WPS8075

Policy Research Working Paper 8075

Predicting Conflict

Bledi Celiku
Aart Kraay

WORLD BANK GROUP
Fragility, Conflict and Violence Cross Cutting Solution Area
&
Development Research Group
Macroeconomics and Growth Team
May 2017

Policy Research Working Paper 8075

Abstract

This paper studies the performance of alternative prediction models for conflict. The analysis contrasts the performance of conventional approaches based on predicted probabilities generated by binary response regressions and random forests with two unconventional classification algorithms. The unconventional algorithms are calibrated specifically to minimize a prediction loss function penalizing Type 1 and Type 2 errors: (1) an algorithm that selects linear combinations of correlates of conflict to minimize the prediction loss function, and (2) an algorithm that chooses a set of thresholds for the same variables, together with the number of breaches of thresholds that constitute a prediction of conflict, that minimize the prediction loss function. The paper evaluates the predictive power of these approaches in a set of conflict and non-conflict episodes constructed from a large country-year panel of developing countries since 1977, and finds substantial differences in the in-sample and out-of-sample predictive performance of these alternative algorithms. The threshold classifier has the best overall predictive performance, and moreover has advantages in simplicity and transparency that make it well suited for policy-making purposes. The paper explores the implications of these findings for the World Bank's classification of fragile and conflict-affected states.

This paper is a product of the Fragility, Conflict and Violence Cross Cutting Solution Area, and the Macroeconomics and Growth Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at at bceliku@worldbank.org or akraay@worldbank.org.

Predicting Conflict

Bledi Celiku (World Bank, FCV)
Aart Kraay (World Bank, DECRG)

JEL Codes: C53, C25, D74, O10

World Bank, 1818 H Street NW, Washington DC 20433, bceliku@worldbank.org, akraay@worldbank.org. We are grateful to Leonardo Garrido for research assistance, and to Ana Areias, Dhruv Grover, Jonathan Hersh, Gary Milante, Nadia Piffaretti, Nicholas Sambanis, Luis Serven, Jake Shapiro, Raju Singh, Roy Van der Weide, Michael Ward, Nils Weidmann, and seminar participants at the World Bank, the 2017 Empirical Studies of Conflict Project Annual Meeting, and the 2017 Barcelona Prediction for Prevention Workshop for helpful feedback. The views expressed here are the authors' and do not reflect those of the World Bank, its Executive Directors, or the countries they represent.

1. Introduction

This paper analyzes alternative models for conflict prediction. We work with a country-year panel dataset covering 114 developing countries since 1977, and define a set of conflict episodes based on the frequency of battle deaths, the presence of UN peacekeeping operations, and the proportion of the population forcibly displaced as refugees across international borders. In this dataset, we assess alternative prediction models that seek to distinguish these conflict episodes from complementary non-conflict episodes, using data on a number of country-level covariates observed prior to the start of the episodes. Our interest in this prediction exercise is motivated by the desire of aid donors and multilateral organizations such as the World Bank and the United Nations for a framework that can be used to anticipate the outbreak of conflict.[1] Ideally, the signals of the risk of conflict generated by such a framework open the possibility of policy interventions to reduce the risk of conflict before it occurs, rather than simply reactions to conflict after the fact.

In this paper, we contribute methodologically and pragmatically to the extensive empirical literature on conflict prediction. Our methodological contribution is based on the observation that conventional approaches to predicting conflict, whether based on fitted values from binary response regression models or machine learning algorithms such as random forests, suffer from an internal inconsistency: the objective function that is optimized when the prediction model is fitted to the data is different from the objective function that typically is used to evaluate the quality of the resulting predictions. Consider for example predictions of conflict based on the fitted values from a probit or logit regression model. Estimating the regression model involves maximizing a likelihood function which typically weights all observations equally, and in particular does not assign any particular priority to fitting conflict or non-conflict observations well. The same is true for random forest algorithms, where the branches in the underlying classification trees are chosen to maximize the homogeneity of observations in the resulting subgroups, again weighting all observations equally.

In contrast, forecast performance typically is evaluated using a prediction loss function that assigns specific weights to Type 1 error rates (i.e. the proportion of conflict observations incorrectly classified as non-conflict) and Type 2 error rates (i.e. the proportion of non-conflict observations

[1] For example, in the latest donor replenishment of resources to the International Development Association (IDA, the part of the World Bank that provides concessional loans and grants to the world's poorest countries), the World Bank committed to "adopt a risk-based approach for identifying fragility", where "fragility" is understood to mean vulnerability to the outbreak of conflict.

incorrectly classified as conflict). This inconsistency between the estimation and prediction objective functions can be particularly acute in applications such as conflict prediction, where conflict is a relatively infrequent event. In this case, the small number of conflict observations in the sample carry only a small weight in the likelihood function that is maximized at the estimation stage, but failure to predict conflict when it does occur may carry a high weight in the prediction loss function.

While some ad-hoc solutions that in effect over-weight the relatively infrequent conflict observations at the estimation stage are available, we suggest an alternative, and in our view more direct, approach: using prediction algorithms that are calibrated to directly minimize the same prediction loss function that is also used to evaluate the quality of the predictions. We implement two such algorithms. The first algorithm is similar to predictions based on fitted values from binary response regression models, in that it predicts conflict if a linear combination of correlates of conflict crosses a given threshold. The difference is that the weights in this linear combination are chosen to directly minimize the prediction loss function, rather than to maximize the likelihood function of the binary response model. We refer to this algorithm as the "linear classifier". The second algorithm chooses a threshold for each of the correlates of conflict included in the model, together with the number of breaches of thresholds that constitute a prediction of conflict. The thresholds and the number of breaches are jointly chosen to minimize the prediction loss function. We refer to this algorithm as the "threshold classifier". We contrast the predictive performance of these algorithms with standard predictions based on probit regressions and random forests, and document improvements in predictive performance that come from our proposed approach of using prediction algorithms directly calibrated to minimize the prediction loss function.

Our pragmatic contribution is based on the observation that in some circumstances there may be a tradeoff between the sophistication of a prediction methodology and its suitability for policymaking purposes. For example, an international organization such as the United Nations or the World Bank might want to use an assessment of the risk of conflict across countries as a tool to persuade donor countries to devote resources to aid programs that could help to mitigate the risk of conflict. Such an assessment could also be a crucial part of the policy dialogue with the countries that themselves are identified as being at risk of conflict, and play a role in the process of persuading them to accept assistance and adopt reforms that might reduce the risk of conflict. In such a setting, the acceptability of the framework for identifying risks of conflict might in part depend on the simplicity and transparency of the underlying model that is used to generate the predictions of conflict.

As the techniques used for conflict prediction have become increasingly sophisticated, evolving from predictions based on simple logit and probit regressions to tools from the machine learning literature such as random forests and neural networks, there is a risk that the predictions generated by more sophisticated models, even though possibly more accurate, may be less useful to policymakers in these circumstances. In contrast, the threshold classifier we propose is arguably one of the most straightforward and intuitive prediction algorithms for policymakers to understand. This is because the threshold classifier has a very simple form: given data on K indicators of conflict, we predict conflict if more than $N \leq K$ of the indicators cross a pre-specified threshold value for each of the indicators.

In fact, the threshold classifier we propose can be thought of as a more rigorously-founded generalization of the rule that the World Bank currently uses to identify countries at risk of conflict. The current rule is based on IDA countries' scores on the Country Policy and Institutional Assessments (CPIA) that are carried out by the World Bank, the Asian Development Bank, and the African Development Bank, as well as the presence of a UN and/or regional peacekeeping or peacebuilding operation. Specifically, countries with CPIA scores below 3.2 (on a 1 to 6 scale), and/or a UN peacekeeping operation in the previous three years, are classified as "fragile situations", and become the focus of conflict-prevention interventions. This rule is a threshold classifier with $K = 2$ indicators and corresponding thresholds (CPIA scores, with a threshold of 3.2, and a dummy for UN peacekeeping operations, with a threshold of 0), together with an aggregation rule that signals risk of conflict if at least $N = 1$ of these two thresholds is crossed.[2] The threshold classifiers we study in this paper are motivated by this simple policymaker-friendly rule, but we generalize it by considering a longer list of explanatory variables, and more importantly, by choosing the corresponding thresholds and the number of breaches required to signal conflict optimally to minimize a prediction loss function.

This threshold-based decision rule for classifying conflict episodes is simpler and more transparent than approaches that base predictions on a weighted average of predictors of conflict, and also is simpler than more sophisticated machine learning algorithms such as random forests. Interestingly, in our dataset we find that this pragmatic advantage of the threshold classifier does not come at the cost of worse predictive performance. In most cases that we consider, the threshold classifier also generates the lowest values of the prediction loss function among the different

[2] Another example of a threshold classifier in regular policy use is the World Bank/IMF Debt Sustainability Framework for low-income countries. This framework features a set of thresholds for five different debt burden indicators, together with a rule that predicts a risk of debt servicing difficulties or "debt distress" if any of these thresholds are breached.

alternatives we consider. This combination of simplicity and good predictive performance makes the threshold classifier an appealing option for operational policymaking purposes.

While we develop the methodological and pragmatic contributions of this paper in the specific setting of conflict prediction, both have wider relevance in other settings as well. In recent years there has been a strong movement among empirical economists to apply predictive tools from the machine learning literature to a variety of economic applications (see for example Mullainathan and Spiess (2017) and Athey (2017) for recent reviews of this trend). In cases where these applications involve binary predictions (e.g. poor vs. non-poor, employed vs. unemployed, etc.), it may also be possible to improve the performance of prediction algorithms by eliminating the inconsistency between the estimation and prediction stages that we emphasize in this paper. Similarly, the point that simple prediction algorithms may be better suited for policy purposes carries over to other settings as well. For example, many anti-poverty programs are intended to be targeted to a particular group of beneficiaries below a certain income threshold. In cases where it is difficult to observe income directly, these programs rely on "proxy means tests" to identify beneficiaries based on a small number of more easily-observable characteristics. Both the practical implementation of such a targeting rule, as well as its political acceptability to those who pay for and benefit from the program, may be enhanced by having a simple and transparent algorithm such as the threshold classifier.

The application of our contributions to conflict prediction builds on a large empirical literature that has analyzed why some countries experience conflict while others do not. Much of this literature is surveyed in Blattman & Miguel (2010), and World Bank (2011) contains an extensive discussion of this literature and its implications for development policy. This literature has emphasized factors such as ethnic tensions (Kalyvas (2008), Fearon and Laitin (2003)), economic greed or grievances (Collier and Hoeffler (1998), Collier and Hoeffler (2002), Collier and Hoeffler (2004)), geographical factors and natural resource endowments (Ross (2004), Fearon (2005)) and even the effects of climate change (Hsiang et. al. (2013)). Hegre and Sambanis (2006) is an important early attempt to assess the robustness of the many findings in this literature, using a variant of extreme bounds analysis to identify a small set of consistently-significant predictors of conflict.

A growing number of papers have built on the insights of this literature to develop models for predicting conflict, and Cederman and Weidmann (2017) provide a recent non-technical overview. Many of these follow the conventional regression-based approach, estimating econometric models of conflict using probit or logit models in panel datasets, and then use the fitted values of these to predict

subsequent conflict events. In this sense, these studies are conceptually similar to the "probit classifier" that we analyze in this paper. Papers in this category include Hegre et. al. (2013) who estimates a multinomial logit model for conflict over the period 1970-2009 and use it to predict conflict forward through 2050; Hegre et al (2016) who offer 100-year projections of conflict under alternative climate change scenarios using a similar methodology; Brandt et. al. (2011) who focus on predicting the time-series behavior of the severity of Israeli-Palestinian conflict; Chadefaux (2014) who emphasizes the importance of high-frequency data on news reports of conflict as a predictor of conflict; Weidman and Ward (2010) who emphasize spatial information as predictors of conflict; and Ward et. al. (2010) who point out that the significance of correlates of conflict in the estimating equation is a poor guide to their ultimate predictive performance, and that the addition of a statistically significant variable can actually reduce the predictive power of these models.

Other papers in this literature have used prediction algorithms from the statistical machine learning literature. An early example is Beck, King and Zeng (2000), who contrast the predictive power of a simple logit model of conflict with that of a neural network, emphasizing the ability of the latter to capture the nonlinearities and interactions driving the incidence of conflict. O'Brien (2010) summarizes the results of a major undertaking by the US military to develop a prediction model for "events of interest" that averages together forecasts based on a wide range of methodologies, including text analytics of news and political leader speeches. Perry (2013) uses random forests and naive Bayes classifiers to develop an early-warning model of conflict, using subnational conflict data in Africa, and finds that both methods improve over predictions based only on lagged conflict. More recently, Muchlinski, Siroky, He, and Kocher (2016) compare the predictive power of alternative logistic models with random forests. Blair, Blattman and Hartman (forthcoming) use variants on logit regressions as well as random forests and neural networks to generate predictions of conflict using subnational data from Liberia. Blair and Sambanis (2017) use random forests to predict conflict, distinguishing between groups of explanatory variables that are more or less aligned with different theories of conflict. Finally, Mueller and Rauh (2016) use text mining tools to generate frequencies of references to topics relating to conflict in newspaper articles, and show that this variable has strong predictive power in a conventional linear regression prediction framework.

We build on this literature by proposing and evaluating the predictive performance of two simple prediction algorithms that are calibrated to directly minimize the prediction loss function of interest – the linear and threshold classifiers. We find that the in-sample predictive power of the

threshold classifier generally dominates that of the other classification rules, particularly as the number of explanatory variables increases. For all classification rules, in-sample predictive power naturally is on average better than out-of-sample predictive power. Here as well we find that the threshold classifier performs better on average than the other classification rules. Interestingly, we find that the out-of-sample predictive power of the probit and linear classifiers worsens as we consider models with more explanatory variables, while it improves, although only modestly, for the threshold and random forest classifiers. While the threshold classifier performs well overall, it is important to keep in mind that the predictive power even of the best-performing model we consider is still modest: while the threshold classifier typically correctly predicts around 90 percent of conflict episodes, at the same time it incorrectly signals conflict in 30-40 percent of non-conflict episodes.

Two key limitations of our approach are also worth noting at the outset. First, a number of papers in the conflict prediction literature have exploited the power of machine learning techniques as a data reduction device, which is particularly beneficial in the contexts where the set of potential explanatory variables can be very large. These include LASSO-regularized logit regression (used in Muchlinski, Siroky, He, and Kocher (2016) and Blair, Blattman and Hartman (forthcoming)), as well as the capacity of random forests and neural networks to handle large numbers of potential explanatory variables. In this paper, we do not address the issue of data reduction, and choose to work with datasets with a modest number of explanatory variables selected in advance. It remains an open question whether simple approaches such as the threshold classifier will continue to perform well when the number of explanatory variables is much larger. A second limitation – shared with other papers in the conflict prediction literature – is that it is difficult to make any claims of external validity of the methodological choices we make here. Papers like ours can only show that a particular classification algorithm works well in a particular dataset, and it is not possible to claim that the preferred approach in this setting – the threshold classifier – would perform as well in other datasets and applications. Nevertheless, the simplicity of the threshold classifier that makes it arguably more suitable for policymakers is an advantage of this approach regardless of the specific dataset to which it is applied.

The rest of this paper proceeds as follows. Section 2 describes the different algorithms for predicting conflict, distinguishing between those that are calibrated to minimize a prediction loss function that those that are not. Section 3 describes the methodology for identifying conflict episodes, and introduces a set of correlates of conflict motivated by the existing literature, that we use to test the

conflict prediction algorithms. Section 4 contains our main empirical findings, and Section 5 offers concluding observations and a discussion of policy implications.

2. Prediction Models for Conflict

Our objective in this paper is to assess models for predicting conflict events. The conflict events we work with are binary -- either conflict occurs or it does not. Predictions of conflict events are often most useful when they also are binary -- either we predict that conflict will occur, or will not occur -- for two reasons. First, a policy rationale for having predictions of conflict is that they can trigger a remedial response on the part of policymakers and aid donors to reduce the risk of conflict. Such mobilization of resources is best done in response to a clear binary prediction, rather than a vague statement that conflict is more or less likely. Second, we will be choosing among different prediction rules based on an evaluation of their predictive performance. While there are tools to evaluate continuous or density predictions of binary variables, the most common approach is to generate binary predictions to match the binary outcomes, and then evaluate how often predictions are "right" and how often they are "wrong".[3]

In Section 3, we define a set of conflict events and complementary non-conflict episodes. Let y_i be a binary indicator taking the value one if episode i is a conflict episode, and zero otherwise; and let y_i^* denote the corresponding prediction of conflict, i.e. y_i^* is equal to one if we predict conflict, and zero otherwise. When evaluating these predictions, we care about both Type 1 errors (failing to predict a conflict episode or "missed calls", i.e. $y_i = 1$ but $y_i^* = 0$) and Type 2 errors (incorrectly predicting conflict when conflict does not occur or "false alarms", i.e. $y_i = 0$ but $y_i^* = 1$)[4]. We evaluate the predictions of conflict using a prediction loss function that penalizes the rate of both types of errors, i.e.

$$L = \omega(Type\ 1\ Error) + (1-\omega)(Type\ 2\ Error) \quad (1)$$

[3] One way to evaluate continuous predictions, such as the fitted probabilities from a probit regression, is to "bin" them into deciles and then plot the mean of the outcome variable within bins against the mean of the fitted probability within bins. A model with strong predictive power should generate a relationship close to the 45-degree line. While this is a valid approach to evaluating continuous predictions, we do not pursue it further here given that most of the literature focuses on evaluating binary predictions based on some combination of Type 1 and Type 2 error rates.

[4] The machine learning literature generally uses the terminology of "sensitivity" – the proportion of conflict episodes correctly predicted to be conflict, i.e.one minus the Type 1 error rate; and "specificity" – the proportion of non-conflict episodes correctly predicted to be non-conflict, i.e. one minus the Type 2 error rate. Yet another equivalent terminology is false negatives for Type 1 errors, and false positives for Type 2 errors.

where the rates of the two types of errors in the sample of interest are $Type\ 1\ Error = \sum_{i=1}^{N} \frac{y_i(1-y_i^*)}{\sum_{i=1}^{N} y_i}$ and $Type\ 2\ Error = \sum_{i=1}^{N} \frac{(1-y_i)y_i^*}{\sum_{i=1}^{N}(1-y_i)}$; and ω and $1-\omega$ are the weights on Type 1 and Type 2 errors. In the empirical results that follow we set $\omega = 0.5$. Although we assign equal weight to the rate of Type 1 and Type 2 errors, since conflict episodes are relatively infrequent, this implies that we are assigning a higher value to correctly predicting conflict episodes than to correctly predicting non-conflict episodes.

We consider four methods for generating predictions of conflict, y_i^*, based on a K x 1 column vector of observable covariates for each episode, x_i. In keeping with the objective of providing predictions of future conflict based on currently-available information, the covariates are observed prior to the beginning of each episode. However, to conserve on notation, we do not explicitly denote this temporal ordering. In the next two subsections we describe the two conventional prediction rules (based on probit regressions and random forests) and the two unconventional prediction algorithms (the linear and threshold classifier calibrated to minimize the prediction loss function) that we consider. Section 2.3 briefly discusses implementation issues for the linear and threshold classifiers.

2.1 Prediction Algorithms Based on Binary Response Regressions and Random Forests

Conflict predictions based on fitted probabilities from binary response regression models are common, and consist of two steps. The first step involves estimating a binary response regression model relating the observed outcome to the explanatory variables, i.e.

$$P[y_i = 1] = F(\gamma_0 + \gamma' x_i) \tag{2}$$

where $F(.)$ is a cumulative distribution function and $(\gamma_0, \gamma')'$ is a vector of parameters to be estimated. The second step is a rule which predicts conflict if the fitted probabilities from the binary response model are greater than some cutoff value p^*, i.e.

$$y_i^* = \begin{cases} 1, & F(\hat{\gamma}_0 + \hat{\gamma}' x_i) \geq p^* \\ 0, & F(\hat{\gamma}_0 + \hat{\gamma}' x_i) < p^* \end{cases} \tag{3}$$

The cutoff value p^* is chosen to minimize the prediction loss function in Equation (1), given the estimated parameters of the probit model, $\hat{\gamma}_0$ and $\hat{\gamma}$. The binary response model is most commonly a probit or a logit specification, corresponding to a choice of Gaussian or a logistic distribution for $F(.)$.

In this paper we report results using a probit specification, and refer to this approach as the "probit classifier". Our findings are virtually identical if we use logistic regressions.

The top panel of Figure 1 illustrates the probit classifier for a hypothetical dataset in which there are $K = 2$ explanatory variables for each episode, x_1 and x_2. The orange round data points correspond to observations with no conflict, i.e. pairs of x_{i1} and x_{i2} for which the corresponding $y_i = 0$, while the blue square data points correspond to pairs of x_{i1} and x_{i2} for which conflict is observed, i.e. $y_i = 1$. The downward-sloping line traces out the boundary between predicted conflict and non-conflict, i.e. the set of points where $F(\hat{\gamma}_0 + \hat{\gamma}_1 x_1 + \hat{\gamma}_2 x_2) = p^*$. Points above this line correspond to pairs of x_1 and x_2 where the probit classifier predicts conflict, while points below the line correspond to predictions of no conflict. Therefore, square data points lying in the upper-right shaded region correspond to correctly predicted conflict events, while the round data points in this region correspond to false alarms. Similarly, in the region below the line, square data points correspond to missed conflict events, while round data points correspond to correctly-predicted non-conflict events. Note that the slope of the boundary line depends only on the ratio of the estimated probit coefficients, $\hat{\gamma}_1/\hat{\gamma}_2$. However, the intercept of the line is determined by the cutoff probability p^*. Optimally choosing the cutoff probability involves parallel up-and-down shifts of this line to find the location which minimizes the prediction loss function in Equation (1).

This prediction rule is very intuitive, and has the added advantage that it connects naturally with the large empirical literature on correlates of conflict that has estimated many different versions of Equation (2). The difficulty with this approach however is that it is not at all clear that it is the best possible prediction rule given the observed data and the loss function in Equation (1). As discussed in Elliott and Leili (2013) and Elliott and Timmerman (2016, Chapter 12), the problem is that the likelihood function that is maximized in the estimation of the binary response model implicitly treats prediction errors differently than the objective function that is optimized in the selection of the cutoff probabilities.[5] Moreover, these two objective functions could very well be contradictory. The key issue

[5] Note that the maximum likelihood problem for probit estimation is:

$$\max_{<\gamma_0,\gamma>} \sum_{i=1}^{N} y_i \, ln\big(F(\gamma_0 + \gamma' x_i)\big) + (1 - y_i) ln(1 - F(\gamma_0 + \gamma' x_i))$$

The term inside the summation is the log-likelihood function for an individual observation. It consists of two terms that are treated symmetrically. The first term rewards values of the parameters that assign a high fitted probability $F(\gamma_0 + \gamma' x_i)$ to conflict episodes where $y_i = 1$, while the second term rewards values of the parameters that assign low fitted probability to non-conflict epsides where $y_i = 0$. The first term is analogous to

is that the likelihood function treats both types of errors symmetrically, and moreover weights each observation in the sample equally. This means that the implicit weights on the rates of Type 1 and Type 2 errors at the estimation stage can be very different from those used to choose the cutoff probabilities that generate the predictions. For example, if the sample happens to contain many non-conflict observations, the estimation procedure will implicitly assign more weight to finding parameter values that correctly predict non-conflict episodes, and will not assign much weight to correctly predicting conflict. As a result, the objective function in the estimation stage is different from the one in the prediction stage, which explicitly chooses the cutoff probability p^* to minimize the prediction loss function. Returning to the top panel of Figure 1, the downward-sloping line that defines the probit classifier has a slope and an intercept that are chosen to optimize different and potentially-contradictory objective functions.

A similar issue arises with predictions based on random forests. As with the probit classifier, predictions based on random forests are a two-step process: (1) fitting a random forest to the data, and (2) retrieving the predictions for each observation and finding a cutoff to classify observations as conflict or non-conflict. Random forests are a collection of classification trees that are fitted to bootstrapped subsamples of the data. Classification trees themselves are sequential binary partitions of the explanatory variables that seek to best discriminate between conflict and non-conflict observations at each partition of the data. The bottom panel of Figure 1 illustrates a two-level classification tree for the same hypothetical dataset as the top panel. In this two-variable example, a classification tree can be constructed by first finding a threshold t_1 for x_1 that provides the best separation between conflict and non-conflict episodes. The next step is to find two distinct thresholds for x_2, conditional on x_1 being above its threshold, t_{2H}, or below its threshold, t_{2L}, that provide the best separation between conflict and non-conflict episodes in these two subsamples. The observations in the resulting four regions, or terminal "nodes", are predicted as conflict or non-conflict based on the majority of observations within the terminal nodes.

Random forests are a technique for improving the predictions of individual classification trees by averaging across a large number of trees fitted to bootstrapped subsamples of the data. The basic idea is to generate independent classification trees in randomly-selected subsamples of the data, and based on randomly-selected subsets of the explanatory variables. The individual classification trees are grown

---

the complement of Type 1 errors, i.e. correctly-predicted conflict episodes or "sensitivity", while the second term is analogous to the complement of Type 2 errors, i.e. correctly-predicted non-conflict episodes, or "specificity".

to be very "deep", so that only a small number of observations remain in each terminal node. Individually these trees are noisy predictors, but averaging across a large number of them results in more stable predictions. Specifically, at each bootstrap iteration, a classification tree is fitted to a subsample of the data, and the fitted classification tree is used to predict the outcome in the remaining data not used for fitting the tree. These "out-of-bag" predictions are aggregated across all trees, and the proportion of trees classifying a particular observation as conflict can be interpreted as a prediction of the likelihood of conflict for that observation. Just as in the binary response regression approach, these proportions or "vote counts" can be compared with some cutoff value above which a prediction of conflict is generated. We select this cutoff value for the vote counts to minimize the prediction loss function.

Classification trees use the criterion of "node purity" to determine the optimal thresholds at which to split the data at each branch of the tree. Most commonly this criterion takes the form $p_L p_H$, where p_L and p_H denote the proportion of conflict observations in the subsamples of observations below and above the threshold. This criterion is equal to zero in the case of a split that perfectly separates conflict and non-conflict observations, since either $p_L = 0$ and $p_H = 1$, or vice versa, corresponding to the case of perfect node purity. Thresholds at each level of the classification tree are set to maximize node purity by minimizing this criterion. Crucially for our purposes, the splitting criterion treats all observations symmetrically, and in particular does not distinguish between Type 1 and Type 2 errors. In this sense, the random forest classifier shares the same potential problem that the probit classifier does – the estimation and prediction stages in the algorithm optimize different and potentially contradictory objective functions.

This shortcoming of the probit and random forest classifiers motivates the two other prediction rules we consider, that are discussed in the following subsections. Following the suggestion of Elliott and Leili (2013) and Elliott and Timmerman (2016, Chapter 12), we focus on prediction rules that are calibrated to directly minimize the prediction loss function of interest. We also note that this potential difference between the objective function at the estimation stage and the prediction loss function is related to the emphasis placed by Muchlinksi et. al. (2016) on the "class-imbalanced" nature of conflict data. Given their definition of conflict, conflict is a very rare event in the sample they study, and they note the benefits of relying on "rare-events" logit estimation, which in effect over-weights the relatively scarce conflict observations, as a way to address this problem. They do not however follow the logic of

this observation to the point of this paper, which is to rely on a prediction algorithm that is calibrated to minimize the prediction loss function of ultimate interest.

2.2 Prediction Algorithms that Minimize the Prediction Loss Function

In this subsection we discuss two prediction algorithms that avoid the potential contradictions between the estimation and prediction stages in the probit and random forest classifiers by instead directly minimizing the prediction loss function. The first is a simple variant on the probit classifier, which bases predictions on a linear combination of the observed covariates that minimizes the prediction loss function in Equation (1). Specifically, we define this prediction rule:

$$y_i^* = \begin{cases} 1, & \beta' x_i \geq c \\ 0, & \beta' x_i < c \end{cases} \tag{4}$$

where, without loss of generality, we can normalize the cutoff value $c = 1$. We then choose the values of β that minimize the prediction loss function in Equation (1).

This prediction rule is similar to the probit classifier in the sense that it bases predictions on whether a linear combination of covariates falls above or below some cutoff value. The difference however is that in the probit classifier, the relative weights assigned to the explanatory variables are pinned down by the estimated probit slope coefficients, $\hat{\gamma}$, while in the linear classifier the relative weights are chosen to directly minimize the prediction loss function that we ultimately care about. To see this more clearly, note that we can re-write the probit classifier as:

$$y_i^* = \begin{cases} 1 \ , & \tilde{\gamma}' x_i \geq 1 \\ 0, & \tilde{\gamma}' x_i < 1 \end{cases} \tag{5}$$

where $\tilde{\gamma} \equiv (F^{-1}(p^*) - \hat{\gamma}_0)^{-1}\hat{\gamma}$ is a simple rescaling of the probit slope coefficients. These rescaled probit coefficients are directly comparable to the relative weights assigned by the linear classifier, β, and to facilitate the interpretation of the results, we report the rescaled probit coefficients in the tables that follow.

The top panel of Figure 2 illustrates the linear classifier for the same hypothetical dataset as before with $K = 2$ explanatory variables, x_1 and x_2. Recall that the orange round data points correspond to observations with no conflict, while the blue square data points correspond to conflict

observations. The downward-sloping line traces out the boundary between predicted conflict and non-conflict based on the linear classifier, i.e. the set of points where $\beta_1 x_1 + \beta_2 x_2 > 1$. As before, points above this line correspond to pairs of x_1 and x_2 where the linear classifier predicts conflict, while points below the line correspond to predictions of no conflict. The main difference between the top panel of Figure 1 and the top panel of Figure 2 is that in Figure 2, both the slope and the intercept of the downward-sloping line are chosen optimally to minimize the prediction loss function. In contrast in the probit classifier in Figure 1, only the intercept of the line is chosen to minimize classification errors, while the slope is pinned down by the estimated coefficients in the probit regression. As discussed in the previous subsection, this means that the slope of the boundary for the probit classifier reflects the maximum likelihood optimization problem, which assigns different implicit weights to Type 1 and Type 2 errors than does the prediction loss function that we ultimately want to minimize.

This prediction rule is related to a class of machine learning algorithms known as "support vector classifiers".[6] In the case of two explanatory variables as illustrated in Figure 2, the basic idea behind these algorithms is again to find a line in x_1, x_2-space that best separates observations into predicted conflict and non-conflict episodes. However, the line is chosen to minimize a different objective function from the prediction loss function in Equation (1) that we use throughout. Typically, the support vector classifier uses a loss function that penalizes the magnitude of misclassifications, defined as the Euclidian distance in (x_1, x_2)- space between each misclassified observation and a band around the separating line. The objective function for the support vector classifier is to maximize the width of this band subject to a penalty for the size of misclassifications. This leads to the same problem as with the probit classifier – the objective function used to calibrate the classifier is not the same as the objective function used to evaluate the predictions. For this reason, we use the linear classifier defined in Equation (4) and the prediction loss function in Equation (1) rather than an off-the-shelf support vector classifier. [7]

[6] For a detailed discussion of these algorithms see Hastie, Tibshirani and Friedman (2015, Chapters 4 and 12)

[7] The linear classifier and the probit regression both also are similar to the venerable naive Bayes classifier that has been used in the machine learning literature since the 1960s, and still is commonly used for text classification schemes. The naive Bayes classifier is also one of the two machine learning models used to predict conflict in Perry (2013). With Gaussian covariates, the naive Bayes classifier generates predictions as linear functions of covariates, and is formally equivalent to a logit regression that also generates predictions based on a linear combination of covariates. It also shares the problem noted in the main text that it is designed to match the observed data as closely as possible, which may imply a different weight on Type 1 and Type 2 errors than the prediction loss function that is used to evaluate the predictions.

Our last prediction algorithm is a "threshold classifier", that consists of a set of thresholds or cutoffs for each of the K explanatory variables, t_k, and a rule that predicts conflict if at least N of the cutoffs are crossed, with $1 \leq N \leq K$. Specifically, the prediction rule for the threshold classifier is:

$$y_i^* = \begin{cases} 1 \ , & \sum_{k=1}^{K} I_{(x_{ik} \geq t_k)} \geq N \\ 0, & \sum_{k=1}^{K} I_{(x_{ik} \geq t_k)} < N \end{cases} \quad (6)$$

where $I_{(x_{ik} \geq t_k)}$ is an indicator value taking the value one if $x_{ik} \geq t_k$ and zero otherwise. In the threshold classifier, the thresholds t_k and the minimum number of breaches required to predict conflict, N, are chosen to minimize the prediction loss function in Equation (1).[8]

The bottom-left panel of Figure 2 illustrates the threshold classifier in the case of $K = 2$ covariates, and where $N = 1$, i.e. at least one threshold needs to be breached to generate a prediction of conflict. This generates an L-shaped shaded region in which conflict is predicted if $x_1 \geq t_1$ or $x_2 \geq t_2$, and a rectangular complementary region in the bottom-left of the graph where $x_1 < t_1$ and $x_2 < t_2$ and accordingly, conflict is not predicted. In the case where $N = 2$ is shown in the bottom-right panel of Figure 2. In this case, the region for predicting conflict would be the rectangular area in the top-right of the graph where $x_1 \geq t_1$ *and* $x_2 \geq t_2$, and the remaining unshaded L-shaped region corresponds to predictions of no conflict. The threshold classifier chooses the thresholds, and the minimum number of breaches required to signal conflict, to minimize the prediction loss function.

The main advantage of the threshold classifier is that it generates a very simple and intuitive method for classifying conflict episodes that is easy for policymakers and other non-technical users of the classification rule to understand. We can think of a covariate of conflict breaching its threshold, i.e. $x_{ik} \geq t_k$ for a given episode i as providing some signal of the likelihood of conflict based on that variable. If sufficiently many such signals (i.e. at least N such signals) are observed, then the observation is classified as a conflict episode. This simplicity comes at the expense that the orientation of each covariate needs to be specified in advance, such that higher values of the variable correspond to a

[8] An early application of this approach is Reinhart et al. (1998) who generated threshold-based leading indicators of currency crises. In contrast with this paper, they considered predictions based on one indicator at a time. In this univariate setting all three algorithms in this paper reduce to simple threshold rules. However, when combining information from different explanatory variables, the three algorithms considered here aggregate information across different explanatory variables in different ways.

greater risk of conflict. Otherwise, exceeding the threshold does not constitute a signal of conflict. In many cases there is a natural intuitive orientation -- for example, it seems reasonable to suppose a priori that a greater incidence of conflict among neighbors raises the risk of conflict at home. But in other cases the orientation might be unclear. To address this issue systematically, we orient each of the explanatory variables to be consistent with the estimated sign in the corresponding probit regression. That is, if a variable enters positively in the probit regression (such that higher values are associated with a higher risk of conflict), then we do not reorient the variable. If on the other hand a variable enters negatively in the probit regression, then we reverse the orientation of the variable. To facilitate the interpretation of results, we reorient variables in this way (i.e. based on the signs of the slopes in the probit regression) for all variables and for all the classification rules, even though this step is of course not necessary for the other prediction algorithms.

2.3 Implementation Issues

Implementing the probit classifier is straightforward, since it simply requires the estimation of a probit regression, followed by a one-dimensional grid search over cutoff probabilities to find the optimal cutoff probability p^* that minimizes the prediction loss function. The grid search itself is automated in the `lroc` command in Stata. To implement the random forest classifier, we rely on the `randomForest` package in R.[9]

Implementing the linear classifier and the threshold classifier is more challenging because it involves minimizing an objective function that is flat in much of the parameter space, and otherwise jumps discontinuously. To see why, consider for example the threshold classifier, and consider what happens when we search over candidate values of the threshold for explanatory variable x_1 that happen to fall strictly between two adjacent observations x_{i1} and $x_{i+1,1}$ (assume observations have been ordered by increasing values of x_1). For any threshold in this range, the conflict/no conflict prediction does not change for any observation, and so the prediction loss function also does not change. However, when the candidate threshold moves just outside this range by crossing either x_{i1} or $x_{i+1,1}$, the classification of observations changes and the objective function changes discontinuously. The size of the "flat" regions of the parameter space and the size of the discontinuous "jumps" in the objective function depend on how large the sample is and how evenly the covariates are distributed over their

[9] https://www.rdocumentation.org/packages/randomForest/versions/4.6-12/topics/randomForest

relevant ranges. In the fairly small samples we work with (around 500 observations), the objective function is sufficiently flat and discontinuous that standard gradient-based minimization algorithms fail.

A natural alternative to gradient-based minimization is to simply do brute-force grid searches over the parameter space. This however becomes increasingly costly as the number of covariates increases, and particularly so for a reasonably fine grid of candidate parameter values. Instead, we rely on the Nelder and Mead (1965) method, which is a derivative-free heuristic minimization algorithm (available as an option within the Mata `optimize` module). Since its convergence properties are not known in general, we first verify that the Nelder-Mead algorithm finds a minimum that is close to the one identified by brute-force grid searches, for several examples with up to four covariates for which grid searches are not too computationally costly. To ensure robustness of the minimum identified by this algorithm to the choice of initial parameters, we minimize the prediction loss function for 100 randomly-chosen sets of initial conditions. The distribution of the initial parameters is centered on the probit weights for the linear classifier, and on the median value of each covariate in the conflict sample for the threshold classifier.

3. Measuring Conflict and Its Correlates

In this section we briefly describe the cross-country panel conflict dataset that we use to evaluate our four prediction algorithms. The first subsection describes our definition of conflict, and the following subsection enumerates the explanatory variables we draw on as predictors of conflict.

3.1 Identifying Conflict Episodes

Our starting point is a panel dataset of country-year observations covering 114 developing countries since 1977. As discussed in more detail below, the sample of countries and the time period are dictated by data availability. Our primary measure of conflict is a binary indicator of whether at least 25 battle-related deaths occur in a country in a given year.[10] This is a standard threshold for "minor armed conflicts" used in the literature to capture smaller conflicts, and naturally also includes much rarer civil war events which typically are measured using a higher threshold of 1,000 battle deaths. We supplement this measure with two additional indicators. The first is a binary indicator measuring whether a UN peacekeeping operation (UNPKO) is present in a given country year. This is intended to

[10] Data comes from Uppsala Conflict Data Program/PRIO Armed Conflict Dataset. It defines armed conflict as "a contested incompatibility which concerns government and/or territory where the use of armed force between two parties, of which at least one is the government of a state, results in at least 25 battle-related deaths."

measure cases in which conflict exists, but is held in check by the presence of a UNPKO, so that battle deaths fall below the threshold for conflict. The second is a dummy variable equal to one if at least 5 percent of the population of a country is involuntarily displaced as refugees across international borders, and zero otherwise. This measure is intended to capture episodes of conflict that may not cross the threshold of battle deaths but nevertheless are sufficiently severe to cause significant involuntary displacement of people. [11] The threshold of 5 percent is somewhat arbitrarily selected, and corresponds to the 99th percentile of all non-zero refugee observations in the data.

Figure 3 helps to visualize the three criteria we use to identify conflict, using the Central African Republic as an example. The dotted lines show the number of battle deaths and the ratio of displaced population as a fraction of the population of the country of origin. The horizontal lines represent the thresholds for battle deaths and total displacement ratio. Whenever the dotted line is above the corresponding horizontal line, or there is a UN peacekeeping mission, the corresponding year is coded as a conflict year.

Figure 4 displays the contribution of individual criteria to identifying conflict country-year observations in our full dataset. Over half of the conflict country-year observations are captured only by the battle deaths criteria, while just 6 percent are captured only by the refugees share of total population (by country of origin). The UN Peacekeeping Mission presence criteria account for a further 19 percent of the conflict-year observations while the remaining 24 percent are identified jointly by at least two of the three criteria.

We next translate these annual country-year observations on conflict into "conflict episodes" that constitute our estimation sample.[12] We are primarily interested in predicting the onset of conflict, rather than its duration. We therefore identify conflict episodes as the first year in which one of the three indicators signals conflict. In many countries, conflict years are immediately preceded by other conflict years. To make sure that only distinct episodes of conflict are identified, we drop conflict episodes that are preceded by conflict in any of the previous three years.[13] Non-conflict episodes are defined as non-overlapping periods of five consecutive years in which none of the three events related

[11] Data comes from UNHCR Population Statistics. Data on IDPs are published by the Internal Displacement Monitoring Centre (IDMC) and have various challenges in collection, hence we do not include these data in the analysis. For an in depth discussion please see World Bank (2016)

[12] This methodology is similar to Kraay & Nehru (2006), who define episodes of debt servicing difficulties in a analogous way.

[13] An important point of the World Development Report 2011, was that modern violence comes in various forms and repeated cycles.

to conflict take place. The non-conflict episodes begin in the first year for which it is possible to identify five consecutive years with no conflict. As with the conflict episodes, we also require the non-conflict episodes to be preceded by three years of no conflict. Figure 3 illustrates how we transform the annual observation on conflict indicators into conflict episodes, using the case of Central African Republic as an example. For this country, we have three non-conflict episodes in 1980, 1985 and 1990, and two conflict episodes, beginning in 1998 and 2006, and shaded gray. In 1998 the conflict episode is triggered by the presence of a UNPKO, whereas in 2006 the conflict episode is triggered by the battle death count criterion crossing the threshold.

We construct the conflict and non-conflict episodes using annual data for 114 countries over the period 1977-2014 for which our core explanatory variables -- discussed in more detail below -- are available.[14] Through this method, we arrive at a sample of 69 conflict events, and 422 non-conflict events. Our dependent variable in the empirical work that follows is a dummy variable taking the value one for conflict episodes, and zero for non-conflict episodes. The 69 conflict episodes identified are listed in Table 1, which identifies the country name and the start year of each episode. The list is dominated by Sub-Saharan African countries, but a number of other conflicts from other regions appear as well. The conflict episodes vary in length, with a median length of two years and an interquartile range from one to six years. Some conflict episodes are very prolonged: the 90th percentile of conflict length is 17 years. In the empirical work that follows, we do not seek to account for this variation in the duration of conflict once it has begun, but rather focus on the prediction of conflict conditional on being in a non-conflict state.

3.2 Correlates of Conflict

We work with a selection of variables that have been emphasized in the existing literature as important correlates of conflict (Fearon & Laitin (2003); Collier & Hoeffler (2004); Hegre and Sambanis (2006); Besley and Persson (2011); Esteban, Mayoral, and Ray (2012); Hegre, Karlsen, Nygard, Strand, and Urdal (2012)). Our goal here is not to innovate relative to this existing literature by isolating novel drivers of conflict. Rather, we wish to have a set of covariates of conflict that is broadly representative of those used in the existing literature, that we can use as a benchmark from which to assess the

[14] The key variable that limits the sample considerably is the World Bank's Country Policy and Institutional Assessment (CPIA) ratings, which starts in 1977 and is available only for developing-country clients of the World Bank.

predictive power of alternative classification rules for conflict. We also do not seek to exhaustively consider all of the variables that have been proposed in this voluminous literature. Doing so would raise a further set of questions relating to model selection – what is the appropriate subset of variables to include in the prediction algorithm? The issue of identifying a robust subset of correlates of conflicts from the many variables proposed in the literature is discussed at length in the robustness analysis of Hegre and Sambanis (2006), and Muchlinksi et. al. (2016) and Blair, Blattman and Hartman (forthcoming), which consider LASSO-regularization as a technique for selecting a subset of explanatory variables in their logit-based predictions. Since we do not seek to innovate on the issue of variable selection, we work with a fixed and relatively small number of potential predictors of conflict to illustrate the performance of the different conflict prediction rules.

Purely for presentational purposes, we organize these variables into three broad categories: latent tensions, shocks, and institutions.[15] Since we are interested in models that can predict conflict based on information prior to the beginning of an episode, we measure all explanatory variables as averages over the three years prior to beginning of the conflict and non-conflict episodes. In the category of latent tensions, we include a number of variables that are intended to capture underlying factors contributing to between-group tensions that may escalate into conflict. We use two innovative recent measures of income inequality, both taken from Alesina, Michalopoulos, and Papaioannou (2016). The first combines satellite night light density -- a spatially-disaggregated proxy for per capita income -- with maps delineating the boundaries of different ethnic groups, to obtain a proxy for income inequality across ethnic groups. The second simply measures inequality across pixels within a country and therefore proxies for overall spatial inequality. Both measures are available only at decadal frequency, and we linearly interpolate the intervening years to obtain annual data that can be averaged over the three years prior to each conflict and non-conflict event. Following a large literature that has documented the importance of conflict over the income generated by natural resources, we include a measure of natural resource rents as a share of GDP. These are constructed in World Bank (2011) and are based on estimates of the difference between commodity prices and unit costs of production, for a large number of commodities. These are aggregated across commodities produced in a country, and expressed as a share of GDP. We also rely on standard measures of ethnic and religious fractionalization that measure the probability that two randomly-selected individuals in a country will belong to different ethnic or religious groups. We also include the share of men aged 15-29 in the total population,

[15] Details on the data sources are in the appendix.

following earlier literature that has documented associations between prevalence of young men and conflict. Finally, in this category of latent tensions, we include the country's history of conflict, defined as the fraction of years since 1970 in which conflict, as defined in our dependent variable, is observed.

In the category of shocks, we gather several types of events that the existing literature has emphasized as increasing the risk of conflict. This category includes two external shocks: the income effect of changes in the terms of trade, which tends to be positively correlated with conflict in our sample, and number of neighboring countries in conflict, which also tends to be positively correlated with conflict in the country of interest. We also include a composite measure of the incidence of natural disasters, taken from Besley and Persson (2011). This is a binary indicator of whether any of the following types of natural disasters occur in a given country and year: extreme temperature events, floods, landslides and tidal waves. Finally, in this category we include per capita GDP growth, as a crude proxy for other shocks hitting the economy that might contribute to the risk of conflict.

The final category is intended to capture institutions that may reduce the likelihood of conflict, by offsetting some of the forces in the two previous categories. We use the World Bank's Country Policy and Institutional Assessment as an omnibus proxy for the overall quality of policies and institutions. In the specific World Bank institutional context, this measure is of particular interest because the Bank's past efforts to classify countries at risk of conflict have relied heavily on selecting countries with low scores on this particular indicator. Turning to measures of political accountability, we include the Freedom House composite indicator of civil liberties and political rights. We also use the Political Terror Scale, which is a long-standing effort to numerically code narrative descriptions of human rights violations in annual reports by Amnesty International, U.S. State Department and Human Rights Watch. Finally, we include in this category the log-level of real GDP per capita, following many papers that have documented a greater prevalence of conflict in low-income countries, and the logarithm of population, following previous papers that have documented an association between country size and conflict.

Table 2 provides descriptive statistics for these explanatory variables in the three-year period prior to the start of conflict and non-conflict episodes in our dataset. A glance at this table reveals many patterns that would be expected from the large empirical literature on conflict. Income inequality, ethnic fractionalization, income inequality across ethnic groups, natural resource rents as a share of GDP, and the history of past conflict are all on average higher prior to conflict events than non-conflict events. Interestingly, however, religious polarization is on average lower in conflict events. Turning to the variables in the institutions group, per capita income and the CPIA indicator are both lower prior to

the start of conflict episodes. The Freedom House and Political Terror Scale variables are both oriented so that higher values correspond to worse outcomes, and we also observe on average higher values of both measures prior to conflict episodes. In the shocks category, growth is lower prior to conflict events, and natural disasters are neighbors in conflict are both higher before conflict. In contrast, positive terms of trade shocks tend to precede conflict events in our sample.

4. Results

Table 3 provides a first look at the univariate predictive power for the various correlates of conflict. The rows of the table correspond to the correlates of conflict discussed in Section 3, grouped into the three categories of latent conflict, institutions, and shocks. The first three columns of the table report the Type 1 and Type 2 error rates, as well as the prediction loss function, for predictions of conflict based on each variable individually. Since the classifiers in this table generate predictions using a single variable, all four classification rules collapse to the same form – setting a threshold value for the single explanatory variable above which the classifier signals conflict.[16] In this table, in all cases the threshold is set to minimize the prediction loss function – and therefore all four classifiers generate the same threshold and predictions.

Table 3 shows substantial heterogeneity across the individual variables in terms of their ability to predict conflict. Type 1 errors or missed conflicts range from 4% to 88% of all conflict episodes, while Type 2 errors or false alarms range from 6% to 83% of all non-conflict episodes. Variables such as the spatial inequality and the young male population share stand out as generating predictions with very low Type 1 error rates, implying that they are consistently high in conflict episodes. In contrast, variables such as the CPIA and natural disasters stand out in terms of having very low Type 2 error rates, i.e. they consistently have good values in episodes with no conflict, but not necessary poor values in conflict episodes. The balance of Type 1 and Type 2 errors reflected in the objective function also shows significant heterogeneity, ranging from 33% to 47%. To put these values of the prediction loss function in context, note that had we simply predicted conflict at random with a probability equal to the unconditional probability observed in the sample, the prediction loss function would be 50%.[17] The two

[16] For the purposes of this table we simplify the random forest to a single classification tree with a single explanatory variable, which corresponds to a single partition of the data.

[17] To see this, note that if the unconditional probability of conflict is p, the probability of correctly predicting a conflict is p^2, while the probability of generating a missed conflict is $p(1-p)$, implying a Type 1 error rate of

variables with the best predictive power, i.e. the lowest value of the prediction loss function, in each group are highlighted (ethnic inequality, natural resource rents, Freedom House, Political Terror Scale, real GDP growth, and neighbors in conflict).

In the next two columns of Table 3, we report the percentile of the distribution of each variable at which the optimal threshold is set, as well as the actual value of the variable at that percentile. In order to interpret the figures in this column, it is important to recall that variables that are negatively correlated with conflict in our sample have been re-oriented to be positively correlated with conflict. These variables (religious fractionalization, the CPIA, per capita income, real GDP growth, and the income effect of changes in the terms of trade) are indicated with a "*" in Table 3. For these variables, values *below* the reported threshold (in percentile and value terms) correspond to predictions of conflict, while for all other variables, values *above* the reported threshold correspond to predictions of conflict. To take a specific example, the threshold for predicting conflict based on the CPIA is set at the 22nd percentile of the distribution of re-oriented CPIA scores. This means that observations in the *bottom* 22 percent of the CPIA distribution would have conflict predicted based on this variable. Interestingly, the value of this threshold of 3.2 is identical to the one used by the World Bank for many years to determine whether a country should be classified as "fragile"[18].

In Table 4 we turn to the more interesting question of assessing predictions of conflict based on multiple explanatory variables. When we consider the joint predictive power of groups of correlates of conflict, the different classification rules combine the information from multiple explanatory variables in different ways, with important implications for overall predictive power. We illustrate the differences between classification rules by considering predictions based on three sets of variables. In Model 1 we examine predictions based on a parsimonious set of six variables, selected as the best two predictors of conflict within the three categories of explanatory variables, i.e. the highlighted rows in Table 3. In Model 2, we consider predictions based on the set of nine explanatory variables that individually are significantly correlated with conflict at the 95% significance level (based on the z-statistics in the last column of Table 3). Finally, in Model 3, we generate predictions based on the full set of all 16 explanatory variables. Naturally, the selection of variables in these three models is somewhat arbitrary,

---

$\frac{p(1-p)}{p^2+p(1-p)} = (1-p)$. Along the same lines, the Type 1 error rate is p, and with a weight on Type 1 errors of $\omega = 0.5$, this implies a prediction loss function of 0.5.

[18] Details on this rule for identifying fragile states can be found at: http://pubdocs.worldbank.org/en/154851467143896227/FY17HLFS-Final-6272016.pdf

and one could in principle generate predictions based on any combination of explanatory variables. We focus on these three models simply for reasons of space, and assess the predictive power of the alternative classifiers in these three models. In Table 4, we focus on in-sample predictions. That is, the classification rules are estimated using the full sample of episodes, and then we examine how well the rules classify observations in the same full sample. Out-of-sample predictive power is assessed in Table 5 and the discussion below.

The three vertical panels of Table 4 correspond to the three models. Within each panel, the columns labelled PC/RF/LC/TC correspond to the probit classifier, random forest, linear classifier, and threshold classifier, respectively. The first three rows of the table summarize the predictive power of the different classification rules in these three models. Moving from left to right, to models with increasing numbers of explanatory variables, we see that predictive power, as measured by the minimized value of the prediction loss function, generally improves. More interestingly, we find that the predictive power of the threshold classifier dominates that of the other three classifiers, and particularly as the number of explanatory variables increases. In the most parsimonious Model 1 with only six explanatory variables, the linear classifier improves slightly over the probit classifier (with the prediction loss function declining from 0.30 to 0.29), and the threshold classifier improves further to 0.27. The loss function is highest for the random forest classifier – however it is difficult to compare this measure of predictive power with that of the other algorithms since it is based on predictive performance in the "out-of-bag" samples rather than in the full sample.[19]

Moving to Models 2 and 3 with more explanatory variables, the threshold classifier performs better in both in absolute terms and relative to the other classifiers. This is most pronounced for Model 3 which includes all 16 explanatory variables, and for which the prediction loss function falls to 20%. In contrast, the addition of more explanatory variables in Model 3 relative to Model 1 improves the predictive power of the probit and linear classifiers only modestly (to 29% and 28% respectively). Interestingly, the linear classifier does not outperform the probit classifier by a very large margin, despite the conceptual differences between the two outlined in the previous section. Overall, however, the combination of the threshold classifier with a fairly large number of explanatory variables has clearly the best performance in terms of in-sample predictive power.

[19] Reporting predictions based on the "in-bag" observations would be equally misleading, as in random forests the individual classification trees are grown to be very deep so that there is only one observation in each terminal node – where by construction the classification tree perfectly fits the data.

The remaining rows of Table 4 document the roles of the different explanatory variables in generating predictions based on these three classification rules. For the probit and linear classifiers, the table reports the weights assigned to each variable in the linear combination on which predictions are based (i.e. Equations (4) and (5) in the previous section). All the underlying variables have been normalized to run from zero to one, so the weights reported in the table can be interpreted as their relative importance in contributing to the predictions of conflict. Given the largely similar predictive performance of the probit and linear classifiers, it is not surprising that they apply broadly similar weights to the different explanatory variables. For example, in Model 3, the correlation across variables of the weights assigned by the probit and linear classifiers is 0.94.

The column for the threshold classifier reports the percentile of the distribution of each explanatory variable at which the optimal threshold is set, i.e. the same as in Column 4 of Table 3. As before, for the variables marked with a "*" that have been reoriented, the threshold corresponds to the percentile below which a signal of conflict is issued, while for the remaining variables the threshold corresponds to the percentile above which a signal of conflict is issued. Note also the bottom row of Table 4, where we report the optimal number of breaches of thresholds required to signal a conflict event, which increases from two in Models 1 and 2, to five in Model 3. Finally, for the random forest predictions, we report the mean improvement in node purity for each variable, a standard measure of predictor importance in a random forest classifier. These are normalized to 1 for the least important predictor, with higher values corresponding to a greater improvement in predictive power. Interestingly, there is some correspondence between the variables identified as important in the threshold classifier and the random forest: in Model 1, the correlation between the measures of importance for the two models is 0.68 (in absolute value). However, this correspondence across methodologies declines as the number of explanatory variables increases, and for Model 3 the same correlation is only 0.13 (again in absolute value).

The analysis in Table 4 focuses on the in-sample predictive power of the different classification rules. In Table 5 we turn to the more important question of out-of-sample predictive power. We do this by dividing the dataset into an estimation sample and a prediction sample, successively using years between 1990 and 2000 as dividing points.[20] The three horizontal panels of the table correspond to

[20] In the machine learning literature, it is common to use cross-validation methods in which the dataset is separated into randomly selected subsamples (or "folds"). The model is estimated sequentially leaving out one fold at a time, and then used to generate predictions in the left-out fold. This is repeated across all folds and the average predictive power across the left-out folds is used as a measure of out-of-sample predictive performance.

predictions based on the same three sets of variables as in Table 4 (i.e. Models 1, 2, and 3), and the rows correspond to different breakpoints. For example, the row for 1990 means that the classification rule was estimated using data between 1977 and 1989, and the predictions are for the period 1990-2014. Finally, the columns of the table report the in-sample and out-of-sample value of prediction loss function, which summarizes the predictive power of the four classification rules. For ease of comparison, in each row the entry shaded in gray corresponds to the model with the best out-of-sample predictive power in that row, i.e. the lowest value of the prediction loss function.

Several observations about this table are of interest. First, for the probit, linear, and threshold classifiers, on average in-sample predictive power is better than out-of-sample predictive power.[21] While for the more parsimonious Model 1 and Model 2, these differences are not large (a few percentage points), they are more pronounced for Model 3 which generates predictions based on 16 variables. This reflects a tendency of these classification rules to somewhat overfit in sample, with adverse consequences for out-of-sample predictive power. The second observation is that for the probit and linear classifiers, out-of-sample predictive power is best in the most parsimonious Model 1, and declines sharply as additional variables are added in Models 2 and 3. In contrast, for the random forest and threshold classifiers, out-of-sample predictive power improves moving from Model 1 with 6 variables to Model 3 with 16 variables, although the improvement is small.

The final observation concerns the relative out-of-sample predictive power of the four classification rules. A glance at the gray-shaded cells in Table 5 indicates that across all three models and all 11 sample splits, the threshold classifier most often has the best predictive power (in 21 out of 33 cases). The probit classifier has the second-best out-of-sample predictive performance, in 8 cases, although as noted abve the probit classifier does poorly in the less parsimonious Model 3. Finally, the random forest and linear classifiers are only rarely the best predictors, at two cases each.

A natural – and difficult to answer – question is the extent to which these findings, particularly on the superior performance of the simple threshold classifier, generalize to other settings. While there

---

The approach taken here is conceptually identical, in that the model is estimated in one subset of the data and used to generate predictions in the remaining subset. However, in this country-year panel setting, there is a natural temporal ordering of the data, and splitting the data by time period is arguably a more relevant exercise, as it answers the question: "how would prediction algorithms have fared ex post if we had used them to predict conflict based on the information available as of 1990, as of 1991, etc.?".

[21] For the random forest classifier, recall that the in-sample measure of goodness of fit is based on the "out-of-bag" subsamples of the training dataset, and therefore is not comparable with the out-of-sample predictions that are based on the entire prediction sample.

is little we can offer in terms of systematic evidence, in Appendix B we document similar, although less stark, evidence on the relative performance of the different prediction algorithms in the well-known Hegre and Sambanis (2006) dataset.

5. Policy Implications and Conclusions

A robust predictive framework for conflict events can provide policymakers with the opportunity to respond more proactively to different risks of conflict, including with policy interventions to address the structural causes or triggers of conflict. This papers aims to strengthen predictions of conflict by evaluating the predictive performance of alternative algorithms that might form the basis for such a framework. We have considered the in-sample and out-of-sample predictive power of four classification rules: two conventional algorithms based on probit regressions and random forests, and two unconventional algorithms designed to directly minimize the prediction loss function – the linear and threshold classifiers. In this particular setting, we find that the threshold classifier dominates the other classification algorithms in terms of in-sample and out-of-sample predictive power, particularly in models with more explanatory variables. Moreover, we have argued that the simplicity of the threshold classifier makes it an attractive tool when conflict predictions are intended to inform policy discussion around aid and conflict-prevention strategies.

While these differences across prediction algorithms are non-trivial, it is important not to lose sight of the fact that the out-of-sample predictive power of these methods is modest. While the best-performing threshold classifier can correctly predict over 90 percent of conflict events, it also incorrectly classifies around 30 percent of non-conflict events as conflict (in Model 3 in Table 4). Whether this rather high rate of "false alarms" is acceptable or not depends in large part on the policy purposes for which these classification algorithms are intended. For example, if resources for conflict prevention programs are scarce, allocating them using a classification rule with a high rate of "false alarms" may result in a highly inefficient targeting of these resources to places where they are most needed. This suggests that before using any of these classification algorithms for policy purposes, careful thought should be given to the weight on Type 1 and Type 2 errors in the prediction loss function.

As noted above, a policy goal of an exercise such as this is to inform frameworks for anticipating the outbreak of conflict. To illustrate how the classification algorithms discussed here might do so in practice, we generate a list of countries that would be considered to be at risk of conflict, based on the most recently-available information in our dataset on the various predictors of conflict, and the four

classification algorithms. Specifically, we take the average of each of the explanatory variables over the most recent period 2012-2014 included in the dataset, and feed these through all four classification rules to generate four alternative lists of countries "at risk" of conflict according to these models. For comparative purposes we also report the list of countries "at risk" of conflict based on the World Bank criteria for being a "fragile or conflict-affected state". As noted earlier, this classification was based on having CPIA scores below a threshold of 3.2, and/or the presence of a UN and/or regional peacekeeping or peacebuilding operation, and we take the most recent classification of countries (i.e. the World Bank's FY17 list).[22] Finally, we focus only on countries in which none of our three annual indicators of conflict is observed in any of the three years 2012-2014. This is to be consistent with the emphasis in our empirics on predicting transitions from non-conflict into conflict.

The results of this exercise are shown in Table 6. The rows of the table correspond to the 87 countries satisfying the criteria outlined above, the columns correspond to the four classifiers as well as the Bank's listing of fragile and conflict-affected situations (FCS). In the table entries, ones indicate predictions of conflict, and zeros predictions of no conflict. The probit, linear, and threshold classifiers all predict conflict in 35, 43, and 37 of these countries respectively, while the random forest algorithm predicts conflict in 26 countries. It is useful to compare these predictions with the World Bank's classification of fragile situations. There are 10 countries that appear on the FCS list and they are highlighted in light gray. Of these, five countries also have conflict predictions from all four classification algorithms (Burundi, Djibouti, Guinea-Bissau, Chad, and Zimbabwe). A further four countries have conflict predictions based on two or three of the algorithms.

Perhaps more interesting are cases where there is disagreement between the classification algorithms and the World Bank list. There is only one country on the World Bank list, Togo, for which none of the classification algorithms predicts conflict. In contrast, there are eight countries not on the World Bank list, but where all four conflict prediction algorithms do signal conflict and they are highlighted in dark gray: Bangladesh, Mauritania, Malawi, Nepal, Niger Senegal, Tajikistan, and Uzbekistan. Interestingly enough, in the latest replenishment round of IDA, an exceptional Fragility, Conflict, and Violence (FCV) Risk Mitigation Regime was established that would make countries who want to address FCV risks to be eligible for up to 1/3 of the country's indicative allocations with a cap of

[22] Note that a country can only appear on the World Bank's list of fragile and conflict-affected states if its CPIA score was publicly disclosed, which is done only for countries eligible for borrowing from IDA. For this reason, countries for which CPIA scores are not publicly disclosed are indicated with as missing values in the fifth column of Table 6.

US$300 million per country per replenishment. The four countries that were selected to be eligible for this replenishment round were Guinea, Nepal, Niger and Tajikistan[23]. For Guinea, three of the four prediction algorithms signal conflict, while for the other three countries, all prediction algorithms signal conflict. This suggests that having such processes in the future be informed by robust classification algorithms would enhance the ability of policymakers to make informed decisions.

It is useful to focus on the fourth column containing the predictions based on the threshold classifier, which as we have shown in the previous section tends to have better predictive power than the other classification rules. The threshold classifier signals a risk of conflict in 37 of the 87 countries reported in Table 6. Interestingly, there is only one case where the threshold classifier predicts conflict but none of the other classifiers do, which is Belarus. Conversely, there are only a handful of cases where the threshold classifier does not predict conflict but at least two out of three of the other classifiers do: Indonesia, Kyrgyz Republic, Sri Lanka, Madagascar, Sierra Leone, and Zambia. Finally, while not explicitly shown in Table 6 for reasons of space, it is also interesting to document the role of the 16 different indicators in triggering signals of conflict for the threshold classifier. There is considerable variation across indicators in this respect. Three of the individual indicators rarely see breaches of their corresponding thresholds: the young male population share, GDP per capita growth, and terms of trade shocks cross thresholds in five or fewer countries. In contrast, spatial inequality, history of conflict, the political terror scale, natural disasters, and neighbors in conflict breach their corresponding thresholds in 30 or more of the 87 countries in Table 6.

While a variable with more breaches of its corresponding threshold is more likely to contribute to predictions of conflict, recall that a minimum of five variables must breach their thresholds in order to generate a conflict prediction. Thus, it also matters how often a given variable crosses its threshold for countries for which at least four other variables cross their own thresholds as well. To capture this notion of specificity, we also calculate the proportion of breaches for each indicator that occur in countries where at least four other variables breach as well, i.e. in countries where the threshold classifier signals conflict. The CPIA and the political terror scale stand out in this respect: although these two indicators individually signal conflict in only 9 and 21 countries respectively, in 8 out of 9 cases for the CPIA and in 18 out of 21 cases for the political terror scale, the threshold classifier also predicts conflict with a total of at least five breaches. At the other extreme, the thresholds for spatial inequality

[23] See details at http://documents.worldbank.org/curated/en/652991468196733026/pdf/106182-BR-IDA18-Fragility-Conflict-and-Violence-PUBLIC-IDA-R2016-0140.pdf.

and natural disasters are low and these two variables generate 47 and 46 signals of conflict, respectively. Yet in only 17 of these cases (for spatial inequality) and 23 of these cases (for natural disasters) does the overall threshold classifier predict conflict.

One final observation suggests directions for future work. While we have documented substantial differences across four classification algorithms in terms of predictive performance, even the best-performing threshold classifier has only modest predictive power. An open question is whether some combination of stronger covariates and more robust prediction algorithms can substantially improve over this predictive performance. In identifying such algorithms, the results of this paper suggest two considerations are important: calibrating the prediction algorithm to directly minimize the prediction loss function, and ensuring that the algorithm features a prediction rule that is simple and transparent if the algorithm is to be used for policy purposes.

References

Alesina, Alberto, Stelios Michalopoulos, and Elias Papaioannou. 2016. "Ethnic Inequality." Journal of Political Economy 124 (2): 428-488.

Athey, Susan (2017). "Beyond Prediction: Using Big Data for Policy Problems". Science. 355:483-485.

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

Besley, Timothy, and Torsten Persson. 2011. "The Logic of Political Violence." The Quarterly Journal of Economics 126 1411–1445.

Blair, Robert, Christopher Blattman, and Alexandra Hartman (forthcoming). "Predicting Local Violence". Journal of Peace Research.

Blair, Robert and Nicholas Sambanis 2017. "Forecasting Civil Wars: Theory and Structure in an Age of Big Data and Machine Learning". Manuscript, University of Pennsylvania.

Blattman, Christopher, and Edward Miguel. 2010. "Civil War." Journal of Economic Literature 48:1, 3-57.

Bodea, Cristina, Masaaki Higashijima, and Raju Singh. 2016. "Oil and Civil Conflict: Can Public Spending Have a Mitigation Effect." World Development (78) 1-12.

Breiman, L. 2001. "Random Forests". Machine Learning 45(1), 5-32.

Cederman, Lars-Erik and Nils B. Weidmann. 2017. "Predicting Armed Conflict: Time to Adjust Our Expectations?" Science 474-476.

Chadefaux, Thomas. 2013. "Early Warning Signals for War in the News." Journal of Peace Research.

Collier, P, and A. Hoeffler. 2002. "On the Incidence of Civil War in Africa." Journal of Conflict Resolution 46,13–28.

Collier, P., and A. Hoeffler. 2004. "Greed and grievance in Civil War." Oxford Economic Papers 56, 563–595.

Collier, Paul, and Anke Hoeffler. 1998. "On economic causes of Civil War." Oxford Economic Papers 50(4), 563-573.

Elliott, Graham and Allan Timmermann. 2016. Economic Forecasting. Princeton: Princeton University Press.

Elliott, Graham and Robert Leili. 2013. "Predicting Binary Outcomes." Journal of Econometrics. 174: 15-26.

Esteban, Joan, Laura Mayoral, and Debraj Ray. 2012. "Ethnicity and Conflict: An Empirical Study." American Economic Review 102(4): 1310-1342.

Fearon, J. D. 2005. "Primary Commodities Exports and Civil War." Journal of Conflict Resolution.

Fearon, J. D., and D. D. Laitin. 2003. "Ethinicity, Insurgency, and Civil War." Am. Polit. Sci. Rev. 97,75–90.

Hegre, Havard and Nicolas Sambanis. 2006. "Sensitivity Analysis of Empirical Results on Civil War Onset." The Journal of Conflict Resolution 50 (4).

Hegre, Havard, Halvard Buhaug, Katherine Calvin, Jonas Nordkvelle, Stephanie Waldhoff and Elisabeth Gilmore. 2016. "Forecasting Civil Conflict Along the Shared Socioeconomic Pathways." Environmental Research Newsletter 11: 1-8.

Hegre, Håvard, J Karlsen, H Nygard, H Strand, and H Urdal. 2013. "Predicting Armed Conflict, 2010–2050." International Studies Quarterly.

Hsiang, Solomon M., Marshall Burke, and Edward Miguel. 2013. "Quantifying the Influence of Climate on Human Conflict." Science.

Kalyvas, S. N. 2008. "Ethnic Defection in Civil War." Comparative Political Studies 41, 1043–1068.

Kraay, Aart, and Vikram Nehru. 2006. "When is External Debt Sustainable?" The World Bank Economic Review 20 (3), 341-365.

Muchlinski, David, David Siroky, Jingrui He and Matthew Kocher. 2016. "Comparing Random Forest with Logistic Regression for Class-Imbalanced Civil War Onset Data." Political Analysis 24: 87-103.

Mueller, Hannes and Christopher Rauh (2016). "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text". CEPR Discussion Paper No. 11516.

Mullainathan, Sendhil, and Jann Spiess. (2017). "Machine Learning: An Applied Econometric Approach." Journal of Economic Perspectives 31 (2): 87-106.

Nelder, John A. and R. Mead. 1965. "A Simplex Method for Function Minimization." Computer Journal 308-313.

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

Perry, Chris. 2013. "Machine Learning and Conflict Prediction: A Use Case." Stability: International Journal of Security and Development 2(3) 1-18.

Reinhart, Carmen, Graciela Kaminsky and Saul Lizondo. 1998. "Leading Indicators of Currency Crises." IMF Economic Review 1-48.

Ross, M. 2004. "What Do We Know about Natural Resources and Civil War? ." Journal of Peace Research 41, 337–356 .

Ward, M, Brian Greenhil, and Kristin Bakke. 2010. "The perils of policy by p-value: Predicting civil conflicts." Journal of Peace Research Vol 47, Issue 4, 363 – 375.

Weidmann, Nils B. and Michael D. Ward. 2010. "Predicting Conflict in Space and Time." The Journal of Conflict Resolution 54 (6): 883-901.

World Bank. 2011. World Development Report 2011: Conflict, Security, and Development. World Bank. Washington, DC: World Bank.

World Bank. 2011. The Changing Wealth of Nations: Measuring Sustainable Development in the New Millennium". World Bank.

World Bank. 2016. Forcibly Displaced: Toward a Development Approach Supporting Refugees, the Internally Displaced, and Their Hosts. Washington, DC: World Bank.

Appendix A: Data Sources

Battle Death Data	UCDP/PRIO Armed Conflict Dataset version 4-2016
Presence of UN Peacekeeping missions	Coded from here
Refugees (by country or territory of origin)	UNHCR population statistics
Ethnic Income Inequality, Spatial Inequality	Alesina, Michalopoulos, & Papaioannou (2016). Ethnic income inequality combines satellite night light density -- a spatially-disaggregated proxy for per capita income -- with maps delineating the boundaries of different ethnic groups, to obtain a proxy for income inequality across ethnic groups. Spatial Inequality measures inequality across pixels within a country and therefore proxies for overall inequality. Data is available only at decadal frequency, and we linearly interpolate the intervening years to obtain annual data.
Natural Resource Rents (% of GDP)	"The Changing Wealth of Nations: Measuring Sustainable Development in the New Millennium" , World Bank (2011). It is based on estimates of the difference between commodity prices and unit costs of production, for a large number of commodities. These are aggregated across commodities produced in a country, and expressed as a share of GDP.
Ethnic and Religious Fractionalization	Alberto Alesina, Arnaud Devleeschauwer, William Easterly,Sergio Kurlat, and Romain Wacziarg. "Fractionalization" *Journal of Economic Growth,* vol. 8, no. 2, June 2003, pp. 155-194.
Male 15-29 Pop. Share	Share of male population aged 15-29 in the total population. From UN POPIN Database
History of Conflict	Fraction of years since 1970 that a country has been in conflict. It is adjusted appropriately for countries that became independent after 1970
CPIA	World Bank's Country Policy and Institutional Assessment
Freedom House	Composite indicator (average) of civil liberties and political rights coming from Freedom House
Political Terror Scale	Composed as an average indicator of the numerically coded indicators from yearly country reports of Amnesty International, the U.S. State Department Country Reports on Human Rights Practices, and Human Rights Watch's World Reports.
Log GDP per capita	Penn World Tables 9.0.
Log Population	World Development Indicators
GDP per capita Growth	Penn World Tables 9.0.
TOT change	Income effect of changes in the terms of trade, comes from Penn World Tables 9.0. It is complemented with data from World Development Indicators for countries with missing data
Natural Disaster	Incidence of natural disasters, constructed from EMDAT similar to Besley and Persson (2011). This is a binary indicator of whether any of the following types of natural disasters occur in a given country and year: extreme temperature events, floods, landslides and tidal waves.
Neighbors in Conflict	Binary indicator of whether a country has any neighbors in conflict

Appendix B: Performance of Classification Algorithms in the Hegre and Sambanis (2006) Dataset

In this annex we briefly describe how the four classifications algorithms described in the main text perform in the conflict dataset in Hegre and Sambanis (2006), (HS). In their seminal contribution, HS were the first to systematically document the robustness of the relationship between civil conflict and many of the correlates of conflict that had been proposed in the previous literature. HS considered a set of 88 such variables, and used a variant of extreme bounds analysis to isolate a subset of 18 variables that were robustly associated with conflict, in the sense of having a weighted average p-value smaller than 0.05. We take these 18 variables, listed in Table 3 of HS, together with two of the three "core" variables included in all of their specifications (logarithm of population and the logarithm of GDP per capita, Table 2 in HS) as our set of candidate explanatory variables.[24]

We use the HS primary definition of civil war as our measure of conflict. As in the original HS contribution, conflict episodes begin in the first year that conflict is observed. HS treat all country-years in which conflict is not observed as non-conflict observations, and then logit regressions to estimate the relationship between conflict and contemporaneous explanatory variables in a country-year panel. Here we use the HS conflict data to create a set of conflict episodes and complementary non-conflict episodes in the same way as in the main text. Conflict episodes begin in a year where conflict is observed, and no conflict is observed in the three previous years. Non-conflict episodes begin in the first of non-overlapping five-year periods in which conflict is not observed, and which are preceded by three years of no conflict as well. As in the main text, and consistent with our emphasis on predicting future conflict based on contemporaneously-available data, we measure the explanatory variables as annual averages in the three years prior to the start of the episode. Transformed in this way, the dataset consists of XX conflict episodes and YY non-conflict episodes.

Annex Table 1 summarizes the predictive performance of the four classification algorithms in the HS data. Model 1 in the left panel considers a model with the top 10 most significant variables from HS, while Model 2 in the right panel considers a model with all 20 variables. The first block of the table considers the in-sample fit of the two models, and is analogous to Table 4 in the main text. Here the threshold classifier performs best in Model 1 with 10 explanatory variables, but the random forest classifier performs best in Model 2. The second and third panels of the table divide the sample at 1980 and at 1990 respectively, and report the in-sample predictive power for the four classifiers estimated in the first half of the sample, followed by the out-of-sample predictive power in the second half of the sample. In Model 2 with 20 explanatory variables, the threshold classifier has the best out-of-sample predictive performance for both sample splits, although extremely close to that of the linear classifier in the sample split at 1980. In the more parsimonious Model 1, the linear classifier edges out the threshold classifier by a tiny margin when the sample is split at 1980, and both perform only marginally better than the random forest classifier. Somewhat surprisingly, the probit classifier has the best out-of-sample performance when the sample is split at 1990. Despite this aberration, overall the results that emerge from Annex Table 1 are broadly similar to those in the main text: the threshold classifier performs relatively well as an out-of-sample prediction rule, and there is little evidence of a penalty in

[24] We do not include the third core variable (number of years at peace) since it was not consistently significant in the HS specifications. We obtained the dataset from the replication materials generously provided by Muchlinsky et. al. (2016) who also work with this dataset.

terms of worse predictive power relying on this simple and transparent classification rule relative to more sophisticated alternatives such as the random forest classifier.

Annex Table 1: Predictive Performance of Classification Rules in the Hegre-Sambanis Dataset

	Model 1 -- HS Top 10 Variables				Model 2 -- HS Top 20 Variables			
	PC	RF	LC	TC	PC	RF	LC	TC
Full Sample								
Type 1 Error	0.073	0.122	0.098	0.134	0.098	0.244	0.122	0.256
Type 2 Error	0.497	0.504	0.439	0.370	0.454	0.244	0.383	0.295
Objective Function	0.285	0.313	0.268	0.252	0.276	0.244	0.252	0.276
Split Sample at 1980								
In-Sample								
Type 1 Error	0.204	0.163	0.082	0.184	0.102	0.347	0.082	0.286
Type 2 Error	0.365	0.494	0.445	0.310	0.449	0.125	0.414	0.208
Objective Function	0.284	0.329	0.263	0.247	0.276	0.236	0.248	0.247
Out-of-Sample								
Type 1 Error	0.333	0.061	0.121	0.303	0.102	0.333	0.082	0.286
Type 2 Error	0.410	0.628	0.549	0.368	0.449	0.261	0.414	0.208
Objective Function	0.371	0.344	0.335	0.336	0.276	0.297	0.248	0.247
Split Sample at 1990								
In-Sample								
Type 1 Error	0.197	0.409	0.076	0.167	0.091	0.152	0.091	0.227
Type 2 Error	0.380	0.242	0.458	0.320	0.434	0.305	0.396	0.354
Objective Function	0.288	0.325	0.267	0.243	0.262	0.228	0.243	0.291
Out-of-Sample								
Type 1 Error	0.188	0.375	0.125	0.313	0.375	0.125	0.250	0.188
Type 2 Error	0.410	0.306	0.528	0.419	0.376	0.520	0.371	0.358
Objective Function	0.299	0.325	0.327	0.366	0.375	0.322	0.311	0.273

Notes: This table reports the Type 1 and Type 2 error rates together with the minimized value of the prediction loss function, for in-sample and out-of-sample predictive performance, splitting the sample at the indicated year.

Table 1: Start Year of Conflict Episodes

Azerbaijan	2005	Egypt, Arab Rep.	1993	Mali	1994	Rwanda	2009
Azerbaijan	2012	Egypt, Arab Rep.	2014	Mali	2007	Senegal	1990
Burundi	1991	Ghana	1981	Mozambique	2013	Senegal	2011
Burkina Faso	1987	Guinea	2000	Mauritania	2010	Sierra Leone	1991
Bangladesh	2005	Gambia, The	1981	Malaysia	1981	El Salvador	1979
Bhutan	1992	Guinea-Bissau	1998	Malaysia	2013	Suriname	1986
Central African Republic	1998	Haiti	2004	Niger	1991	Togo	1986
Central African Republic	2006	Indonesia	1997	Niger	2007	Togo	1993
China	2008	Iran, Islamic Rep.	2005	Nigeria	2009	Thailand	2003
Cote d'Ivoire	2002	Kenya	1982	Nicaragua	1978	Tajikistan	2010
Cameroon	1984	Lao PDR	1989	Nepal	1996	Trinidad and Tobago	1990
Congo, Rep.	1993	Liberia	1980	Panama	1989	Tunisia	1980
Congo, Rep.	1997	Liberia	1989	Peru	1982	Turkey	1984
Comoros	1989	Sri Lanka	1984	Peru	2007	Ukraine	2014
Comoros	1997	Lesotho	1998	Paraguay	1989	Uzbekistan	1999
Djibouti	1991	Mexico	1994	Romania	1989	Uzbekistan	2004
Djibouti	1999	Mali	1990	Rwanda	1980	Venezuela, RB	1992
Algeria	1991						

Notes: This table lists the conflict episodes studied in the paper, identified by the country and the start year of the episode.

Table 2: Summary Statistics

Average before episode of		
Latent Conflict		
Ethnic Income Inequality	0.472	0.578
Natural Resource Rents (% of GDP)	8.291	13.099
Ethnic Fractionalization	0.480	0.549
Religious Fractionalization	0.445	0.387
Spatial inequality	0.456	0.534
Male 15-29 Pop. Share	0.136	0.137
History of Conflict	0.077	0.116
Institutions		
CPIA	3.749	3.393
Freedom House	3.730	4.739
Political Terror Scale	2.265	2.685
Log GDP per capita	8.392	7.873
Log Population	15.542	15.954
Shocks		
GDP per capita Growth	0.024	0.016
TOT change	0.012	0.010
Natural Disaster	0.487	0.551
Neighbors in Conflict	0.490	0.720
No. obs.	422	69

Notes: This table reports means of the correlates of conflict considered in the paper, separately by conflict and non-conflict episodes.

Table 3: Univariate Predictions of Conflict

	In Sample Predictions			Estimated Thresholds		Area Under	`| Z-statistic |
	T1	T2	Obj	Percentile	Actual Value	ROC Curve	in Probit
Latent Conflict							
Ethnic Income Inequality	0.145	0.661	0.403	0.312	0.413	0.610	3.24
Ethnic Fractionalization	0.493	0.339	0.416	0.624	0.629	0.634	2.12
Religious Fractionalization*	0.609	0.239	0.424	0.542	0.525	0.516	1.81
Spatial inequality	0.043	0.827	0.435	0.153	0.157	0.586	2.42
Natural Resource Rents	0.377	0.400	0.389	0.566	6.156	0.653	2.95
Male 15-29 Pop. Share	0.130	0.777	0.454	0.208	0.127	0.506	0.58
History of Conflict	0.522	0.296	0.409	0.678	0.044	0.568	1.78
Institutions							
CPIA*	0.623	0.194	0.409	0.224	3.200	0.617	3.36
Freedom House	0.145	0.521	0.333	0.420	3.333	0.664	4.39
Political Terror Scale	0.290	0.405	0.348	0.552	2.500	0.670	4.26
Log GDP per capita*	0.275	0.448	0.362	0.489	8.390	0.658	4.22
Log Population	0.406	0.455	0.430	0.352	15.043	0.567	1.91
Shocks							
GDP per capita Growth*	0.623	0.182	0.403	0.210	-0.009	0.555	1.20
TOT change*	0.145	0.749	0.447	0.705	0.024	0.484	0.56
Natural Disaster	0.884	0.057	0.470	0.935	1.667	0.506	0.79
Neighbors in Conflict	0.290	0.469	0.380	0.483	0.667	0.618	3.60

Notes: This table summarizes predictions of conflict based on each of the individual explanatory variables, taken one at a time. T1 and T2 refer to Type 1 and Type 2 error rates, and Obj refers to prediction loss function. The last two columns report the percentile rank and the value of the threshold that separates conflict and non-conflict episodes. Since with a single explanatory variable the probit, threshold, and linear classifiers are identical, results are reported for the probit classifier only.

* indicates that orientation of variables has been reversed so that higher values are associated with a greater risk of conflict. Values of these variables *below* the indicated threshold correspond to predictions of conflict. For all other variables, values *above* the indicated threshold correspond to predictions of conflict.

Table 4: Multivariate Predictions of Conflict

	Model 1				Model 2				Model 3			
	PC	RF	LC	TC	PC	RF	LC	TC	PC	RF	LC	TC
Predictive Performance												
Type 1 Error	0.174	0.087	0.116	0.130	0.188	0.246	0.101	0.101	0.174	0.290	0.101	0.072
Type 2 Error	0.429	0.630	0.464	0.408	0.431	0.396	0.462	0.398	0.405	0.334	0.460	0.310
Objective Function	0.301	0.359	0.290	0.269	0.310	0.321	0.282	0.250	0.290	0.312	0.281	0.191
Explanatory Variables												
Latent Conflict												
Ethnic Income Inequality	0.187	5.950	0.175	0.788	0.13	5.59	-0.04	0.98	0.101	4.635	0.125	0.632
Ethnic Fractionalization*					0.02	4.66	-0.15	0.95	0.056	3.652	0.077	0.890
Religious Fractionalization*									0.184	4.376	0.071	0.747
Spatial inequality*					0.256	6.548	0.156	0.989	0.210	4.990	0.186	0.493
Natural Resource Rents	0.405	5.907	0.634	0.142	0.372	7.232	0.596	0.205	0.315	5.713	0.371	0.165
Male 15-29 Pop. Share*									0.063	4.247	0.051	0.947
History of Conflict									0.064	3.115	0.161	0.088
Institutions												
CPIA*					0.204	6.342	0.375	0.581	0.081	5.811	0.107	0.594
Freedom House	0.724	3.921	0.763	0.666	0.131	5.333	0.300	0.858	0.119	4.761	0.152	0.741
Political Terror Scale	0.373	6.400	0.317	0.532	0.467	4.378	0.629	0.661	0.387	4.007	0.415	0.508
Log GDP per capita*					0.539	7.261	0.320	0.603	0.393	6.136	0.414	0.591
Log Population*									0.110	5.556	0.082	0.786
Shocks												
GDP per capita Growth*	0.373	6.400	0.317	0.532					0.079	5.536	0.097	0.539
TOT change									0.002	4.378	-0.019	0.958
Natural Disaster									0.041	1.840	0.076	0.191
Neighbors in Conflict	0.242	1.000	0.258	0.999	0.164	1.000	0.380	0.664	0.103	1.000	0.138	0.921
Number of Breaches				2				2				5

Notes: PC=Probit Classifier, RF=Random Forest, LC=Linear Classifier, TC=Threshold Classifier. Top panel reports Type 1 and Type 2 error rates and the minimized value of the prediction loss function. Remaining rows report measures of importance in predicting conflict for individual explanatory variables. Columns for PC and LC report weights on indicated variables in linear combination of variables used to predict conflict. Columns for TC report the percentile of the distribution of the indicated variable where the optimal threshold is set -- lower values imply more breaches and therefore a greater role in predicting conflict. Columns for RF report mean improvement in node purity, normalized to 1 for the lowest improvement -- higher values imply greater importance. Number of breaches refers to number of thresholds that must be crossed to signal conflict. Variables indicated with * have orientation reversed, i.e. variables that are negatively correlated with conflict in our sample have been re-oriented to be positively correlated with conflict.

Table 5: In-Sample and Out-of-Sample Predictive Power

	Probit Classifier		Random Forest		Linear Classifier		Threshold Classifier	
Cutoff	In-Sample	Out-Sample	In-Sample	Out-Sample	In-Sample	Out-Sample	In-Sample	Out-Sample
Model 1								
1990	0.342	0.430	0.405	0.375	0.274	0.418	0.312	0.335
1991	0.324	0.392	0.404	0.381	0.296	0.380	0.296	0.330
1992	0.340	0.404	0.376	0.390	0.309	0.407	0.292	0.345
1993	0.360	0.327	0.366	0.354	0.341	0.357	0.309	0.296
1994	0.353	0.297	0.381	0.367	0.341	0.333	0.310	0.284
1995	0.354	0.289	0.385	0.324	0.332	0.345	0.282	0.352
1996	0.338	0.286	0.358	0.350	0.327	0.306	0.300	0.299
1997	0.334	0.290	0.368	0.336	0.329	0.305	0.276	0.327
1998	0.331	0.299	0.367	0.380	0.326	0.295	0.275	0.358
1999	0.359	0.231	0.385	0.363	0.334	0.258	0.286	0.380
2000	0.354	0.280	0.379	0.343	0.328	0.238	0.288	0.305
Average	0.344	0.320	0.379	0.360	0.322	0.331	0.293	0.328
Model 2								
1990	0.321	0.438	0.371	0.427	0.290	0.362	0.211	0.309
1991	0.336	0.444	0.404	0.394	0.276	0.422	0.208	0.303
1992	0.329	0.414	0.373	0.400	0.267	0.389	0.218	0.322
1993	0.346	0.368	0.404	0.378	0.287	0.317	0.224	0.287
1994	0.348	0.312	0.384	0.328	0.296	0.341	0.212	0.299
1995	0.345	0.338	0.386	0.380	0.284	0.330	0.214	0.271
1996	0.339	0.369	0.360	0.341	0.293	0.385	0.202	0.313
1997	0.320	0.308	0.360	0.355	0.287	0.315	0.217	0.345
1998	0.323	0.309	0.369	0.386	0.297	0.288	0.211	0.325
1999	0.339	0.279	0.350	0.338	0.320	0.308	0.238	0.343
2000	0.325	0.266	0.349	0.348	0.296	0.272	0.236	0.327
Average	0.334	0.350	0.374	0.370	0.290	0.339	0.217	0.313
Model 3								
1990	0.324	0.486	0.429	0.406	0.285	0.361	0.204	0.347
1991	0.292	0.472	0.448	0.367	0.260	0.432	0.196	0.357
1992	0.326	0.443	0.472	0.365	0.273	0.355	0.195	0.262
1993	0.323	0.422	0.425	0.383	0.267	0.374	0.195	0.351
1994	0.325	0.506	0.410	0.347	0.260	0.370	0.201	0.312
1995	0.307	0.516	0.416	0.330	0.245	0.408	0.201	0.244
1996	0.300	0.449	0.386	0.356	0.251	0.370	0.189	0.302
1997	0.287	0.499	0.380	0.278	0.245	0.443	0.204	0.236
1998	0.282	0.364	0.347	0.326	0.257	0.324	0.178	0.324
1999	0.279	0.444	0.362	0.317	0.252	0.394	0.202	0.332
2000	0.281	0.446	0.375	0.328	0.242	0.337	0.200	0.367
Average	0.302	0.459	0.405	0.346	0.258	0.379	0.197	0.312

Notes: This table reports the minimized value of the prediction loss function, for in-sample and out-of-sample predictive performance, splitting the sample at the indicated year. Model 1, 2, and 3 are as defined in Table 4.

Table 6: Conflict Predictions Based on Alternative Classification Algorithms

	Classifications Based on 2012-2014 data (1=Conflict, 0 otherwise)				
	Probit	**Random Forest**	**Linear**	**Threshold**	**FCS List**
Angola	1	1	1	1	..
Albania	0	0	0	0	..
Argentina	0	0	0	0	..
Armenia	0	0	1	0	..
Burundi	1	1	1	1	Y
Benin	0	0	1	1	N
Burkina Faso	1	0	1	1	N
Bangladesh	1	1	1	1	N
Bulgaria	0	0	0	0	..
Bosnia and Herzegovina	0	0	0	0	..
Belarus	0	0	0	1	..
Belize	0	0	0	0	..
Bolivia	0	0	0	0	N
Brazil	0	0	1	0	..
Botswana	0	0	0	0	..
Chile	0	0	0	0	..
China	1	1	1	1	..
Cameroon	1	0	1	1	N
Congo, Rep.	1	1	1	1	N
Comoros	0	1	0	1	Y
Cape Verde	0	0	0	0	N
Costa Rica	0	0	0	0	..
Djibouti	1	1	1	1	Y
Dominican Republic	0	0	0	0	..
Ecuador	1	0	1	1	..
Fiji	0	0	0	0	..
Gabon	0	0	0	0	..
Georgia	0	0	1	1	..
Ghana	0	0	1	0	N
Guinea	1	0	1	1	N
Gambia, The	1	0	1	1	Y
Guinea-Bissau	1	1	1	1	Y
Equatorial Guinea	1	1	1	1	..
Guatemala	0	0	0	0	..
Honduras	0	0	0	0	N
Croatia	0	0	0	0	..
Indonesia	0	1	1	0	..
Jamaica	0	0	0	0	..
Jordan	1	0	1	1	..
Kazakhstan	0	1	0	1	..
Kenya	1	0	1	1	N
Kyrgyz Republic	1	0	1	0	N
Cambodia	1	0	1	1	N
Korea, Rep.	0	0	0	0	..
Lao PDR	0	0	1	1	N
St. Lucia	0	0	0	0	N
Sri Lanka	0	1	1	0	N
Lesotho	0	0	0	0	N

Note: Table continues on next page

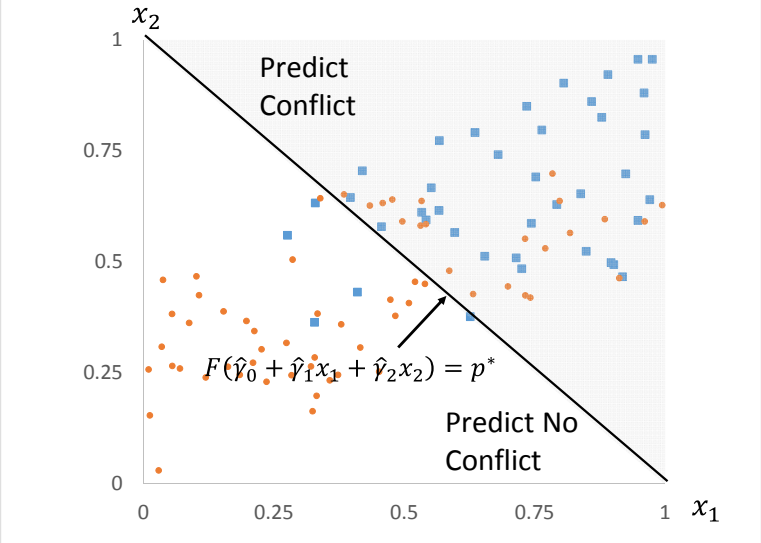
Table 6, Cont'd: Conflict Predictions Based on Alternative Classification Algorithms

	Classifications Based on 2012-2014 data (1=Conflict, 0 otherwise)				
	Probit	**Random Forest**	**Linear**	**Threshold**	**FCS List**
Morocco	1	1	1	1	..
Moldova	0	0	0	0	N
Madagascar	1	1	1	0	Y
Mexico	1	0	1	1	..
Macedonia, FYR	0	0	0	0	..
Mongolia	0	0	0	0	N
Mauritania	1	1	1	1	N
Mauritius	0	0	0	0	..
Malawi	1	1	1	1	N
Namibia	0	0	0	0	..
Niger	1	1	1	1	N
Nicaragua	0	0	0	0	N
Nepal	1	1	1	1	N
Panama	0	0	0	0	..
Peru	1	1	1	1	..
Poland	0	0	0	0	..
Paraguay	0	0	0	0	..
Romania	0	0	0	0	..
Senegal	1	1	1	1	N
Sierra Leone	1	0	1	0	Y
El Salvador	0	0	0	0	..
Swaziland	0	0	0	0	..
Seychelles	0	0	0	0	..
Chad	1	1	1	1	Y
Togo	0	0	0	0	Y
Tajikistan	1	1	1	1	N
Turkmenistan	1	1	1	1	..
Trinidad and Tobago	0	0	0	0	..
Tunisia	0	0	0	0	..
Tanzania	1	0	1	1	N
Uruguay	0	0	0	0	..
Uzbekistan	1	1	1	1	N
St. Vincent and the Grenadines	0	0	0	0	N
Venezuela, RB	1	1	1	1	..
Vietnam	0	0	0	0	N
Vanuatu	0	0	0	0	N
South Africa	0	0	0	0	..
Zambia	1	0	1	0	N
Zimbabwe	1	1	1	1	Y
Predict Conflict	35	26	43	37	
Predict No Conflict	52	61	44	50	

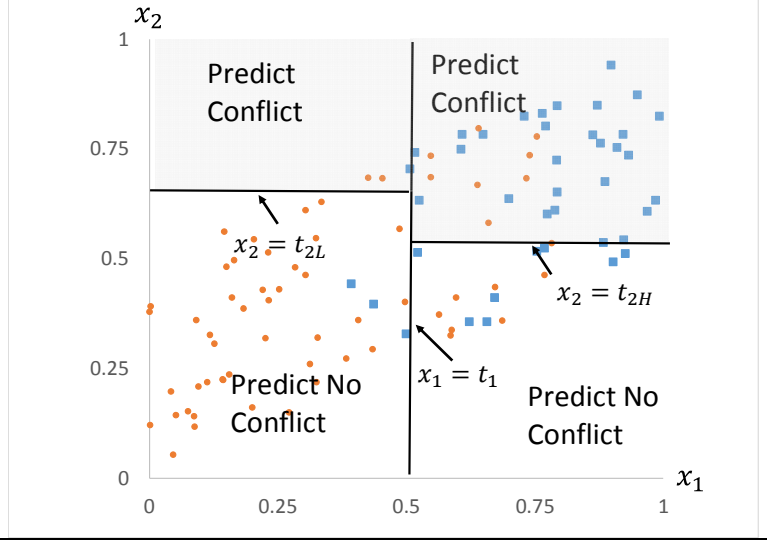
Notes: This table reports predictions of conflict based on the most recently-available data 2012-2014 for each country for which complete data on all explanatory variables is available and no conflict signals are observed over this period. Predictions are based on model 3. FCS List refers to the World Bank classification of "fragile situations" as of FY2017, and is based on a country having a CPIA score below 3.2 and/or a UN Peacekeeping Operation.

Figure 1: Probit Classifier and Classification Tree

Probit Classifier



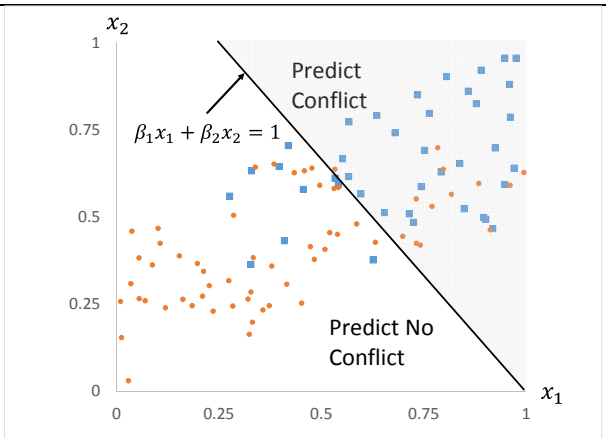
Classification Tree



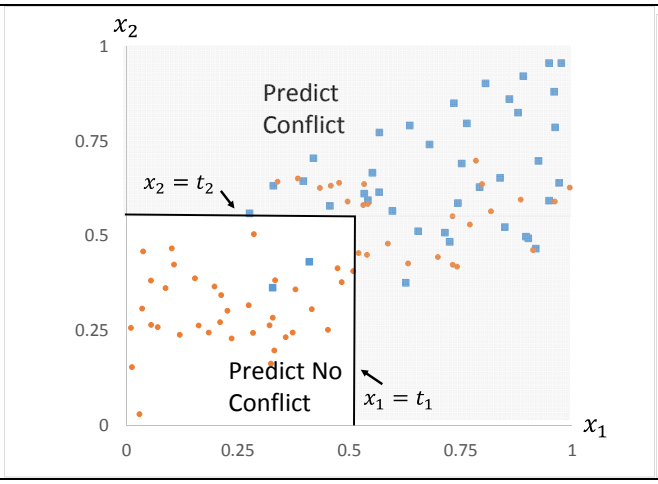
Notes: This figure illustrates the two conventional classification rules studied in the paper, for a hypothetical dataset with two explanatory variables. Conflict observations are indicated as blue squares and non-conflict observations are orange circles. The shaded region in each graph identifies the region in the data for which conflict is predicted.

Figure 2: The Linear and Threshold Classifiers

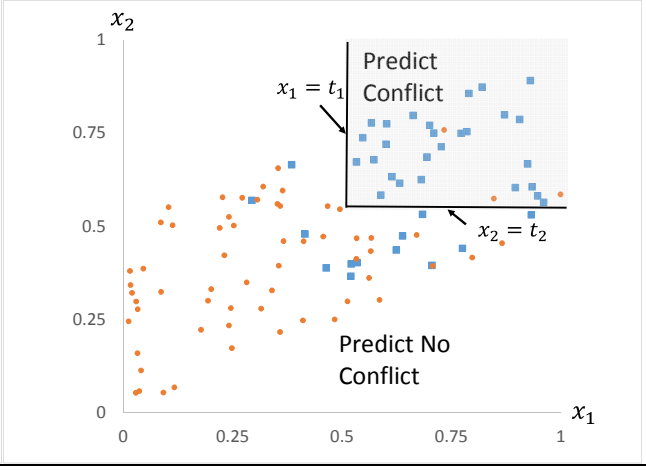
Linear Classifier



Threshold Classifier ($N = 1$)

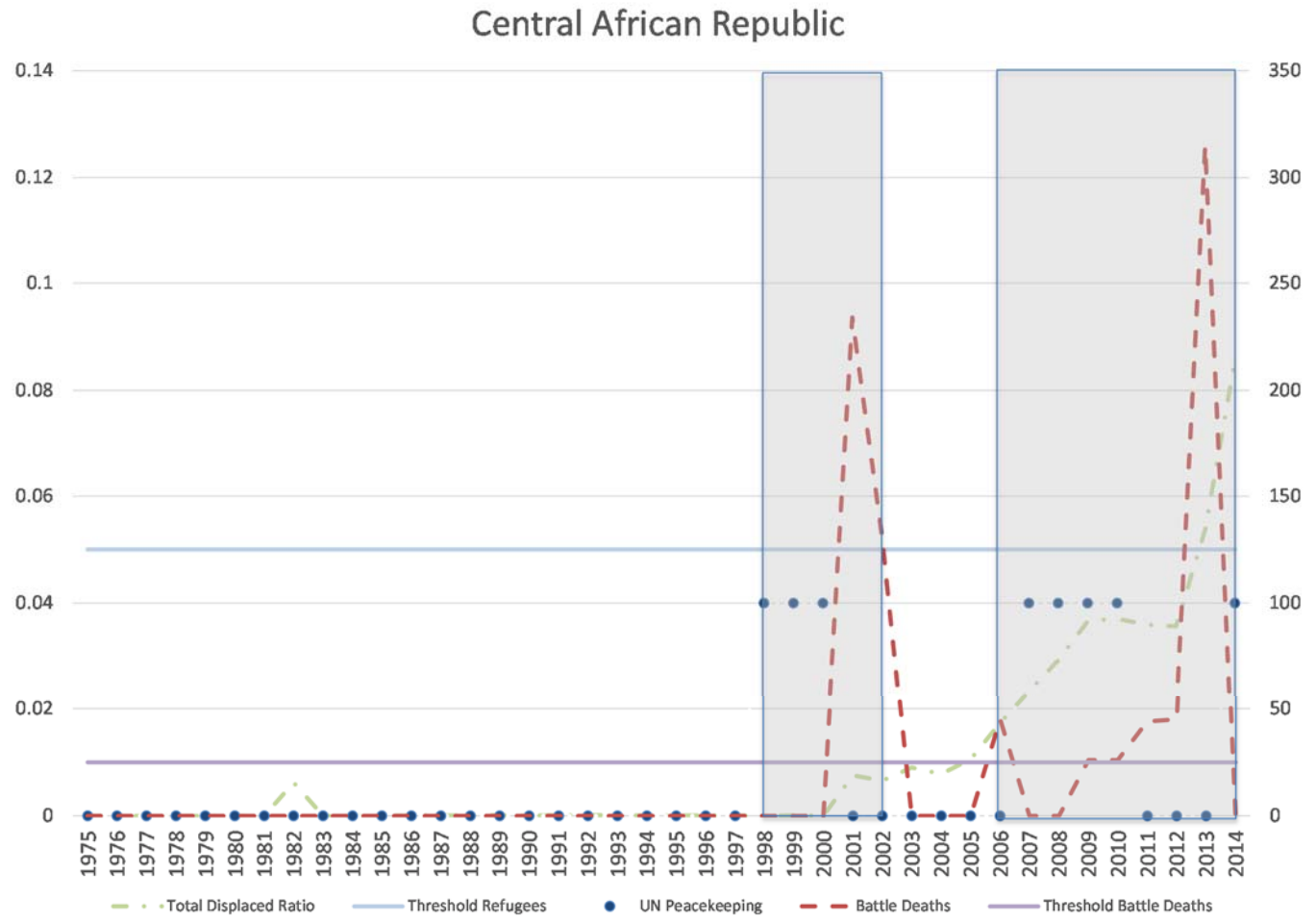


Threshold Classifier ($N = 2$)



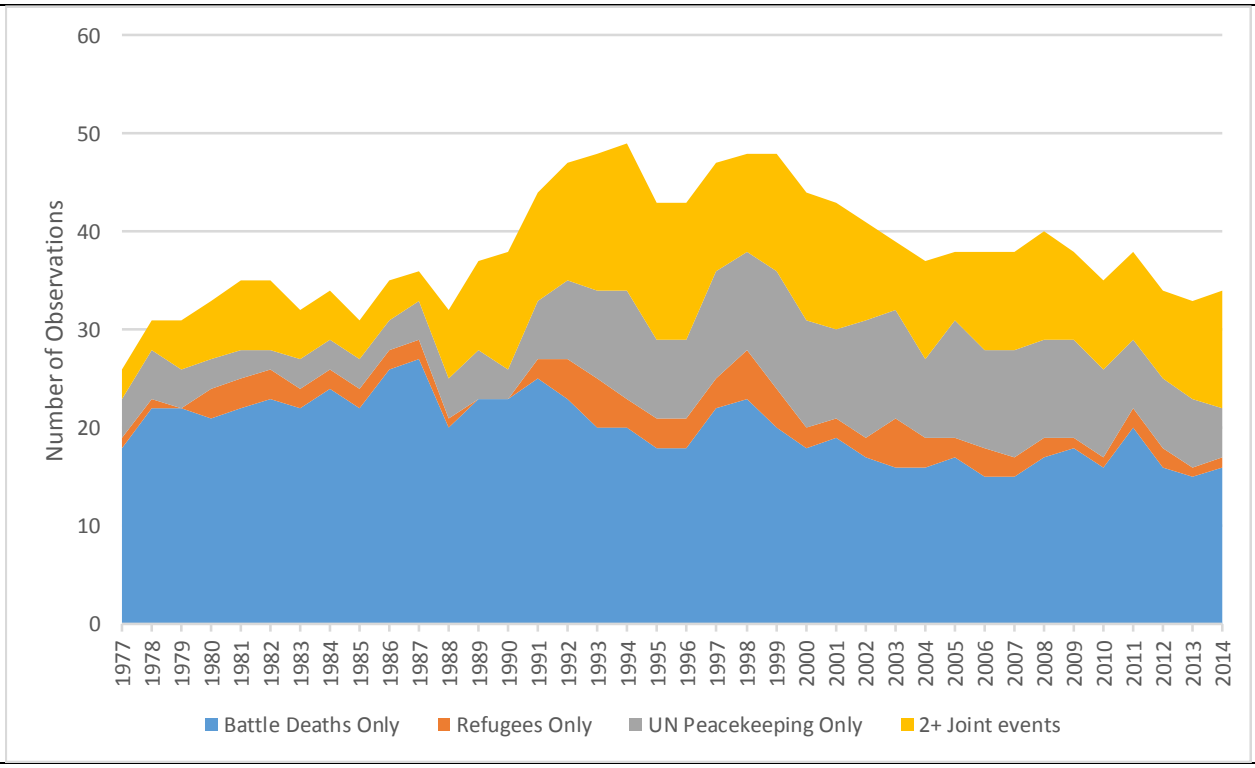
Notes: This figure illustrates the two unconventional classification rules studied in the paper, for a hypothetical dataset with two explanatory variables. Conflict observations are indicated as blue squares and non-conflict observations are orange circles. The shaded region in each graph identifies the region in the data for which conflict is predicted.

Figure 3: Identifying Conflict Episodes: Central African Republic Example



Notes: This graph illustrates the mapping from conflict-years to conflict events. The graph plots the number of battle deaths and the corresponding threshold; the number of displaced persons and the corresponding threshold; and an indicator for UN Peacekeeping Operations. The years corresponding to conflict episodes are shaded.

Figure 4: Contribution of Indicators to Conflict Years



Notes: This graph shows the contribution of each of the indicators of conflict to the total number of observed conflict episodes for each year in the dataset.