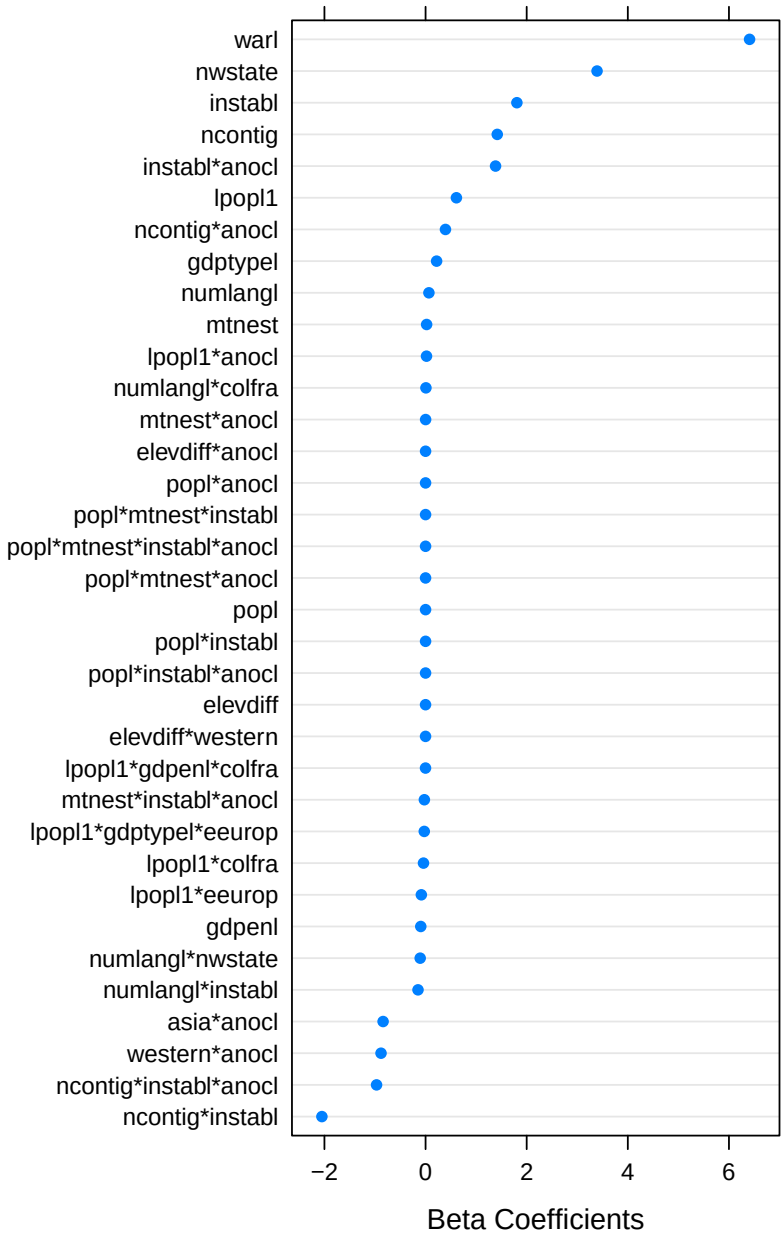

Figure 4: **Beta coefficients for parameters retained in CV PR Logistic LASSO**

Results of Prediction

As a means of comparison, I use Fearon and Laitin's selected variables to run a logistic regression ("FL Logistic") as well as a logistic LASSO ("FL Logistic LASSO")[8] and compare error testing

[8]Fearon and Laitin selected variables are found in Table 1 of their 2003 paper: prior war, per capita income, log(population), log(% mountainous), noncontiguous state, oil exporter, new state, instability, democracy, ethnic

rates across the 3 models (see Figure 2). Because the data is highly unbalanced (testing sets each have 120 cases of civil war and 720 cases of no civil war, or about 14.3% civil wars), blindly predicting that no civil war will occur in a given country-year gives an accuracy rate of 85.7%. Therefore, an highly desirable classifier should be able to not only correct predict no civil war occurring in a country-year but, more importantly, be able to predict a civil war *occurring* in a given country-year.

From Table 2 the logistic LASSO that uses PR identified variable sets does substantially better with an average testing error rate of 2.4% (thus correctly predicting the outcome 97.6% of the time). More importantly, however, we have to make comparisons between correct prediction rates of true positives (correctly guessing that a civil war occurs).

Error rates	**FL Logistic**	**FL Logistic LASSO**	**PR Logistic LASSO**
Fold 1	0.14	0.14	0.021
Fold 2	0.14	0.14	0.025
Fold 3	0.14	0.14	0.013
Fold 4	0.14	0.14	0.023
Fold 5	0.13	0.13	0.02
Average Error Rate	**0.138**	**0.138**	**0.024**
Average Correct Prediction Rate	86.2%	86.2%	97.6%

Table 2: **5-fold CV prediction error rates on testing sets.** FL models using variables from Fearon & Laitin's 2003 main regression. PR Logistic LASSO uses variable sets retained from the Partition Retention method. All error rates are from testing sets.

In Table 3, I break down the errors into false positives and negatives and the correctly predicted cases into true positives and negatives. False positives occur when the model predicts a civil war occurring when the true outcome was no civil war, while false negatives occur when the model predicts a civil war-free year when in reality a civil war did occur. Most importantly, while the PR approach results in slightly more false positives than the FL models, the number of false negatives are drastically cut by roughly 75%. More civil wars are correctly predicted as occurring under the PR approach. A false negative (predicting that a civil will not occur and actually seeing a civil war) is perhaps more normatively problematic than a false positive (predicting a civil war and not seeing one).

I also provide statistics on true positives (correctly predicting a civil war) and true negatives

fractionalization, religious fractionalization, anocracy, democracy.

(correctly predicting no civil war). While FL Logistic and FL Logistic LASSO predict roughly 6-7 civil wars occurring, the PR Logistic LASSO correctly predicts 108 out of the total 120 actual civil war years in each of the testing sets. This boost in true positives is from 5.67% to 90% accuracy. True negatives under the PR logistic LASSO dips slightly below the FL logistic and FL logistic LASSO percentages (99.75)%, at 99.28%. In other words, while the FL models are high in specificity[9] (both 99.83%), they are quite low in sensitivity[10] (5.65% and 5.32%). The PR logistic LASSO however is more balanced between specificity and sensitivity, at 99.36% and 84%, respectively. I remark here that using the logistic LASSO with variables selected through significance testing (FL logistic LASSO) provides little improvements to error rates overall. Thus, it appears the variables selected as influential in the PR logistic LASSO are exerting most of the prediction work, and not the model or model specification.

Average Across Folds					
	False Positives	False Negatives	Total # of Errors	True Positives	True Negatives
FL Logistic	1.2	113.6	114.8	6.8 (5.67%)	718.2 (99.75%)
FL Logistic LASSO	1.2	114	115.2	6.4 (5.33%)	718.2 (99.75%)
PR Logistic LASSO	4.6	20.6	25.6	108.2 (90%)	714.8 (99.28%)

Table 3: **False Positives, False Negatives, True Positives & True Negatives under 3 models**. Total positives (wars) in the testing set = 120. Total negatives (no wars) in the testing set = 720. True positive and true negative percentages of total positives and negatives are given in parentheses.

Conclusion

Predicting civil wars requires both identifying highly influential variables and variable sets and selecting good models that minimize prediction errors in out-of-sample testing. Since the main goal is minimizing error rates, and not establishing significance of key independent variables as

[9] Specificity= $\frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives}$

[10] Sensitivity= $\frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives}$

in the case of theory-building and testing, we should approach predicting civil wars differently. Current efforts towards prediction often turn to theoretical work on the causes of civil war as a method of identifying variables to use in models for prediction. Selecting significant variables for prediction does not seem to automatically lead better predictions, however. I argue in this paper that it is in fact the universe of collected independent variables — *and all of their possible interactions* — that compose the full set of possible variables to consider for prediction. This atheoretical approach to finding variables and variable sets is more appropriate for an atheoretical goal.

However, given the large set of variables and their interactions, selecting for influential variables a) without using significance as a criterion and b) taking into account higher order interactions is a difficult feat. This paper proposes applying the Partition Retention method as a possible candidate for these two goals and illustrates the method using Fearon and Laitin's 2003 civil wars data. Results from the analysis are promising. Compared to similar models that use significance-based selection of variables, the model that uses variable sets returned through PR searching improves prediction rates by 11.4% (from 86.2% to 97.6%). In particular, false positives, are drastically lower, while true positives are roughly 92%.

This analysis only tackles the prediction of a civil war occurring given a country and a year, however. Predicting civil war onsets, duration, and termination are all highly important as well and require further investigation. In addition, the analysis here is meant to demonstrate the applicability of the PR method. It is highly likely that important existing variables not collected in the Fearon and Laitin dataset may capture important influence for the prediction of civil war. Further work needs to use a more exhaustive set of data to retrieve influential variables and variable sets. The analysis conducted here requires a strong assumption of one-period dependent Markov chain exchangeability. Next steps include incorporation of a cross-validation technique with a moving window of training and testing sets, designed for time series cross-sectional data like that within this paper.

Another avenue of further research is to demonstrate application of the PR approach when we are seeking for variable sets that explain civil war, that, due to their higher orders, might evade traditional significance-based testing. It may be beneficial to have a variable set selection stage

prior to a researcher's theory-building stage in order to identify key groups of variables that are significantly associated with the outcome variable of interest. This may facilitate the researcher's development of the causal theory behind how the identified group of variables interact to affect the outcome. The researcher may then decide to collect data to test the developed theory. The variable set selection stage would entail calculations of the I-scores of the variable sets and creating permutations of null distributions of I-scores. Variable sets that have I-scores that are significant at specified alpha levels when compared with the null distributions can be deemed statistically significant and kept aside for consideration in the theory-building stage.

Finally, part of the ongoing collaborated work related to this project is in the creation of a user-friendly R package that can allow easy implementation of the PR approach.

Additional Tables

	Mean	Standard deviation
year	1977.30	9.60
popl (lagged population)	31999.8	103237.6
lpopl1 (lagged log population)	9.044	1.479
polity2l (lagged polity II score)	-1.17	7.58
gdpenl (lagged GDP)	3.825	4.507
lgdpenl1 (lagged log GDP)	7.722	1.028
mtnest (% mountainous terrain)	17.44	21.07
lmtnest (lagged mtnest)	2.107	1.420
elevdiff (elevation difference)	3114.72	2005.54
ethfracl (lagged ethnic fractionalization)	0.4038	0.2905
efl (lagged EF score)	0.4751	0.2726
plurall (lagged plurality)	0.6401	0.2510
secondl (share of 2nd largest ethnic group)	0.1598	0.1103
numlang (# of languages)	7.36	7.46
relfracl (lagged religious fractionalization)	0.3734	0.2193
plurrel (size largest confession)	72.20	20.46
musliml (lagged % muslim population)	25.91	37.13
minrelpcl (size second largest confession)	18.66	13.04
gdptypel (lagged GDP type)	0.423	1.144
western	0.1659	0.3720
eeurop (Eastern Europe)	0.0760	0.2650
lamerica (Latin America)	0.1784	0.3829
ssafrica (Southern Africa)	0.2862	0.4520
asia	0.1640	0.3703
nafrme (North Africa/Middle East)	0.1296	0.3359
colbrit (former British colony)	0.303	0.460
colfra (former French colony)	0.192	0.394
oill (lagged dummy >33% oil exports)	0.1336	0.3402
ncontig (noncontiguous state)	0.1675	0.3734
nwstate (new state)	0.0146	0.1201
instabl (lagged instability)	0.1310	0.3375
anocl (lagged anocracy)	0.1807	0.3848
deml (lagged democracy)	0.310	0.462
warl (lagged war)	0.1373	0.3442
war	0.1431	0.3502

Table 4: **Descriptive statistics of Independent variables used in Partition Retention**. Variable explanations in parentheses.

Fold 1					
	False Positives	False Negatives	Total # of Errors	True Positives	True Negatives
FL Logistic	4	112	116	8	716
FL Logistic LASSO	4	112	116	8	716
PR Logistic LASSO	8	10	18	110	712

Table 5: **Fold 1 Testing Set: False Positives, False Negatives, True Positives & True Negatives under 3 models**. Total positives (wars) in the testing set = 120. Total negatives (no wars) in the testing set = 720.

Fold 2					
	False Positives	False Negatives	Total # of Errors	True Positives	True Negatives
FL Logistic	1	115	116	5	719
FL Logistic LASSO	1	115	116	5	719
PR Logistic LASSO	5	16	21	104	715

Table 6: **Fold 2 Testing Set: False Positives, False Negatives, True Positives under PR identified variable sets**. Total positives (wars) in the testing set = 120. Total negatives (no wars) in the testing set = 720.

Fold 3					
	False Positives	False Negatives	Total # of Errors	True Positives	True Negatives
FL Logistic	2	115	117	5	718
FL Logistic LASSO	2	117	119	3	718
PR Logistic LASSO	4	7	11	113	716

Table 7: **Fold 3 Testing Set: False Positives, False Negatives, True Positives under PR identified variable sets**. Total positives (wars) in the testing set = 120. Total negatives (no wars) in the testing set = 720.

Fold 4					
	False Positives	False Negatives	Total # of Errors	True Positives	True Negatives
FL Logistic	2	113	115	7	718
FL Logistic LASSO	2	113	115	7	718
PR Logistic LASSO	4	14	19	106	716

Table 8: **Fold 4 Testing Set: False Positives, False Negatives, True Positives under PR identified variable sets**. Total positives (wars) in the testing set = 120. Total negatives (no wars) in the testing set = 720.

Fold 5					
	False Positives	False Negatives	Total # of Errors	True Positives	True Negatives
FL Logistic	0	111	111	9	720
FL Logistic LASSO	0	111	111	9	720
PR Logistic LASSO	5	11	17	108	715

Table 9: **Fold 5 Testing Set: False Positives, False Negatives, True Positives under PR identified variable sets**. Total positives (wars) in the testing set = 120. Total negatives (no wars) in the testing set = 720.

References

[1] Predicting the influence of common variants. Nat. Genet. 45, 339 (2013).

[2] Aas Rustad, S. C., Buhaug, H., Falch, a., & Gates, S. (2011). All Conflict is Local: Modeling Sub-National Variation in Civil Conflict Risk. Conflict Management and Peace Science, 28(1), 15–40. doi:10.1177/0738894210388122

[3] Braumoeller, B. F. Hypothesis Testing and Multiplicative Interaction Terms. Int. Organ. 58, 807–820 (2004).

[4] Brambor, T. Understanding Interaction Models: Improving Empirical Analyses. Polit. Anal. 14, 63–82 (2005).

[5] Buhaug, H. & Gates, S. The Geography of Civil War. J. Peace Res. 39, 417–433 (2002).

[6] Burke, M. B., Miguel, E., Satyanath, S., Dykema, J. A. & Lobell, D. B. Warming increases the risk of civil war in Africa. Proc. Natl. Acad. Sci. U. S. A. 106, 20670–20674 (2009).

[7] Chernoff, H., Lo, S.-H. & Zheng, T. Discovering influential variables: A method of partitions. Ann. Appl. Stat. 3, 1335–1369 (2009).

[8] Clayton, Govinda and Kristian Skrede Gleditsch. 2014. "Will we see helping hands? Predicting civil war mediation and likely success." Conflict Management and Peace Science, 3(3): 265-284.

[9] Collier, P. & Hoeffler, A. On the Incidence of Civil War in Africa. J. Conflict Resolut. 46, 13–28 (2002).

[10] Collier, P. Greed and grievance in civil war. Oxf. Econ. Pap. 56, 563–595 (2004).

[11] Fan, R. & Lo, S.-H. A robust model-free approach for rare variants association studies incorporating gene-gene and gene-environmental interactions. PLoS One 8, e83057 (2013).

[12] Fearon, J. D. & Laitin, D. D. Ethinicity, Insurgency, and Civil War. Am. Polit. Sci. Rev. 97, 75–90 (2003).

[13] Fearon, J. D. Primary Commodities Exports and Civil War. J. Conflict Resolut. (2005). doi:10.1177/0022002705277544.

[14] Geisser, Seymour (1993). Predictive Inference. New York, NY: Chapman and Hall. ISBN 0-412-03471-9.

[15] Goldsmith, B. E., Butcher, C. R., Semenovich, D., & Sowmya, a. (2013). Forecasting the onset of genocide and politicide: Annual out-of-sample forecasts on a global dataset, 1988-2003. Journal of Peace Research, 50(4), 437–452. doi:10.1177/0022343313484167

[16] Gurr, T. R. 1970. Why men rebel. Princeton: Princeton University Press.

[17] Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning: Data mining, Inference and Prediction. Second Edition. Springer Science+Business Media (2009).

[18] Hazlett, Chad. 2011. "New lessons learned? Improving genocide and politicide forecasting." Paper presented at the United States Holocaut Memorial Museum, Washington, DC, October 2011.

[19] Hegre, H., Karlsen, J. & Nygård, H. Predicting Armed Conflict , 2010 – 2050. Typescript, Univ. . . . 250–270 (2009). at <http://folk.uio.no/hahegre/Papers/PredictionISQ_Final.pdf>

[20] Hirose, K., Imai, K. & Lyall, J. Can Civilian Attitudes Predict Civil War Violence? (2013). Working paper.

[21] Kalyvas, S. N. Ethnic Defection in Civil War. Comp. Polit. Stud. 41, 1043–1068 (2008).

[22] Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12): 1137–1143.(Morgan Kaufmann, San Mateo, CA)

[23] Lo, A, Chernoff. H., Zheng, T., & Lo, S. Making good prediction: a theoretical framework. (2015). Working paper (under review).

[24] Lo, A, Chernoff. H., Zheng, T., & Lo, S. Why aren't significant variables automatically predictive? (2015). Working paper (under review).

[25] Lo, S.-H., Chernoff, H., Cong, L., Ding, Y. & Zheng, T. Discovering interactions among BRCA1 and other candidate genes associated with sporadic breast cancer. Proc. Natl. Acad. Sci. U. S. A. 105, 12387–12392 (2008).

[26] Lu, Henry Horng. "Discovering Influential Variables: A General Computer Intensive Method for Common Genetic Disorders." Handbook of statistical bioinformatics. Berlin: Springer, 2011. Print.

[27] Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. Journal of Econometrics, 99, 39–61.

[28] Ross, M. L. What Do We Know about Natural Resources and Civil War? J. Peace Res. 41, 337–356 (2004).

[29] Sambanis, N. A Review of Recent Advances and Future Directions in the Quantitative Literature on Civil War. Def. Peace Econ. 13, 215–243 (2002).

[30] Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. J. R. Stat. Soc. Ser. B 58, 267–288 (1996).

[31] Wang, H., Lo, S.-H., Zheng, T. & Hu, I. Interaction-based feature selection and classification for high-dimensional biological data. Bioinformatics 28, 2834–42 (2012).

[32] Ward, M. D., Greenhill, B. D. & Bakke, K. M. The perils of policy by p-value: Predicting civil conflicts. J. Peace Res. 47, 363–375 (2010).

[33] Weidmann, N. B., & Ward, M. D. (2010). Predicting Conflict in Space and Time. Journal of Conflict Resolution, 54(6), 883–901. doi:10.1177/0022002710371669.