

The perils of policy by p-value: Predicting civil conflicts

Journal of Peace Research
 47(4) 363–375
 © The Author(s) 2010
 Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
 DOI: 10.1177/0022343309356491
jpr.sagepub.com



Michael D Ward

Department of Political Science, Duke University

Brian D Greenhill

Department of Political Science, University of Washington

Kristin M Bakke

Department of Political Science, University College London

Abstract

Large-n studies of conflict have produced a large number of statistically significant results but little accurate guidance in terms of anticipating the onset of conflict. The authors argue that too much attention has been paid to finding statistically significant relationships, while too little attention has been paid to finding variables that improve our ability to predict civil wars. The result can be a distorted view of what matters most to the onset of conflict. Although these models may not be intended to be predictive models, prescriptions based on these models are generally based on statistical significance, and the predictive attributes of the underlying models are generally ignored. These predictions should not be ignored, but rather need to be heuristically evaluated because they may shed light on the veracity of the models. In this study, the authors conduct a side-by-side comparison of the statistical significance and predictive power of the different variables used in two of the most influential models of civil war. The results provide a clear demonstration of how potentially misleading the traditional focus on statistical significance can be. Until out-of-sample heuristics – especially including predictions – are part of the normal evaluative tools in conflict research, we are unlikely to make sufficient theoretical progress beyond broad statements that point to GDP per capita and population as the major causal factors accounting for civil war onset.

Keywords

civil conflicts, cross-validation, prediction, statistical models

Introduction

It is not uncommon in conflict research to conduct statistical analyses and then draw policy inferences from the statistical information, without ever trying to make specific predictions. However, probability statements are post-dictions and often are used to prescribe what should happen in the future. Consider Paul Collier's award winning volume, *The Bottom Billion*, in which he notes:

We were also asked to use our model to predict where the next civil wars would be.... But we were never that foolish. Our predictions might have been used as labels and thus likely to damage the very countries I was concerned to help; they might even have become self-fulfilling prophecies. More fundamentally, our model can not be used for prediction. It *can* tell you what typically are the structural factors underlying proneness to civil war.... From this it can tell you the sort of countries that are most at risk. But

it cannot tell you whether Sierra Leone will have another civil war next year. That depends upon a myriad of short-term events. (Collier, 2007: 19)

He goes directly on to note that 'if a country grows at 3 percent, the risk [of civil war] is cut from 14 percent to 11 percent; if its economy declines at 3 percent, the risk increases to 16 percent' (Collier, 2007: 20). Presumably this also applies to Sierra Leone.¹

In another recent study, Collier, Hoeffler & Söderbom (2008: 10) note that 'severe autocracy appears to be highly successful in maintaining the post-conflict peace.... if the polity is

Corresponding author:

Michael D Ward, mw160@duke.edu.

¹ In a recent review of *The Bottom Billion*, Easterly (2008) provides a critique of some of the potentially dangerous causal inferences that are drawn from its statistical analysis.

highly autocratic the risk is only 24.6%, whereas if it is not highly autocratic the risk more than doubles to 62%. Our point is not that Collier (or anyone) should necessarily be required to generate predictions. Our point is that basing policy prescriptions on statistical summaries of probabilistic models (which *are* predictions) can lead to misleading policy prescriptions if out-of-sample predictive heuristics are ignored. Often such policy prescriptions might be based on factors that are important in one sample, but not very salient in the out-of-sample cases to which they are implicitly being applied. It is quite possible to focus on statistically significant results that are artifacts in the sense that they do not generalize beyond the specific cases studied. This happens if we focus only on statistically significant relationships and may actually hinder our ability to generalize to out-of-sample situations, such as the future!

This article makes the argument that scholars need to make and evaluate predictions in order to improve our models. We have to be willing to make predictions explicitly – and plausibly be wrong, even appear foolish – because our policy prescriptions need to be undertaken with results that are drawn from robust models that have a better chance of being correct. The whole point of estimating risk models is to be able to apply them to specific cases. You wouldn't expect your physician to tell you that all those cancer risk factors from smoking don't actually apply to you. Predictive heuristics provide a useful, possibly necessary, strategy that may help scholars and policymakers guard against erroneous recommendations.

Things we know about what causes civil war

It is by now widely recognized that civil wars are increasingly prevalent and lethal, providing an important threat to human security. An estimate of the total number of battle-deaths in civil wars can be found in the PRIO Battle Deaths dataset, which reports a figure of 2–3 million since 1946 (Lacina & Gleditsch, 2005).² For students of civil wars, the list of cases to study is long, as is the list of explanatory factors. Indeed, there seems to be both a scholarly and a popular sense that we now know what causes civil wars. In this article, we re-analyze the influential and oft-cited articles by Fearon & Laitin (2003) and Collier & Hoeffler (2004) and demonstrate that while these studies have established a number of statistically significant findings, they do a poor job of predicting civil war onsets. We argue that the search for statistically significant relationships – in these and other studies – may not be the strategy best suited for evaluating our models' ability to explain real world events, and we suggest using out-of-sample heuristics to obtain a better sense of the uncertainty implied by the in-sample statistical results.

Fearon & Laitin's article 'Ethnicity, insurgency, and civil war' (2003) and Collier & Hoeffler's 'Greed and grievance

Table I. Variables included in the Fearon & Laitin model

<i>Variable</i>	<i>Statistically significant at 0.05 level</i>
Prior War	Yes
GDP per capita	Yes
Population	Yes
Mountainous Terrain	Yes
Non-contiguous State	No
Oil Exporter	Yes
New State	Yes
Instability	Yes
Democracy	No
Ethnic Fractionalization	No
Religious Fractionalization	No

* based on Fearon and Laitin, 2003: Table 1, Column 1.

in civil war' (2004) have both made important contributions to the study of civil wars, and most other large-n empirical analyses are now expected to test their claims against the statistically significant findings of these studies. The main lesson from these studies is that, contrary to long-held beliefs, civil wars are not driven by collective grievances stemming from ethnic diversity and/or income inequality. A large literature has posited that ethnic diversity or territorially concentrated ethnic groups somehow contribute to the likelihood of intrastate conflicts.³ Likewise, a longstanding body of work of a more materialistic nature suggests that it is not identities but access to wealth that causes conflict. Income inequalities may create economic grievances and mobilization on the part of the poorer party (Gurr, 1970; Horowitz, 1985; Gurr, 2000; Stewart, 2005). Neither Fearon & Laitin (2003) nor Collier & Hoeffler (2004) find much empirical support for these identity or grievance-based arguments. Rather, they argue, civil wars are largely driven by the opportunities for insurgency: intrastate conflicts emerge in states that are not strong enough to control potentially rebellious groups and where such groups can thrive thanks to access to resources and funding.

A summary of the statistically significant findings for the Fearon & Laitin model is given in Table I; Table II provides the same information for the model estimated by Collier & Hoeffler.

³ The ethnic conflict literature has proposed a number of mechanisms for how and why ethnicity causes or somehow contributes to conflict. For instance, some argue that the link between ethnicity and conflict rests with emotions, including longstanding hatreds (Kaplan, 1993), resentment towards ethnic groups other than one's own (Petersen, 2002), or fear-driven attempts at protecting the existence of one's group (Posen, 1993; Lake & Rothchild, 1996). Ethnicity may also contribute to conflict, others suggest, through power-seeking political entrepreneurs who manipulate collective identities (e.g. Gagnon, 1994; Kaufman, 2001). Others maintain that ethnic conflicts are about social psychology and favoritism for one's own group (Hewstone & Greenland, 2000) or economic, social, or political discrepancies among different groups (Horowitz, 1985; Gurr, 2000). According to Toft (2003), violent ethnic conflicts are particularly likely to occur when ethnic groups are territorially concentrated in an area they consider to be their homeland.

² <http://www.prio.no/CSCW/Datasets/Armed-Conflict/Battle-Deaths/>

Table II. Variables included in the Collier & Hoeffler model

Variable	Statistically significant at 0.05 level
Commodity Dependence	Yes
Squared Commodity Dependence	Yes
Male Secondary Schooling	Yes
GDP Growth	Yes
Peace Duration	Yes
Geographic Dispersion	Yes
Population	Yes
Social Fractionalization	Yes
Ethnic Dominance	No

*based on Collier and Hoeffler, 2004: Table 5, column 5.

Collier & Hoeffler (2004: 564) argue that civil wars and rebellions are explained 'not by motive, but by the atypical circumstances that generate profitable opportunities'. This suggests that civil wars occur where and when rebel groups have the opportunity to raise revenues, most commonly where and when these groups are able to exploit (loot) natural resources; where and when they can take advantage of high levels of unemployment and poverty and, thus, readily available rebel recruits; and where and when they have ethnic diasporas willing to financially support them. In this view, rebels are rational agents, driven by opportunity and greed, rather than grievances. Likewise, Fearon & Laitin (2003) also find that a dependence on oil exports is positively correlated with intrastate conflict, but they reject the resource-predation hypothesis and suggest that the causal link here is about state weakness. Abundance of oil and dependence on oil exports, they argue, encourage weak state institutions. While Fearon & Laitin share Collier & Hoeffler's rejection of grievance-based arguments, they propose more of a state-centric perspective. Their central argument is that conflicts in the post-World War II era are a result of favorable insurgency conditions, by which they refer to circumstances that are hypothesized to ease mobilization by limiting the central state's ability to control its territory. Such conditions include mountainous terrain, large populations, political instability, the newness of the state, and low levels of economic development. Notably, although Fearon & Laitin's argument concerns state strength, the study includes no indicators for state institutions besides democracy, which has no statistically significant relationship to conflict. Their main indicator of state strength is GDP per capita. Whereas they interpret the negative and statistically significant relationship between GDP per capita and conflict onset as supportive of their argument that state weakness encourages civil wars, Collier & Hoeffler take the same finding to indicate that low income encourages civil wars.

In policy circles, the World Bank certainly was supportive of and influenced by Collier & Hoeffler's 'economics of civil war' approach, which promotes economic growth as the cure for preventing civil wars (see Collier et al., 2003). In the

popular press, Fearon & Laitin's argument about weak states creating favorable insurgency conditions has been featured in news outlets such as *The New Yorker*, *The New York Times*, *The New Republic*, and even the popular online magazine *Slate* (see Bass, 2006a,b; Diamond, 2006; Lemann, 2001). Importantly, and of some concern to several scholars, a key policy implication of these studies is that resolving the stated grievances over which civil wars are fought, such as discrimination and repression of certain ethnic groups, does not necessarily help. Indeed, in a hearing of the Subcommittee on National Security, Emerging Threats, and International Relations, James Fearon expressed his doubts about Iraq working its way out of a civil war through constitutional negotiation and agreements about the distribution of scarce resources such as oil. If there is a way towards a stable Iraq, he argued, it is based on economic efficiency and a central government able to exploit the country's oil resources (House of Representatives, 2006).

Despite their influence in scholarly, policy, and popular circles, Collier & Hoeffler and Fearon & Laitin's models prove ill-equipped to predict civil war onsets. As we demonstrate in the cross-validation exercise below, the models' statistically significant relationships do not necessarily make significant contributions to their predictive ability. To the degree that we want academic work to have policy implications, such lack of predictive ability is both a problem and an opportunity. There are three possible reasons for this absence of predictive power. First, it is possible that the models are misspecified. Second, it is possible that the models' explanatory factors are measured at too high a level of aggregation; most violent conflicts are local and not nationwide, yet most statistical studies seek to explain variation based on national-level variables. Third, it is also possible that the lack of predictive power is a result of research design. In this study, we focus on this latter point. We argue that the search for statistically significant relationships may not be the strategy best suited for evaluating our models' ability to explain real world events, and we suggest using out-of-sample heuristics as a complementary strategy.

What statistical significance does not tell us

The traditional approach to the quantitative study of civil wars involves testing the statistical significance of variables that are thought to be theoretically interesting. Once one or more of these variables is found to be significant in both a statistical and a substantive sense, the tendency is to conclude that we have come one step closer to discovering the true logic of civil wars. However, as we discuss in the following section, a variable that might at first be thought to represent an important conceptual breakthrough in our understanding of conflicts, owing to its statistical significance, often only leads to a very modest improvement in our ability to predict the onset of civil wars.

What is often overlooked is that despite their many strengths, existing models of civil war do a surprisingly poor

Table III. Number of correctly predicted onsets and false positives at varying cut-points

Threshold	<i>Fearon & Laitin model</i>		<i>Collier & Hoeffler model</i>	
	<i>Correctly predicted</i>	<i>False positives</i>	<i>Correctly predicted</i>	<i>False positives</i>
0.5	0/107	0	3/46	5
0.3	1/107	3	10/46	20
0.1	15/107	66	34/46	110

job of correctly predicting the onset of civil wars. Table III provides an illustration of the ability of the Fearon & Laitin and Collier & Hoeffler models to predict the onset of civil wars within the particular country-year samples used in the estimation of these models. Because both studies use logistic models to estimate the probabilities of civil wars occurring, we need to evaluate their predictive power by applying a rule that converts these probabilities into a dichotomous war/no war outcome. Typically, such an evaluation involves comparing each predicted probability to some arbitrary threshold above which civil wars are deemed to occur. The first row of the table evaluates the predictive power of these models by considering all probabilities greater than or equal to 0.5 to mean that a civil war is expected to occur. Although a threshold of 0.5 is entirely arbitrary, it seems a natural place to start, given the dichotomous nature of the dependent variable. Under these circumstances, we can see that the Fearon & Laitin model fails to predict any of the 107 civil war onsets that actually occurred in their dataset. The Collier & Hoeffler model does only slightly better, predicting three out of the total of 46 civil wars included in their dataset.

In the second row, we lower the threshold to 0.3. This now enables the Fearon & Laitin model to correctly predict one out of the 107 wars, but this comes at the cost of three false positives (i.e. non-events incorrectly labeled as occurrences of war). The Collier & Hoeffler model does somewhat better, predicting 10 out of 46 actual wars, but this too comes at the cost of 20 false positives. In the third row, the threshold is set to 0.1. This leads to an increase in the number of correctly predicted onsets for both models – 15 of 107 for the Fearon & Laitin model and 34 of 46 for the Collier & Hoeffler model – but again this comes at the cost of a much greater number of false positives for both models.

Given the obvious trade-off between the proportion of correctly predicted onsets and the number of false positives at different thresholds, a more robust measure of a model's predictive power would be one that is not sensitive to the user's choice of threshold. This is usually achieved with a Receiver Operator Characteristic (ROC) plot, which illustrates the relationship between the rate of false positives (defined as the number of incorrectly predicted onsets of civil war divided by the total number of cases where civil war did not occur) and the rate of true positives (defined as the number of correctly predicted onsets divided by the total number of cases where civil war did occur) over the entire range of possible thresholds (from 0 to 1).

ROC plots were initially adapted from signal detection theory during the Second World War, but have become a gold-standard for epidemiology, medical research, and the machine-learning community in statistics and computer science. The ROC curve is a type of classifier that compares four groups of predictions: true positives, false positives, false negatives, and true negatives. Each prediction is made by taking the predicted probabilities from a model and establishing a threshold above which the event is predicted to occur, and below which the event is predicted not to occur. Rather than having a large number of 2×2 tables (sometimes called confusion tables) for each possible value of the threshold, the ROC curve shows the ratio of these four components for *all possible values of the threshold* by plotting the true positive rate (typically on the *y*-axis) against the false positive rate (on the *x*-axis).

The concept of the ROC curve can be illustrated using the Fearon & Laitin data in the following way (see Figure 1). When the threshold is set to a value of 0.1, the model incorrectly identifies 66 out of the 6295 cases of peace as cases of civil war, resulting in a false positive rate of 0.01. At the same time, the model correctly identifies 15 of the 107 actual cases of civil war, giving a true positive rate of 0.14. Thus a false positive rate of 0.01 is associated with a true positive rate of 0.14 for the Fearon & Laitin model (represented by point A on the figure). If we then try to increase the true positive rate by lowering the threshold to, say, 0.01, we arrive at a false positive rate of $3326/6295 = 0.53$ and a true positive rate of $90/107 = 0.84$. This is represented by point B on the figure. (The comparable positions of the 0.1 and 0.01 thresholds are also indicated on the Collier & Hoeffler ROC curve.) This demonstrates that an increase in the number of true positives comes at the cost of a vast increase in the number of false positives. By repeating this procedure for every possible threshold within the 0 to 1 range, we can generate a curve that represents the trade-off between true positives and false positives over all possible combinations of the two.

As a general rule, a decrease in the threshold will lead to an increase in both the true positive rate and the false positive rate. However, models with more predictive power will tend to generate true positives at the expense of fewer false positives than those with less predictive power. In an ideal case, a perfectly predictive model will correctly identify all actual cases of civil war and never generate false positives. At the other end of the spectrum, a model with no predictive power whatsoever would, on average, generate one incorrect prediction for every correct prediction at all thresholds. This suggests that the

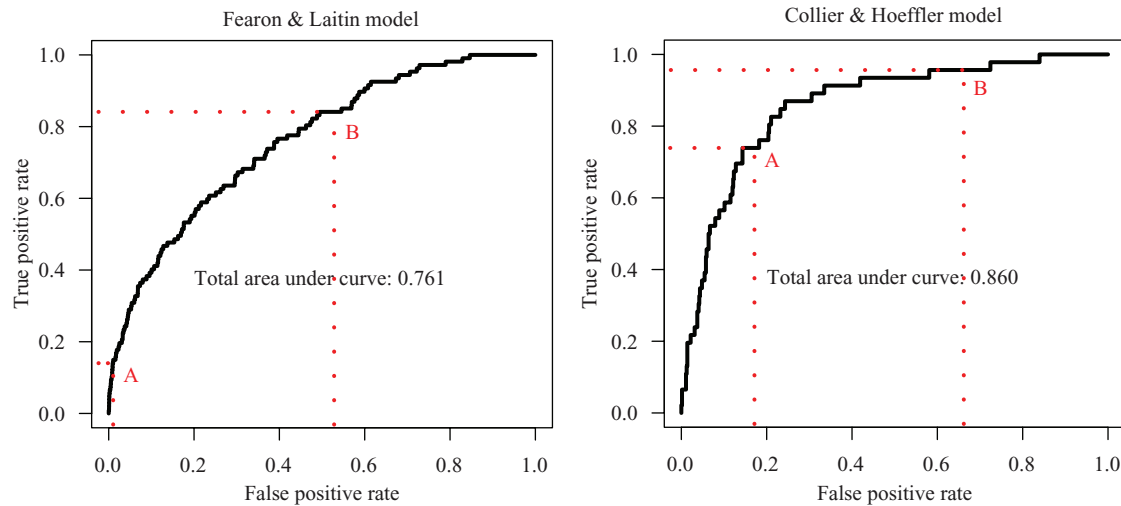


Figure 1. ROC plots

overall predictive power of a model (across the full range of possible thresholds) can be inferred from the size of the area between the x -axis and the ROC curve, which ranges from a minimum of 0.5 (in the case of the entirely non-predictive model that is no better than chance) to 1.0 (in the case of the perfectly predictive model). Better models will have ROC curves that cover more area and will have the lowest false positive rate coupled with the highest true positive rate. The area under the ROC curve is therefore often used to generate a single statistic that summarizes the model's overall predictive power (Fawcett, 2006). Many such statistics are available, including the Gini coefficient, Mann-Whitney U, Somer's d , and others, but frequently the area under the curve is used. This area is the probability that the model will have a higher predicted probability for a randomly chosen positive event than for a randomly chosen non-event. We use this single number summary of the ROC, along with the presentation of the entire curve, recognizing that single number summaries discard information about the entire ROC.

The ROC curve is dimensionless in the sense that it can be used to compare the predictive accuracy of different models which may even use different data. It is not dependent on the base rate of onsets, but only depends on the accuracy of the model. It is entirely possible that a good model for an outcome variable with a 10% base rate of events will be better (the same, or worse) than another model for a different outcome variable that has a much higher base rate. What is essential is how well the model can distinguish the events from the non-events, no matter what their ratio to one another is.

Figure 1 shows that the Fearon & Laitin and Collier & Hoeffler models can be said to have predictive powers (in terms of the area under the ROC curve) of 0.761 and 0.860 units, respectively. Although this difference between the two models is quite substantial, our primary interest here is not in pitting these models against one another in the hope of declaring one to be the better overall predictor, but rather to

evaluate the theoretical contribution of these models in terms of the incremental effect that each of their key independent variables has on the models' predictive power. These two ROC plots therefore represent the baseline for our subsequent analysis.

Statistical significance vs. predictive power

Showing that an individual variable is statistically significant in a given model does not always imply that the variable is associated with a significant improvement in the model's predictive power. In this exercise, we assess the predictive power of each of the key independent variables included in the Fearon & Laitin and Collier & Hoeffler models. We do this by deleting one independent variable from the model at a time, and then measuring the effect that the deletion has on the resultant model's ability to make in-sample predictions. The contribution that each variable makes to the overall predictive power of the model can be assessed by comparing the predictive power of the model without our variable of interest with that of the original model. We measure predictive power in terms of the area under the ROC curve, often denoted 'AUC'.

Figure 2 enables us to make a side-by-side comparison of the statistical significance of each variable (when included in the original model specification) with its contribution to the model's in-sample predictive power. The position of each point along the x -axis represents the variable's statistical significance measured in terms of the absolute value of its z -score. The position of each point along the y -axis represents the marginal contribution the variable makes to the original model's overall predictive power, measured in terms of the difference between the area under the ROC curve calculated for the full model and the corresponding area calculated for a model that lacks that particular variable. For example, excluding the *Democracy* variable from the Fearon & Laitin model causes the

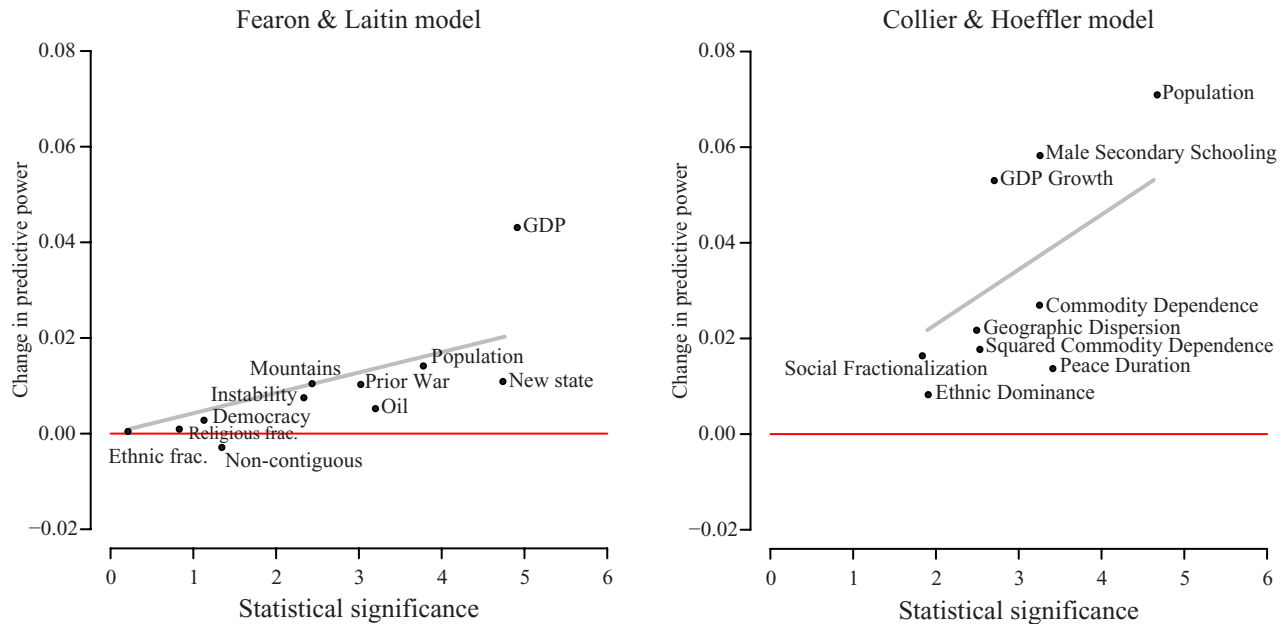


Figure 2. Comparison of predictive power and statistical significance

model's predictive power to fall from a value of 0.761 to 0.759. We can therefore say that, at the margins, the *Democracy* variable makes a contribution of 0.002 units to the overall predictive power of the Fearon & Laitin model.

The grey line represents the relationship between statistical significance and predictive power estimated using OLS regression with no intercept. This draws attention to the fact that the relationship between statistical significance and predictive power is not nearly as close as many people assume: the correlation coefficient is only 0.75 for the Fearon & Laitin model and 0.71 for the Collier & Hoeffler model. Note that each graph also includes a horizontal line plotted at $y = 0$ to indicate the predictive power of the original model specification. Points that lie above this line suggest that the variable of interest makes a positive contribution to the model's overall predictive power, while points that lie below the line make a negative contribution.

When we consider the graph for the Fearon & Laitin model, we can see that one variable in particular, *GDP per capita*, makes a much larger contribution to the model's predictive power than the variable's level of statistical significance would suggest. In other words, excluding *GDP per capita* from the model results in a substantially larger reduction in the model's predictive power than the deletion of any other variable. What is also striking is the fact that the dummy variable for non-contiguous states actually makes a *negative* contribution to the in-sample predictive power of the model. Here we can see that its position on the y -axis lies slightly below the $y = 0$ line, indicating that the model that lacks the *Non-contiguous* variable has better predictive power than the full model.

In the graph for the Collier & Hoeffler model, we again see evidence of the disconnect between the measures of statistical

significance and predictive power. This is particularly noticeable when we compare variables such as *Peace Duration*, *Commodity Dependence*, and *Male Secondary Schooling*. These three variables are all statistically significant at the usual 0.05 threshold and have very similar positions along the x -axis (representing z -scores of 3.40, 3.23, and 3.23, respectively), yet their marginal contributions to the model's predictive power differ widely. For instance, the inclusion of *Male Secondary Schooling* increases the predictive power by 0.058 units, whereas the inclusion of *Peace Duration*, which has slightly higher statistical significance, only increases the model's predictive power by 0.015 units.

In Figure 3, we extend our analysis of the models' predictive power by considering the consequences of deleting more than one variable at a time. The black dot on the far left-hand side of each of the panels represents the predictive power of the original Fearon & Laitin and Collier & Hoeffler models (again expressed in terms of area under the ROC curve). The 'fan' of solid circles to the right of that dot indicate how the in-sample predictive power changes as a result of deleting each one of the variables included in the original model specification. For example, when *GDP per capita* is dropped from the original specification of the Fearon & Laitin model, the predictive power of the model falls from 0.761 to 0.717 (as indicated by the vertical position of the circle labeled 'GDP').

To the right of this point, we can explore the effects of dropping additional variables. The next 'fan' shows the predictive power that results from dropping both *GDP per capita* and one of each of the other remaining variables from the model. For example, the position of the lowest circle in this 'fan' represents the in-sample predictive power of a model that lacks both *GDP per capita* and *Instability*.

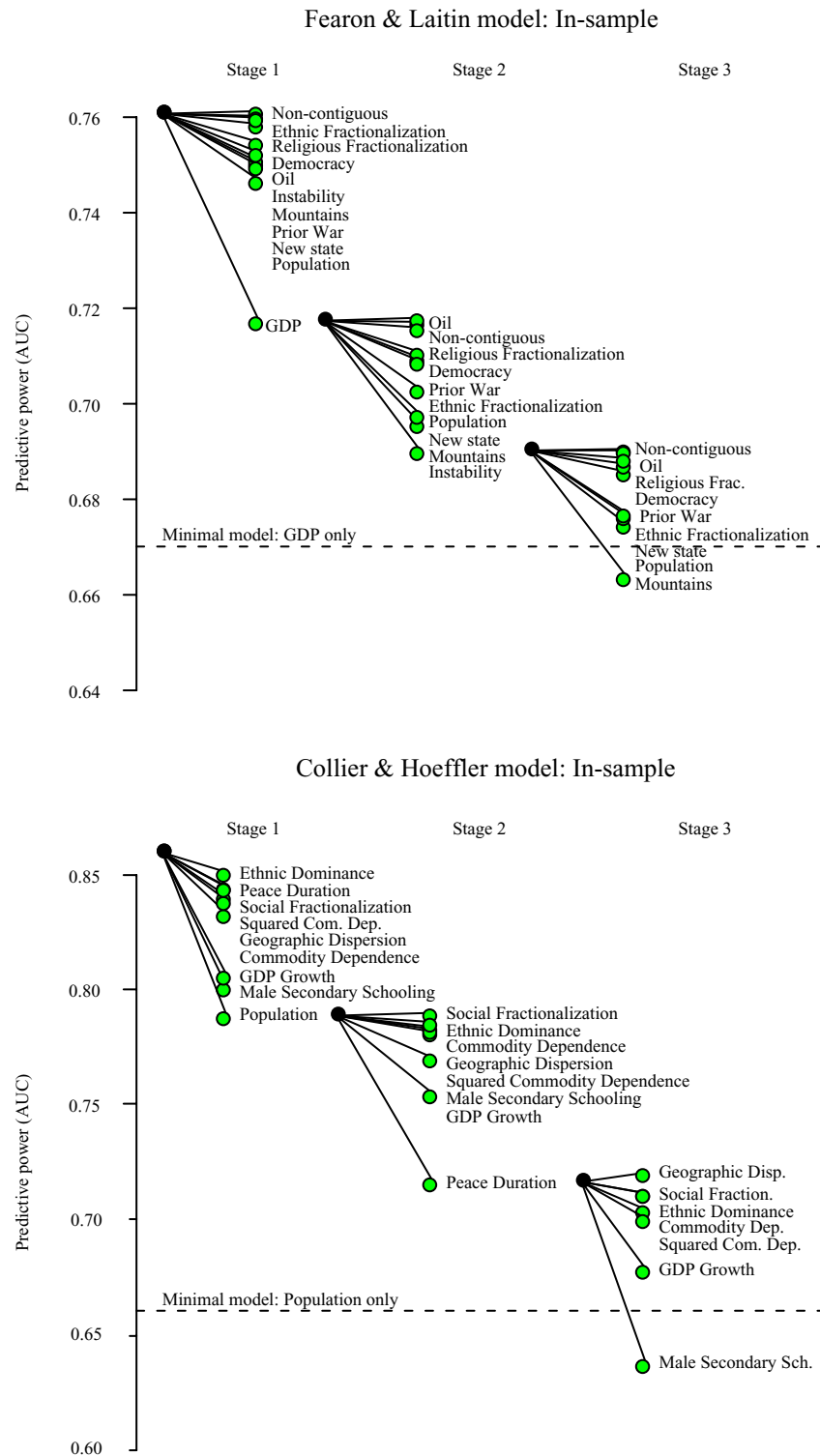


Figure 3. In-sample predictive power

Its predictive power falls from 0.717 when only *GDP per capita* is excluded to 0.690 when *Instability* is also excluded. The effect of a third round of deletions is shown on the far right-hand side of the panel. Here we can see that deleting *Mountainous Terrain* from a model that has

already had both *GDP per capita* and *Instability* excluded results in the AUC decreasing to a value of only 0.664. What is particularly interesting about this finding is that this model, which still retains eight of the original 11 variables that were included in the Fearon & Laitin model,

performs worse in terms of its predictive power than a minimal model that has *GDP per capita* as its sole covariate (represented by the position of the dotted line labeled 'Minimal model'). This decline in performance occurs in spite of the fact that this particular model still includes four of the covariates that were found to be statistically significant in the original model, namely *Prior War* ($z = -2.97$), *Population* ($z = 3.77$), *Oil* ($z = 3.17$), and *New State* ($z = 4.71$).

The results for performing this exercise on the Collier & Hoeffler model are similar. We can again see that in one particular instance, the deletion of only three variables can cause the resultant model's predictive power to fall below that of a minimal model that includes *Population* as its only covariate. Once again this minimal model outperforms the reduced model in spite of the fact that the latter retains a number of variables that were found to be statistically significant in the original model specification (*Commodity Dependence*, *Commodity Dependence Squared*, *Geographic Dispersion*, and *GDP Growth*). What this exercise demonstrates is that when it comes to choosing a model that best predicts the occurrence of civil war, a very parsimonious model can often fare better than one that contains a relatively large number of statistically-significant variables. This finding further emphasizes the need to look beyond statistical significance in assessing the relative importance of each variable.

Cross-validation: Out-of-sample predictive power

A harder test for the predictive power of these models comes from assessing their ability to make out-of-sample predictions. The foregoing discussion of predictive power has referred only to the models' in-sample predictive power – in other words, the ability of these models to correctly predict outcomes within the very same set of data that was used to generate the models in the first place. What is far more difficult to achieve, however, is correctly predicting events in a previously unseen dataset. If the models have succeeded in capturing the underlying relationship between the independent and dependent variables, then the models should continue to perform well when presented with a new set of data. If, however, the models merely provide a detailed description of the relationships that happen to exist in the original dataset without capturing their underlying causal relations (in other words, if the models suffer from overfitting), their ability to make correct predictions in a new dataset will turn out to be much poorer (Beck, King & Zeng, 2000).

The main difficulty with conducting out-of-sample tests arises from the fact that most models of events of interest to conflict scholars have been estimated using all available country-year cases. Unlike in the case of studying, say, individual voting behavior, there simply aren't any more previously unexamined cases of civil war (or peace, for that matter) that can be used for further model testing. Most quantitative models of civil war were estimated using all available data on every case of civil war occurrence or non-occurrence within the modern state system. One possible solution to this problem

is to perform a cross-validation exercise by re-estimating the models using only a subset of the available country-year cases. In this case, a small number of observations can be set aside for an out-of-sample test of the models' predictive abilities. By rotating through the various ways of dividing the country-year observations between the so-called 'training' and 'test' sets, it is possible to arrive at an overall estimation of the out-of-sample predictive power of a given model without having to find new data.⁴

In this section, we use this technique to assess the out-of-sample predictive power of the Fearon & Laitin and Collier & Hoeffler models. By examining the effect that successive deletions of individual variables have on the models' out-of-sample predictive power, we are able to provide a far more stringent test of the contribution that each variable makes to the models' predictive power.

We performed a 4-fold cross-validation exercise in the following way: all of the country-year observations used in each model were randomly divided into four segments. Three of these four segments were pooled together to create a 'training set' that was used to re-estimate the model, a step known as the 'learning step'. The fourth segment (the 'test set') was kept aside for assessing the predictive power of the model estimated using the training set. Thus, the trained model is used to generate predictions that apply to the test set. The out-of-sample predictive power was then measured by calculating the area under the ROC curve for the test set. This procedure (known as a k(4)-fold cross-validation) was repeated four times, such that in each iteration a different combination of three segments was used to estimate the model, while the remaining fourth segment was kept aside for assessing the model's predictive power. Because the results of this process are sensitive to the way in which the database was divided into four segments in the first place, this whole cycle of four-way cross-validation was repeated for ten different random partitions of the database. To summarize, we undertook a 4-fold cross-validation quasi-experiment, which was repeated 10 times. The results reported represent an average of these quasi-experiments.⁵

Figure 4 provides an indication of the contribution that each of the variables in the Fearon & Laitin and Collier & Hoeffler models makes to the overall out-of-sample predictive power of each model. As before, the starting point in each of these diagrams is indicated by the solid black dot at the upper left-hand side, which represents the average

⁴ Geisser (1975) is often credited with the early development and promotion of (k-fold) cross-validation predictive statistics. Efron (1983) remains a classic improvement.

⁵ The question of the optimal choice of k in such experiments arises often. In the machine learning community, for example, 10 seems to be the norm. If k = number of observations, the technique is called leave-one-out cross-validation. There is no evidence that the choice of k makes much difference in most applications, as long as the subsample in each fold is large enough to calculate the usual statistics. Specifically, classification error under the k-fold approach behaves approximately like a test on a sample of size n.

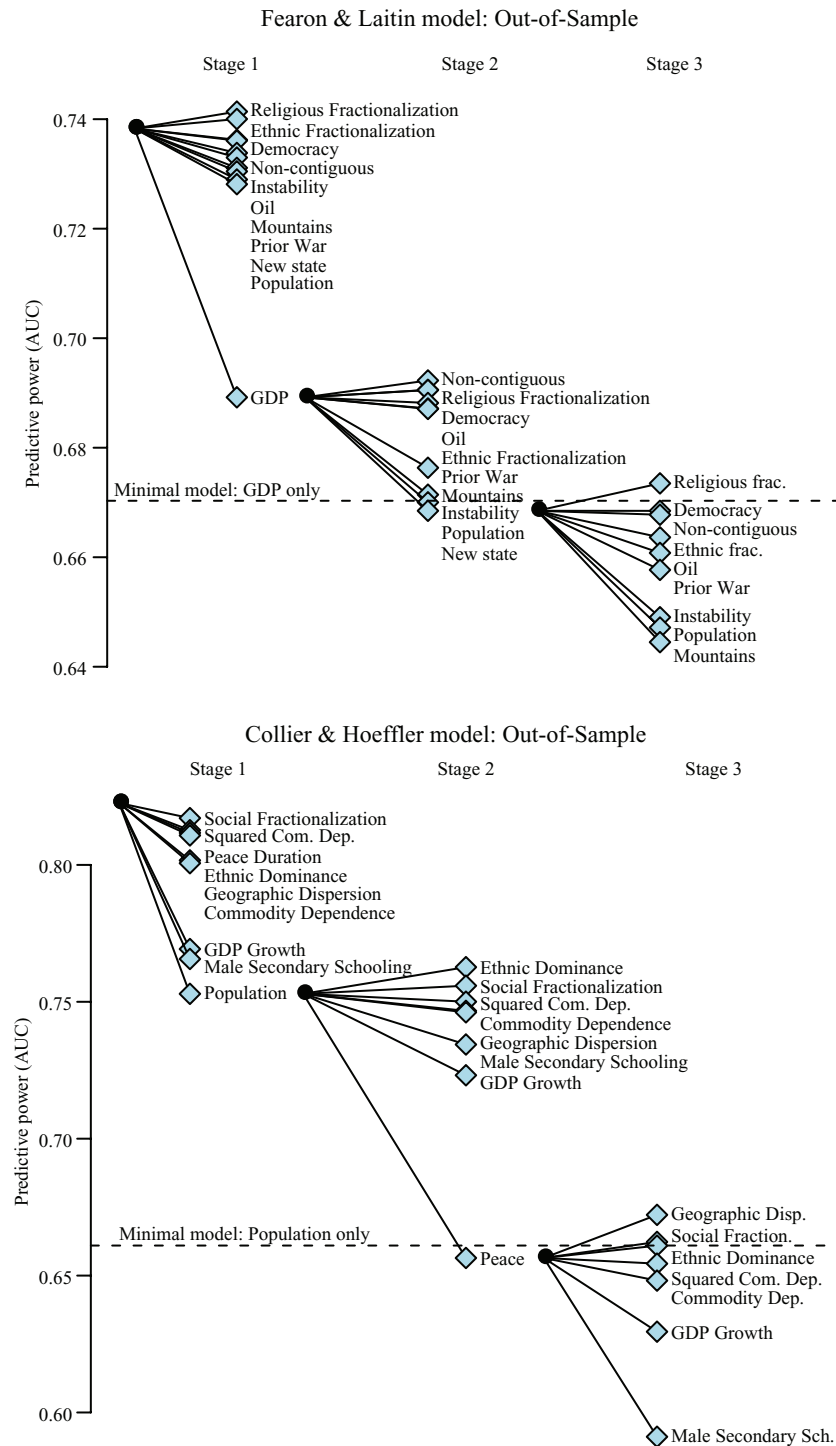


Figure 4. Out-of-sample predictive power

out-of-sample predictive power for the original model specifications (i.e. before any variables have been deleted). As one would expect, these values are less than they were for the in-sample estimates shown in Figure 3. The out-of-sample predictive power for the Fearon & Laitin and Collier & Hoeffler models (again measured in terms of area

under the ROC curve) are 0.738 and 0.823 respectively, which compares to values of 0.761 and 0.860 when all available data are used. In proportional terms, both models show a decrease in predictive power of three to four percent when we compare their out-of-sample predictive power to their in-sample predictive power. While this figure suggests

that the two models on the whole do a reasonably good job of making out-of-sample predictions (and therefore do not suffer excessively from overfitting), what it does not tell us is how much of a contribution each individual variable makes to that overall measure. To assess this contribution we need to again consider the change in predictive power that results from deleting individual variables, as we did in the in-sample exercise above.

In the Fearon & Laitin model, we can again see that the variable that makes by far the greatest contribution to its out-of-sample predictive power is *GDP per capita*. Excluding *GDP per capita* from the original model results in an even greater decrease in the model's predictive power than it did in Figure 3, where we had considered the effect of its exclusion on in-sample predictive power. The figure also shows that in some cases, excluding GDP per capita along with only one other variable results in a model whose out-of-sample predictive power is worse than that of a far more parsimonious model that includes *GDP per capita* as its only covariate. We can also see that the exclusion of certain variables (albeit not the ones that Fearon & Laitin claimed to be most important) can, in several cases, lead to an improvement in the model's predictive power. Once again, this exercise demonstrates that including a variable to improve a model's fit can sometimes have a negative effect on the model's overall predictive power.

The results for the Collier & Hoeffler model shown in Figure 4 are broadly similar. Here we can see examples of model specifications where the exclusion of a variable that was statistically significant in the original model can actually lead to an increase in the model's predictive power. For example, when we compare the predictive power of the model that excludes *Population* to the one that also excludes *Social Fractionalization*, the predictive power of the latter turns out to be higher. As was the case with the Fearon & Laitin model, this figure also illustrates the disproportionately large impact that some variables have on the model's predictive ability. The figure shows that a highly parsimonious model whose only covariate is *Population* performs better in terms of its predictive power than a model that excludes *Population* and *Peace Duration* while retaining all of the others. What is important to note from this analysis is that a relatively uninteresting variable such as *Population* appears to do much of the heavy lifting when it comes to assessing the model's overall predictive power.

Conclusion

Large-n studies of conflict have produced a large number of statistically significant results, but as yet little accurate guidance in terms of anticipating the onset of conflict. There are a number of different methodological issues that could be responsible for this problem. An obvious candidate is overfitting, whereby the results are tuned to the observed data. Overfitting is an important threat to validity in studies undertaken on a population, not a sample, of the data. In such studies, the estimated standard errors of coefficients, which are typically

used to gauge which variables are important, lose much – if not all – of their meaning. Standard errors are estimates of the sampling variability of the results, but if there is no sampling that is undertaken, little is to be gained by confusing estimates of non-existent sampling error with estimates of uncertainty.

However, the results of the cross-validation performed in the previous section suggest that overfitting is not a major cause of the lack of predictive power among the Fearon & Laitin and Collier & Hoeffler models. Instead, it appears that rather than suffering from overfitting, the lack of predictive power is the result of too much emphasis having been placed on finding statistically significant variables, which may be overdetermined. Statistical significance is generally a flawed way to prune variables in regression models based on observational data, even though we recognize that the studies examined here – and many other studies – have used statistical significance as a major tool. The estimated standard errors of coefficients tell us something about the observed fit of the regression to the data, but they do not reflect uncertainty about the 'true parameters'. Indeed, prediction may be a better way to evaluate models that are constructed without having their findings annealed with out-of-sample validation data. The inclusion of statistically significant variables can actually reduce our ability to make correct predictions. This finding seems counter-intuitive and should serve as a heuristic that something is amiss. We also believe an avoidance of prediction as a goal of research can lead to a lot of results but few powerful guides to policy. Predictive validity, even in-sample, is a useful heuristic. Out-of-sample predictive validity is not the only way to understanding, but it is a very useful heuristic along that path.

Hegre & Sambanis (2006) used Leamer's extreme bound analysis to re-examine empirical results for civil war onsets, analyzing the sensitivity of some seven dozen potentially important variables drawn from the leading studies on civil war onset. To accomplish their analysis, they undertook approximately 5,000,000 logistic regressions, drawing on statistical significance to adjudge the basic results. They found that population and GDP per capita appear to have a robust, statistically significant relationships with conflict onset (in in-sample calculations). They note for example that a 1% increase in GDP per capita reduces the odds of civil war onset by about 0.5% (Hegre & Sambanis, 2006: 524). Many other findings are shown by Hegre & Sambanis to be highly dependent on the specification of the model, a point in harmony with our demonstration herein. However, the extreme bounds analysis does not address the inferential problem posed here with a definitive answer but is limited to probabilistic inferences in a classical framework. Moreover, it is based on all the available data. Thus, it is still possible to find combinations that are robust and significant in-sample but do not add to our ability to predict outcomes in new or different cases. That said, a Leamer-type analysis is preferred to just gazing at the significance stars. Model dependency also can confound inferences based on statistical results that are not robust. King & Zeng

(2007) argue that sensitivity analyses and extreme bounds analysis may not be sufficient to guard against extrapolating beyond the empirical reach of a particular model. They offer suggestions about how to avoid drawing conclusions that are based on the degree of model dependency.

What we have tried to show is that the power of a model to generate accurate predictions in data on which it was not initially estimated is not necessarily the same thing as the statistical significance of the model and its subcomponents (see Gill, 1999). Statistically significant variables may actually degrade the predictive accuracy of a model. This point is at once simple and profound. If we base policy on models that are constructed on the basis of pruning that is undertaken with the shears of statistical significance, it is quite possible that we are winnowing our models away from predictive accuracy. This might not really make much difference, except to other research, unless the results are used to guide policy, or discussion of policy.

Another reason for lack of predictive power is model misspecification. Importantly, although several extant studies have simple controls for neighbors, none takes into account the dependencies of the observations. This shortcoming is difficult to remedy, but important. In recent years, scholars such as Gleditsch (2002) and Salehyan & Gleditsch (2006) have demonstrated that understanding intrastate conflicts, including ethnic conflicts, may benefit from taking into account regional and diffusion effects. Similarly, Beissinger (2002) has demonstrated the diffusion effects of nationalist mobilization within one (large) country, the Soviet Union. Raleigh (2005) has shown that instability caused by one country's war may impact the democratic institutions in neighboring countries, which in turn may affect those countries' internal stability. The idea that the neighborhood is important has wide currency, even in popular writings (Diamond, 2005) about the reasons for societal as well as institutional collapse. A growing body of research has suggested that an additional shortcoming of existing studies of civil wars rests with the level of analysis. While most large-*n* studies of intrastate conflicts use country-level data, most conflicts are more local than that (e.g. Varshney, 2002; Kalyvas, 2006). By disaggregating the focus from the country-level to regions or localities, further progress can be made. A good example can be found in the article by Buhaug & Rød (2006), which also uses an out-of-sample heuristic. Yet at the same time, even in more disaggregated studies of civil conflict, it will be important to take into account the potential bias introduced by the nature of our non-experimental and non-sampled data. In so doing, we have to find more clever ways to recognize the interdependency of these cases over space and time, as well as ways to overcome our reliance on statistical significance as a tool to gauge our models.

Schrodt (2002) also argues against the notion that political science should focus on explanation and avoid trying to predict the unpredictable. Indeed, to the extent that empirical research will be especially informative, it will have to embrace

forecasting not only to get essential feedback. To do so may require the restraint of analyzing only some of the available data, and doing so with an eye towards generating not only statistical models that are tractable, but also models that produce valid inferences, even on data that have not (yet) been analyzed. The study of civil conflict has made dramatic advances in recent years, and the studies analyzed in this article are key contributions. There are a few studies that do employ forecasting and prediction heuristics to examine facets of conflicts. Notable among them are O'Brien (2002), Schrodt & Gerner (2000), Gurr & Lichbach (1986), Ward & Gleditsch (2002), and some of the work of the Political Instability Task Force (Goldstone et al., 2005). However, until out-of-sample heuristics – especially including predictions – are part of the normal evaluative tools in conflict research, we are unlikely to make sufficient theoretical progress beyond broad statements that point to GDP per capita and population as major explanatory variables.

Data

Fearon and Laitin To address missing data in the population series, we examined those observations with missing data. Both the Federal Republic of Germany and the German Democratic Republic were missing data for early years in the 1950s. The missing data in part come from the splitting-up of the Third Reich during the occupation of Germany and annexed Austria after the Second World War. Both East and West Germany were deleted prior to 1951, owing to their post-war occupation status. One could make the argument to delete them until about 1955. Estimates of post-1950 East and West German populations are, however, available at www.gesis.org/en/social_monitoring/social_indicators/Data/System/keyindic/population.pdf. For the Federal Republic, 68,377 and 68,879 are the population estimates in millions. Figures for East Germany were given as 18,338, 18,351, 18,328, and 18,178 in millions annually from 1950 through 1953. Both of these sets of numbers include the population of Berlin, East and West. Austria was excluded before 1951, on the same grounds. Japan is a similar case, though the Japanese Federal Statistical office www.stat.go.jp/data/chouki/zukyou/ has a long series on Japanese population figures. We excluded Japan prior to 1954. Since Libya was not independent until 1952, it was not included until 1953. Syria is missing data for 1960; we interpolated between adjacent years and utilize 4.307 million as its 1960 population. Finally, Laos was not independent until 1954; we use the figure of 2.103 million for its 1953 population.

Fearon and Laitin (2003: 81 and note 18) point out that instability is assessed on the basis of a change of at least three points on the polity democracy score in any of the prior three years. Foreign occupations are considered to be transitions, but foreign occupation is treated as missing, which causes the observation to be deleted from the analysis. Where possible, the Polity data were updated from more recent releases and

corrections of the Polity project. Countries in transition were coded as -10, per the suggestions in Fearon and Laitin. Thus, codes of -10 were assigned to the following: GDR (1990), Hungary (1957), Bosnia (1996–99), Uganda (1980), Syria (1959–61), Lebanon (1991–99), China (1945–46), Cambodia (1980–88), Laos (1954), and South Vietnam (1954–55). Ghana (1957–60) was rated -8; Zimbabwe (1965–70): 4; Tunisia (1956–59): -9; Kuwait (1961–63): -8; Kuwait (1991): -10; and India (1947–50): 9. By implication, the following have instability: Ghana (1959–60), Zimbabwe (1967–70), Tunisia (1958–59), Syria (1961), Kuwait (1963), Japan (1952), and India (1949–50).

Many countries are missing values for the *GDP per capita* variable. About one-fifth of these can be easily imputed from one-period lagged values of GDP, which are almost perfectly correlated ($R^2 = .99$). The equation used was $y_t = -0.075 + 1.0048 \cdot y_{t-1}$. This still left about 200 country-years for which there were no data on GDP per capita. After experimenting with multiple imputation based on other available data, no satisfactory set of imputations was available, and these cases were omitted from further analysis. This resulted in a slightly smaller sample of onsets, 106 versus 110 in the original data.

Collier & Hoeffler These data are provided in Stata[®] format at <http://users.ox.ac.uk/~ball0144/g&g.zip>. No special handling was necessary to conduct the analyses reported herein.

Replication data

This analysis was conducted using the statistical programming language R. Full replication data are available at www.prio.no/jpr/datasets and dvn.iq.harvard.edu/dvn/dv/mward.

Acknowledgements

This article was presented at the Conference on Disaggregating the Study of Civil War and Transnational Violence, University of California Institute of Global Conflict and Cooperation, San Diego, CA, 7–8 March 2005. Special thanks are due to Kristian Skrede Gleditsch and Håvard Strand for their guidance on this project. We would like to thank Idean Salehyan, Lars-Erik Cederman, Jay Ulfelder, Kosuke Imai, Susan Woodward, and James Fearon for their comments, suggestions, and encouragement. Thanks too to the reviewers and editors of *JPR*.

Funding

This project was partially supported by a grant from the Methods, Measurement, and Statistics Program at the National Science Foundation, grant number SES0417559, and by a Grant from the Human and Social Dynamics cross-directorate program of the National Science Foundation, HSD-04333927.

References

Bass, Gary J (2006a) Memo to Iraq: Four strategies for averting civil war. *Slate*, 28 March.

- Bass, Gary J (2006b) What really causes civil war? *The New York Times*, 16 August.
- Beck, Nathaniel; Gary King & Langche Zeng (2000) Improving quantitative studies of international conflict: A conjecture. *American Political Science Review* 94(1): 21–36.
- Beissinger, Mark R (2002) *Nationalist Mobilization and the Collapse of the Soviet State*. Cambridge: Cambridge University Press.
- Buhaug, Halvard & Jan Ketil Rod (2006) Local determinants of African civil wars. *Political Geography* 25(3): 315–335.
- Collier, Paul (2007) *The Bottom Billion*. Oxford: Oxford University Press.
- Collier, Paul & Anke Hoeffler (2004) Greed and grievance in civil war. *Oxford Economic Papers* 56(4): 563–595.
- Collier, Paul; V L Elliot, Håvard Hegre, Anke Hoeffler, Marta Reynal-Querol & Nicholas Sambanis (2003) *Breaking the conflict trap: Civil war and development policy. A World Bank policy research report*. Washington, DC: World Bank.
- Collier, Paul; Anke Hoeffler & Måns Söderbom (2008) Post-conflict risks. *Journal of Peace Research* 45(4): 461–478.
- Diamond, Jared (2005) *Collapse: How Societies Choose to Fail or Succeed*. New York: Viking.
- Diamond, Larry (2006) What civil wars look like. *The New Republic*, 13 March.
- Easterly, William (2008) Foreign aid goes military! *New York Review of Books*, 6 December (available at <http://www.nybooks.com/articles/22126>).
- Efron, Bradley (1983) Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* 78(382): 316–331.
- Fawcett, Tom (2006) An introduction to ROC analysis. *Pattern Recognition Letters* 27(8): 861–874.
- Fearon, James D & David D Laitin (2003) Ethnicity, insurgency, and civil war. *American Political Science Review* 97(1): 75–90.
- Gagnon, V P (1994) Ethnic nationalism and international conflict: The case of Serbia. *International Security* 19(3): 130–166.
- Geisser, Seymour (1975) The predictive sample reuse method with applications. *Journal of the American Statistical Association* 70(350): 320–328.
- Gill, Jeff (1999) The insignificance of null hypothesis significance testing. *Political Research Quarterly* 52(3): 647–674.
- Gleditsch, Kristian Skrede (2002) *All International Politics is Local: The Diffusion of Conflict, Integration, and Democratization*. Ann Arbor, MI: University of Michigan Press.
- Goldstone, Jack A; Robert H Bates, Ted Robert Gurr, Michael Lustik, Monty G Marshall, Jay Ulfelder & Mark Woodward (2005) A global forecasting model of political instability. Paper prepared for presentation at the Annual Meeting of the American Political Science Association, Washington, DC, 1–4 September.
- Gurr, Ted Robert (1970) *Why Men Rebel*. Princeton, NJ: Princeton University Press.
- Gurr, Ted Robert (2000) *Peoples versus States*. Washington, DC: United States Institute of Peace Press.
- Gurr, Ted Robert & Mark Irving Lichbach (1986) Forecasting internal conflict: A competitive evaluation of empirical theories. *Comparative Political Studies* 19(1): 3–38.

- Hegre, Håvard & Nicholas Sambanis (2006) Sensitivity analysis of empirical results on civil war onset. *Journal of Peace Research* 50(4): 508–535.
- Hewstone, Miles & Katy Greenland (2000) Intergroup conflict. *International Journal of Psychology* 35(2): 136–144.
- Horowitz, Donald (1985) *Ethnic Groups in Conflict*. Berkeley, CA: University of California Press.
- House of Representatives (2006) Panel II of a Hearing of the Subcommittee on National Security, Emerging Threats, and International Relations of the House Government Reform Committee. 15 September. Federal News Service.
- Kalyvas, Stathis N (2006) *The Logic of Violence in Civil War*. New York: Cambridge University Press.
- Kaplan, Robert D (1993) *Balkan Ghosts: A Journey Through History*. New York: Vintage Books.
- Kaufman, Stuart J (2001) *Modern Hatreds: The Symbolic Politics of Ethnic War*. Ithaca, NY: Cornell University Press.
- King, Gary & Langche Zeng (2007) When can history be our guide? The pitfalls of counterfactual inference. *International Studies Quarterly* 51(1): 183–210.
- Lacina, Bethany & Nils Petter Gleditsch (2005) Monitoring trends in global combat: A new dataset of battle deaths. *European Journal of Population* 21(2–3): 145–166.
- Lake, David A & Donald Rothchild (1996) Containing fear: The origins and management of ethnic conflict. *International Security* 21(2): 41–76.
- Lemann, Nicholas (2001) Letter from Washington: What terrorists want: Is there a better way to defeat Al Qaeda? *New Yorker*, 29 October.
- O'Brien, Sean P (2002) Anticipating the good, the bad, and the ugly: An early warning approach to conflict and instability analysis. *Journal of Conflict Resolution* 46(6): 791–811.
- Petersen, Roger D (2002) *Understanding Ethnic Violence: Fear, Hatred, and Resentment in Twentieth-Century Eastern Europe*. Cambridge: Cambridge University Press.
- Posen, Barry (1993) The security dilemma and ethnic conflict. *Survival* 35(1): 27–47.
- Raleigh, Clionadh (2005) A discussion of the role of neighbors in conflict, development, and democracy. Paper presented at the annual Norwegian Political Science Conference, Hurdalsjøen, Norway, 5–7 January (available at <http://www.statsvitenskap.uio.no/konferanser/nfkis/cr/Raleigh.pdf>).
- Salehyan, Idean & Kristian Skrede Gleditsch (2006) Refugees and the spread of civil war. *International Organization* 60(2): 335–366.
- Schrodt, Philip A (2002) Forecasts and contingencies: From methodology to policy. Paper presented at the Annual Meeting of the American Political Science Association, Boston, MA, 29 August–1 September.
- Schrodt, Philip A & Deborah J Gerner (2000) Cluster-based early warning indicators for political change in the contemporary Levant. *American Political Science Review* 94(4): 803–817.
- Stewart, Frances (2005) Policies towards inequality in post-conflict reconstruction. CRISE Working Paper No. 7, Oxford University.
- Toft, Monica Duffy (2003) *The Geography of Ethnic Violence: Identity, Interests, and the Indivisibility of Territory*. Princeton, NJ: Princeton University Press.
- Varshney, Ashutosh (2002) *Ethnic Conflict and Civic Life*. New Haven, CT: Yale University Press.
- Ward, Michael D & Kristian Skrede Gleditsch (2002) Location, location, location: An MCMC approach to modeling the spatial context of war and peace. *Political Analysis* 10(3): 244–260.

MICHAEL D WARD, b. 1948, PhD in Political Science (Northwestern University, 1977); Professor, Department of Political Science, Duke University (2009–). Current main interest: dynamic analysis of international networks.

BRIAN D GREENHILL, b. 1977, MA in Political Science (University of Washington, Seattle, 2008); PhD Student, Department of Political Science, University of Washington (2005–). Current main interest: quantitative models of human rights practices.

KRISTIN M BAKKE, b. 1977, PhD in Political Science (University of Washington, Seattle, 2007); Lecturer, Department of Political Science, University College London (2009–); Research Associate, the Peace Research Institute, Oslo, Norway (2008–). Current main research interest: separatist conflicts and post-conflict societies.