The trouble with herding cats

Country effects on the accuracy of conflict forecasting

Martin Smidt

Master's thesis in Political Science
Department of Political Science

UNIVERSITY of OSLO

Spring 2015

The trouble with herding cats

Country effects on the accuracy of conflict forecasting

2015

The trouble with herding cats: Country effects on the accuracy of conflict forecasting

Martin smidt

http://www.duo.uio.no

Trykk: Reprosentralen, Universitetet i Oslo

Abstract

Armed conflict theory has in recent years seen an increase in the use of forecasting models. These have brought with them a shift from the use of explanatory power to predictive power when evaluating model performance (Gurr et al., 1999; Goldstone et al., 2010; Hegre et al., 2013). As methods of evaluation change, so must our diagnostic tools. Tests for statistical outliers are common, but so far little has been done to adapt such tests to the use of predictive power. In order to improve our understanding of theory, and ultimately to be able to give better advice to policy makers, it is important to investigate the effects of single countries on our model's forecasts.

In this thesis I present a method of testing for statistical outliers for forecasting models using common measures of predictive power. By applying the method to a forecasting model I attempt to uncover any patterns among the outlying countries that could help further the theoretical understanding of armed conflict occurrence.

I utilize a dynamic forecasting model developed in Hegre et al. (2013) and a cross-sectional time-series dataset containing 162 countries observed between 1950 and 2013. The model is repeated once for every country, each time dropping one of them from the estimation and evaluation process. The results are compiled into evaluation sets, and these are then used to estimate each country's influence on model accuracy. Four measures of predictive power are used to evaluate this: ROC AUC, PR AUC, F-score and Brier score.

I find that effect on coefficients is only partially related to effect on predictive power. By examining the outliers in detail I illustrate differences in how the measures weigh predictions, and how this affects the overall score. I also show how cross-validation using cross-sectional time-series data is problematic and greatly influenced by choice of evaluation period.

Acknowledgements

I must start by thanking my thesis advisor, Håvard Mokleiv Nygård, for his immense support throughout this ordeal. His advice made this thesis possible, and his comments have been of immeasurable help as I have struggled through. I would also like to thank my assistant advisor, Håvard Hegre, for his help and comments. I also wish to express my gratitude to both Håvards for giving me the opportunity to write my thesis at PRIO, and I extend my thanks to all PRIOites for being extremely welcoming and helpful, especially Jonas Nordkvelle for helping me with PRIOsim/Stata/R/++, and the inhabitants of the Learner's Loft for keeping me company over the last months.

I must of course also thank my fellow students on the 9th floor. These last two years have been a great experience, and I have enjoyed sharing it with you all.

Thanks to Linn Hege, Haakon and Vegar for proof reading. They have also contributed through discussions, suggestions and cooking, which has been of great help. Despite him fleeing the country when it was time to write our theses, I want to thank Aasmund for spiritual guidance these last six years. A thanks also to mom and dad for all the proof readings and other forms of support these past years, without which I could not have come this far.

Despite the best efforts of all the above, some of my errors may still remain. These are my responsibility alone.

Word count: 34,335

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The study of armed conflict has in later years shifted its focus from international wars between nations to internal conflicts between governments and rebel groups. Such conflicts have long since become the most numerous, and are arguably a much larger problem in today's world. They have wide ranging impacts for the further development of the countries where they occur (Collier et al., 2003), as well the stability of their neighbors (Salehyan and Gleditsch, 2006; Gleditsch, 2007; Buhaug and Gleditsch, 2008). Internal armed conflicts are also more likely to occur in poorer countries that already suffer from poor standards of living, exacerbating conditions for the population further. Ongoing conflict makes it difficult for local governments and international aid organizations to build the institutions and infrastructure necessary to maintain law and order, and to create higher standards of living. Recognizing this, aid organizations have in recent years shifted from being purely reconstruction-based to taking a preventive stance (Collier and Sambanis, 2005).

In order for prevention to be possible it is necessary to know how and why the conflicts arise. A wide reaching literature has been created that seeks to explain this. By studying historical records of armed conflicts, researchers have over the last decades identified variables that correlate with conflict occurrence (Collier and Hoeffler, 2004; Fearon and Laitin, 2003; Hegre et al., 2001). Others have in turn taken the step from pure empirical analysis to attempting conflict forecasting. Goldstone et al. (2010) build a model that they use to infer which countries are likely to experience political instability. Their aim is to predict incidents of several types of instability two years before they occur, and claim they "have substantially achieved that objective" (Goldstone et al., 2010, p. 204). Hegre et al. (2013) take the predictions further, producing forecasts as far as forty years ahead of their data. Their model predicts the likelihood of conflict, and can also cover transitions between conflict intensity (Hegre et al., 2013, p. 252).

1.1 Motivation, goals and methods

Both Goldstone et al. (2010) and Hegre et al. (2013) build their models on global data, meaning that every nation is taken into account. This means that every country affects their estimates, and therefore their forecasts. Single countries could potentially have great effects on these forecasts, skewing the results and directing our attention in the wrong direction. Testing for the effects of influential outlying units on coefficients and measurements of model explanatory power is common practice. Few attempts have been made so far to adapt such tests to predictive power, and applying them to forecasting models. As Ward et al. (2010) have shown, statistically significant variables do not necessarily add any predictive power, which makes it highly likely that tests of influence on predictive power will return different countries compared to those of explanatory power. Little is known about the degree to which a single outlying instance of conflict can disturb the estimation and forecast processes. As scientists are resorting to comparing their models using measures of predictive power, it important to know what fluctuations can be expected to arise from dropping units.

The aim of this thesis is threefold: first to examine to what degree single countries affect our predictions. In doing this I will identify those countries that affect estimations the most. If models are evaluated based on their predictive power, it is important to understand how outliers affect a model's performance by such measures. The second aim is to identify common features among these countries in order to uncover important factors that could affect how the models are specified. This could bring to light new variables, or new ways of approaching existing variables. If there are systematic errors, these could be taken into account to improve our models and forecasts. The third aim is to examine how the results vary depending on what measure of predictive power is used. As many measurements are currently in use, variations in how these respond to units could have implications for the conclusions drawn by the researchers using them. In short the aim is to improve forecasting models, to improve the understanding of their output, and ultimately to make scientists better equipped to advise and assist policy makers.

My research questions are as follows:
- Which countries are outliers by effect on predictive power, and are these the same as outliers by conventional standards?
- Do these divergent conflicts have a common denominator?
- How do different measures of predictive power differ in their reactions to the dropping of countries?

To accomplish these goals I will use as a starting point an unpublished forecasting

model that is a further development of the forecasting theory and techniques developed in Hegre et al. (2013). The model uses a combination of existing theories to build a model that predicts the onset, incidence and termination of armed conflicts. Like Hegre et al. (2013) the model is also used to simulate a forecast of future conflicts based on projections of relevant predictors. Using a method similar to jackknife resampling I intend to identify those conflicts and countries that do not adhere to this model. By examining the effect of single countries on the precision of conflict probability estimates, I will uncover the countries that have the greatest effect on the model's predictive power. These countries are then subjected to closer scrutiny in order to ascertain whether they are linked by common traits or not, and to look more closely at how they affect the predictive power.

I do this by running the model estimation once for every country in the dataset, dropping one country with each iteration. In this way I emulate more conventional test of outliers, such as tests of unit influence on β-coefficients. By comparing the results of a control model with the country-drop iterations I can calculate the effect each country has on the predictive power of the model. I then extract those countries that appear as having either a very strong negative or positive effect on the model's predictive power. Two measurements are used, and the results are compared, both in the values of the outlying countries on the predictors and their individual predicted probabilities. To see whether the countries have indirect effects I examine their effect on coefficients and on their neighborhoods.

1.2 Thesis structure

In Chapter 2 I will present the theoretical background of the thesis. The first part of the chapter is devoted to the armed conflict literature. Here I examine the different aspects of armed conflict and present how it will be defined in this thesis. I then present the major findings in the literature over the last decades. Theoretical insights gained and variables found to be reliably correlated with conflict are presented. I then argue for the further use of out-of-sample cross-validation, and in its extension the use of forecasting to provide testable predictions for unseen data. The second part of the chapter provides the statistical theory to support the use of such forecasts. I provide a summary of criticism of relying purely on significance based analysis, and present alternative methods of evaluation. In the final section of the chapter I provide the theoretical basis for my research design by drawing parallels to existing tests of unit influence.

Chapter 3 presents my dataset and the variables I will be using. By combining a number of data sources, as well as imputations where necessary, the utilized dataset has complete information on the relevant variables for 162 countries from 1950 to 2013. I then present

the simulator design, which is an advanced version of that used in Hegre et al. (2013). Lastly, I show how the data extracted from the simulator is evaluated using a number of R-packages to calculate measures of predictive power.

Chapters 4 and 5 show the results and analysis of the simulations. In chapter 4 I first present the outcome of a more conventional test of unit influence involving unit effects on β-coefficients. I proceed to present individual unit effects on the model's predictive power using Receiver Operator Characteristics (ROC) and Precision-Recall (PR) curves, as well as their respective Area Under Curve-measures (AUC). I then group the countries with the most extreme effects on PR AUC, naming these my predictive outliers. These outliers are split in two groups depending on whether their effect on predictive is positive or negative. These results are then compared to those of the test of effect on β-coefficients in order to establish whether the tests return the same outliers.

In order to uncover whether the members of the predictive outlier groups have any shared attributes that could be the cause of their deviance, the average values on important predictors are compared between the outliers and the remaning countries. This could uncover flaws in the model's specification, which can be taken into account in future research.

In order to explain how the countries affect the predictive power I examine their predicted conflict probabilities from a control model containing all countries. These probabilities are compared to their conflict history, and any deviance between predicted and observed values will determine the direct effect a country has on predictive power. To uncover indirect effects I also examine countries' effects on coefficients as well as their effect on their neighbors through neighborhood variables.

Chapter 5 is structured almost identically to Chapter 4. It differs in that it does not contain the test of unit influence on coefficient effects, and it includes a robustness test for an erroneous lag in the model. While Chapter 4 has its own summary, the added findings from Chapter 5 are discussed in Chapter 6. In the final chapter I summarize the findings and discuss their implications. I also discuss weaknesses in the design and recommend alterations that would address these in future research.

1.3 Thesis findings

In summary I find that there are great variations in the effect that countries have on model accuracy, showing that some countries do have greater impacts than others. How extreme the effects are vary depending on the measure and evaluation period used, but there are clear outliers regardless. An important point is that tests of unit effect on β-coefficients do not return the same units as tests of effect on model predictive power. As forecasters

evaluate models using predictive power it is important to also examine how such statistics can change by the presence, or absence, of single units in the dataset.

The countries with the most detrimental effect on predictive power have conflict histories that follow two main patterns. The first pattern is a shift from a largely peaceful period to one consisting mainly of conflict, or a similar shift from conflict to peace, that occurs near the split in data between estimation and evaluation sets. This shows that choosing where to split data for cross-validation has implications for how the model performs in evaluation. The second pattern is a series of transitions in rapid succession between conflict and peace. This creates data that it is impossible for a statistical model to predict with reasonable accuracy. I find that the conflict definition is largely to blame for the data, and recommend that the conflict definition based on a strict battle death threshold should be modified.

I also find differences in how the measures of predictive power react to country drops. The two measures have their advantages and disadvantages depending on what is more important to the researcher. I find that the PR AUC is better at assessing the overall model performance. The Brier score is however better at returning countries with individually poor predictions, and its results are easier to decompose making it easier to establish exactly how countries affect the predictive power.

Chapter 2

Background and theory

Having established the context and goal of this thesis, I will now describe my starting point in greater detail. I will provide a summary of the study of armed conflict, describing its evolution over the last decades. I summarize what can be described as the commonly accepted findings in the field, and place my own work as a continuation of existing work. The theoretical framework is built around key variables like wealth, population, regime type and conflict history. Neighborhood variables are also an important component, with spill-over effects playing a key role in modeling.

The second part of the chapter is the statistical basis behind my research design. First, I describe the significance based approach that has become the norm for quantitative political science. I then present criticisms of this approach, especially its application on the study of armed conflict. Following this comes a review of alternative solutions to the tasks performed by significance testing, such as measures of predictive power and cross-validation methods. Lastly I will discuss influential units, as this is central to the research design.

2.1 Armed Conflict Research

The literature on armed conflict is diverse and covers many topics, and only the most relevant parts will be covered in this section. I will first define the unit of study, as there are discrepancies between studies as to what constitutes an armed conflict. I then summarize some important theoretical contributions and the most widely used predictors. I conclude by introducing conflict forecasting and argue for its further use.

2.1.1 Defining Armed Conflict

The conflicts of interest for my purposes are only those that occur between a state and one or more non-state actors (intrastate conflicts). These civil conflicts have become more

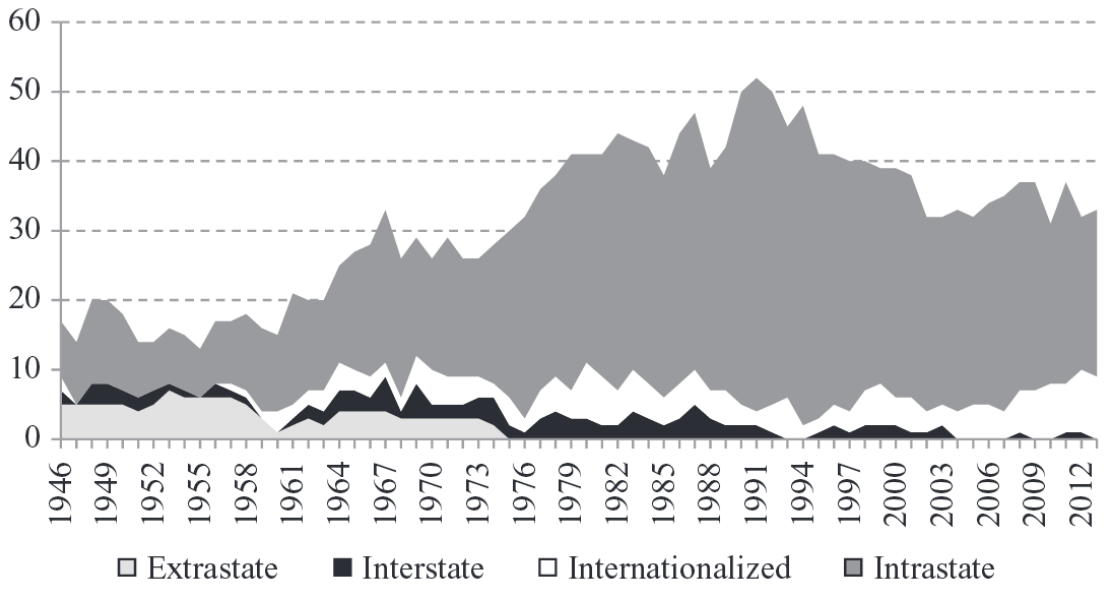


Figure 2.1: Number of armed conflicts by type, 1946-2013 (Themnér and Wallensteen, 2014, p. 544).

prevalent since the end of the Cold War, while the number of interstate conflicts has been declining (Themnér and Wallensteen, 2014). As seen in Figure 2.1, the trend since the end of the Cold War has been an overall reduction in conflict numbers. Interstate conflicts have almost vanished, while internationalized conflicts have increased.

There are a number of definitions of civil conflict in the current literature, with variations between researchers and projects (Sambanis, 2004). I follow the definition used by Gleditsch et al. (2002), which is also used in Hegre et al. (2013). This definition states that "Internal armed conflict occurs between the government of a state and internal opposition groups without intervention from other states." (Gleditsch et al., 2002, p. 619). A distinction between minor and major conflicts is used, where a conflict that causes more than 1000 casualties per year is labeled a war, while those between 25 and 1000 are labeled minor conflicts (Gleditsch et al., 2002, p. 619).This distinction between high and low intensity conflict can be useful, as some variables have been shown to correlate with only one of two conflict levels (Hegre and Sambanis, 2006).

Further distinctions can be made in research between finding the correlates of conflict onset, incidence, termination, duration and severity. The most used are onset and incidence, of which the latter will be used in this thesis. Most datasets are in the country-year format, with each country observed once per year. Studying conflict incidence includes all conflict years, regardless of whether it is the first or last year of conflict. Incidence studies are directed at the basic, underlying factors that determine whether a country is conflict prone.

Conflict onsets are the first country-year units after a peace year, with any following

years of conflict not included as positive outcomes. This approach seeks to understand more specifically what leads to the outbreak of war, rather than just the underlying factors that make conflict more likely. Such studies need to include variables that have the potential to change rapidly, and with a distinct effect on the political climate. An example of this is the use of economic growth rather than just GDP, as the variable can change dramatically from year to year. Depending on the theoretical approach, a drop in national income can be either the last blow a weakened state can take before losing control, or the drop in expected income that drives people to rebel rather than work.

The study of incidence and onset can be seen as studying respectively *where* and *when* conflicts occur. Incidence is focused on revealing the slow moving factors that create an environment where conflict becomes possible. This will tell us *where* conflicts are likely to occur. Onset will also tell us this, but here the focus is more on the changes that occur just before conflicts that act as triggers, telling us *when* conflicts are likely to occur.

Studying conflict termination focuses on the other end of the conflict, attempting to find the correlates of peace. Changes that occur directly before the end of conflict may be interpreted as necessary preconditions for a stable peace agreement. Studying both onset and termination, in other words the duration of a conflict, gives us insight into factors that prolong conflicts once started. Some of the variables I will put forward in the next section affect both conflict incidence and duration. This includes geographic and economic variables, but not all of these have the same effect on duration as they do on incidence. While the possibility of recruiting a well trained and well armed group increases the chance of rebellion, rebel force strength has been found to shorten the duration once a conflict is initiated (Buhaug et al., 2009, p. 561). Factors such as these are important when simulating incidence forecasts too. Forecasts have to predict both onsets, terminations and renewed conflict. It is therefore important to look to studies of not just incidence, but also onset and duration for guidance when building forecasting models.

Lastly the severity of a conflict is also the subject of many studies. Severity is often measured in number of deaths. What qualifies as a relevant death can vary, but the dataset used in this thesis utilizes a battle death definition where only casualties in armed fighting between a government and a rebel force are counted. Civilian casualties that are a direct result of fighting are also included, but indirect deaths from starvation or lack of basic services are not. One sided violence, where a state or group assault an unarmed party, are also dismissed. Severity can also be measured using different casualty definitions, such as the victims of one sided violence, or violence between rebel groups without the involvement of government forces. Another approach is indexes that combine several factors, such as weaponry used in the conflict, destruction of property and more (Pfetsch, 2015). As with duration, what determines severity needs not be the same as incidence or onset.

The different studies can be used in conjunction to give more detailed pictures of what is at risk. Incidence risk can tell us where we are most likely to see conflicts occur in the long term. Onset studies can add to this by telling us what short term factors create the most risk at any given time. Duration and severity studies can then tell us how long and severe these potential conflicts are likely to be, giving us the possibility to estimate which will be the most costly, both economically and in terms of the number of lives lost. In this way conflict research can give crucial policy advice on where to implement counter measures, such as aid programs focused on either food, education or governmental assistance.

In this thesis I will only examine conflict incidence, but I will be using a conflict variable which divides conflict into two categories by their severity. The model is still aimed at explaining incidence rather than severity; it simply seeks to explain the incidence of two conflict categories that happen to be divided by severity.

2.1.2 Correlates of war

Much of the recent quantitative studies of armed conflict have focused on the motives and opportunities for rebellion. Motives are the driving forces that push groups into rebellion, while opportunities are factors that make such a rebellion a feasible option (Fearon and Laitin, 2003; Collier and Hoeffler, 2004). Traditionally, political science has focused on grievances as the main motive and driving force behind civil conflicts. Ethnic and religious tensions, as well as economic inequalities, have been seen as the main culprits (Gurr, 1970, 1993, 2000). These factors have faced considerable scrutiny, and studies have cast doubt on their relevance (Fearon and Laitin, 2003; Collier and Hoeffler, 2004). Further research has shown that while individual economic inequalities may not be robust, horizontal inequalities between ethnic groups do lead to an increased risk of conflict (Cederman et al., 2011). Both richer and poorer ethnic groups are more involved in conflict than groups with wealth on par with the national average. Correlations have also been shown between conflict and political exclusion along ethnic lines. Discrimination against certain ethnic groups is linked with greater risk of separatist rebellions (Cederman et al., 2010).

Some scholars criticizing the grievance based approach shift the focus from ethnic grievances between groups to personal economic gains. Collier and Hoeffler (1998, 2004) hypothesize that groups are more likely to rebel if they expect to profit from such action. They find that both lower *GDP per capita* and access to *natural resources* lead to a greater risk of war, although the effect of resources is the opposite in exceedingly wealthy countries. Their interpretation is that poverty increases the risk of rebellion, as less is at stake and more is to gain from taking up arms. It also means recruiters can offer lower wages, as competing modes of income have less to offer. Natural resources are seen as a commodity that can easily be looted by rebel groups, and this is believed to increase the

risk of conflict through economic incentives for rebellion. Collier and Hoeffler (2004) find support for this hypothesis using primary commodities export as a proxy for access to such resources. While the opportunity approach is also compatible with grievances, Collier and Hoeffler (2004) find that economic factors have more explanatory power.

Regime types have also been found to be linked with the risk of conflict, with the theory being that different types have differing degrees of control over their territories and populations. Hegre et al. (2001) find that coherent democracies and authoritarian states are much less prone to conflict than intermediate regimes. Changes in regime type, in either a more or less democratic direction, are found to be associated with conflicts. Their findings are supported by Fearon and Laitin (2003), who find that anocracies are more prone to conflict. Vreeland (2008) disputes their findings, pointing out problems with the measure of democracy used. The PolityIV index includes a measure of political instability. This means an anocratic score can be the result of, not the cause of, political violence. Vreeland does however note that a change of regime type remains significantly correlated with conflict even when the potentially self-fulfilling element is removed from the index (Vreeland, 2008, p. 403). By estimating regime survival times, Gates et al. (2006) find that the anocratic regimes are the least durable. Authoritarian regimes have established a repressive power base, and democracies have strong institutions that enforce laws and regulations. The intermediate regimes, on the other hand, lack both the repressive power and the institutions, and thus also lack the ability to uphold a monopoly on violence. Their results were tested with Przeworski (2000)'s measure of regime type, and found to be robust. Goldstone et al. (2010) also decompose the Polity data set to create their own measure of democracy, with which they find that pure democracies and authoritarian states are less at risk than partial regimes.

Geographic variables are a major part of the field, featured in most major studies, as well as being important in the historical study of conflict (Buhaug and Gates, 2002). Natural resources, either in abundance or in scarcity, is one of the subcategories. While there are differing opinions, some consensus is appearing in the literature as to the effect of such resources on conflict. Onshore *oil* is found by some to have a positive effect on the risk of conflict, and on the duration if located within the conflict zone (Lujala, 2010). The effect of *diamonds* is less clear, but a correlation is found both with incidence and with certain types of conflict onset (Buhaug and Rød, 2006). While natural resources are seen as a more or less robust variable, there are many caveats and complex interactions, including with wealth as previously mentioned. Corruption has been found to dampen the effect of resources, as profits can be made without resorting to violence. Also, the impact seems to be U-shaped, meaning that the effect dissipates with extreme levels of abundance.

The negative effect of natural resources at higher levels of wealth is interpreted as the

effect of increased security that a richer state can afford, and that this effect is simply amplified by natural resources. Fearon and Laitin (2003) argue that the opportunities approach should be focused on these state capacities rather than on the individual's motive to rebel. Rather than interpreting GDP as a measure of potentially lost income, they believe it should be seen as a proxy of state capacity to uphold basic services as well as law and order. A poorer country will have less ability to keep its citizens pleased, and more crucially it will not have the law enforcement capabilities to prevent rebellions, nor the military capacities to fight them. While GDP per capita and governance are often highly correlated, closer study has shown that government capacity, rather than cheap labor, is the cause of increased risk (Fearon, 2011, p. 4).

Another factor found to be correlated with conflict onset is a country's proportion of mountainous terrain (Fearon and Laitin, 2003, p. 85). Inaccessible areas give rebels areas of operation that are out reach of government forces. Government reach is also affected by their capabilities, with richer countries being better equipped to go after groups seeking refuge in this terrain. Studies using disaggregated data have not found the same correlations as country level studies, but they also point out that this is not necessarily contradictive, as rebels may use such areas as bases while fighting occurs elsewhere (Buhaug and Rød, 2006, p. 327).

Apart from local terrain and resources, geographical factors also include distances. Buhaug (2010) finds that conflicts tend to occur further from the capital in weaker states. There is also some support for distances to borders playing a role in the occurence (Buhaug and Rød, 2006, p. 325) of conflict, with stronger support for an effect on conflict duration (Buhaug et al., 2009). The theory is that rebels take advantage of porous borders to evade government forces. By operating in border areas they can slip away from pursuers by crossing into neighboring countries where government forces cannot follow. This border activity is related to a further geographic aspect of conflict: neighborhood spillover effects. As groups operate on both sides of borders, the risk of conflict spreading to the neighbor increase. Sambanis (2001, p. 268) finds that a country is more prone to conflict if it has *neighbors that are experiencing conflict*, or if it is in a *neighborhood that is conflict prone*.

Neighborhood factors are also relevant for other variables than just conflict. Hegre et al. (2013) include the neighborhood average of a number of variables, including male secondary education, infant mortality rate and youth bulges. Such variables represent the potential of conflict spreading across borders, although there are differing theoretical approaches to the causal mechanisms. For the conflict variable, Salehyan and Gleditsch (2006) argue that conflict diffusion is due to the movement of refugees from neighboring conflicts. While the refugees do not necessarily fight, they can bring with them arms and ideology. They also affect their new location by changing the ethnic make up and economic situation,

possibly creating ethnic tension and food shortages. Other linkages include the increased availability of arms in the region, making it easier to equip rebel groups. Alternative causes may be more direct, such as neighboring states intervening in local disputes, or provide support for rebel groups (Gleditsch, 2007). Such contagion is most likely to occur where there are transnational ethnic ties, and where there are secessionist struggles (Buhaug and Gleditsch, 2008). There are also indirect effects through decrease in trade. As conflict is found to harm the economy, it will also harm the flow of trade. If the economy is weakened, there will be less flow of trade, both due to lack of supply and to obstruction of lines of communication. A great deal of trade happens between neighbors, and local conflicts will therefore have a detrimental effect on the economy of any neighbors as well as the country experiencing the conflict. Murdoch and Sandler (2002, 2004) find that these economic effects are the cause of increased neighborhood risk.

Neighboring effects can be coded several ways. A neighbor can be defined as a country that shares borders, or that is within a given distance. Conflict variables are often coded as dummies where a positive value is given when one of the neighboring countries has experienced conflict, often lagged by one year. The geographically larger regional variables are also intended to pick up on many of the same effects as direct neighbors, but they are also interpreted to include effects of "ethnic makeup, resource endowments, and geography" that are not given neighborhood variables (Sambanis, 2001, p. 268). These variables also include the effects of a regions collective level of wealth, development and other variables. The criteria for choosing regional borders vary, but examples are UN standards, cultural regions or entire continents.

Unsurprisingly, conflicts have been shown to be contagious not only in space, but also over time. While an unstable history should be removed from the measure of democracy, it makes sense to include such a measure of instability in our analysis on its own. Conflict history can be an indication of both instability and an increased opportunity to rebel. Collier and Hoeffler (2004) use *time since last conflict* as a proxy for easier access to weaponry. The more recent the conflict, the more guns are in circulation locally. Another interpretation is that not only are guns available, but so is the manpower to use them. Previous conflicts would result in a supply of trained veterans who would make it easier to recruit a group capable of waging a war. It may also take time to demobilize rebel groups after ceasefires or peace agreements, making it easier to restart conflicts. The detrimental effects that conflicts have on the economy, health and other factors also increase the risk of war. This creates a circle of violence, known in the literature as the conflict trap, where ongoing conflict creates an environment more and more prone to further conflict (Collier et al., 2003). Empirical support for the existence this effect has been found by, among others, Hegre et al. (2013). They include dummy variables for conflict state the previous

year, and variables for the time spent in the current state. They find that the longer a country is in a state, the less likely it is to change.

The aforementioned suspected correlates of conflict have been studied intensively, but the results are not always convergent. Different studies often come to opposing conclusions, and there are a number of reasons as to why this is. Various datasets record different conflict data, creating insecurities as to whether or not conflicts have occurred. As various studies apply different conflict definitions to recorded data, another layer of uncertainty is added. Further, the operationalizations of variables can also vary between different studies. The result is a myriad of studies with varying support for different hypotheses and variables.

To test the robustness of the most common variables, Hegre and Sambanis (2006) apply a global sensitivity analysis. Their study includes over 4 million regressions that test different model specifications and variable operationalizations on two conflict datasets. A number of variables are found to be very robust, with some being only partially stable. Population and per capita income are confirmed as robust, and the two variables are perhaps both the most used, and the most consistently significant variables found in the literature. Unfortunately, their robustness sheds little new light on the causal mechanism involved. Inconsistent democratic institutions are also found to be robust, supporting Hegre et al. (2001)'s U-curve hypothesis. Rough terrain and weak militaries are found robust, supporting the theory that rebellions are more likely to occur where rebels can evade numerically superior government forces or where the government is incapable of restricting rebel activities. This lends credibility to Fearon (2011)'s theory that state capacity is crucial. Neighborhood effects are also among those found consistently significant, along with regional dummies for undemocratic areas. This supports the theory that geographical clustering of conflicts is caused both by bordering conflict areas as well as by regional attributes. Some variables, such as oil exports, are found to be robust only for lower levels of conflict, but not for a more severe definition of civil war (Hegre and Sambanis, 2006, p. 531-533). I will be using many of the variables found robust by Hegre and Sambanis (2006) in my model, along with interactions between them. Not all the aspects discussed in this chapter will be included, as I also attempt to be parsimonious. For further review of the theory see Blattman and Miguel (2010).

2.1.3 Forecasting conflict

> *Only to the extent that we are able to explain empirical facts can we attain the major objective of scientific research, namely not merely to record the phenomena of our experience, but to learn from them, by basing upon them theoretical generalizations which enable us to anticipate new occurrences and to control, at least to some extent, the changes in our environment (Hempel and Oppenheim, 1948, p. 138).*

Quantitative armed conflict research is a social science, but as social sciences go it is certainly one of the more positivistic. I use the term positivistic in the sense that it is focused on emulating the natural sciences and their law based form. Quantitative armed conflict research aims to find laws that govern the nature of human society, more specifically those aspects that lead to its breakdown. By statistically analyzing correlations we seek to uncover the underlying causal relationships that lead to the occurrence of armed conflict. The implicit goal is to understand conflict as a phenomenon, not to describe historical events. An important aspect of understanding is to test the theories to see how our understanding matches reality. Natural sciences rely on experiments for such tests. A theory can gain support or be weakened depending on how well its predictions conform with observations. The norm in conflict literature is to test theories by seeing how well statistical models fit recorded data. This is different in that data is only examined and an explanation is made to match it. While examining data is crucial in gaining any knowledge, relying on it completely as a means of validating our broader theories has been criticized. A vocal critic of established norms, argues that "explanation in the absence of prediction is not scientifically superior to predictive analysis, it isn't scientific at all!"(Schrodt, 2014, p. 290). This is perhaps an aggressive statement, but it echoes the point made by Hempel. What sets social scientists apart from meticulous historians is the creation of theory and generalization. If the theories actually explain typical behavior they should be able to predict it.

A way of incorporating this aspect into research is cross-validation using existing data, a process where models are estimated on parts of data and tested using the remaining units that are then "new" as far as the model is concerned. Out of sample evaluation is a very good way of counteracting overfitting. Overfitting is when a model is specified with so many variables that added explanatory power comes at the expense of increased multicollinearity. This issue will be discussed further in Chapter 2.2. While out of sample cross-validation goes some way towards addressing the issue, it is not perfect. The scientists developing the model have still seen the data, and even if they do not estimate on the whole set they will be aware of where conflicts have occurred. Such knowledge may influence their research and give an unrealistic advantage compared to attempting to predict unknown future conflicts. The advantage of forecasting ahead in time is that it allows for truly independent predictions that can then be evaluated.

There are also practical arguments for forecasting. If our models can provide risk assessments, efforts can be directed towards those countries that are most at risk of experiencing conflict. We can also get an insight into which factors are creating the risk so that these problems can addressed directly. Using statistical models for such purposes is not a complete novelty. Both Collier and Hoeffler (2004) and Fearon and Laitin (2003) estimate risks

for hypothetical countries, and the latter study gives advice regarding how policy makers can reduce conflict risk. If such advice can be based in the models, then it is a small step to simply calculate the estimated risks for real countries, creating short term forecasts.

A pioneering study into the feasibility of creating such forecasts was done by the State Failure Task Force (Gurr et al., 1999). Their dataset only covers the 1980-1992 period, meaning they have little data for both estimation and out-of-sample evaluation. The data also suffers from limited data on variables of interest, resulting in their conclusion that more data is needed before advancements can be made. They also make the important note that not all variables that have been found statistically significant add to the predictive power of their model (Gurr et al., 1999, p. 66). This provides support for the claim that lack of out of-sample-evaluation leaves previous research vulnerable to overfitting.

A following study by O'Brien (2002) has a larger data set, covering 1975 to 1999. The results are an improvement on the accuracy of the forecasts, yet he still deems the project exploratory. Like Gurr et al. (1999), O'Brien (2002) comments that more data is required for forecasting to become feasible. He also comments that while his forecasts do well in anticipating "the oiliness of the rags", the underlying risk of conflict a country is at, they need to include factors that could act as sparks (O'Brien, 2002, p. 807). Such a spark may have been found by Goldstone et al. (2010). Goldstone and colleagues classify countries by a regime type variable of their own construction, and they find that changes in this variable predicts instability. Their model is parsimonious, using only regime types, the infant mortality rate and a binary variable for both neighboring conflicts and state-led discrimination. In a comparison with Fearon and Laitin (2003)'s model they find their own to be considerably more accurate despite its simplicity (Goldstone et al., 2010, p. 204).

Moving beyond predicting only a few years ahead, Hegre et al. (2013) include projections of several key variables to forecast over several decades. Using a dynamic multinomial logit model they simulate several scenarios, based on different projections of the independent variables. Like previous forecasts they use historical data for model selection and evaluation. They test several combinations of baseline variables before arriving at a parsimonious base. Using a pool of previously statistically proven variables, they then test a multitude of expanded models. The end result is a variation of models with varying combinations of variables and interactions (Hegre et al., 2013, p. 256-257). These few top performing models include different combinations of main variable groups, focusing each one on different aspects. These models are all re-estimated on the complete data set, and are then used to simulate. The simulation predicts probabilities for the first year of data, before drawing realizations of these probabilities. The conflict history and neighborhood conflict variables are then updated to take into account any positive draws. This is repeated for each year, and the whole process is repeated 2,000 times and averaged to create

forecasts. Like Goldstone et al. (2010), they find that a simple model performs just as well as the more complex models. However, to make meaningful forecasts they needed to include more variables for which long term projections exist. The fact that very simple models perform well leads back to the problem of overfitting. The minimal improvements that can be found by adding additional variables do not necessarily justify their inclusion. In the case of conflict forecasting, the choice of variables to include in forecasts determines the direction they take. Two conclusions can be drawn from this; firstly that the principle of parsimony is if anything even more important when forecasting. Second, to make forecasts beyond predicting a continuation of the status quo, we need to include more variables. Simple models appear to perform well, but they cannot test theories beyond the variables they include.

2.2 Statistics

> *All opinions are not equal. Some are a very great deal more robust, sophisticated and well supported in logic and argument than others (Adams, 2002).*

In this section I will be explaining the theoretical base for my research design. I start by reviewing and questioning the traditional approach that is most used in modern quantitative political science. Many of the techniques that appear most often are are either misused or not at all applicable to conflict prediction. As a replacement I present techniques originally developed for use in other scientific fields, which have in recent years been applied to political science. The main points are the use of predictive power, rather than explanatory power. This includes the use of Receiver Operator Characteristics (ROC) curves, Precision-Recall (PR) curves, and the Area Under Curve (AUC) metrics of the two. The F-scores that can be derived from PR are also mentioned, and I present the Brier score as an alternative to probability-threshold based measures. Lastly, I attempt to link old and new by presenting my approach as an analogue of established methods of testing for influential units.

2.2.1 The significance based approach

The null hypothesis significance test is the main judge of merit for quantitative political science (Gill, 1999, p. 647). The test is a mix of Fisher's test of significance and Neyman and Pearson's hypothesis test.

The Fisher test includes only a single hypothesis, the null hypothesis, or H_0. The null hypothesis is any hypothesis to be disproven. This is not necessarily a hypothesis stating that there is no relationship between dependent and independent variables, the null simply

means it is to be nullified. The procedure is a simple test of whether the data exhibits the properties that are expected given H_0. A test statistic is calculated from your data, and compared to the expected distribution of the statistic providing H_0 is true. If this is our expected distribution, and we know our calculated test statistic, we can find the *p-value* by calculating the area under the curve to the right of our value. If your p-value is sufficiently small you can reject the null hypothesis, if it is not you cannot draw any conclusion (Gill, 1999, p. 648-649).

As an example, Figure 2.2 shows a standard normal distribution. In this example our calculated test statistic has a value of 1.84, and the area under the curve to right of this value, and thus our p-value, is 0.033. If we are conducting a two-tailed test, as is usual, we have to take into account the fact that the error can go in both directions. This means we also have to include the corresponding area on the left hand side of the curve, giving us a p-value of 0.066. This is when the arbitrary nature of the test reveals itself: Where do we draw the line between significant and not? Fisher used .05 and .01 as thresholds for low N agricultural experiments, and these levels have become convention. These levels may not be appropriate for larger samples, but no clear guidelines for appropriate levels have been agreed upon (Raftery, 1995, p. 114).

Neyman and Pearson's hypothesis testing sets two hypotheses up against each other. Like with the Fisher test we first identify our hypotheses and an appropriate test statistic. The distribution of the test statistic given that the first of our hypotheses is true, and a critical value of the test statistic at a chosen significance level is determined. The test statistic of our data is then calculated, and depending on whether it reaches a critical value or not we decide to accept one of our two hypotheses. The test results not just in the rejection of one, but in the acceptance of another hypothesis. The power of the test can be determined, and is interpreted as the probability that the test correctly rejected the null-hypothesis (Gill, 1999, p. 651-652).

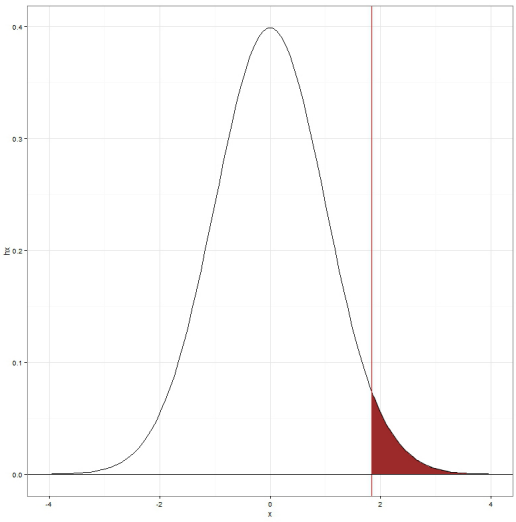


Figure 2.2: A normal distribution. The red vertical line marks the value 1.84 on the x axis, and the colored area is 3,3% of the area under the curve.

The two types of test have been mixed by the social sciences into a null hypothesis significance test. Fisher applied his test to a hypothesis of interest, and took the derived p-value as a measure of its strength. The null hypothesis significance test applies the

test to a null hypothesis, one that says there is no correlation between dependent and independent variables, and then interprets the p-value as the strength of a competing hypothesis. Each variable is given a p-value which determines whether or not we trust its effect. The interpretation of the p-value is the same as that of Neyman and Pearson's meaure of test power, the probability of the alternative hypothesis being rejected over time. This mix goes against the purpose of both original tests (Gill, 1999, p. 652-653). The p-value is not related to the odds of results being a result of chance. The p-value is the odds of your results given that they are random (Lambdin, 2012, p. 74-80). The only thing that can be proven is that the data does not conform to a completely random distribution. That is, the correlation between dependent and independent variables is not non-existent. The p-value is not a measure of the confidence you can put into the coefficient being the true effect, only that the coefficient is not 0. This might be useful, but as Bakan (1966, p. 426) points out: "There is really no good reason to expect the null hypothesis to be true in any population."

This simplification of interpretation, giving each variable a significance level represented by stars, is a shortcut that has become very common. Three stars is taken as a sign that the variable is worth keeping, excusing a scientist from having to argue further for its inclusion. The significance level diverts attention from problems such as miniscule or wide ranging effects. An effect that is minutely small will be included due to its statistical significance, yet may have little to no effect in practice (Lambdin, 2012, p. 72). Ziliak and McCloskey (2008, p. 44) argue that an effect that is very powerful, but where the confidence interval happens to cross below zero, should not be ignored on grounds of insignificance. This focus on significance over effect has been challenged numerous times, to no avail. A comprehensive list of articles of this nature, including works by Fisher, Neyman and Pearson, are largely ignored by the mainstream literature, as noted with frustration by Ziliak and McCloskey (2008, p. 57-58). Despite what they find to be insurmountable evidence proving common practice to be at best misleading, there are few signs of change. They argue that significance is a test of how well the model describes data, but that this is not necessarily what we want to do. A simple example of theirs uses cutlery to explain their argument. A spoon and a fork can be identical apart from the forked end, and even there the two are somewhat similar. A significance test may tell you that the two are not significantly different based on their looks. The handle is exactly the same, and the outline of the head is very similar. Significance testing the difference would show that they are significantly similar, but putting them to work will however instantly reveal which is the better at scooping up soup. Having tested the scooping-power, one can conclude with ease that the spoon outperforms the fork, despite their similar appearance. Similarly one should not accept fork-shaped models simply because their appearance is very close to that of the

spoon-shaped model. The two should be tested, and the one that performs best should be chosen (Ziliak and McCloskey, 2008, p. 49). The metaphor may not be perfect, but it does not weaken the argument that description is not the same as practical use.

Another issue is that p-value is often misinterpreted. Lambdin lists the four most problematic and common misconceptions as "(a) the odds your data are due to chance (b) the odds your research hypothesis is correct, (c)the odds your result will replicate, and (d) the odds the null is true" (Lambdin, 2012, p. 74). These variations lead to results meaning very different things depending on the person evaluating them. As we have seen, the p-value is a metric to be used when dealing with a sample taken from a larger population. This means using it to analyze conflict datasets becomes problematic, even if the correct interpretation is used. This is because the datasets are complete, encompassing the whole population of cases. Intensive and systematic data gathering has resulted in datasets that include, as near as makes little difference for this issue, every country in the world for the time period being studied. As the p-value in a null hypothesis test tells us the likelihood of our sample given a population where the null hypothesis is true, estimating a p-value on a population is meaningless (Bakan, 1966, p. 428; Schrodt, 2014, p. 297). We do not need to know whether our sample is representative of a population; our sample is the population.

Other critics point out that model selection based on p-values is highly susceptible to tweaking (Raftery, 1995; Gill, 1999). A variable's p-value is entirely dependent on the other variables included, as illustrated by Hegre and Sambanis (2006). Tweaking the operationalization also affects the p-value. These factors can be exploited to achieve significant results (Gill, 1999, p. 656). By testing every possible combination of variables it is possible to find models with the same explanatory power, but with very different specifications, and thus theoretical implications (Raftery, 1995, p. 120).

Another method is choosing variables based on their contribution to R^2. Measures of explanatory power such as R^2 are based on fit to data, and attempting to maximize them can lead to overfitting (King, 1989, p. 24,33). By adding more variables your model is adapting to the data at hand, gaining explanatory power for each one added. This explanatory power comes at the cost of being able to generalize your model to new data (Hastie et al., 2009, p. 220). While parsimony can ease the symptoms, the underlying issue of explanation versus generalization remains. Maximizing R^2 is therefore a problematic strategy if you wish to predict.

Despite these many arguments against significance based research, it remains the mainstay of conflict research. While it may not be interpreted correctly and it has no real theoretical meaning when dealing with our data, it serves a role as the universal measure of fit. It remains a useful tool for conveying certainty in variable correlation, as it sums up a number of factors in a single figure. While the methods described in the next section are

great at evaluating whole models, they are less precise when it comes to single variables. It would be foolish to disregard the statistical significance of variables completely, but it would be equally foolish to rely on significance alone.

2.2.2 Predictive power

In this section I will describe the methods that I will apply when evaluating the effect of countries on the performance of my model. I will present the most common measures used for model selection and evaluation in the forecasting literature. Hegre et al. (2013) utilize the ROC AUC in their design, and this is perhaps the most prevalent measure at this time. While I will include ROC, more weight will be put on the similar PR curves and their AUC. A derivative of the PR curves is the F-score, which I will include for comparison. A last measure that I use is the Brier score, which differs from the previous three in many aspects of its calculation.

All four methods are based on predictive power, rather than explanatory power. As discussed in section 2.1.3, models such as those of Fearon and Laitin (2003) and Collier and Hoeffler (2004) are used to estimate risks in a manner that assumes they apply as well to future events as to the data they are fitted on. This is problematic as their models are built and evaluated using measures based on statistical fit, which is not the same as ability to predict events. Predictive power is an alternative to explanatory power, and one more suited for evaluating forecasts of future risks.

Regression models usually have outputs that include the predicted probabilities of each possible outcome for each unit. In this paper I will be operating with a multinomial response, but I will be collapsing the results into a dichotomous conflict or no conflict response. This means that the output will consist of only a single probability. In this case the probability is the model's estimate of a country's risk of experiencing conflict in a given year.

The two main qualities that interest us the most when evaluating the model are calibration and sharpness (Gneiting et al., 2007). Calibration is how well a model output corresponds with the observed events. For my conflict forecasting model the output is a predicted conflict probability, and the observed event is a conflict occurrence. My model will be well calibrated if it predicts higher probabilities for country-years with observed conflict occurrences than for those with no observed conflict. Sharpness is how well clustered the predicted probabilities are, and is independent of observed data. A sharp forecaster predicts probabilities that are tightly concentrated, which is a trait that is positive subject to calibration (Gneiting et al., 2007, p. 246). I will be focusing mainly on calibration.

In order to evaluate calibration and sharpness, we start by classifying each prediction based on two factors: The first is whether or not an event was predicted, making the

prediction positive or negative. The second is whether the prediction matches the observation, giving either a true or a false response. The four possible outcomes are shown in a 'confusion matrix' as seen in Table 2.1. It consists of *true positives* (TP), where a conflict was both predicted and observed. *False positives* (FP), where a positive was predicted but not observed. *True negatives* (TN), where negatives are both predicted and observed, and lastly *false negatives* (FN), where an observed event was not predicted. To convert probabilities into clear yes or no responses we need to apply thresholds. Thresholds are critical values above which a probability is deemed to be a predicted event, while those below are deemed to predict a non-event.

		Observed	
		Event	non-event
Predicted	Event	TP	FP
	non-event	FN	TN

Table 2.1: Confusion Matrix

Having done this at any given threshold, we calculate ratios to use as comparable summary statistics. There are several that can be used, each focusing on different aspects. The true positive rate (TPR), also known as sensitivity or recall, is the proportion of actual conflicts that is correctly predicted. A parallel can be drawn to the calibration term, as this is how well the model is able to pick up on positive outcomes. The false positive rate (FPR), or specificity, is the proportion of correctly predicted non-events. Both these deal with a proportion of the total number of observed events or non-events. Precision focuses on a proportion of the predicted outcomes, namely the proportion of predicted positives that are correctly predicted. While TPR and FPR say something about a model's ability to correctly classify a set, precision tells us something about how much noise is included in the predictions.

$$\left(\ TPR \ \right) = \frac{TP}{TP+FN} \qquad \left(\ FPR \ \right) = \frac{TN}{FP+TN} \qquad \left(\ Precision \ \right) = \frac{TP}{Observedpositives}$$

These statistics can be calculated for every threshold, from 0 to 1. Choosing a single threshold is problematic. How well a model is able separate conflicts varies greatly between thresholds, and a trade-off must be made between being able to predict all conflicts and being not returning too many false alarms. Replacing the 0.05p-value threshold with a an arbitrary probability threshold would be far from an ideal solution. It is however possible to assign loss-functions to the results, which again makes it possible to calculate the total

cost at any threshold. The best possible outcome, the one with the lowest cost, can then be found by cycling through all thresholds. There are also ways of evaluating models over all thresholds without assigning a loss-function.

ROC

A Receiver Operator Characteristics (ROC) curve is a plot of the sensitivity and specificity of a model over all thresholds, as shown in Figure 2.3. The ROC curve shows the trade off between being able to correctly identify all conflicts in the set and the proportion of non-events that are misclassified. At the top right corner every event is correctly labeled, but every non-event is wrongly labeled. The bottom left is where every non-event is correctly labeled, but every event is wrongly labeled. A perfect model would have a curve going vertically up from the bottom left to the upper left corner, where all events and non-events are correctly classified, and from there to the upper right corner.

While the curve itself is a good illustration of model performance, the information it provides can also be compressed. The proportion of the ROC plot that is under the curve is known as the Area Under Curve (AUC), and has a value ranging from 0 to 1. The ROC AUC is interpreted as the likelihood that your model will give a random event a higher probability than a random non-event. An AUC of .5 would mean your model is no better at predicting than chance, while an AUC of 1 is a perfect predictor. Values under .5 are worse than chance, but occur rarely. The better a model is at correctly classifying outcomes, the higher the curve and greater the AUC. An advantage of the AUC is that it is comparable between models and datasets (Ward et al., 2010, p. 366-367).

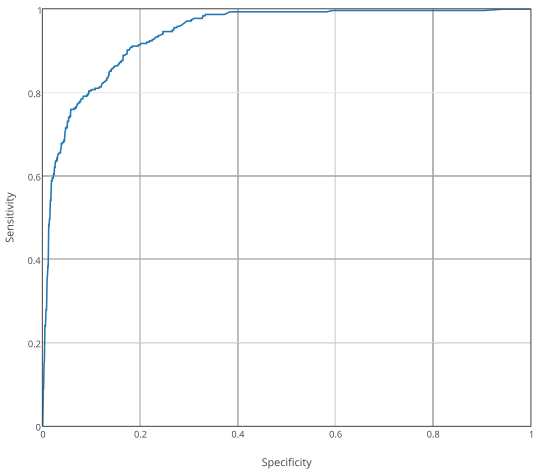


Figure 2.3: An example of an ROC curve. The y-axis is the sensitivity of the model, while the x-axis is the specificity.

The ROC curve and AUC have some disadvantages. Comparing curves is often difficult, as they can be indistinguishable, or without one being obviously better than the other. The AUC can also be misleading, depending on your data and goal. Two models with very different predictive characteristics can achieve the same AUC, and thus appear similar despite producing very different results (Kuhn and Johnson, 2013, p. 264). When faced with skewed data, meaning data containing few units with positive outcomes compared to

the number of negative outcomes, it may give a misleadingly positive image of a model's capabilities. This is because of the inclusion of specificity, or false positive rate, in the ROC curve. When the number of non-events in data outnumber the events by a large margin, the specificity can become inflated. A model that is completely incapable of distinguishing two equally sized groups of events and non-events will still score well with ROC AUC if the dataset also contains a much larger group of correctly predicted non-events.

Precision-Recall

The precision-recall (PR) curve is an alternative that can sometimes distinguish between two models that appear to have identical ROC curves. Precision is the proportion of units predicted as positive that are actually positive. Recall is another name for sensitivity (Davis and Goadrich, 2006, p. 233). The curve illustrates the model's ability to keep the level of false alarms to a minimum as it correctly identifies all real conflicts. Whereas the ROC curve is optimal if it follows the left and upper sides of the plot, the PR curve is optimal when it follows the upper and then right sides of the plot. Note that while recall and sensitivity are the same thing, and appear in both curves, they are normally not given the same axis. In the PR curve shown in Figure 2.4, recall/sensitivity is on the x-axis, as opposed to the y-axis in ROC plots.

By replacing specificity with precision, the PR curve is much more sensitive to false positives than the ROC. This means that true negatives are no longer given the same weight, which is useful when dealing with skewed data. While the ROC curve can remain virtually undisturbed by a huge increase in false positives (as long as the number of non-events is large enough), the PR curve will experience much greater impacts.

In a large set of 10,000 units with only 100 positive outcomes, increases in false positives will affect the false positive rate negligibly. Specificity, on the other hand, will have noticeable differences by even a single false positive.

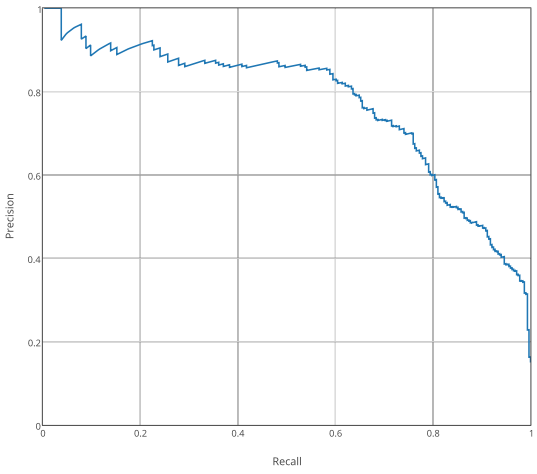


Figure 2.4: An example of a PR curve. The y-axis is the precision of the model, while the x-axis is the recall.

As an example we can imagine a model using this example data that at a given threshold predicts half of the positive cases correctly and predicts 50 false positives.

$$\left(TPR \right) = \frac{50}{50+50} = 0.5 \qquad \left(FPR \right) = \frac{50}{50+9950} = 0.005 \qquad \left(Precision \right) = \frac{50}{100} = 0.5$$

The results show that when half the events are correctly predicted, the false positive rate is still almost perfect. This would lead us to believe that the model is doing exceedingly well if only the ROC is used. Should we however be using PR we would see that already here we find that half of our predicted events are false. PR is showing us interesting details that we would miss using ROC, making it easier to distinguish between models.

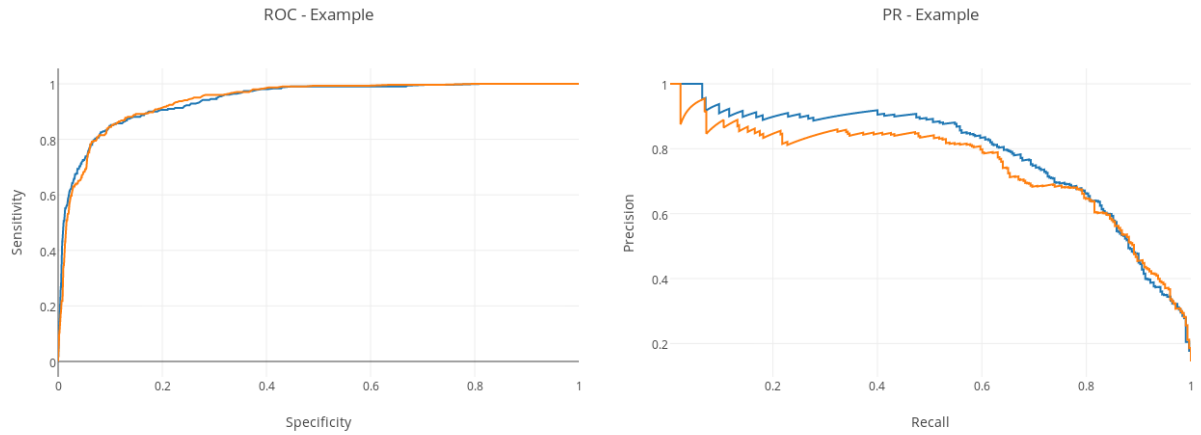
The differences become even more stark if we now compare the previous model with one that performs worse. Below are the results of a model that predicts the same proportion of the positive cases correctly, but also predicts 200 rather than 50 false positives.

$$\left(TPR \right) = \frac{50}{50+50} = 0.5 \qquad \left(FPR \right) = \frac{200}{200+9800} = 0.02 \qquad \left(Precision \right) = \frac{50}{250} = 0.2$$

Both models will have the same sensitivity or recall of .5, so for both ROC and PR one axis will remain unmoved. The difference in specificity will be a minute 0.015. Judged by the criteria used in an ROC curve they are basically identical, but we know that the latter model will include three times as many irrelevant units. If we use a PR curve the difference in performance will be clear, with a change in precision from .5 to .2.

Conflict datasets are not necessarily as skewed as this, but even though this is an exaggerated example the principle remains the same; differences between models using conflict data will be much more easily distinguished using PR. Figure 2.5 shows the ROC and PR curves of two drops, the Democratic Republic of Congo and the Philippines. The two ROC curves are virtually the same, with minor differences but with AUC scores that are close to identical. The two PR curves on the other hand display a distinct difference, with the model that drops the DRC having a curve that is clearly superior. As I intend to drop one of 162 units, the difference between estimations might not be very great. PR will therefore be the better choice for evaluation, as it is more likely to reveal potential differences.

In the analysis I will be focusing on the AUC of both types of curve rather than the curves themselves. Both the ROC and PR curves and AUCs are based on probability thresholds, and calculate the models performance over all of these. Collapsing the curves into simple AUC figures discards much of the detail, but makes comparison over several cases, or very similar cases, easier. As mentioned, the ROC AUC has a simple interpretation, namely the probability that a randomly chosen event has been given a higher



(a) ROC examples (b) PR Examples

Figure 2.5: PR and ROC comparisons of two models. While the two have almost identical ROC curves, they have clearly differing PR curves.

probability than a randomly chosen non-event. The PR AUC has no such theoretical definition, making the metric merely a way of saying which curve covers the greatest area in PR space.

Scoring rules

Another single figure metric that can be extracted from PR curves is the F-score, which is the harmonic mean of precision and recall at any given level of either.

$$F - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Unlike the AUC measures, which summarize performance over all thresholds, the F-score reported is the highest achieved F-score over all levels of PR. The score spans from 0 to 1, where higher is better.

The previous three measures have been based on thresholds, and on the concept of true or false predictions at these thresholds. A metric that is independent of this is the *Brier score*, which is the mean of squared differences between predictions and observations.

$$Brierscore = \frac{1}{N}\Sigma(Pred - Obs)^2$$

This makes the nature of the Brier score somewhat different from the curve based measures. PR and ROC throw all predicted probabilities in a pool and then go through them thresholdwise. This means all units are held to the same standards when it comes to what probabilities constitute conflict. Whether the probabilities given are close to 0 or 1 becomes less important as long as they are in the correct place with regard to the group.

This means low probability conflict years in some countries create problems not only for the country itself, but also for other countries that have high predicted probabilities for peace years.

The Brier score is less concerned with the ranking of predictions, as it evaluates each observation on its own merits and then averages. The score spans from 0 to 1, but it is worth noting that the Brier score punishes a complete miss more than a near miss. The increase in Brier score when going from a perfect prediction to a miss by 0.1 is very small. The perfect prediction is scored 0, and a miss by 0.1 is only scored 0.01. However, a further increase of 0.1 in predicted probability, taking us to 0.2 in total, will result in a score of 0.04, an increase in 0.03 compared to the first 0.01. This exponential increase means that the same incremental increase in probability is punished more heavily the further from the observation it is. A possible weakness of the Brier score is the averaging nature, where each unit counts the same. A skewed dataset with mostly peace will automatically be given a decent score if it predicts a low score for all units. For a dataset of 1000 cases with 100 events and 900 non-events, a constant prediction of 0.1 would get a Brier score of $\frac{((900*(.1^2))+(100*(.9^2)))}{1000} = 0.09$.

A predicted probability of 0.2 for a non-event year would get a Brier score of $0.2^2 = 0.04$.

Application

Beyond theoretical evaluation, predictive power can also be used to evaluate a model's practical value. By calculating risks for all countries a list of those most likely to experience conflict can be made. These can then be individually examined to see which variables are causing the risk. Humanitarian operations can be implemented to counteract the problem. As with any problem we would want a model that predicts as many conflicts as possible without raising any false alarms. By adding a loss function, a penalty for predicting the wrong response, we can find the model that is the most economical. We can also determine whether following the model is more economical than simply doing nothing. To do this we need to put a price on failure. Armed conflict has been shown to reduce both GDP, growth and trade for the countries involved and their neighbors (Gates et al., 2012, p. 1719; Collier, 1999, p. 175; Murdoch and Sandler, 2002).

Kennedy (ming) provides a framework for evaluating whether a model is better than simply expecting no conflict. By assigning assumed costs to false positives and false negatives, and assuming that correct predictions are preferred over incorrect, they arrive at a method of calculating the cost of decisions. With the data and model used by Goldstone et al. (2010), he shows how a decision to simply do nothing can be compared to following a model with a given decision threshold. Goldstone et al. (2010) use a threshold that splits the data into quintiles, where the lower four quintiles are predicted as non-events and the

uppermost quintile are predicted to be state failures. By comparing the resulting confusion matrices of the "see-no-evil" approach and that of Goldstone et al. (2010), they find that the cost of an intervention must be less than 7.7% of the cost of allowing a state failure.

Collier (2004) calculated the cost of a conflict to be $50 billion, while Collier et al. (2003) find an average cost of $14.1 billion for a group of armed interventions in conflict zones. With these numbers, an intervention costs 28.2% of allowing the conflict to happen. This makes Goldstone et al. (2010)'s model less economic than simply allowing all conflicts to occur. It should be noted that these numbers are taken from military interventions in conflict zones, and that preemptive humanitarian aid missions need not be as costly. Regardless, the method shows how forecasting models can actually be assessed by criteria that aren't reliant on arbitrary choices of significance levels or thresholds. Goldstone et al. (2010) are here judged on their one chosen threshold, but the method can also be used in reverse. If the cost of both conflicts and interventions are given beforehand, and each cell is given a cost, it is easy to calculate the ratio that is needed between cells of the confusion matrix in order to outperform doing nothing. Any model can then be tested at all thresholds to see if it can achieve the required values to be less costly than the control. By undertaking a detailed cost-benefit analysis of preemptive humanitarian operations one could establish a rough guide to how well a model should perform to be worth implementing as a policy-guiding tool.

2.2.3 Cross-validation

As mentioned in section 2.2.1, a problem of measures of explanatory power is that they are susceptible to overfitting. Replacing explanatory power with predictive power does not solve this problem on its own. If predictions are compared to the same data that the model was fitted to we are estimating its *training error*. Like with R^2, the training error is maximized at the cost of ability to generalize. The *test error* of a model measures how well it predicts on data the model has not been exposed to before validation. Figure 2.6 shows the trade-off between the two as model complexity changes. The reduction in training error increases along with model complexity, but the test error follows a u-curve. The test error is obviously of more interest, and it can be estimated through cross-validation.

Cross-validation requires us to partition our available data, using parts of for estimation and parts for validation. One way is to simply leave a unit out of the estimation and then predict on it, repeating this for each unit and then averaging the results. This is called leave-one-out cross-validation (LOOCV). Another common and more effective way of doing this is by the k-fold approach, where data is first split into k-parts. Each part is then in turn left out of the fit and used to estimate the prediction error. The average error over all k-parts is the model's prediction error (Hastie et al., 2009, p. 241-242).

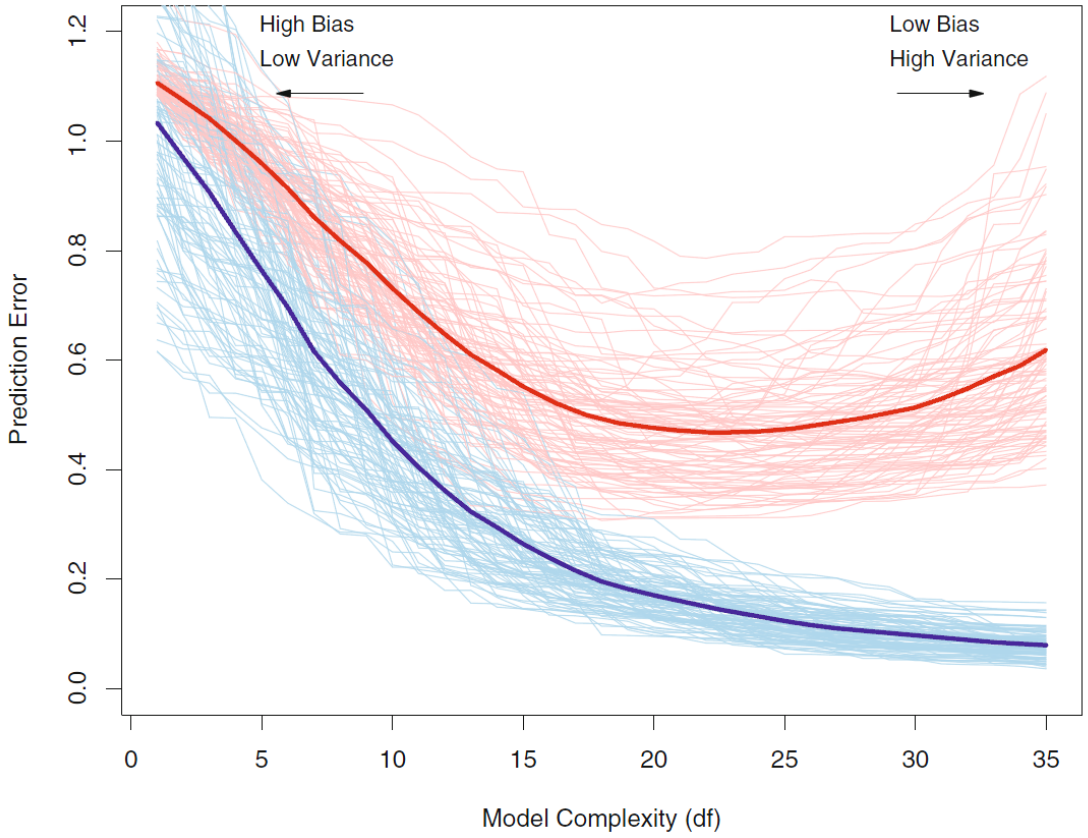


Figure 2.6: Curves showing how training error (blue) decreases, while test error (red) increases as variables are added. (Hastie et al., 2009, p. 220)

The conflict literature tends to split the data chronologically. Hegre et al. (2013, p. 256) fit their models to data from 1970 to 2000 and validate using data from 2001 to 2009. Similar approaches were also used by Weidmann and Ward (2010, p. 896), where data was split by months, and O'Brien (2002). Only fitting and validating along a single split in the data could result in a skewed estimate of prediction error. In theory, if we are seeking to uncover the underlying factors common to all conflict, any set of conflicts should be able to predict other conflicts. Limiting ourselves to only validating on a single group would therefore be an unnecessary shortcut that leaves estimates vulnerable to bias that can change depending on where we choose to partition data.

The structure of conflict history data is however of such a nature that applying k-fold cross-validation is not without issues. The strength of k-fold validation lies in it splitting the data randomly into groups. The UCDP/PRIO conflict dataset that I use is in a time-series-cross-section format, where countries are repeatedly observed over time. This means that each unit is no longer independent (Beck, 2001). The models I am using also have neighborhood effects, making the data even more dependent on its spatial and temporal neighbors in the dataset. Removing random units would have such an effect on those around it in time and space that a k-fold cross-validation would not function as intended.

Removing a conflict unit would not only have the effect of removing that single unit, but it would remove the neighborhood effects on those around it and would distort the conflict history of those coming after it.

A last argument for a single split is practicality. I will be cross-validating for each country dropped, and each of these will be averaged over a large number of simulations. With a single split the process takes several days to complete, and a 10-fold cross-validation would probably take several weeks.

2.2.4 Unit influence

The concept of studying the effect of single units on models is nothing new. It is a common test of robustness performed to ensure that results are stable, as opposed to being created or disproportionately affected by some units rather than the data as a whole. My research design is in essence a complex test of unit influence, where I adapt the goal of the test to the new methods of evaluation.

There are numerous ways of detecting influential units or outliers in existing literature. Extreme values on predictors could lead to an unusually large effect on its coefficient. Similarly extreme values on the dependent variable could lead to similar effects on predictor coefficients. Leverage statistics can also be calculated to identify units that drive results. Commands to calculate such statistics are built into commonly used statistics packages, such as the DFBETA function for Stata. This is based on how units affect β-coefficients and their standard errors. Others tests examine unit influence on summary statistics of model fit. For example, Menard (2010, p. 134-136) describes how the change in a model's χ^2 can be used to determine unit effect on the models overall fit. Units that cause particularly large changes in χ^2 when dropped are considered influential units.

The earlier methods are based largely on how units affect explanatory power. As research shifts from utilizing explanatory power to predictive power, so too should the diagnostic tools. As I have shown, explanatory power is not necessarily the same as predictive power. If variables can be shown to increase significance without increasing predictive power, then surely the same can be true of influential units.

A method more suited to diagnosing forecasting models is using the change in predictive power. Beck (2001, p. 283) describes a method of specification comparison similar to LOOCV, only using countries rather than units. If, rather than averaging and comparing across specifications, we extract the results of each iteration we can compare the effect each country has on predictive power. In this way we can identify which countries are shaping the model, and which countries need a different model specification. My main method of choosing outliers is most similar to the test of change in χ^2 mentioned above, with a single figure measure of model predictive power replacing χ^2. I also present a method that

measures influence by how much a unit affects the predicted probabilities of other units. As predictive power is measured in how well the probabilities match observations, this is an important aspect. A test of influence on a summary statistic will tell us how much the unit influences this particular metric. This is an important characteristic as models are evaluated by these metrics, but it provides no information on the scale of changes in the results on a lower level.

I will be using PR AUC and Brier scores as my summary measures of predictive power. I will also include two additional measures, ROC AUC and F-scores, for comparison. The change in these measures as a country is dropped will determine whether or not a it is an influential unit. If the predictive power increases as a country is dropped, then it does not fit the pattern set by the rest of the data set. These units are of interest as they could reveal problems with the model specification. Inversely, if predictive power is lost by dropping a unit then the unit adheres to the same behavior as the majority of the data. The latter units will also be of some interest as they have the most appropriate attributes for the model to predict precisely. The predicted probabilities of these units could also reveal the type of predictions that are the most harmful to each measure of predictive power.

Having identified a group of outliers it is possible to examine what makes them have a greater influence than others. It is important to differentiate between the direct and indirect effects on predictive power that result from dropping a country. The direct effect is the effect of the model not having to explain the country that has been dropped. We can get an idea of how great this effect is by examining how well a complete model predicts each country. Poor predictions will have a detrimental effect on the model's overall predictive power, and so countries whose removal leads to an increase in predictive power are likely to have poor predictions.

Due to the inclusion of neighborhood variables the removal of conflict ridden countries will also have an effect on the status of neighbors. This could lead to the model having trouble predicting conflicts that are the result of spillover from the dropped country. Should a dropped country have conflicts in the period, then the result of dropping it will lead to lower risk estimates for their neighbors. Depending on whether the neighbors experienced conflict this could have a positive or a negative effect on predictive power. If the dropped country does not have conflict, then dropping it will not affect the neighborhood conflict variable. Apart from conflict, I will also be including neighborhood variables that measure wealth and political stability. Due to the effect of these variables, some countries that are predicted well could lead to a loss overall when removed. I am here referring to the change in variable values, and it is important to note that this is different from the effect that is made on the variables' β-coefficients.

Countries also affect the predictive power through indirect effects on units. One channel

for such effects is the model's coefficient estimates, $\hat{\beta}$. When a country is dropped before estimation, $\hat{\beta}$ will change as the model no longer has to take into account the data the country provided. These changes will in turn change the probability estimates of other countries, which in turn will affect the model's overall predictive power. Estimating the effect countries have on $\hat{\beta}$ is done by simply comparing the $\hat{\beta}$ of the model when estimated with and without the country included in data (Long, p. 100). If a country has accurate predictions and no adverse effects on its neighborhood's predictions, then the effect on predictive power is likely to be caused by an effect on $\hat{\beta}$-estimates leading to global changes in predictions.

As I will later illustrate, examining the change in predictive power is not the same as merely looking at which countries have the worst predictions. These countries could be found by simply estimating the model and examining the predictions, but this would not reveal the true extent of their effect on predictive power. It is to be expected that such countries are among those that will lead to the greatest improvement in overall predictive power when dropped, but indirect effects may also play important roles. The nature of how PR AUC and the Brier score are calculated will also decide how great an effect predictions have on the estimated predictive power.

2.3 Summary

The armed conflict literature has more or less arrived at an agreement that a set of variables appear to be linked to conflict. Among these are GDP per capita, population, regime type and neighborhood spillover effects, which will feature in the research design of this thesis. I have also argued that quantitative political analysis, and especially the armed conflict field, has much to gain in using forecasting and out-of-sample evaluation using predictive power. I have presented four measures of predictive power: PR AUC and Brier score, which I will be using to select outliers, as well as the ROC AUC and F-score. The first two will form the backbone of my analysis, while the latter are included to examine how other measures respond to the same changes in data. I have also presented the rationale behind my research design, comparing it to established test of unit influence. By dropping countries on by one I will find those with the strongest effects on model predictive power, which I can then examine in detail to uncover commonalities.

Chapter 3

Research Design

In this chapter I will present the data that is used, as well as the statistical model and the forecasting simulator setup. The dependent variable, conflict, is gathered from the UCDP/PRIO conflict dataset. As mentioned in the previous chapter, I will be using only the conflicts that are classified as internal armed conflicts. The variable has three outcomes: no conflict, minor conflict and major conflict. The independent variables are gathered from several sources, combined in an the as yet unpublished Shared Socio-Economic Pathways (SSP) dataset in development at PRIO. Conflict history variables are a crucial part, both internally and in the neighborhood. These are also interacted with each other, and with the other predictors, to take into account effects changing under different circumstances. Other variables measuring political stability are neighboring regime changes and time since the country gained independence. The basic variables for most conflict models, GDP and population size, are also included. The Polity IV-index is included as a measure of democracy despite its issues. Lastly there are a number of variables to take into account unobserved spatial and time effects.

The simulation setup is developed from the simulator presented in Hegre et al. (2013). I use a more recent version, modifying it to suit the nature of my study. The simulation process is in essence a multinomial logit model that is allowed to update its own dataset when predicting outcomes, creating a truly dynamic model. Random country effects are added through the use of a multilevel model, while the independent variables' effects are estimated with a multinomial logit model. While the coefficients are estimated using Stata, the process of estimating probabilities is done by a separate simulation program.

3.1 Data

The SSP-dataset combines information from several data sources to create a set of variables with no missing values for the whole period from 1950 to 2013. The data is in a country-

year format, with each unit representing a country for a specific year. As well as historical data the set contains projections of five different future scenarios. The data used here includes 162[1] countries for the time period from 1950 to 2013, and only uses recorded data and imputations.

As mentioned, all units are extremely dependent on units in spatial and chronological proximity. It is therefore crucial that there are no missing units. The SSP-data has a very low amount of missingness, although this comes at a cost. Data has been merged under various assumptions to fill in missing units. These issues are taken into account when discussing the predictive outliers. Some countries are removed from the analysis, either wholly or partially, for various reasons. Some countries have ceased to exist before, or not yet gained independence by, the start of the evaluation period. Others have problematically high levels of missingness, rendering them useless.

3.1.1 Dependent variable

The dependent variable is a three leveled measure of internal armed *conflict incidence* taken from the UCDP/PRIO Armed Conflict Dataset (Gleditsch et al., 2002; Themnér and Wallensteen, 2014). Using incidence rather than onset gives a larger set of conflict years to estimate and evaluate on. This leads to single country-years having less influence, making evaluation less vulnerable to random errors.

The reference category is no observed conflict, coded as 0. The alternative outcomes are minor and major conflicts. Minor conflicts, coded 1, are conflicts with 25 or more battle related deaths. Major conflicts are coded 2, and have 1000 or more battle related deaths. For the 162 countries included in my data there are 21345 peace years, 920 minor conflicts and 416 major conflicts. This gives a 4.1% probability of randomly selecting a minor conflict, and a 1.8% chance of randomly selecting a major conflict.

For simulation I will be using the conflict variable as is, with two conflict levels. Having run the simulations I will dichotomize the variable, giving it the outcomes no conflict and conflict. This results in some loss of data, but makes the process of evaluation much simpler.

3.1.2 Independent variables

The model I am using is from Hegre and Nygård (ming), and is similar to the baseline model in Hegre et al. (2013). It includes a few basic predictors and several interactions,

[1]Due to the simulator crashing if Burma is dropped, I am only able to present the effect of 161 countries on predictive power. Burma is still included in the estimations of the other countries, providing both data for estimation and evaluation. All the remaining 161 countries are named in the tables reporting the results of my analyses, found in Appendix A.4 and A.5.1.

as well as a few regime variables. The simplicity of the model suits the diagnostic nature of my study. How countries behave using this model will have implications for how one is to interpret more complex models, and more reliably understand other variables when included.

Conflict history

Conflict risk has been shown to increase drastically in countries with recent conflict history (Collier et al., 2008). To take this into account the model includes dummy variables for both minor ($c1_{t-1}$) and major ($c2_{t-1}$) conflicts in the previous year, $t-1$. As countries are expected to become more stable over time, conflict history beyond the previous year is also included. This stability is represented by the period of peace prior to $t-2$. The variable,, $\ln(t)_0$, is operationalized as the logged number of years since last conflict, measured at $t-2$. All conflict history variables are calculated from the dependent variable

Independence

A further measure of stability is the age of the state in question. Newer states are more at risk of conflict as their institutions and governments have yet to establish themselves (Fearon and Laitin, 2003; Hegre and Sambanis, 2006; Hegre et al., 2013). The variable, *ltimeindep*, is coded as logged years since independence, or since 1700 if they have been independent since before the 18th century.

GDP per capita

The data on GDP per capita is gathered from the World Bank's World Development Indicators (WDI) (World Bank Group, 2013), Maddison Working Paper 4 (Bolt and van Zanden, 2013) and Penn World Tables v8.0 (Feenstra et al., 2013). The projected data is taken from the OECD ENV-Growth projections (Chateau et al., 2012) which uses the same PPP adjusted 2005 USD as the WDI. The WDI is therefore the main source, with data from the other sets adjusted to be consistent. Conversion rates were arrived at by averaging the ratio where the sets overlap. Some countries are subject to more substantial assumptions when merging and imputing data. An example is that the conversion rate into PPP was lacking for Somalia, and so the conversion rate of Ethiopia was used. When such substitutions are used, the countries involved are always similar in geographical and economic terms, and it should not cause any major issue with data validity.

The variable is log-transformed as the effect is expected to diminish with extreme values. Interactions with ($c1_{t-1}$) and $\ln(t)_0$ are included.

Population

Population is included through the logged total population of each country in thousands. Population data is gathered largely from United Nations World Population Prospects (United Nations, 2013). This set contains both historical data and future projections for use in forecasting. The variable is log-transformed as the effect is expected to diminish with extreme values. The population variable, *lpop*, is also interacted with both local conflict history dummies.

Polity

The Polity2 variable from the PolityIV-project (Marshall and Jaggers, 2002) is also included, both in a linear and a squared form. The linear variable measures the level of democratic development, and spans from -10 for autocratic regimes to 10 for fully democratic regimes. The squared will pick up the anocratic U-curve effect, and has values from 0 for completely anocratic to 100 for fully democratic or authoritarian. Both the linear and squared polity-variables are interacted with the local conflict history dummies. As discussed in section 2.1.2, there are problems associated with using the squared polity variable. This variable is used as it is the only one that has complete data for the whole period and for all the countries involved. While another measure would be preferred, pragmatism dictates the use of an inferior variable as it enables the use of much more data.

Neighborhood effects

The effect of neighborhood conflicts is included by a dummy, *nc*, for an ongoing conflict in any neighboring country the previous year. Neighboring country is here defined as a country sharing a land border, or a border over an inland sea. This means that Spain and Morocco are not neighbors across the Straits of Gibralter, but Tanzania and the Democratic Republic of Congo are neighbors across Lake Tanganyika. The neighborhood conflict variable is also interacted with the local conflict dummies. As with country stability, a stable neighborhood should create more peaceful conditions. The logged number of years all neighboring countries have experienced peace, *ltnsc*, is therefore included. An interaction is also added between neighborhood peace and time since last conflict, *ncts*0. Beyond conflict, a neighborhood instability is modeled by adding a variable for recent regime changes. The dummy variable nb_TSRC_5 measures whether a regime change has occurred in neighboring countries in the last 5 years.

Time effects

To capture unobserved global effects that are time-specific, dummies are included for each decade. Each decade, the 1950s, 1960s, 1970s, 1980s and 1990s is given a dummy variable, leaving all years from 2000 and onward as the reference category.

Spatial effects

Unobserved country specific effects are modeled by the inclusion of random effects for each country. These are derived from a multilevel mixed-effects model using the same variables as the main model and a random intercept. The random intercept of each country for both outcomes is extracted and included here as the $random_1$ and $random_2$ variables.

The intercepts are not simply directly extracted from the estimated β-coefficients. Under the assumption that the β-coefficients of the random intercepts are at the middle of a probability density function(pdf) that is normally distributed, with the standard errors as the span from mean to the confidence intervals. The random effects to be used in the simulations are drawn at random from the pdf, and then stored for later use.

All variables and their shortened names are listed in Table A.1.

3.2 Multinomial logit model

While a multilevel model would be ideal, the simulation software can only handle a multinomial model [2]. As the dependent conflict variable has three outcomes, a multinomial logit model is used. The model includes lagged independent variables, including lagged variants of the dependent variable. This type of model is known as a "dynamic model" (Greene, 2003, p. 558). From this model we can extract the transition probabilities between conflict states. Table 3.1 shows the probabilities of a transition from any of the three states to any other in the data set. If we were to include conflict state at $t-1$ as the only predictor then these are the probabilities it would predict for any unit.

The multinomial model is estimated using the mlogit-function in Stata. While Stata can be used to estimate the probabilities of all outcomes, this would be done using the recorded conflict history data in the dataset. To create true simulations of conflict probabilties for each year, the conflict history variables need to be updated depending on the conflict state the previous year to the one being estimated. To do this the *mlogit* is merely used to estimate coefficients, and these are then handed on to a simulator program along with the random effect variables from the multilevel model.

[2] Which is why the random intercepts are extracted from a multilevel model beforehand and converted to variables.

The β-coefficients of the multinomial are treated in the same way as the random intercepts from the multilevel model. The actual coefficients used in the simulations are drawn at random from a probability density function defined by the multinomial β-coefficients and their standard errors. The simulator then uses the random and multinomial coefficients to estimate the transition probabilities to each conflict state based on the conflict state in the prior year. The simulation process is described further in the next section.

		Conflict level at t		
		None	Minor	Major
Conflict at $t-1$	None	96.8	2.8	0.4
	Minor	21	69.8	9.2
	Major	9.3	20.1	70.6

Table 3.1: Transition probability matrix. The matrix shows the probabilities of a country transitioning from one state to another.

3.3 Simulation procedure

To estimate the out-of-sample predictive power I will utilize methods developed by Hegre et al. (2013) to produce predictions. There are some modifications, most notably the repeat over all units and the inclusion of random country effects which also necessitates an extra layer of loops. The simulations are important as they allow conflict history variables to be updated. Simply predicting each year individually would not be a realistic evaluation, as the model would be "cheating" by checking the dataset for the correct conflict status in the previous year.

In practice, the simulation consists of a master Stata .do-file that calls a number of subscripts and datasets, before data and parameters are fed to a *C#* library that runs the actual simulation. The master .do-file determines several aspects of the simulation:

- The parameter files to be used, determining model specification.
- The data to be used.
- The number of simulations in each step.
- The years that are to be used for estimation and simulation.

Neighborhood variables are added to the data, estimated from the conflict data and from a database of distances between countries. In the country-by-country loop, the countries are dropped from the data set at the start of this script. This leaves the countries' neighborhood effects out of both the historical estimations as well as the simulated predictions. Conflict

history variables are also calculated using the conflict variable before the data is passed to the simulator.

The simulator first estimates a multilevel model from which the β-coefficients and standard errors of the individual country intercepts are extracted. The multinomial model is then estimated, and the β-coefficients of all variables are extracted along with their standard errors. A realization of the random country effects is then drawn. These realizations are stored as variables, one for minor conflict outcomes and one for major conflict outcomes. A realization of the multinomial coefficients is then drawn. Using these and the realization of the random coefficients, the predicted probabilities for the first year of simulation is calculated. These probabilities are then used to draw conflict outcomes for the year in question, which are then used to update the conflict history variables for the next year. Having updated the relevant variables, the simulator then proceeds to the next year, repeating the process for each year of simulation.

The year-by-year process is then repeated, creating 50 sets of predicted outcomes for each country in each year. The inner loop of 50 is then repeated with a new draw of multinomial coefficients, creating 5 sets of 50. This is then repeated with new draws of random country coefficients, creating a total of $5*5*50 = 1250$ simulations. The outcomes are then averaged, giving a single predicted probability for each country for each year. When the first year is completed, the conflict history variables are updated for that year before simulations start on the second year. This set will provide a benchmark. The process is then repeated with each country dropped before neighborhood variables are created. This results in $50 * 5 * 5 = 1250$ simulations for for a control set, and $50 * 5 * 5 * 161 = 202,500$ simulations for 161 country-drop sets.

The 161 data sets, each with predictions for 161 countries over 13 years, contain a total of 336,973 predictions. The simulator calculates the probabilities of all conflict states. The probabilities for minor and major conflicts are combined to create a variable that is the probability of one or the other occurring, as opposed to the unit experiencing peace. This variable is stored as an individual Stata .dta-file. The final output from the simulator is thus 163 datasets of predicted conflict probabilities.

The process is shown in Figure 3.1[3] and the process is explained below:

1. Specify the statistical model.

2. Load dataset and drop country i.

3. Estimate the multilevel model and extract the random intercepts' β-coefficients and their standard errors.

[3]The flowchart is a general illustration of the simulator. The country drop (step two on the list below) would occur before the random effect model is estimated.

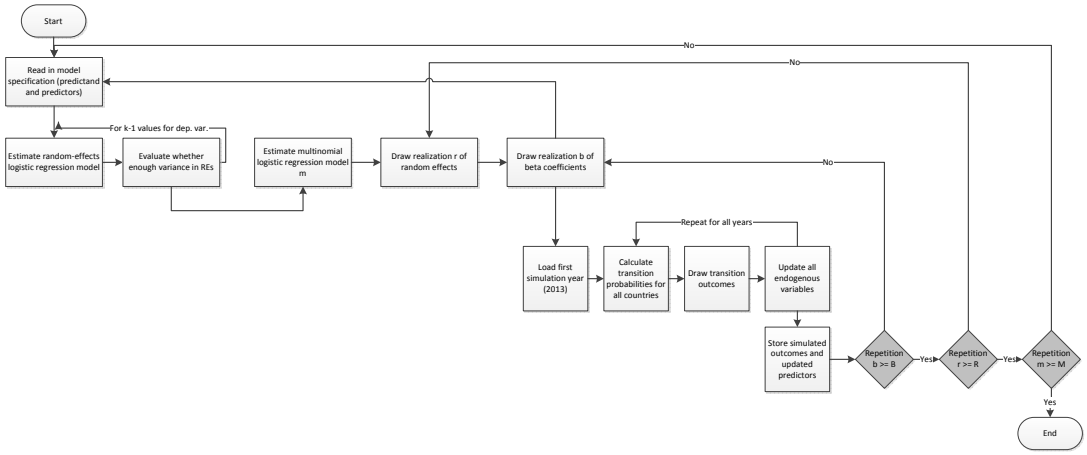


Figure 3.1: Flowchart showing the simulation process. The whole simulation is repeated M times for each country drop, in this case 162 iterations. The random effect loop is repeated for R realizations of random effects, here 5 iterations. The inner loop is repeated for B draws of multinomial coefficients, here 5 iterations. The transition outcomes are drawn 50 times for each country year.

4. Estimate the multinomial model and extract the variables' β-coefficients and standard errors.

5. Draw a realization of the random country effects.

6. Draw a realization of the multinomial β-coefficients.

7. Start simulation in first year, for my simulations 2001.

8. Calculate the probabilities of transition between levels for all countries for the first year based on the realized coefficients and the observed values for the predictor variables.

9. Randomly draw whether a country experiences conflict based on the estimated probabilities, and then update the explanatory variables that include conflict history and neighboring conflict for the year in question.

10. Repeat steps 8 and 9 for each year of the simulation period, in my case 2001 to 2013, and record the simulated outcomes.

11. Repeat step 8 to 10 fifty times to even out the impact of individual draws.

12. Repeat steps 6 to 11 five times to even out the impact of individual realizations of the multinomial logit coefficients.

13. Repeat steps 5 to 12 five times to even out the impact of country random effect draws.

14. Repeat steps 2 to 13 for each country in the dataset, meaning 161 repetitions and one control run where all countries are kept.

3.4 Evaluation

For evaluation, the data is transferred to R where it is cleaned of noise and sorted. The ROC curve coordinates are calculated using the performance function found in the ROCR-package (Sing et al., 2005). ROC AUC is calculated using the colAUC function found in the caTools-package (Tuszynski, 2014). PR curve coordinates are calculated using the precision.at.al.recall.levels-function of the PerfMeas-package (Valentini and Re, 2014), and AUC is calculated using the prauc function in the same package.

Unit influence on coefficients is estimated by dropping each unit from the control set before the multinomial coefficients are estimated. The difference between control and drop is taken as the country's influence on the coefficient estimate. Estimates are done in Stata, using the mlogit-function and the dataset from the control simulation.

Stata .do-files and R-scripts are available online at `http://tinyurl.com/ottdz3q`, or available on demand.

Chapter 4

PR outliers

This chapter presents the results of the simulations. I will use the evaluation methods described in Chapter 2.2.2 to evaluate country influence, ranking countries by the difference they make on the model's overall predictive power when they are dropped from the data. As the literature cited asserts that PR is better suited to the data than ROC, my focus will be on PR with ROC included for comparison. I examine the distribution of ROC and PR curves, as well as the ROC and PR AUC. Having identified a set of predictive outliers, I will examine their characteristics. Their values on dependent or independent variables may reveal commonalities that could reveal weaknesses in the model specification. The outliers' effects on the predicted probabilities of other countries reveals to what degree their influence is indirect or caused simply by their own predicted probabilities.

Another point of interest is also the degree to which the measures agree on outliers. If there is no agreement between the four measures, then the choice between them will have an even greater impact on results.

This chapter and the next will be similar in structure. In this chapter I start by performing a test of unit influence as measured by their effect on β-coefficients, and a test of influence on predicted probabilities. These tests are presented in this chapter but are referenced throughout both chapters. Apart from this, both chapters follow the same structure. Before going into detail, I present summary results for the results as a whole. The ROC and PR curves of each iteration is compared with that of a control model where all countries are included. I then show how the countries affect the ROC AUC and PR AUC. The countries with the greatest negative and positive effects are selected for further examination. Their values on important predictors are compared to reveal any common issues that could be the results of model misspecification. The predicted probabilities of the outliers, taken from a control model, are then compared to their conflict histories. This is to see whether they create problems because the model has to predict the country itself. To examine how they potentially affect the model indirectly through other countries

I then examine their effects on specific variables. I also examine how they might affect their neighbors through the neighborhood variables and their values. Countries whose effect on predictive power can be explained early on will receive less attention as I progress. Before continuing to the next chapter I give a brief summary of what I have found.

4.1 Coefficient effects

Before looking at the predictive outliers, I present the results of a more conventional test of unit effect on coefficients. To test for this I extract the dataset that is used in my control case after the estimation of the random effects. This makes it a test of what countries you would find with a conventional test for outliers, rather than how much the coefficients change with each iteration of my main loop. Each country is dropped before the model is estimated, and the absolute differences for all coefficients are calculated by subtracting the control case coefficient from the drop case. As the coefficients vary greatly in strength, the effect on each should be weighted to account for this. A difference of 0.2 means much more if the coefficient is 0.4 than if it is 2.2. By dividing the absolute difference with the value of the coefficient I derive the relative differences. The possible range of the variables also varies. Some are dichotomous, making their possible effect simple to deduce as it is the same as the coefficient. Other variables are continuous, and their potential effect can be anything from near zero to infinity. To account for this I also weigh by the variables potential effect on probability estimates. This is done by multiplying the relative differences with the maximum possible effect of each variable. This maximum effect is the maximum absolute value for each variable found among the units in the data set multiplied by the variable's coefficient. Table 4.1 shows the top ten least and most influential countries. The multinomial coefficients from the control case including all countries, and their minimum and maximum values from the country drops, can be found in Table A.2.

There are few surprises in the least influential countries, which are mostly smaller and wealthier countries. In the most influential we find countries with some, by theory, contradictive attributes. The US has a high GDP and large population, two variables that pull in opposite directions. China has a low GDP and a large population, but very little conflict. The others are mostly conflict prone nations, with for example Laos having several major conflict years. As the values here are somewhat arbitrary, setting a strict cutoff point for what is deemed a more influential unit is unreasonable. Considering the distribution of differences, any unit below 100 would be considered as without any noticeable effect. Units below 400 have some effect, but not enough to stand out. From 400 to 1050 are influential units, while those over 1050 stand out as very influential. These are approximate values, and which limit to choose depends on how many countries you believe can count

Rank	Weighted diffs	Country
1	2.19	Slovenia
2	3.10	Qatar
3	3.91	Norway
4	4.00	Bahrain
5	4.92	Luxembourg
6	5.40	New Zealand
7	5.52	Lithuania
8	6.02	Australia
9	6.13	Singapore
10	6.40	Bhutan

(a) Least influential.

Rank	Weighted diffs	Country
153	1366.15	China
154	1436.02	Nicaragua
155	1462.12	Indonesia
156	1501.44	Colombia
157	1577.27	Congo
158	1820.34	France
159	1889.56	DRC
160	2027.80	Lebanon
161	3645.44	Laos
162	5050.84	USA

(b) Most influential.

Table 4.1: The least and most influential countries by effect on coefficients.

as exceptional in a group of 162.

Measuring influence through effect on coefficient estimates is only one way of measuring unit influence. As I am more interested in predictions than on the coefficients, it would be more useful to examine country effect on the estimated probabilities of other units. By comparing the predicted probabilities of conflict of all countries in the control with those in country drop models I can calculate the effect the dropped country has on the remaining countries' probabilities. Table 4.2 shows the top and bottom ten countries by absolute difference in country probabilities. Dividing these differences by number of units gives us the average effect of the unit for each prediction. For example, Liberia has an average effect of $\frac{Diff}{Countries*Years} = \frac{42.01}{161*13} = 0.02$, or 2% on each unit's predictions.

Comparing Tables 4.1 and 4.2 we find that while Norway, Bahrain and Luxembourg are among those with the least effect on coefficient estimates, they are among the countries with the greatest effect on predicted probability estimations. The complex structure of the model may be why these countries could make such differences without there being visible effects on coefficient estimates. To reiterate, the coefficient test above is not taken from the simulation loop, but from looping a simple multinomial model over country drops using the control case data from my simulations. This means that the differences in coefficients of each country could be greater in the actual simulations than what is observed here. Also unobserved is the effect each country has on the multilevel model and the extracted random effects of each country. Adding to this is the effect on neighborhood variables, not through coefficients but through presence in a neighborhood.

The full list of countries and their effect on coefficients can be found in Table A.3, while the full list of effect on predicted probabilities can be found in Table A.4.

Rank	Country	Difference
1	Bosnia-Herzegovina	39.18
2	Mauritius	40.26
3	Gabon	41.25
4	Azerbaijan	41.71
5	Liberia	42.01
6	Slovenia	42.15
7	Netherlands	42.22
8	Madagascar	42.35
9	Niger	42.47
10	Fiji	42.49

(a) Least influential.

Rank	Country	Difference
152	El Salvador	60.65
153	Sierra Leone	63.00
154	Turkey	65.59
155	Finland	69.60
156	Kenya	70.32
157	UAE	70.80
158	Luxembourg	73.68
159	Bahrain	74.02
160	Armenia	78.87
161	Norway	86.93

(b) Most influential.

Table 4.2: The least and most influential countries by effect on coefficients.

4.2 Outlier scores and groups

A quick look at the bigger picture is necessary before going into details. The results are divided in 4 groups, divided by period and evaluation method. Hegre et al. (2013) argue that models with history variables tend to be unfairly rewarded for predicting a continuation of the status quo. They therefore calculate evaluations both from 2001 to 2009, and also on only the last three years of the evaluation period, from 2007 to 2009. By leaving a gap between the last estimation year and the first evaluation year, the first "free" years are removed from the process. The model has to rely on its own estimation of the country's evolution over the gap period to give it a correct starting point for the evaluation period. Dividing the data like this is highly sensitive to the conflict history in the gap period. If a peaceful country has a short period of conflict that coincides with this gap then the gap goes from handicap to performance boost. It is however likely that the gap will have the expected effect for most countries. To test whether it is a reasonable assumption that countries perform better on the whole period, I calculate results using both the whole and the latter part of the evaluation period.

To get a visual impression of the results, I plot the curves of each iteration of the country drop loop starting with ROC. Figure 4.1 shows the distribution of ROC curves for all country drop models and the control. Figure 4.1a shows the ROC curves for the whole 2001-2013 period, while Figure 4.1b shows the 2006-2013 period. The higher curves are those of countries who affect the model adversely, resulting in a boost to performance when they are dropped. Inversely, the lower curves are units whose removal reduce the accuracy. The variations are not extreme, but there are substantial differences between

the control and the best and worst performing iterations. Considering that the difference is merely the removal of one unit, not a respecification of the model, large changes are not to be expected. By eyeballing the curves, there appears to be change of a magnitude that is interesting to examine further.

A few curves appear as distinct outliers in the top left part of the plot. Finland, Norway and Bahrain fall below the others by a small margin in both time periods. Among those that appear above the others briefly are Cambodia and Thailand. These curves are not however consistently outside the main cluster, so picking outliers from this plot is difficult. [1] The variation between countries appears to be slightly greater when the evaluation period is limited to 2006-2013, while the control appears to be very similar.

Having established that ROC differences are small but great enough for analysis, I move on to PR. Figure 4.2a shows the distribution of PR curves for all iterations. It is immediately clear that there is a greater variation than with the ROC curves. This supports the claim that PR AUC is the better metric for assessing models using this data. As with the ROC curves, the variation among PR curves appears to increase in the limited period compared to the full. There are some curves that appear to stand out, but the cluster is still too thick to reliably discern outliers.

To more easily differentiate between the models I calculate the AUC of both ROC and PR curves. The control model's AUC is subtracted from the AUC of the country drop models. As for the curves, positive values indicate that removing the country leads to an increase in predictive power, and that the countries are problematic to include. Negative values indicate a decrease in predictive power, and that the country is supportive of the model. Table 4.3 shows the ten countries whose drop result in the greatest increase in AUC. In other words these are the countries that are the most detrimental to the model's predictive power. The table contains the 10 worst for both measures and both periods. Full tables of all country effects on PR AUC can be found in Tables A.5 and A.6. To avoid cluttering and an overwhelming amount of tables I focus on PR and on the countries with negative effects on model accuracy.

Figure 4.3 shows the full sets for both periods and both measures in a scatter plot form. The variation in PR is somewhat greater than that of ROC, as shown by the curves. For the whole period the PR-differences span from -0.063 to 0.023, and ROC spans -0.015 to 0.004. The variation increases for both PR and ROC when moving to the limited evaluation period. Here the PR AUC difference spans from -0.084 to 0.018 and from -0.017 to 0.005 for ROC. Both measures are reduced for the control case when limiting the period, supporting the handicap theory.

[1] The curves in the plot are not marked, but by following the URL in the caption you will find an interactive plot. In this you can zoom in on parts of the curve, and identify individual curves by hovering over them. Links to online plots are provided for all plots where this function is of use.

	Country	ROC AUC
1	China*	0.004
2	Nigeria*	0.003
3	Morocco	0.003
4	Cote D'Ivoire	0.003
5	Malaysia*	0.003
6	Spain	0.003
7	Syria*	0.003
8	DRC*	0.003
9	Sierra Leone	0.003
10	Hungary	0.002

(a) ROC 2001-2013 - Control: .954

	Country	ROC AUC
1	Jordan	0.005
2	Malaysia*	0.005
3	China*	0.005
4	Syria*	0.005
5	DRC*	0.004
6	Nigeria*	0.004
7	Libya	0.004
8	Denmark	0.004
9	Angola	0.004
10	Venezuela	0.004

(b) ROC 2006-2013 - Control: .944

	Country	PR AUC
1	Djibouti*	0.023
2	Spain*	0.016
3	Angola*	0.011
4	Qatar	0.011
5	Macedonia	0.010
6	DRC*	0.010
7	Syria	0.008
8	Slovenia	0.008
9	Uzbekistan	0.008
10	China	0.007

(c) PR 2001-2013 - Control: .820

	Country	PR AUC
1	Angola*	0.018
2	Tanzania	0.008
3	Djibouti*	0.007
4	Austria	0.005
5	Spain*	0.005
6	DRC*	0.004
7	Cyprus	0.004
8	Mali	0.003
9	Canada	0.003
10	Lithuania	0.001

(d) PR 2006-2013 - Control: .781

Table 4.3: ROC and PR AUC differences from control. Asterixes mark those that appear in both periods.

The countries in the upper right quadrants of Figure 4.3a and 4.3b are those that have a negative effect on both ROC and PR AUC when dropped. Conversely, the lower left quadrant contains those that have a positive effect on both when dropped. The upper left contains those with a positive effect on PR but negative on ROC, while the lower right contains those that have a positive effect on ROC but negative on PR. Some of the outliers are shared, but we also see a great deal of outliers on one axis that sits within the main cluster on the other. Only a few countries appear distinctly in the top right, whereas the lower left is more scattered and shows several countries outside the tighter main cluster. The center of the cluster is somewhat into the negative side the first period

for both measures.

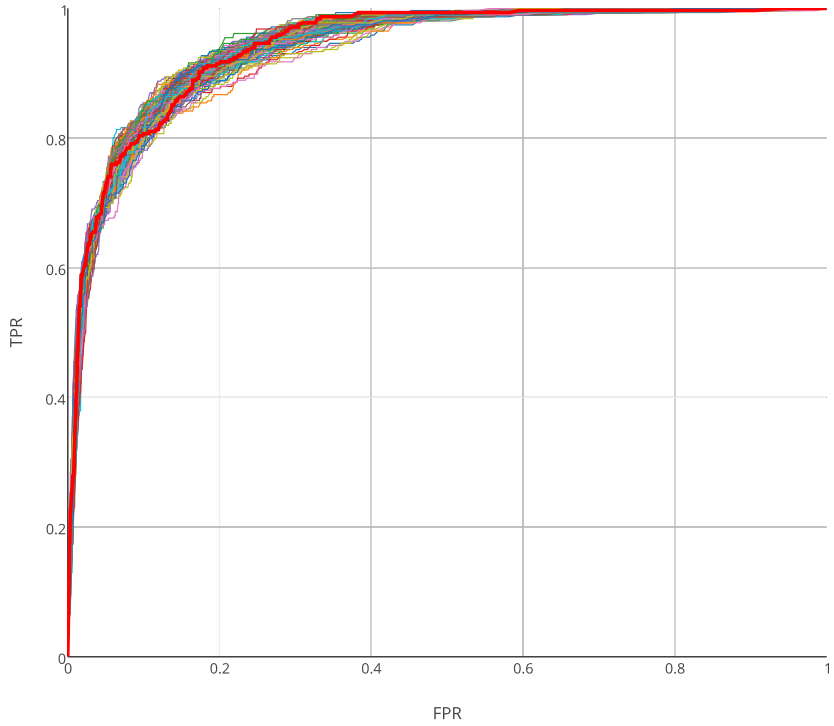
Comparing the two periods reveals that limiting the evaluation period has indeed increased variation, although the difference is much more distinct for PR. For PR AUC the cluster shifts towards negative values, with only a few units left on the positive side. For ROC AUC the opposite appears to happen, with more units moving up into positive values. While there appears to be a relationship between ROC and PR, the two measures do not share all the same outliers. China, Syria, the DRC, Spain and Angola are some of the countries consistently in positive space, strongly indicating a poor fit with the model.

On the negative side of the scale, the two measures appear to agree to a greater extent. Turkey, Bahrain, Luxembourg, the Philippines and Finland being among those most frequently observed in extreme positions.

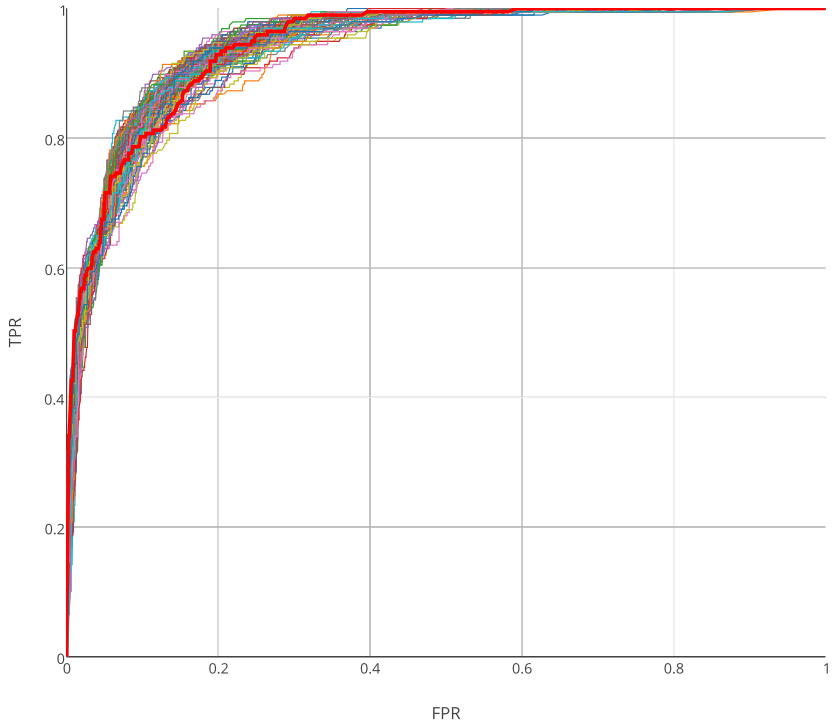
Limiting the evaluation period has also lead to a decrease in the control model AUC. The reduction in ROC AUC is small, from .954 to .944, but the PR AUC has a more drastic decrease from .82 to .781.

Having established a list of countries of interest, I will now examine these in greater detail in an attempt to uncover why they deviate. To reiterate, the countries from Tables 4.3c and 4.3d are those countries whose removal leads to an increase in overall model accuracy. These are collected in the first column of Table 4.4, and are from here on referred to as the *destructive countries*. When the two periods are combined there are a total of 16 destructive countries. I also separate those that have the most positive effect by the same rule. These are countries that appear among the 10 with the most negative effect on PR AUC when removed for either period, from here on referred to as the *reinforcing countries*. A list of 15 reinforcing countries is found in the second column of Table 4.4.

Comparing this list to the results of the coefficient influence test, we find that several countries from Table 4.4 appear among those that have the greatest effect on coefficients. On the destructive side, the DRC, Angola, Syria and China make an absolute weighted relative difference of over 400 to the coefficients. This means that the remaining twelve would most likely not have been found to be influential units by conventional methods. The DRC and China are well above 1000, making them among the most influential by effect on coefficients. It is also clear that coefficient influence does not equal a negative effect on predictive power; from the reinforcing column we find Indonesia at almost 1500, making it the eighth most influential country by effect on coefficients. Cambodia and Sudan are also far up the list, while Yemen has levels approaching interesting. This leaves eight more countries in the reinforcing group that would not be picked up by the coefficient test, for a total of twenty for both groups combined.

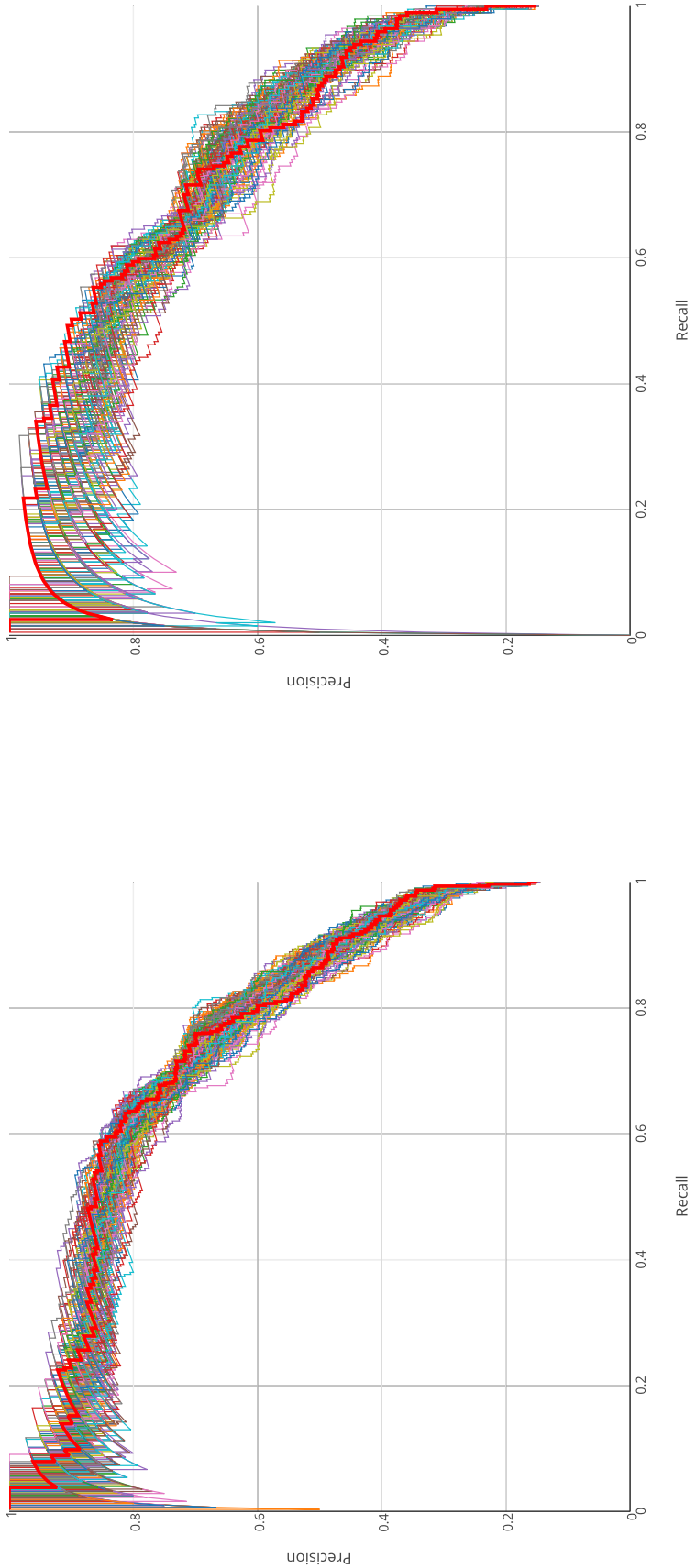


(a) 2001-2013. Interactive plot: `https://plot.ly/~msmidt/791`



(b) 2006-2013. Interactive plot: `https://plot.ly/~msmidt/785`

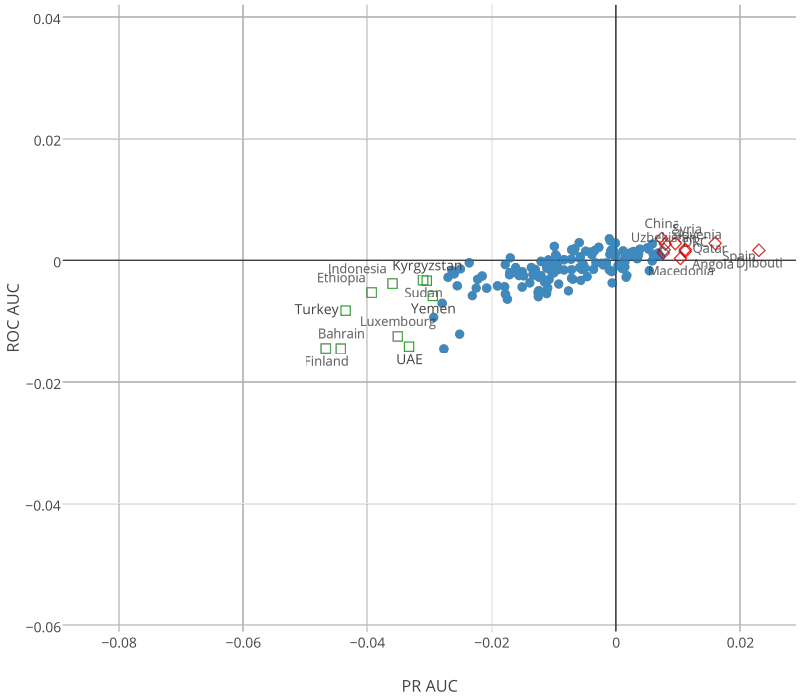
Figure 4.1: ROC curves for all drops. The thicker red curve is the control case.



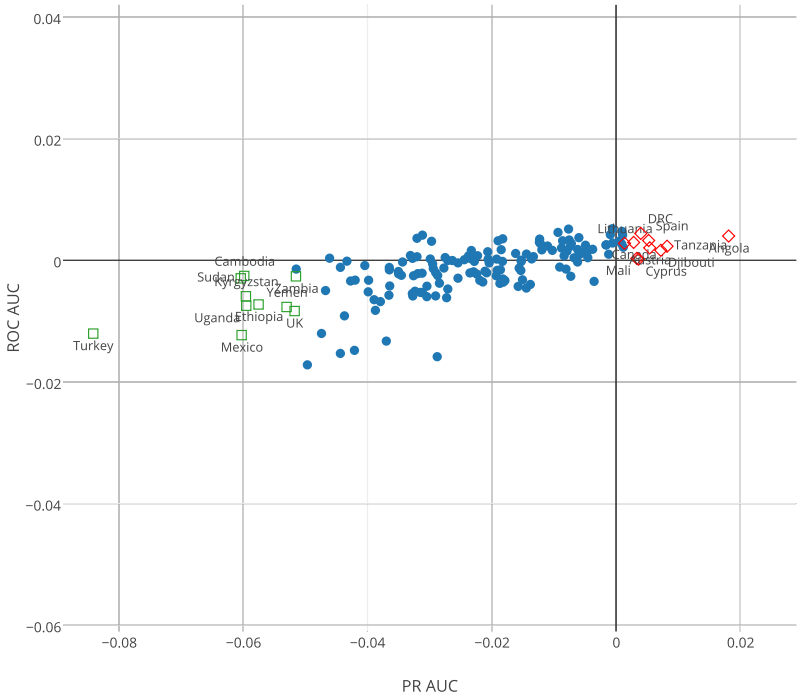
(a) 2001-2013. Interactive plot: `https://plot.ly/~msmidt/1006`

(b) 2006-2013. Interactive plot: `https://plot.ly/~msmidt/1028`

Figure 4.2: PR curves for all drops. The thicker red curve is the control case.



(a) Period: 2001-2013. Control ROC AUC: .942. Control PR AUC: .767. Interactive plot: `https://plot.ly/~msmidt/764`



(b) Period: 2006-2013. Control ROC AUC: .944. Control PR AUC: .781. Interactive plot: `https://plot.ly/~msmidt/766`

Figure 4.3: The difference in AUC between country drops and the control case. PR outliers marked in green squares and red diamonds.

	Destructive	Reinforcing
1	DRC	Finland
2	Angola	Bahrain
3	Spain	Turkey
4	Slovenia	Ethiopia
5	Syria	Indonesia
6	China	Luxembourg
7	Djibouti	UAE
8	Qatar	Sudan
9	Macedonia	Kyrgyzstan
10	Uzbekistan	Yemen
11	Tanzania	Uganda
12	Lithuania	Mexico
13	Austria	Cambodia
14	Cyprus	Zambia
15	Canada	UK
16	Mali	-

Table 4.4: The outlier groups by effect on model accuracy.

It would be logical to expect to find countries from the reinforcing group among those with the least effect on coefficients; countries who adhere perfectly to the remaining groups coefficients should increase overall predictive power by adding well predicted results. It is therefore interesting to see that the country with the least effect on coefficients is Slovenia, followed closely by Lithuania and Canada. These are all in the destructive group, meaning they have a strong negative influence on the model's predictive power despite not influencing coefficient estimates. Reassuringly there is also some agreement between the tests on this end, with Bahrain, Finland, Luxembourg, the UAE, Kyrgyzstan and Zambia having minimal effect on coefficients and a positive effect on model predictive power.

Continuing to the effect on individual predicted probabilities in Table A.4, it becomes clear that Slovenia's results are hard to explain. The country had the smallest coefficient effect, and also has the sixth smallest effect on predicted probabilities. Austria, Lithuania and Canada are also ranked very far down by effect on probabilities while having negligible effects on the coefficients. On the other hand we also find that the UAE, Luxembourg, Bahrain and Finland have some of the greatest effects on probabilities. These countries had very low effects on coefficients, yet somehow affect other countries' predicted probabilities almost twice as much as India, which had a great effect on coefficients.

The three measures, effect on coefficients, effect on predicted probabilities, and effect on

PR AUC, appear to have some correlation but also display a great degree of disagreement. The disagreement is likely to be amplified by the complex structure of the research design. The random effects and numerous neighborhood variables are likely to create various dynamics that are hard to predict, and which are not necessarily logical at first glance.

Going forward I will compare descriptive statistics to look for any differences between outliers on either side and the remaining group of countries. I go on to examine whether the countries are outliers simply because their own predicted probabilities in the control simulation do not match observations. I then examine how the countries affect others by looking at their effect on predicted probabilities, effect on coefficients and how they affect their neighborhoods. I will attempt to explain any country whose effect cannot be explained by either its own predicted probabilities or effect on coefficients by examining its neighborhood. If there is an effect on others that cannot be accounted for through these channels, a last possible way the countries could effect others is the effect that a country has on the multilevel model that the random effects are taken from. This effect is harder to track, as it isn't enough to examine the difference in the two random effect variables' coefficients. Each country is also given values on the variables, and tracking the change in these values for 161 countries over 161 models and for two variables is an expansive task.

As the list of countries is large and there are multiple pathways, I will not be going into the full details of each. Countries that can be explained by the first step will receive less attention in the next.

4.3 Group attributes

In this section I will go through some basic descriptive statistics, comparing destructive and reinforcing outliers with the other countries in the dataset. There are 16 countries in the destructive group, 15 in the reinforcing group, and 146 in the remaining group. Tables providing descriptive statistics for some variables are provided in Appendix A.4.2. The outliers have an average of 46.6 years of data, compared with 56 for reinforcing and 53.6 years for the others. The destructives are slightly younger nations, with the average time in independence being 16 years shorter than the other groups. The destructive outliers have roughly the same proportion of major conflicts as the main group, 5% compared to 4%. They do however have fewer minor conflicts. 7% of the destructive country units are minor conflicts compared to 10% in the remaining countries. The reinforcing outliers are clearly more conflict prone, with 20% of the units in minor conflict and 11% in major conflict.

The destructive outliers have shorter peaceful periods than the other groups. As they also experience less conflict, they would appear to have their conflict years interspersed with

peace years, while the main group has longer continuous periods of peace and conflict. The reinforcing countries have even shorter periods of peace, but this could be explained by the much heavier presence of conflict.

Population and GDP per capita characteristics for the groups appear similar. The populations of the destructive outliers are both approximately 1 million greater on average than the main group, with the reinforcing countries 2.4 million over this. The destructives are slightly wealthier on average, while the reinforcing countries are poorer than the main group by a small margin. The neighborhood variable ranks the reinforcing group as the wealthiest, closely followed by the destructives and then the main group. The differences in logged numbers are miniscule, which means they have little impact on calculations of predicted probabilities.

The Polity variables are hard to interpret from the descriptives. The destructive outliers are slightly more democratic on average, and have lower average levels of anocracy than the other groups. The reinforcing countries appear similar to the main group, yet there are some differences. The top line of Figure **??** shows histograms of the different groups' distributions on the Polity variable. The destructive countries appear to have no real pattern, although there is a slightly denser area around -5 on the scale. The reinforcing group has similar grouping tendencies around very weak autocratic values, but there are also groups at either end of the scale. The reinforcing group is somewhat concentrated at a point slightly left of the middle, but also has countries at either extreme. The lack of a clear pattern in the destructive group and the low N of both outlier groups makes it hard to conclude whether there is a connection between the Polity variable and effect on model performance. The reinforcing group hints at the model being better at handling either extremes or slightly autocratic anocracies, but as the destructive group has similarities such a conclusion is tenuous.

Other variables, such as ncts0, ltsnc and nb_TSRC_5 shows miniscule differences, if any, between groups. This is an indication that they are behaving as theoretically expected. The most interesting difference between groups is in conflict occurrence and length of peaceful periods. The destructive countries are more peaceful, with conflicts spread out over time. The reinforcing units are conflict prone with even shorter peaceful periods. The conflict history must be examined in more detail to uncover whether this is due to a problem with the model or with the conflict definition.

4.4 Predicted and observed values

To get a more intuitive understanding of the conflict history, we must look at the data in another way. Figure 4.4 gives a more intuitive view of the conflict occurrences than

summary tables. We can see that seven countries, Canada, Cyprus, Austria, Tanzania, Qatar, Jordan and Slovenia are without any conflict. The rest vary from a single conflict, as in Macedonia, to nearly only conflict, as in Angola. As expected, the countries have short periods of conflict interrupting mostly peaceful timelines. Angola and the DRC are exceptions, with especially Angola seeing high levels of conflict. Already it is clear that some countries are poorly predicted due to the way I have delimited estimation and evaluation periods. Two years into the evaluation period, Angola ends its 27 year conflict streak and becomes more peaceful. The model will need time to adapt to such a change, and is hampered further by the conflict recurring every few years.

The majority of conflicts occur in the estimation period, with Mali being the only country to experience more conflict as of 2001 than before. Mali is mostly peaceful, but the limited evaluation period sees more conflict than peace, putting Mali in an opposite position of Angola. We can also see that for this data, having a gap between estimation and evaluation could possibly give the model an advantage rather than a disadvantage. Apart from China, each country has a more uniform distribution after 2006. For example, should the model predict only peace for Macedonia then the limited set would be perfectly predicted, as opposed to the full period that contains a conflict. A model predicting only conflict for Mali could also possibly benefit from using the limited set, as the removal of five peace years would lead to a great reduction in false positives. Though, as we shall soon see, Mali is predicted as continued peace, and limiting the evaluation period only damages its results further.

The reinforcing group is presented in Figure 4.5. As expected we see much more conflict, but there are also perfectly peaceful countries. Uganda, Sudan, Ethiopia and Turkey are all almost entirely conflict stricken after 1980. The UK, Cambodia and Indonesia also have long periods of conflict, but these cease towards the end of the period. Mexico stands out as it has conflict, but very little. Yemen also has less than the other conflict stricken countries in the group, but its conflicts are most often grouped.

To see whether the countries' effect on the model's total predictive accuracy is due to the model having to predict the country's own outcomes, I examine the results of the control model. If the predicted probabilities of a country do not match observations, then it will have a negative effect simply due to the model having to predict their conflict histories. Figures 4.6, 4.7 and 4.8 show the predicted and observed values for Uzbekistan, Angola, Sierra Leone, the DRC, Djibouti, Mali, Syria, China, Macedonia, Tanzania, Sudan, Ethiopia, Uganda and Yemen. The remaining countries have no observed conflicts and have low predicted probabilities. Austria, Canada, Qatar and Slovenia vary between perfect zeroes and 0.6% predicted probability. Cyprus barely exceeds 1% at its maximum, while Spain and Lithuania vary between 1% and 2%. Jordan, with 1% to 3.3%, is the most at

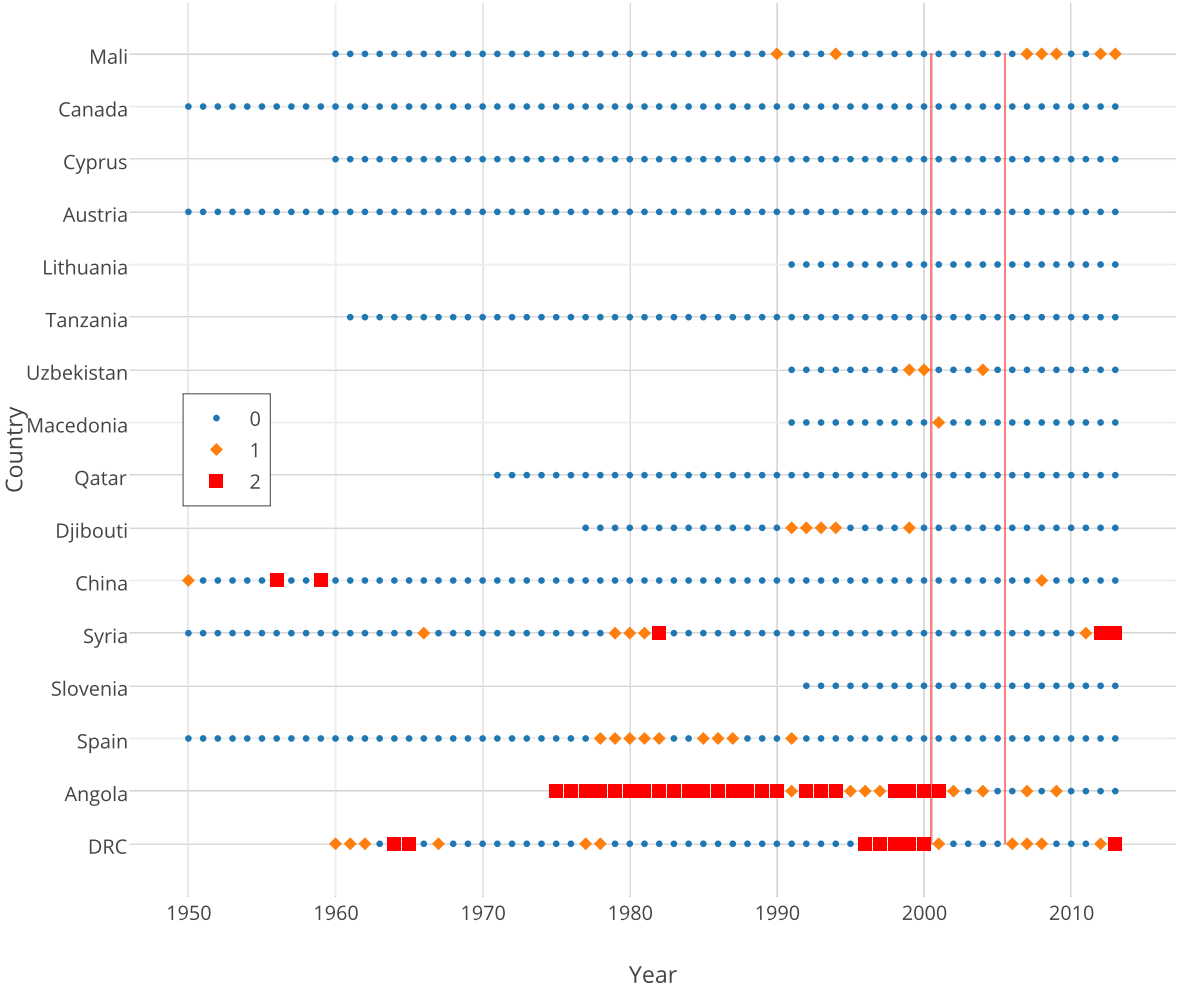


Figure 4.4: Conflicts over time for the destructive group. Blue dots represent peace, yellow diamonds are minor conflicts and the red squares are major conflicts. The first vertical line marks the cut point between the estimation and evaluation period, and the second line marks the start of the limited evaluation period.

risk. The average predicted probability for all units in the control set is 15.1% for the full period and 14.5% for the limited period.

The first three countries of Figure 4.6 are the destructive countries with the most extreme predicted probabilities. Uzbekistan has a single minor conflict, while the DRC and Angola have several minor conflicts and one major each. Uzbekistan's high starting values are related to the two consecutive conflicts prior to the evaluation period. The neighboring conflict variable is also active for all years except 2002. A high population and middling GDP also contributes to high estimates, which go down over time as conflict history variables change from recent conflict to lasting peace. Just as the predicted values fall, a minor

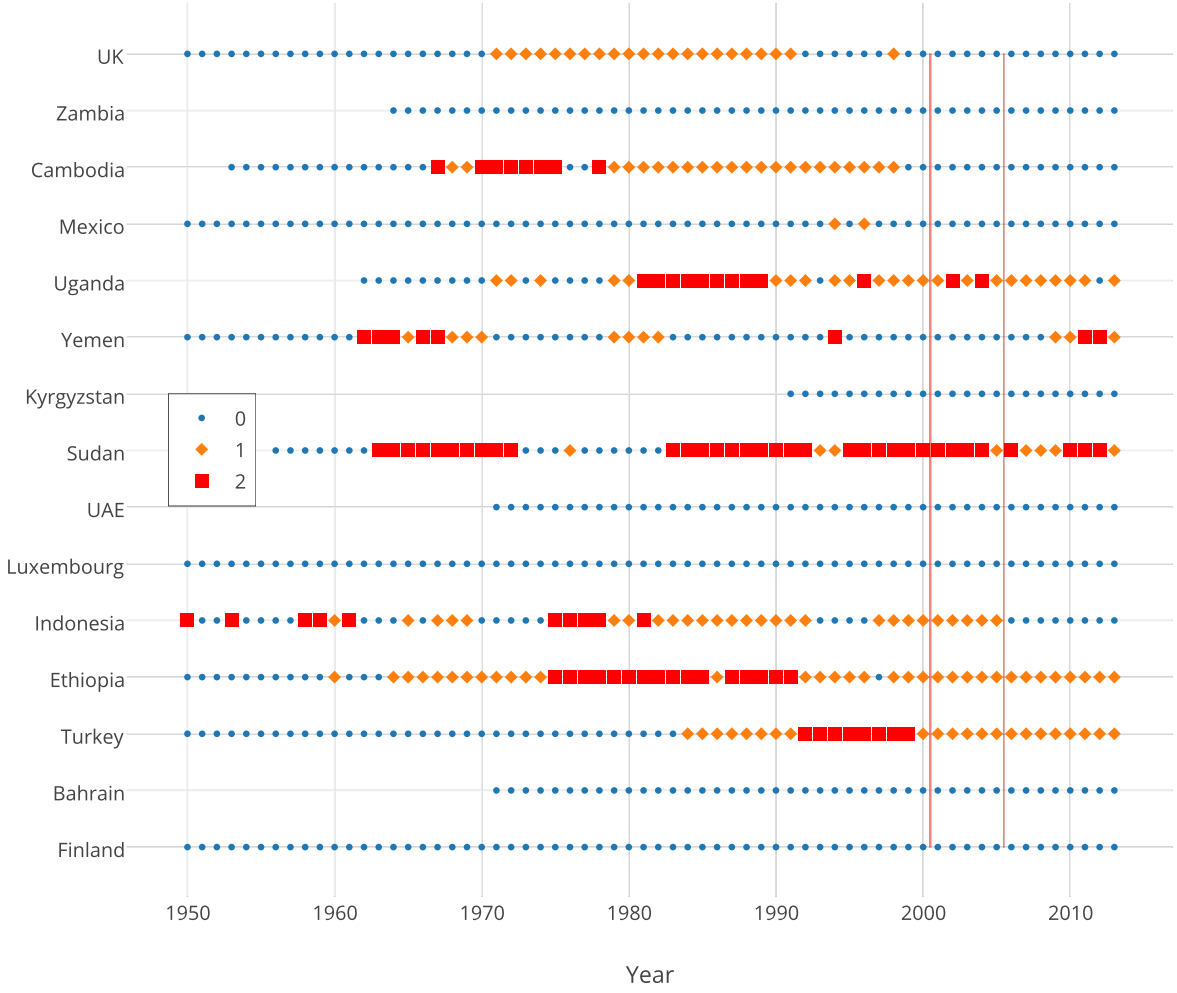


Figure 4.5: Conflicts over time for the reinforcing group. Blue dots represent peace, yellow diamonds are minor conflicts and the red squares are major conflicts. The first vertical line marks the cut point between the estimation and evaluation period, and the second line marks the start of the limited evaluation period.

conflict occurs, after which the curve start to flatten. Uzbekistan was a predictive outlier when evaluating on the whole period, however the effect on model precision is reversed when using the limited period. As we can see from Figure 4.6a, there are no conflicts after 2006. The predicted probabilities are lower in this period, but they remain at a level above the global average.

Angola is the third most destructive country by effect on PR AUC using the entire evaluation period, and the third most destructive using the limited period. Figure 4.6b shows the predicted probabilities dropping from over 90% to 60% over the full period. The constant drop is due to the steady increase in GDP and neighborhood GDP, which

counteracts the increase in population and time in independence. The Polity interactions and neighboring conflicts weaken the impact of the conflict variables, which is why there are no spikes after conflicts. The recurring conflicts keeps the time in peace variable from building up, which is why the probabilities remain high. These rapid transitions are problematic, but may also be the result of the chosen conflict threshold. A closer look reveals that there were 20 battle related deaths in 2003, meaning it is coded as peace despite there being a clear case for arguing that the conflict was merely in a lull in this year[2]. There are also a small number of battle deaths reported for several of the other peace years, while the conflict in 2007 is at exactly 25 battle related deaths. Using a country-year based dichotomous conflict variable based only in battle related deaths becomes problematic in these circumstances, as the model has no chance of reacting to a situation that is moving up and down at the threshold.

Two of the five conflicts remain when limiting the evaluation period, along with six peace years. Angola had experienced only conflict prior to the evaluation period, and the model continues to predict high probabilities in the evaluation period. This gives the model an advantage when using the full period, as three of the first five years are conflicts. Skipping these years leaves the model predicting more conflict for a period that is considerably more peaceful than previously. The average predicted probability to average conflict occurrence ratio is much higher in the limited period. For the whole period, the average predicted probability is 75% and occurrence is 38%. This equals a $\frac{.75}{5/3} = 2$ ratio of predicted conflicts to observed conflicts. For the limited period the predictions are 66% and the occurrence 25%, resulting in a ratio of $\frac{.66}{2/8} = 2.6$. This means that the model predicts 2 conflicts in Angola for every observed conflict using the whole set, and 2.6 using the limited set.

The Democratic Republic of Congo displays much the same characteristics as Angola. It has five years of continuous conflict before the evaluation period, and has very high predicted probabilities that are reduced over time. Six of the thirteen years are conflict years, the most of all the outliers. There is little economic growth to account for the fall in probabilities, but the population growth is also very low. As with Angola, the neighboring conflict weakens the impact of minor conflicts.

Djibouti does not have the same extreme predicted values as the previous countries. The first year is very high, followed by a sharp drop that stabilizes at around 20%. The initial high values are the result of a conflict prior to the evaluation period, and the sheer drop is attributed to the sharp rise in the time since conflict variable and its GDP interaction.

Mali has five conflict years but low predicted probabilities. Mali has two conflicts in the estimation period, but these are separated from both each other and the evaluation period

[2]All battle death numbers from country-years coded as peace are taken from the UCDP GED (Sundberg et al., 2010)

by several years. There is no increase in predicted probability leading up to the conflict in 2007, and only after three years of conflict is there a marked increase in predicted risk. This hints at a problem with the lagging of variables which will be explored further later. The constant presence of neighboring conflict reduces the impact of the minor conflicts, and the probabilities remain at a low level even when the conflict variables take effect. Being a small and poor country one would expect much higher probabilities, especially considering the neighboring conflict. The time since previous conflict variable would keep the estimations low, but the increase after the first domestic conflict occurrence should be greater than what we observe.

Syria experiences an almost flat probability curve over the period, with a long period of peace being interrupted by a sharp conflict escalation in the last three years. The model does not react to the increased conflict level even after two conflict years.[3] Being a moderately strong autocracy with a long period of peace keeps the probabilities very low, but the outbreak of conflict should give a marked increase. Syrias conflicts are all in the 2006-2013 period, yet Syria is an outlier when using the whole period but not the limited. This is somewhat counter-intuitive, as the peace years with a low predicted probability should help increase the country's overall predictive power.

Figure 4.7 shows the remaining destructive outliers. China has a completely flat risk curve, with single occurrence mid-period. The single conflict does not lead to a surge in predicted probabilities, despite the resetting of a very high value on the time since conflict variable. The random effect variables, coupled with strong economic growth, appear to keep the probabilities low despite a high population.

Macedonia also has low probabilities for the whole period, with a small increase for 2004, and a single experienced conflict in the first year. The conflict results in a small spike, albeit too late, but the level quickly returns to normal as the time since peace variables return to positive values. Tanzania has no observed conflict, but sees a steady rise in predicted probabilities through the period. This is the result of growth in the population, neighborhood time in conflict, and time since independence somehow outmatching the time since conflict. The baseline is above the minimum due to a high population and low GDP combined with an anocratic Polity score.

The remaining countries are among the countries on the other end of the scale when it comes to effect on model PR AUC. If they are dropped from the model, then the PR AUC will decrease, meaning their presence strengthens the model. Cambodia has predicted values that increase over time despite there being no conflict. This could be because of neighboring conflict, or changes in other variables. The predicted values exceed those of Tanzania, and yet Cambodia has a positive effect on model PR AUC while Tanzania has a

[3]This increases the suspicion that the simulator does not handle the lagging of conflict history properly.

negative effect. Indonesia has a very high predicted risk for the first years, which then drops rapidly towards low values. This coincides with conflicts occurring the first five years, with peace for the rest of the period. This apparently quick reaction to a halt in conflict, with an almost perfect separation of conflict and peace by predicted probabilities, is probably as good as a structural model can get. The sudden drop is helped by Indonesia's moderately democratic Polity score, which amplifies the effect of the conflict history variables. The predicted probabilities respond very well to the change in observed conflict, and correcting for an assumed erroneous lag the results become even better.

Turkey has minor conflict occurrences in every year of the evaluation period. The predicted probabilities are high, but go down over the period. They are kept up by the combination of conflict and Polity, as well as Turkey's random effects. A curiousity is that Turkey's effect on model PR AUC increases in the limited period, where predictions are worse. While high, the probabilities drop below 70% by the end of the period. This puts them below many peace units in other countries examined here. Uganda follows the same pattern, and both countries appear to see little change over the period apart from a slow growth in both populations and GDP. Uganda especially fits the model by being poor and conflict stricken in a poor and conflict stricken neighborhood.

Sudan and Ethiopia have the same amount of conflict as Turkey, and also start at very high levels of predictied conflict probability. For them the predictions remain above 80% for the entire period. Ethiopia and Sudan are also in the same position as Uganda when it comes to GDP, but they have even greater populations.

Yemen is a country where the model appears to predict conflict before its first occurrence. The predicted probabilities rise steadily before the outbreak, reaching over 30% at the first conflict unit. Curiously there is no marked increase in probabilities after conflict onset. Yemen has a rich neighborhood and stable period that keeps the probabilities down leading up to the evaluation period, but the underlying poverty, high population for a country of its size and anocratic regime makes it conflict prone. This is reflected in the random effects that also pull Yemen towards higher probabilities.

The countries from Figure 4.6 have predictions that alone could account for the countries being outliers. Whether the same is true for the countries in Figure 4.7 is not as certain. While Syria has three unpredicted conflicts, the remaining three are more precise. China and Macedonia miss a single conflict each, and Tanzania has a relatively low level of predicted probabilty and no conflict. These are more likely to indirectly affect the predictive power, as their direct effects appear to be small. The remaining outliers with a negative effect on PR AUC have no experienced or predicted conflict, which should leave them predicting perfectly. Any effect on AUC should therefore be through indirect effect on other countries' predictions. The two positive countries shown here are more uncertain.

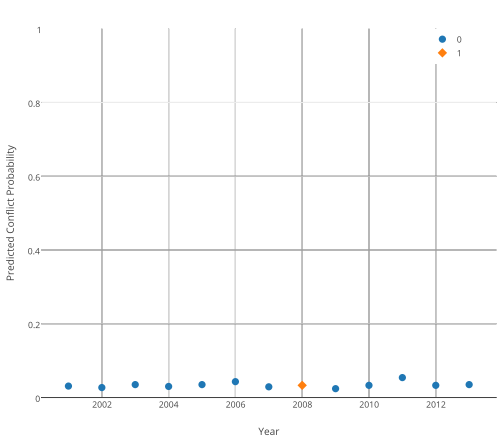
By predicted results alone, Cambodia should be much more destructive than Tanzania, yet it is on the opposite end of the scale. Indonesia appears to be the model country, with predictions rapidly changing in a pattern that follows conflict occurrence. Compared with the negative outliers Indonesia is clearly better at distinguishing conflict years from peace. There are several years with very poor predictions after the transition to peace, but the predictions adapt relatively fast compared with Sierra Leone. The other countries that have a great positive effect on model PR are countries with low predicted probabilities that experience no conflict, and countries with very high probabilities that experience only conflict, or mostly conflict with single peace years. The latter have the same error rate as China or Macedonia, even with less extreme values in the correct direction. This shows that the PR AUC rewards correct predictions of positive outcomes more than correct negative outcomes.

The observed conflict pattern found in Angola, transitioning back and forth between conflict and peace, is problematic for the model to adapt to. Slow moving structural variables have no way of picking up such fluctuations unless the war is of a scale that it affects the GDP on a major scale. While this is true, it is more reasonable to look at the dependent variable in this case. Are the fluctuations in our coding of conflict based on sound reasoning? Looking more closely at the data, we find that the two coded conflicts in 2007 and 2009 both are recorded at 25 battle deaths, exactly the threshold to be counted as conflict. 2003 is coded as a peace year, but there are 21 recorded battle deaths, as well as numerous victims of one sided political violence that do not count towards the conflict status. As the same groups are involved, it is only our choice of time unit that creates breaks in conflict or peace. This is an underlying problem that arises when using a binary definition of conflict with a strict threshold; any country transitioning from one side to the other will cause issues if the transition is gradual.

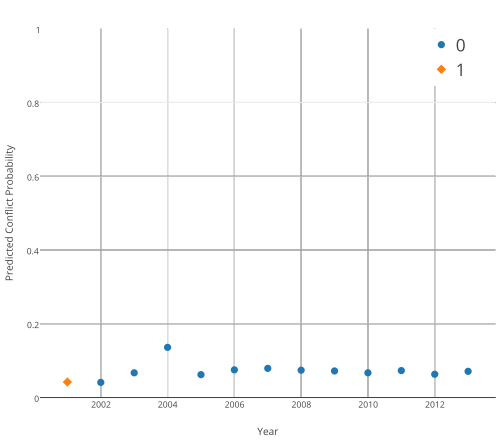
(a) Uzbekistan (b) Angola (c) DRC

(d) Djibouti (e) Mali (f) Syria

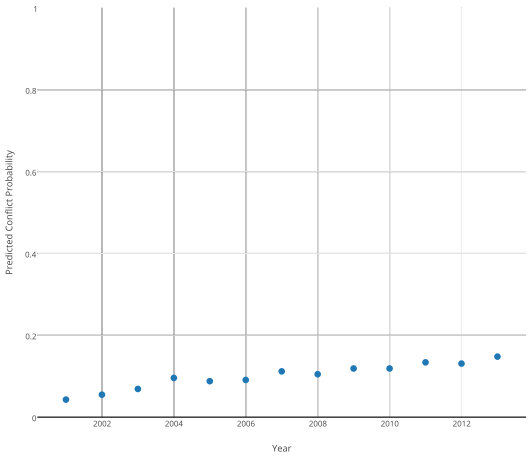
Figure 4.6: Predicted conflict probabilities versus observed conflict over evaluation period.



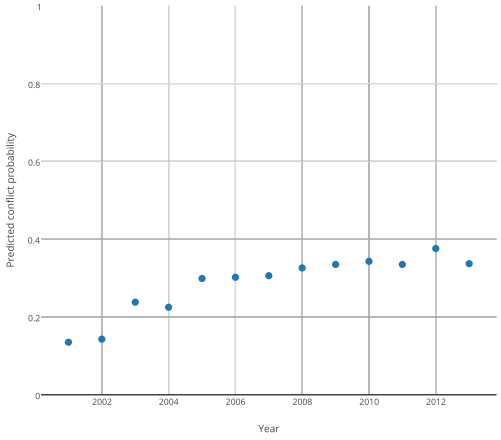
(a) China



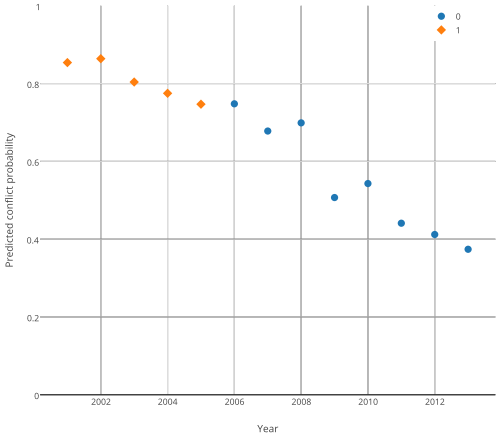
(b) Macedonia



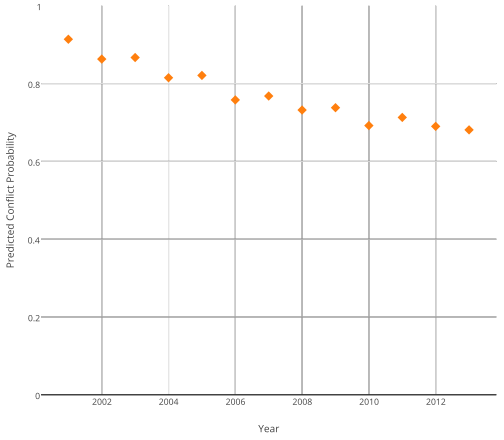
(c) Tanzania



(d) Cambodia

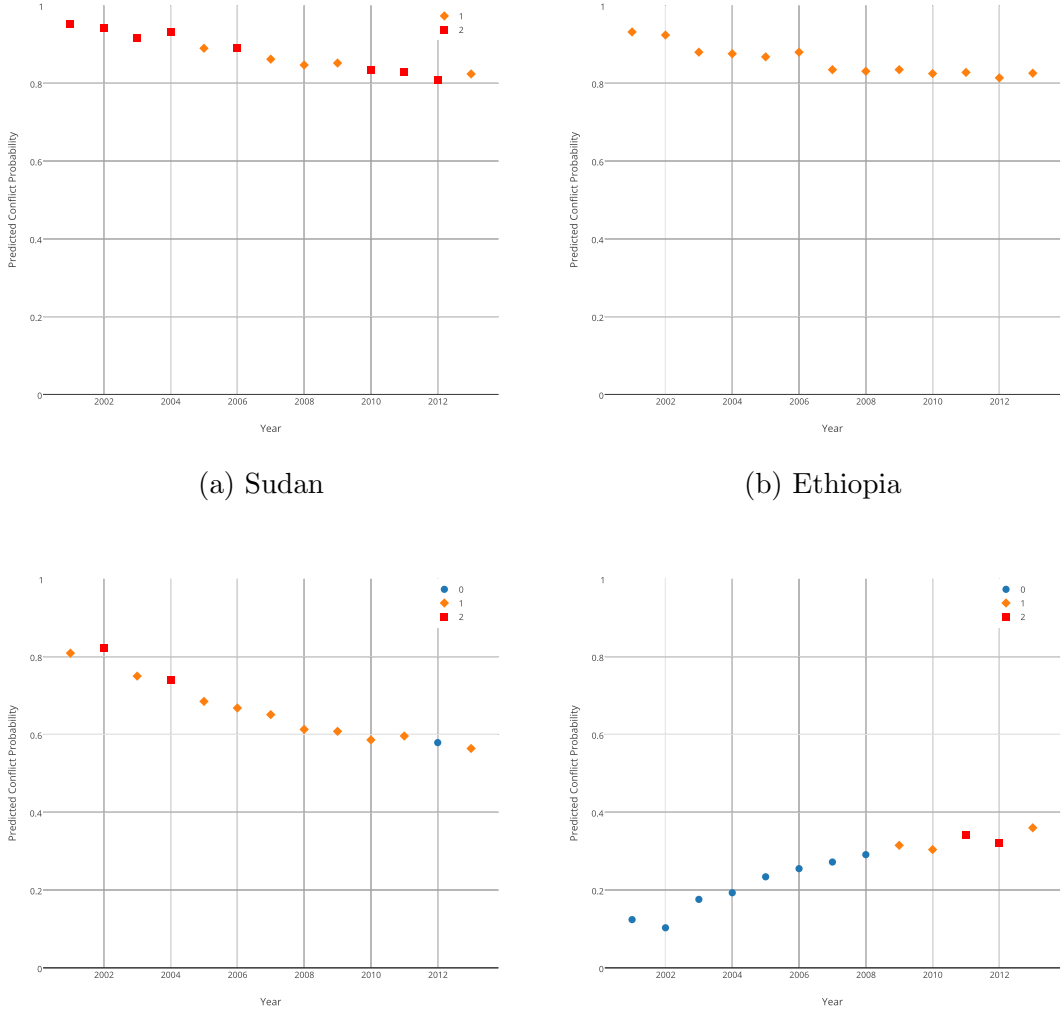


(e) Indonesia



(f) Turkey

Figure 4.7: Predicted conflict probabilities versus observed conflict over evaluation period.



(a) Sudan (b) Ethiopia

(c) Uganda (d) Yemen

Figure 4.8: Predicted conflict probabilities versus observed conflict over evaluation period.

4.5 Indirect effects through coefficient effects

A large portion of the indirect effect of units is likely to be through their effect on the model's coefficients. The effect of different variables vary depending on the units involved in the model's estimation, and this in turn affects the predicted probabilities of all units involved. As we have seen previously, many of the predictive outliers have very little effect on coefficients. A.3 Figures B.1 and B.2 shows the distributions of country effects on key variables with outliers on individual variables highlighted. Recurring countries found in many of the figures are Iraq, Indonesia, the USA, France and Pakistan. This is in line with the previous test of effect on coefficients.

From the destructive outlier group, only China, Angola, Syria and the DRC are among

those that clearly disturb the model. China is seen to reduce the effects of conflict history, as well as increasing the effect of Polity and reducing that of squared Polity. Syria has less influence, but the effects are on the important conflict history variables. Angola causes large increases in many conflict variables as well as the neighboring regime change variable. It also drastically reduces the effect of GDP.

Figures B.3 and B.4 highlight the predictive outliers, and we see that Spain is also outside the main group on several variables. Spain has a positive effect on the neighboring conflict variable, *nc*, and the related *ncts*0 variable. This is a very common variable, and would explain why Spain has a large effect on PR despite not having a large effect on coefficients in total. Spain also negatively effects the squared Polity variable, a powerful and theoretically crucial variable.

The other destructive countries are however not of great influence on any variable. Tanzania and Sierra Leone are seen as distinct from the group, but not by a noteworthy margin. The remaining predictive outliers are in the main core of units, with weak effects. This, as well as the presence of a myriad of extreme outliers that are not on the list of predictive outliers, indicates that coefficients are not as important to the effect on other units as first expected.

Further, there is no uniformity among the destructive countries when it comes to effect on the variables. While the DRC and China have some similarities, they are often on opposite ends. Only on $c2$, the effect of a major conflict the previous year, does there appear to be a uniformly negative effect.

Indonesia and Cambodia are among the countries that have great effects on coefficients. These show that units affecting the coefficients is not the same as being detrimental to predictive power. Indonesia has an effect on a great deal of coefficients, with the strongest being on several conflict history variables, GDP, and the neighborhood regime change variables. Indonesias long conflict stretches means it increases the power of conflict in the previous year, which could create better results overall as we have seen that the increase in probabilities in response to a conflict is often too small. Cambodia affects the Polity and squared Polity variables the most. While the unaltered Polity variable has a relatively weak effect, the squared variable has five times its potential for major conflict outcomes and ten times the potential for minor conflicts. While Cambodia has a large effect on coefficients overall, it is likely that it is this effect on the Polity variables that is crucial in making it an influential unit. Yemen is also clearly visible as having a large effect on conflict history variables, reducing the effect of $c1, c2$ and nc on minor conflicts. It also affects *ncts*0 negatively for both outcomes. This is somewhat hard to understand when looking at Yemen's conflict history, as it often has minor conflicts following both minor and major conflicts. The effect on ltsc0 is easier, as Yemen does have longer periods of

peace between conflict instances.

4.6 Indirect effects through neighborhoods

This leaves several countries that have no effect on any coefficients, with perfect predictions, that still have strong negative effects on the model's overall PR AUC. Austria and Canada for example both appear in the top ten list for the limited period. I will now examine whether these countries have any effect on their neighbors by influencing their neighborhood variables' values. There are only two neighborhood variables in the model, and these cover GDP and regime change. The regime change variable records any regime change occurring in the neighborhood the past five years, and any country without a regime change will not change the neighbors if removed. The GDP variable averages the neighboring countries logged GDP per capita.

The removal of Canada would for example lead to a sharp reduction in the US' value on the neighborhood GDP variable, as Mexico with its much lower GDP would be the only remaining neighbor. Comparing the predicted probabilities of the US from the control to the Canada drop reveals that removing Canada has no clear effect. Some years see an increase, including 2001, when the US experienced a conflict. Other years see a reduction in predicted probability when Canada is not present. In conclusion it would appear that Canada has no concrete effect on predictions as a neighbor.

Lithuania is very similar to Estonia and Latvia, and yet only Lithuania appears to have a negative effect on PR. There is no apparent reason why Lithuania should stand out. All three share borders with Russia, and no other countries with conflict. All three are perfectly predicted, peaceful countries with no strong effects on coefficients. Its border with Poland cannot account for the great differences, and since it should have no greater effect on Russia than its Baltic neighbors it is hard to explain Lithuania's effect on PR through neighborhood effects.

Austria is surrounded by perfectly predicted and peaceful neighbors, and while Austria is richer than some of its neighbors, they all have at least three other neighbors. This dilutes the small negative effect that Austria might have. Austria has not had any regime changes in the evaluation period, leaving this variable unchanged in its neighbor if Austria is removed. As all of Austria's neighbors are peaceful, a negative effect on neighbors predictive precision would have to be through a negative effect on the neighborhood GDP per capita. These would increase the neighbors predicted probabilities, making the predictions less accurate. However, none of Austria's neighbors has a neighborhood that is noticeably wealthier if Austria is removed.

Cyprus is an island, and as such has no neighbors that it can affect. Qatar only borders

Saudi Arabia, and has an extreme GDP. This effect is diluted, as Saudi Arabia has eight neighboring states. As Saudi Arabia also has perfect predictions for the whole period, it is unlikely that neighborhood effects can account for Qatar being destructive.

The effect of countries on their neighborhood gets more complicated with many neighbors and differing conflict levels among these. Angola is richer by far than most countries in Central Africa, and as such it has a positive effect on the neighborhood GDP variable of its neighbors. One neighbor is the DRC, with a high conflict level. The average predicted probability of the DRC is increased by 2.5% when Angola is removed, which could account for some of the difference in PR AUC. Angola also has a negative effect on other neighbors that have no conflict, and as such its removal would also be destructive for the model's PR AUC. Attempting to calculate the exact effect on the model's PR or ROC AUC caused by neighborhood variables in this fashion would be an extremely intricate task.

4.7 Summary

There are few common denominators for the destructive outlier group. Disregarding Angola, the DRC and the perfectly predicted countries, we find that one repeating feature is short conflict spells interrupted by short peaceful periods. This is hard for the model to quickly adapt to, and results in poor precision. The first years of conflict are often completely unexpected, and when they fail to reappear the next year there is the added penalty of the conflict history variables giving increased risk predictions for a peace year.

Angola and the DRC could be outliers merely because of their predictions, which contain several misses due to the patchy conflict history in the evaluation period. This is a symptom of their conflicts being coded as peace when they could arguably be described as dormant. As I have also pointed out there is also a problem with the casualty numbers being near the threshold in the case of Angola. The two also have strong effects on a number of variables. Angola is reasonably wealthy by regional standards, something that should have reduced conflict levels.

For Mali, the model fails to respond quickly and forcefully to the onset of conflict. The first year comes as a surprise, but the conflict variables are not strong enough to create the needed increase in probabilities to predict the following years of conflict. A quicker response from the conflict history variables could have helped, but the problem lies more with the power of the effect than the response time. The predictions for Syria do not respond at all to conflict, although this is probably due to a faulty lag. It is still reasonable to assume that the effect of the conflict variables would, as in Mali, be to weak to predict the continuing conflict in Syria.

Uzbekistan appears to be predicted rather well, with a steep downward curve after

conflicts just prior to the evaluation period. There are however still many false positive years, which is probably the reason behind its detrimental effect on the model.

China is very well predicted apart from a single conflict year. China is a special case, having an enormous population and low GDP, yet barely any conflict. The random effects push China to a reasonable conflict probability level despite its poor starting point, but China's effect on several variables gives it influence over other units' predictions, and this influence manifests itself negatively.

Spain is present because of its conflict history, where all its conflict take place after Spain has transitioned to an almost perfect democracy. This causes Spain to affect the Polity variables enough to cause disturbances in other countries.

Djibouti and Tanzania could be explained by their high probabilities and lack of conflict. Djibouti has had a conflict history that the model spends the first years of the conflict period adapting to. This brings Djibouti down to a lower probability level, but it still stays at a suspiciously high level for a country without conflict. This is likely due to Djibouti being in a very poor and conflict stricken neighbourhood, driving its own predictions up. Tanzania also suffers from a problematic neighborhood, and is itself a poor nation. Their lack of conflict therefore goes against some of the core principles of the theory and the model.

The results appear to indicate that PR does indeed punish false positives quite harshly. Some countries have poor predictions that will lead their inclusion in the data to have a negative effect on the model's overall PR. As an example, the conflict risk predictions for the DRC are lower over time as conflict occurrence increases. This results in the years with conflict occurrences having on average 6% lower predicted risk than the peace years. —

While many destructive outliers can be explained, and the reinforcing countries could simply be the result of their fit to the model, there is a surprising amount of countries with no apparent reason to be on the list of destructive countries. Canada, Cyprus, Austria, Lithuania, Qatar and Slovenia are all almost perfectly predicted. They have virtually no effect on coefficients. While Lithuania might have some effect on Russia, the others have no detrimental effects on their neighborhood of any consequence. It should be worth mentioning again that the effect on coefficients was measured from a control model, and not the actual simulations, which means there could be effects on coefficients in the simulations that I have not uncovered here. However, the most likely reason they stand out is that they somehow affect the multilevel model and the random effects estimations.

It is important to note that the effect on coefficients is not straight forward to interpret. Reinforcing countries like Indonesia and Cambodia often appear very close to destructive countries like the DRC and China in the plots of effects on coefficients. This makes it hard to conclude whether the effect is positive or negative, as two countries with the same effect

on a variable end up having opposite effects on predictive power.

Chapter 5

Brier outliers

In this chapter I will repeat the process in the last chapter, replacing PR AUC with Brier scores when evaluating predictive power. This is done to compare how the measures differ in their responses to influential outliers, and see if the results support the findings in the previous chapter regarding outliers and their attributes.

To reiterate, I will first identify the groups of destructive and reinforcing outliers by measuring each country's effect on predictive power when dropped from the dataset. I then examine the groups, comparing their attributes, conflict history and individual performance. I then examine to what degree they have indirect effects, and whether there are any countries than can only be explained through their effect on the random country effects.

While both measures are likely to have some of the same outliers, it is also to be expected that there are differences. As Brier judges predictions individually it is likely to not treat some of the PR outliers as harshly. Macedonia and Tanzania are examples of countries where the probabilities are essentially quite low and correct, but the countries end up being destructive outliers by PR nonetheless. This could be because their peace years are given higher probabilities than other countries' conflict years, thus making the predictions for Tanzania and Macedonia seem erroneous when compared to the entire model. Brier score should disregard this, as these low probabilities would be judged on their own merit rather than how they fit in with other countries' predictions.

I also display country distribution by effect on F-score. The F-score is also in use as a measure of predictive power (D'Orazio et al., 2011; Hegre et al., ming), and it is included for comparison and to illustrate how the measures differ, and I will not be discussing it to a great degree.

5.1 Outlier scores and groups

First off is a look at the bigger picture. Table 5.1 shows the worst performers by both scores. Several other countries that were found to be inexplicable destructive outliers by PR are also destructive by Brier, but they remain outside the top ten. Tables A.11 and A.12 show the full list of countries by effect on Brier score for both periods. Both measures score better using the full period than the limited, and by a decent margin. The differences in outlying units between periods are very small, with only three countries being different for Brier score and two countries for F-score. The two measures have few destructive outliers in common, with only Slovenia appearing in the lists for both Brier and F-score. Slovenia and Jordan are in fact the only destructive countries that F-score has in common with any of the other measures. The Brier score list shows a greater degree of convergence with PR. Angola, the DRC, Syria, Mali and Slovenia are shared as top ten countries with a destructive effect. Figure 5.1 shows that while there is little agreement between Brier and F-scores in the destructive end, the reinforcing countries in the lower left are shared. Among these we find Turkey and Finland, who were also among the countries with the most reinforcing effect for both PR and ROC AUC. The reinforcing group shows an even greater overlap between Brier and PR. Bahrain, the UAE, Luxembourg and Cambodia are among the most reinforcing for both measures. Without further study it is already clear that the Brier outliers are mostly countries one might expect to find. Only Poland and Slovenia stand out as completely unexpected, whereas the F-score outliers include Bulgaria, Japan, Slovenia and France. France is not inconceivable, but the other three are not those you would expect to find. Turkmenistan and Kazakhstan are also slightly odd, as they are peaceful dictatorships with relative wealth.

Table 5.1 shows the most destructive countries by effect on Brier score. The Brier results are clearly less affected by the choice of evaluation period, showing a greater degree of overlap. This is likely due to the fact that PR compares pooled results, and the units included in the limited group change the characteristics of the group in such way that the effect of individual predictions changes. Brier predictions are judged the same in either period, and the only change comes from which units are included in the country itself.

There is also a solid overlap with the outlier groups by PR, but at first glance the destructive countries appear to have fewer inexplicable members. Slovenia remains a curiosity. As I have shown previously, it has perfect predictions and does not stand out by its effect on others. The other perfectly predicted western democracies are gone, but Poland steps in as a replacement. The remaining countries are African and Asian countries with a history of conflict.

Comparing the scores between periods reveals that, as with the AUC measures, limiting the period leads to a decrease in accuracy. It also leads to increased effects and scattering

	Country	Brier
1	DRC*	-0.003
2	Angola*	-0.002
3	Somalia*	-0.002
4	CAR*	-0.002
5	Iran	-0.002
6	Pakistan*	-0.001
7	Poland*	-0.001
8	Sierra Leone	-0.001
9	Chad	-0.001
10	Syria*	-0.001

(a) 2001-2013 - Control: .055

	Country	Brier
1	Angola*	-0.003
2	Pakistan*	-0.003
3	Poland*	-0.002
4	Somalia*	-0.002
5	DRC*	-0.002
6	Mali	-0.002
7	Syria*	-0.002
8	Niger	-0.001
9	CAR*	-0.001
10	Slovenia	-0.001

(b) 2006-2013 - Control: .064

	Country	F-Score
1	Bulgaria*	0.021
2	Dominican Republic*	0.017
3	Jordan*	0.016
4	Kazakhstan*	0.015
5	Kenya*	0.013
6	Japan*	0.012
7	Slovenia*	0.010
8	Haiti*	0.010
9	Papua New Guinea	0.009
10	France	0.009

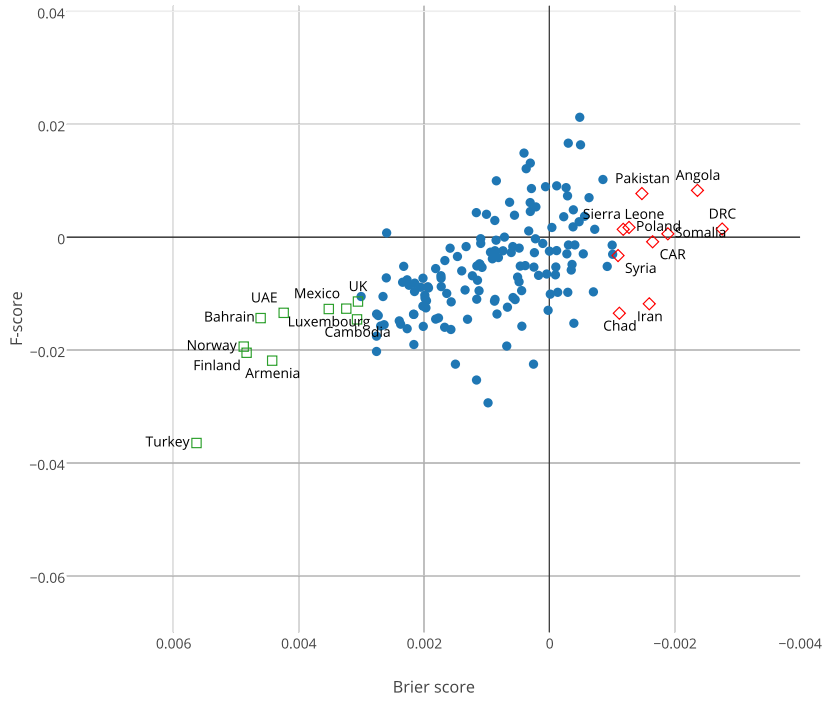
(c) 2001-2013 - Control: .758

	Country	F-Score
1	Bulgaria*	0.040
2	Nepal	0.029
3	Jordan*	0.028
4	Kazakhstan*	0.027
5	Haiti*	0.027
6	Dominican Republic*	0.026
7	Japan*	0.026
8	Kenya*	0.024
9	Slovenia*	0.023
10	Turkmenistan	0.023

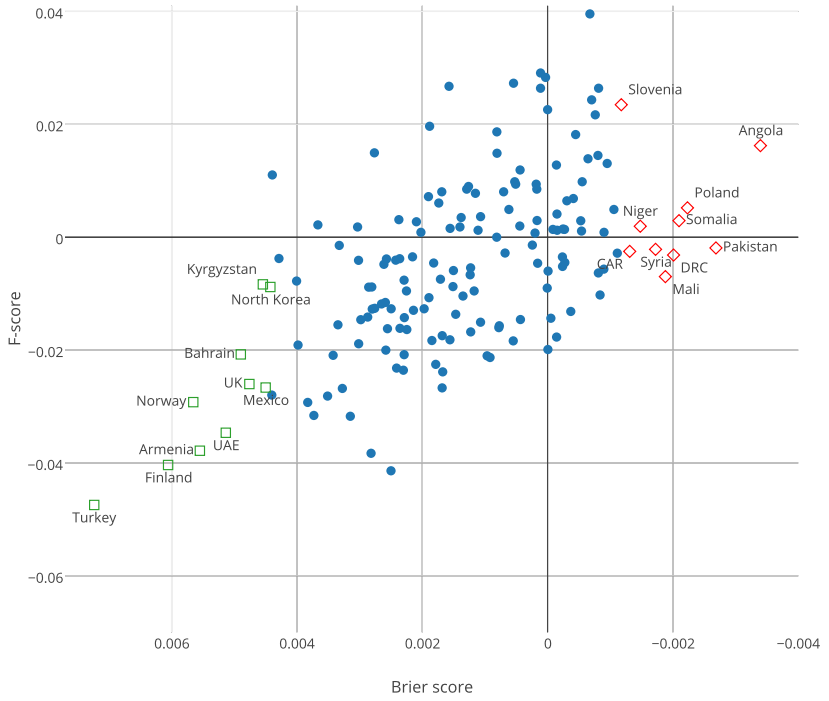
(d) 2006-2013 - Control: .717

Table 5.1: F score and Brier score differences from control. Asterixes mark those that appear in both periods.

of countries. While many countries appear to remain in the same positions relative to each other, the space between them is inflated.



(a) Period: 2001-2013. Control Brier 0.055. Control F: .758.
Interactive plot: `https://plot.ly/~msmidt/872`



(b) Period: 2006-2013. Control Brier: .064. Control F: .717.
Interactive plot: `https://plot.ly/~msmidt/884`

Figure 5.1: The difference in Brier and F-scores between country drops and the control case. Outliers marked in green and red. NOTE: Y and X axes are scaled differently to account for smaller absolute changes in Brier scores.

	Destructive	Reinforcing
1	DRC	Finland
2	Angola	Bahrain
3	Somalia	Turkey
4	Slovenia	Norway
5	Syria	Armenia
6	CAR	Luxembourg
7	Iran	UAE
8	Pakistan	North Korea
9	Poland	Kyrgyzstan
10	Sierra Leone	UK
11	Chad	Cambodia
12	Mali	Mexico
13	Niger	

Table 5.2: The outlier groups by effect on model accuracy.

Using the same selection criteria as with PR, all countries that appear in either period are selected as outliers. Table 5.2 shows the destructive and reinforcing groups. Among the reinforcing countries there are only three countries that did not appear as reinforcing for PR: Norway, Armenia and North Korea. These are all perfectly predicted, peaceful countries. Norway is obviously among the well fitting cases, with a small population, high GDP and strong democracy in a rich neighborhood. Norway happens to be the country with the greatest effect on predicted probabilities, and also the third smallest effect on coefficient estimates. These results seem somewhat contradictive, but again it is possible that Norway has an effect through the random effects. North Korea and Armenia have average effects on coefficients. While North Korea has a median effect on predicted probabilities, Armenia is second only to Norway.

The newcomers among the destructive countries are Somalia, CAR, Iran, Pakistan, Poland, Sierra Leone, Chad and Niger. Apart from Poland, these are all countries with well known conflict histories. Iran is high on the list of countries by effect on coefficients, with Chad not far behind. Pakistan and Somalia are also above the 400 mark, while Niger lurks just below. While these countries would be picked up by a test of effect on coefficients, Sierra Leone would probably not be. CAR would not have attracted any attention, and Poland barely has any effect. Sierra Leone is on the other hand high on the list by effect on predicted probabilities, with Chad being a bit over average. Niger is ninth least influential, while the rest are all on the middle of the scale.

5.2 Group attributes

While the analysis of descriptives on PR outlier groups revealed few differences beyond conflict history, the Brier groups show more contrast. There are 13 countries in the destructive group and 12 in the reinforcing, leaving 136 in the remaining group. Tables containing the full descriptive statistics by group on the variables discussed below are found in Chapter A.5.2.

As with PR, the destructives are still younger countries. For Brier score the difference is 23 years on average. Data availability is just over 53 years for all groups. The conflict pattern is reversed compared with the PR outliers, with the destructive group having almost twice the amount of minor conflicts compared to the other two groups. 19% of the destructive units are minor conflicts, while the reinforcing and remaining groups have 11% and 10%. Major conflicts are even more skewed, with 9% of the destructive country units compared to 2% among the reinforcing, and 5% among the remaining. The destructive group again has shorter periods of peace than the main group, and the reinforcing countries has the longest periods of continuous peace. While the short periods in the destructive unites could be explained by the sheer volume of conflict, this also looks like what was found in the last chapter, where units featuring quick transitions back and forth between peace and conflict were found to be destructive. The reinforcing countries' longer periods also supports this, and the topic will be explored further when I present the conflict history of the two outlier groups.

The destructive countries are situated in more conflict prone neighborhoods, as 68% of their units have neighboring conflicts, compared to the 40% and 43% of the other groups. The interaction between neighborhood conflicts and time since conflict interaction shows that the destructive group has a much lower average than the other groups. This means that the conflicts in the destructive group are more synchronized with neighboring conflicts. The *ncc1* and *ncc2* variables have a negative effect. This means the model assumes the effect of neighboring conflict weakens if the country has itself experienced conflict the previous year. It could appear that this assumption does not hold for the destructive outliers here, as there are many consecutive conflict years where the predictions underestimate the risk in the presence of neighboring conflict.

The destructive group has a somewhat higher population than the other two groups. A clear trend can be seen in the GDP variable, where the destructive countries are poorer on average than the main group, and the reinforcing group is richer. Both these variables have coefficients that should increase the probabilities of the outliers, which would mostly be beneficial. Their interactions are mostly pulling in the same direction, but it appears that population interacted with major conflict in the previous year slightly diminishes the overall risk of conflict.

Moving on to the Polity variables we find interesting differences. The destructive outliers are far less democratic and more anocratic than the others. The reinforcing countries are more democratic and less anocratic, while the remaining group is more evenly distributed with a slight democratic shift. This suggest that the model is better at handling countries at either extreme.

5.3 Predicted versus observed values

Brier has an advantage over PR when it comes to disentangling the effects of countries on model score. Unlike with PR AUC it is easy to calculate whether or not a country's Brier scores pull the model's overall Brier score up or down. The Brier score is averaged over all units regardless of data structure. By calculating the Brier score of the country and comparing it to that of the control model we see whether it has had a positive or negative effect. The control model Brier score for the whole period is 0.061, and 0.064 for the limited period. Any country whose own average is above this will have a negative effect on the model average, and so here it will be easier to determine to what degree each country's predictions are responsible, rather than their indirect effects.

Figures 5.2 and 5.3 present the conflict histories of the destructive and reinforcing groups. We can see that the destructive countries in Figure 5.2 do indeed have great deal of conflict. Poland and Slovenia stand out as the only countries without any conflict in the dataset. Mali, Niger, Syria and CAR have lower levels of conflict, all of it scattered in smaller groups along the timeline. Such grouping tendencies create problems for cross-validation. If only a single split between estimation and evaluation is used, the point in time that data is divided by dictates the results of the evaluation. When conflict periods coincide with the division of data, the evaluation can return results that are more pessimistic than is fair regarding model accuracy.

An example is CAR, that has all of its conflict in the evaluation period. With no conflict in the estimation set, CAR's particular attributes will not be taken into account in the model. When including random country effects these could also suppress the probabilities as the model is only concerned with keeping probabilities low in the estimation period, ignoring the conflict that awaits in evaluation. Mali is in a similar position, and as we have seen already the model fails to react to the sudden conflict surge there. Pakistan also appears to be experiencing the same problem, although it does see some conflict in the last decade of the estimation period. Sierra Leone has a continuous period of conflict that ends just into the evaluation period. The same is true for the DRC, but here the conflict reignites. Angola is again present with its many fluctuations. Iran, Chad and Somalia are all mostly conflict, but with a few peaceful years in and around the evaluation period.

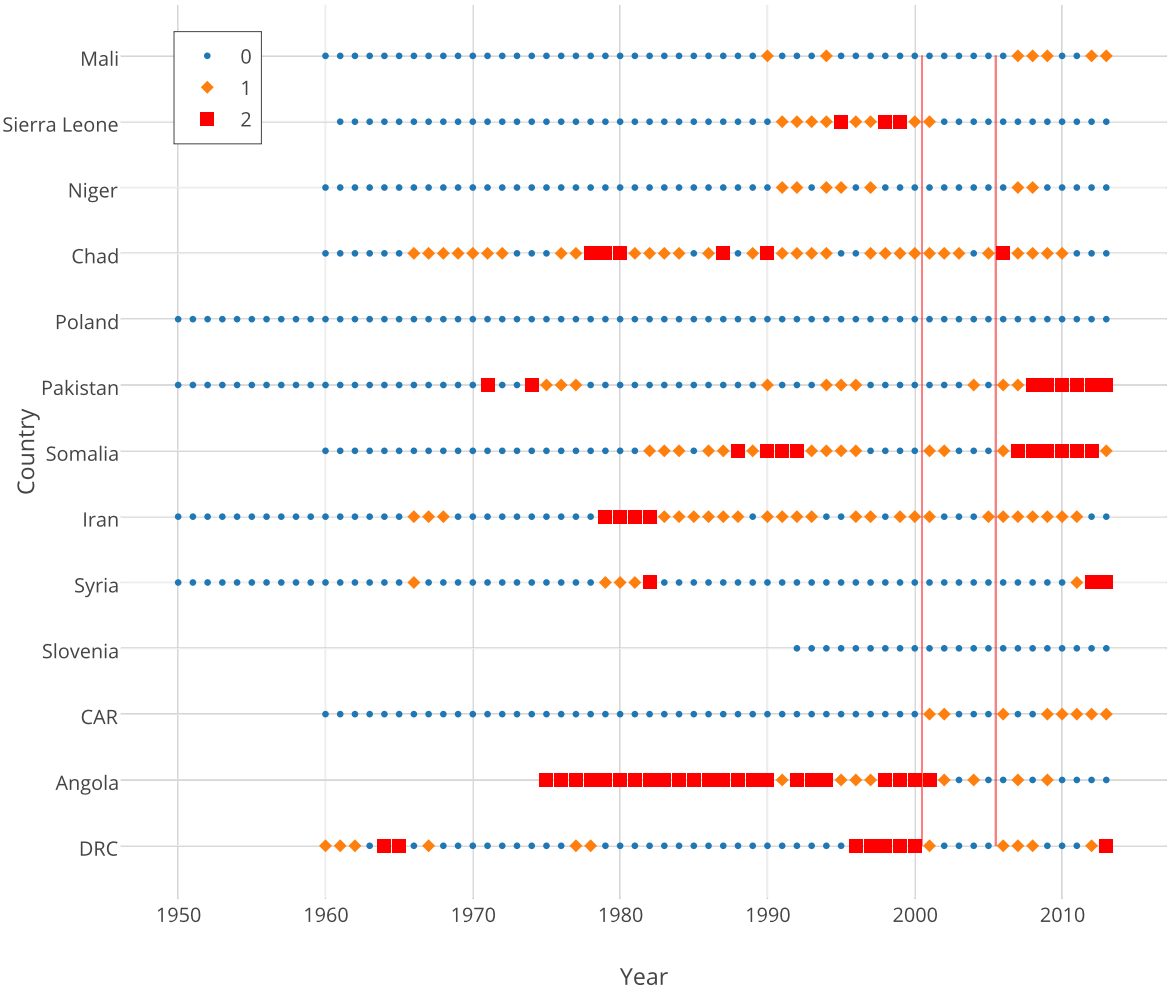


Figure 5.2: Conflicts over time for the destructive group. Blue dots represent peace, yellow diamonds are minor conflicts and the red squares are major conflicts. The first vertical line marks the cut point between the estimation and evaluation period, and the second line marks the start of the limited evaluation period.

Overall it would appear that the problem is again that there are abrupt changes in conflict status, often going back and forth between states, which creates problems when classifying.

The reinforcing countries do have a lot less conflict, and more importantly the conflicts occur in only four of the twelve countries. They are further clustered so that there are four major clusters and only three lone conflict years in the whole group. Mexico stands out with two lone conflict years, while the UK has one lone conflict outside its main conflict period. This lone conflict is also the results of the battle death threshold, as the Troubles continued in the period here coded as peace. Cambodia's conflicts are split by two peace years. These years happen to coincide with the Khmer Rouge genocide, which means the

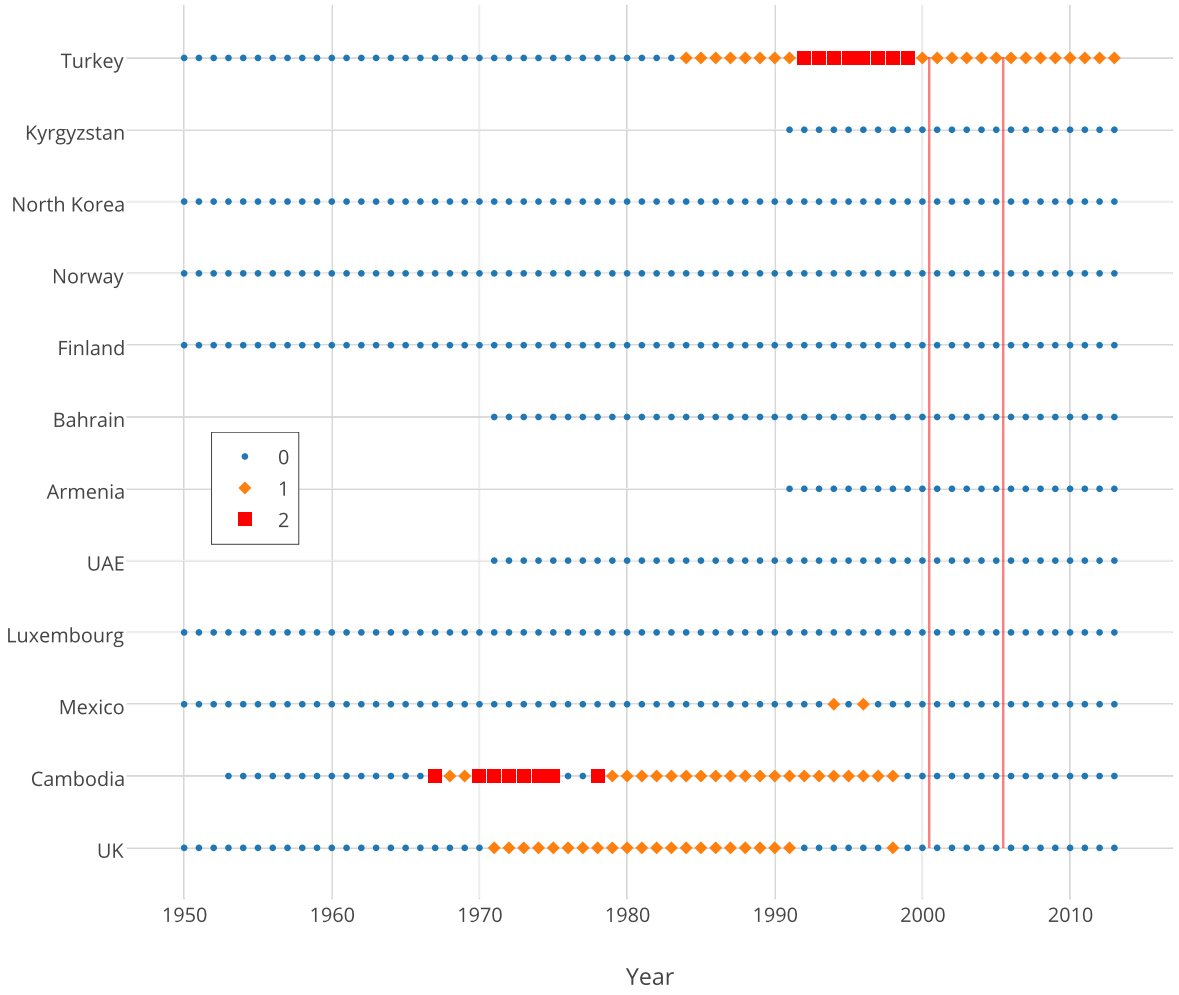


Figure 5.3: Conflicts over time for the reinforcing group. Blue dots represent peace, yellow diamonds are minor conflicts and the red squares are major conflicts. The first vertical line marks the cut point between the estimation and evaluation period, and the second line marks the start of the limited evaluation period.

political situation was far from peaceful. Turkey has continous conflict from the mid 80's until the end of the dataset. All other countries in the group are conflict free for the entire period. The countries with conflict have either no conflict in the evaluation period or only conflict.

As previously mentioned, Slovenia has very low predicted probabilities. Slovenia has near perfect Brier scores, and has no reason to be a destructive outlier based on these. The same is true for Poland, with similarly minute probabilities.

Apart from Turkey, all the reinforcing countries are well predicted and peaceful. Kyrgyzstan has the highest probabilities, but these do not exceed 7%. The other countries

with interesting predictions are shown in Figure 5.4 and 5.5. Beginning with Syria, the last three years are still problematic. The previous years have so low probabilities that the Brier scores are still minute. The last three years on the other hand are punished harshly, so the period as a whole gets a poor score.

CAR also has low probabilities, but these climb slowly over the period in response to increased conflict levels. Conflicts occur at the start and end of the period, with a lone conflict breaking up a peaceful stretch mid-period. Again we see the slow response to the conflict, and that the response is too weak to clearly predict the following conflict. The lone conflict mid-period is not far over the threshold, and there are continued hostilities in the peace years despite the battle deaths remaining under the threshold. CAR has a low GDP, but also a small population. Combined with a Polity score of 5 CAR manages to keep its probabilities low, especially in the first year when the conflict history variables have yet to set in. Once they do set in there is an increase in probabilities that climbs above 20% towards the end of the period. The results would not be as bad had the lag been properly timed, but the peace years would still be a problem. The effect is also somewhat weak. It is clear that the model would not have been able to respond to the peace years, as probabilities go up much faster than they come back down afterwards.

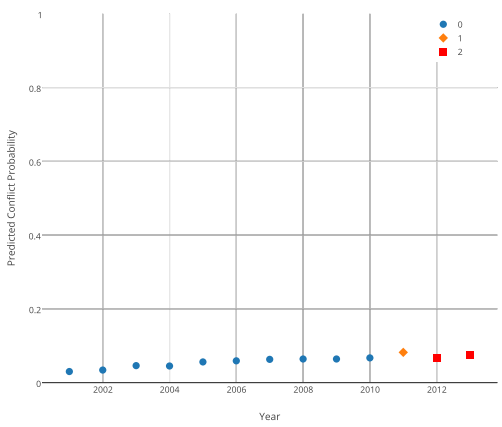
Niger has probabilities similar to those of CAR, between 10 and 20% for the whole period. Niger does not experience any conflict until late in the period, and then only experiences two years of conflict before reverting to peace. The faulty structure means the second year of conflict is not given a higher probability. The effect is also weak when it sets in, meaning a correct lag would only marginally improve the overall results. Niger has very high random effects keeping its probabilities low, combined with relatively high scores on Polity 2 for such a poor country. Continuous neighborhood conflicts also dampens the effect of the minor conflict variable. As with Syria, it is the two conflict years that increases the Brier score to damaging levels. Niger scores 0.135, while Syria and CAR score 0.2 and 0.451 respectively. This means they all have scores above the model average, although CAR is far worse than the other two. All three countries's predictions are so poor that they could account for the destructive effect on the model. Niger is however likely to have an effect through other channels as well, as its predictions are not as poor as the other two.

Sierra Leone has extreme values at the beginning of the period that drop quickly before flattening out just over 50%. The conflict in the first year is well predicted, but the rest of the period is peace and the predictions are far off. The probabilities come steadily down as the GDP and time in peace variables increase. A problem for Sierra Leone is that the ncts0 and ltsnc variables are positive. As the neighborhood becomes peaceful in 2004, the effects of these keep the probabilities from falling further, leading to overestimation of the risk.

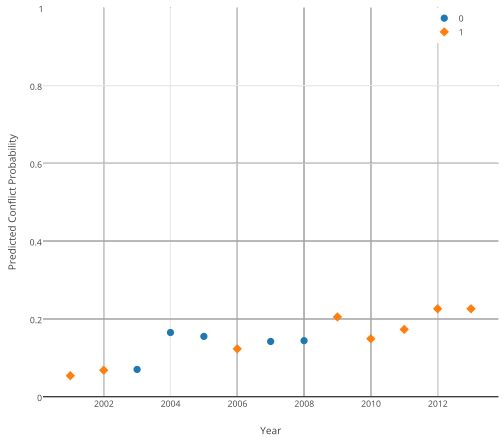
Comparing the countries in the lower row of Figure 5.4 shows that Sierra Leone has predicted probabilities strikingly similar to those of the DRC and Angola. Compared to Angola we see only a slightly lower average predicted probability of 64.7%. As only a single conflict occurs, the observed average is 7.7%, giving a predicted to observed rate of 8.4 over the whole period. This is drastically worse than Angola's ratio of 2, and the ratio for the DRC is even better at 1.5. Calculating the Brier score for the two countries' predictions alone shows that Angola has scores of .32 and .34 depending on period, while Sierra Leone has scores of .37 and .32. For the full period the DRC has a Brier score of .34, which improves to .23 in the limited period. These are very similar scores, especially for the full period, despite the fact that Sierra Leone's predictions are much worse.

In the full period, Angola and the DRC have effects on the full model's Brier score that are greater than that of Sierra Leone, even though their own scores are lower. Sierra Leone has a far larger effect on predicted probabilities, but a smaller effect on the coefficients. This means that the indirect effects of Angola and the DRC are of a more harmful nature than those of Sierra Leone, and that the conflicts in Sierra Leone are of a nature that fits better with the global model than those of Angola and the DRC.

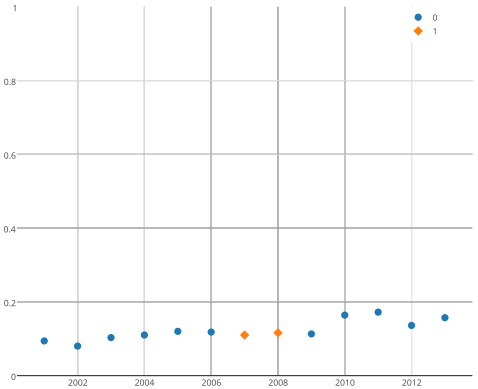
While experiencing a great deal of conflict in the eighties and nineties, Somalia was peaceful in the last half of the 2000's. We can see that the conflict probabilities are going down at the start of the period before increasing in response to conflict. They gradually increase, but even after many years they flatten and hover around 40%. The problem is therefore not a lack of response to conflict, but rather its lack a power. The many conflicts that are predicted at only 40% would be enough to cause major problems regardless of whether the peace years were better predicted. Somalia is poor, but the population is not so large that it becomes very problematic. The neighborhood and a poor Polity score should result in high scores, but apparently not high enough to account for all the observed conflict. The increase in the early period is helped by Somalia's economy stalling, going virtually without growth from 2000 to 2005. As the economy starts growing again, the GDP variable assists in flattening the probability curve. Somalia's probabilities alone are enough to account for it being a destructive unit by Brier score.



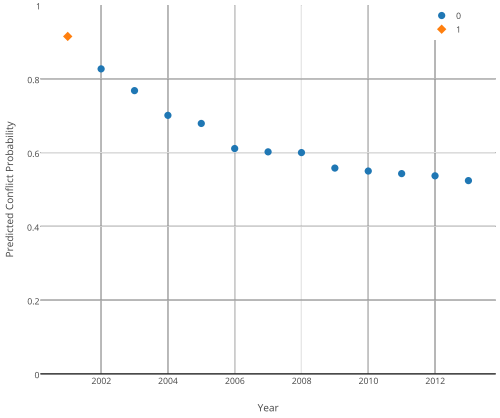
(a) Syria - Brier score .2



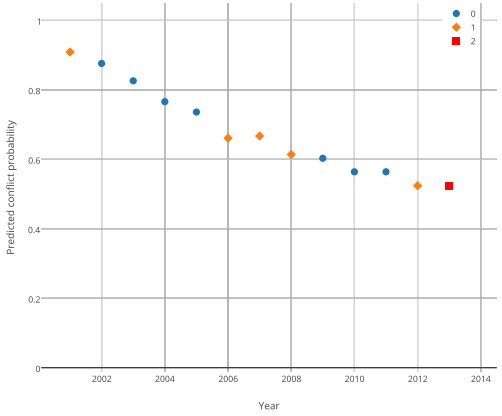
(b) CAR - Brier score .451



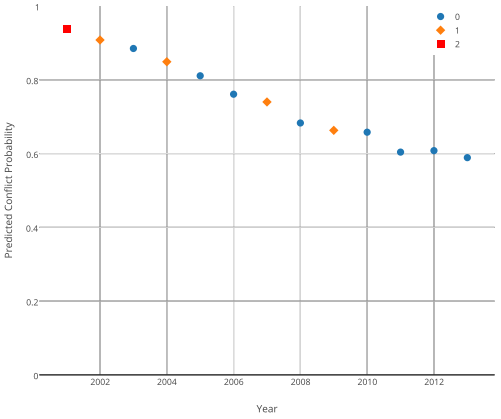
(c) Niger - Brier score .135



(d) Sierra Leone - Brier score .369

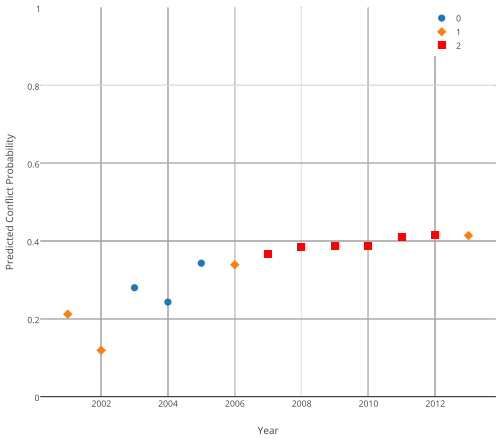


(e) DRC - Brier score .34

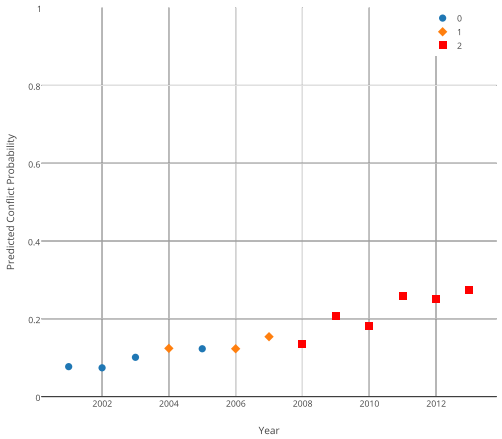


(f) Angola - Brier score .324

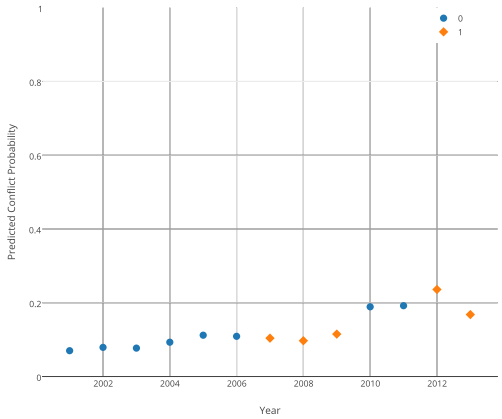
Figure 5.4: Predicted conflict probabilities versus observed conflict over evaluation period.



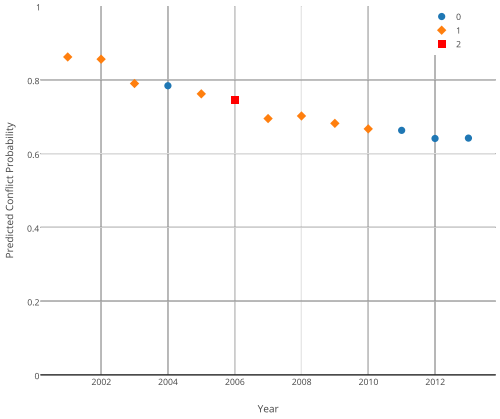
(a) Somalia - Brier score .358



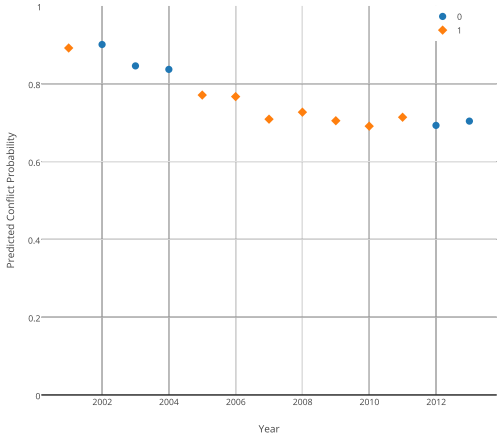
(b) Pakistan - Brier score .46



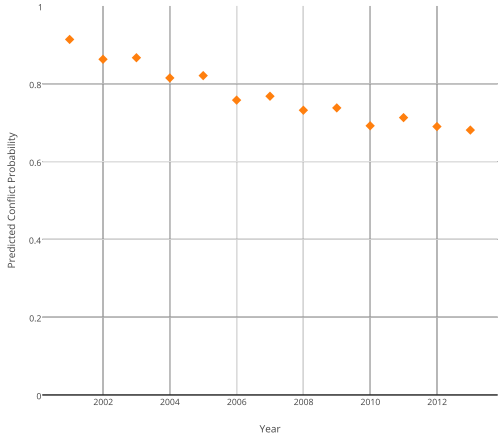
(c) Mali - Brier score .292



(d) Chad - Brier score .19



(e) Iran - Brier score .288



(f) Turkey - Brier score .057

Figure 5.5: Predicted conflict probabilities versus observed conflict over evaluation period.

Pakistan and Mali are both countries that start the evaluation period in peace, and countries where the model fails to respond swiftly to conflict. In Pakistan we can see that the probabilities do rise towards the end, but considering that the last eight years are conflict and probabilities do not reach 30% the response is slow and inadequate. The probabilities do divide the peace and conflict years, which could explain why the country is not as damaging when using PR. A high population and conflict stricken neighborhood pushes Pakistan up, but the poverty is not extreme. While negative, the Polity score is high enough to give the squared variable a high enough score to reduce the conflict probability considerably. For both Mali and Pakistan, the problem becomes many conflict years with low probabilities. Mali was also destructive by PR, but as mentioned Pakistan may have escaped this due to the results being well sorted.

Iran and Chad are on the opposite end, being punished for having false positives. They are also denied full reward for the true positives as many of these are predicted at probabilities below 80%.

All the destructive countries, apart from Poland and Slovenia, could be destructive simply because of their own predictions. Their individual Brier scores show that they are clearly higher than that of the control, giving their inclusion a negative effect on full model Brier score. Syria, Niger and Chad show slightly lower scores than the others, which could be an indication that their indirect effects are of a more harmful nature with regards to model Brier score than the others.

5.4 Indirect effects through coefficient effects or neighborhoods

While all destructive countries apart from Poland and Slovenia can be explained by their predictions, there are a few that also have effects on coefficients. I hypothesized that Syria, Niger and Chad would be likely to have noteworthy effects, which they do on some variables. Syria has effects on the important conflict history variables, which easily account for the added destructive effect on model Brier score.

Niger has a relatively extreme positive effect on the *lGDPcap*, and a strong positive effect on *ltsc0*. These are both important variables, and as the two are interacted any effect is also multiplied. Chad also has a positive effect on *ltsc0*, and as there are no reinforcing outliers with similar effects on the variable this could be an indication that a positive effect on *ltsc0* is particularly damaging to the model as a whole.

Among the remaining countries we find that Pakistan has a strong effect on a number of variables, including most conflict history variables. It does however have similar effects on many variables as Indonesia, so it is hard to say whether the effects are of a detrimental

nature overall. The other countries appear to have their effect scattered among all variables, with no particular variable being affected to a great degree.

Poland and Slovenia are now the only destructive country yet to have an explanation, and I have already shown that Slovenia is inexplicable. While Poland does have a slightly lower GDP than its western neighbors, there is in no way a gap that could alone account for Poland's effect on model Brier score. This leaves the destructive group for Brier score with two countries that can only be explained through their unknown effect on the random effects.

5.5 Robustness when correcting conflict lag

Some of the predicted probabilities appear to show a problem with variable lagging, as the model at times reacts in inexplicable ways to conflict history. Uzbekistan is a prime example, where the predicted probabilities shoot up in 2007 following a conflict in 2004, two years after they should have. The same pattern is found in other countries, with varying importance for the results. In Mali the effect is not very severe, as the predictions are very low even when the conflict variables take effect. In Indonesia the faulty lag causes the drop in predicted probabilities after conflict cessation to come later than it should. Here in the middle spectrum the difference made by changes in logit are the greatest, and the effect is that 2007 and 2008 are predicted approximately 20% above the correct levels.

An erroneously lagged conflict variable is most likely to have cause the issue, but the parameter files[1] and input data appear to be correct. Also, not all countries seem to be affected to the same extent. Djibouti appears to have a different problem. A conflict occurred in 1999, which by the pattern established by the previous cases should have lead to a jump in probabilities in 2002. Instead there is a peak in 2001 and a small spike in 2003. This is probably a country specific issue, as the pattern of a two year gap between event and reaction is followed by all other cases where an effect is visible.

The faulty lag will have an effect on the validity of the results. As the input data is correct, the cause of the problem is probable in the simulator itself. It is hard to tell whether only the conflict variables are affected, or whether the effect of other variables are also wrongly timed.

To test how badly the results are affected by the faulty lag I have shifted the predictions two years backwards and rerun the analysis. The predictions here go from 2001 to 2011. The ROC results become much more similar to those of PR. China remains the worst

[1]The parameter files are text files that contain information passed to the simulator. They contain the names of the variables that are to be included and instructions for how they are to be lagged. The parameter files and input data are saved by Stata before being sent to the simulator, making it possible to see exactly what the simulator receives.

offender, with the DRC and Spain climbing up the list. Nigeria also remains in the top ten, but the rest are changed. Syria, now with only one wrongly predicted conflict, disappears along with the others. Angola and Tanzania are among those that replace them, backing the earlier PR finds.

The PR list for the full period remains largely undisturbed, with only minor changes in order and magnitude. All countries have slightly lower effects on the PR AUC than before. Macedonia climbs over Angola and Qatar, but the order of the top six is otherwise the same. The Brier results are also very stable, with the appearance of Djibouti high on the list as the only major change. Djibouti is one of the outliers by PR, and its predictions are such that its addition to the list comes as no surprise.

Correcting the lag appears to make only minor differences in the results, and all major points made so far are still valid.

Chapter 6

Conclusion

In this chapter I will first summarize the findings from the previous chapter and discuss these in light of the results from Chapter 4. Both periods show clearly that poor predictions is the main reason for countries becoming destructive outliers, while effects on β-estimates are less important. They also show that the threshold based conflict definition is problematic, and that defining the divide between estimation and evaluation periods has important consequences for the evaluation outcome.

The first part of the chapter presents the summary of Chapter 5 and the discussion of the findings. I then present the main conclusions that can be drawn, and present possible solutions to problems encountered. I also reiterate some weaknesses in my research that must be taken into account when assessing the impact of the results.

6.1 Summary and discussion

The results from the analysis of the Brier scores show differences from the PR results, but there are consistencies as well. All four measures applied support the theory that leaving a gap between estimation and evaluation reduces accuracy. There are however countries that benefit from the limited period, and the choice of period limits may affect these findings. Both the destructive and reinforcing groups show a great degree of overlap between measures. With Brier score, as with PR AUC, the groups are not the same as would be expected from the test of effect on coefficients. There is considerably less change between periods, showing that an evaluation gap becomes less important when using Brier score. I believe this is because the Brier score evaluates country-years individually, without concern for the data set, while PR AUC is dependent on all the predictions of the model. The PR AUC is not only affected by the dropping of years for the individual country being dropped, it is also affected by the overall change in the prediction pool for the whole model. As other countries' predictions are dropped with the period limitation, the data that each

country is compared to changes in nature. This is why the Brier scores are relatively unchanged in comparison with what I found using PR AUC.

The destructive countries as a group are distinct in a number of ways with regards to variable values. Like with PR they are younger nations, which indicates that the time since independence variable may be a problem for the model. The group sees more conflict while the reinforcing countries have less, the opposite of what was found with PR. The destructive group has shorter periods of peace, backing the finding from PR. They also experience far more neighboring conflicts, an indication that the neighborhood variables may not behave as expected. The groups are also split by GDP, with the destructive group being poorer than the main group of countries, while the reinforcing group is richer. It could be that the log function is not the correct function for GDP, creating problems for some poorer countries. A dichotomous variable based on a threshold could be considered as a replacement, or perhaps a categorical variable based in a qualitative assessment of development. Lastly the destructive countries are more autocratic and anocratic. This indicates that the squared Polity variable is problematic. While not necessarily discrediting the anocracy theory, this fits well with the criticism of the Polity scale mentioned in the theory chapter. Applying a different measure of democratization or using categorical regime type variables are possible alternatives.

Many of the destructive outliers are countries with underestimated conflict probabilities who are also located in neighborhoods with high levels of conflict. Several of these countries are placed in a crescent following the "spine" of the African continent. This results in a great number of neighbors and an almost guaranteed presence of neighborhood conflict at any given time. It could be that the constant presence of neighboring conflict in these countries is problematic due to the interactions between neighborhood conflict and other variables.

The observed conflict history shows that the destructive group does indeed have more conflict, and that the conflict periods are broken up by short periods of peace. This is similar to what was found with PR, in that the problem is fluctuations between peace and conflict that the model has no time to react to. The reinforcing group shows a high degree of clustering in its conflicts which further backs this conclusion. The issue appears to be a symptom of an underlying problem with the conflict definition, and the accuracy could be greatly increased by using a more flexible conflict definition that is based on more factors than yearly battle related deaths.

The conflict history also reveals that how the cross-validation groups are split is important. The reinforcing group is clearly split with the evaluation group being either only peace or only conflict. CAR is the prime example, with only peace right up until the evaluation period starts. Sierra Leone is the opposite, with a long period of war that ends

one year into the evaluation period. As the model is slow to react, these sudden changes are likely to be followed by several years of conflict probability estimations that slowly transition from one end of the spectrum to the other. Mali is not on the top list of destructive countries by PR or Brier for the full period, but enters both lists when the period is limited. Looking at the conflict history it is obvious why this happens. The evaluation gap is peaceful, just as Mali's history, but the limited period has more conflict than peace. When the evaluation period is no longer "watered down" by the peace years in the gap period, Mali's low probabilities suddenly score much worse.

The destructive Brier group shows different patterns in its predictions compared to the destructive group by PR AUC. As mentioned, Brier score judges countries individually rather than by their place in the dataset as a whole. This means that countries such as Pakistan, with predictions that separate conflicts from peace years and give them probabilities that are high enough relative to the rest of the data, are not outliers by PR. They are however outliers if the predictions are judged on their own, as numerous years of conflict have predicted probabilities below 30%.

Because PR pools all years together when evaluating, some years with reasonable probabilities can end up being more harshly judged than is perhaps appropriate. An example is Tanzania, which becomes the second most destructive unit by PR in the limited period, but only the fifteenth worst by Brier. This is most likely due to the presence in the set of several conflicts predicted at below 20%, something that makes the probabilities in Tanzania more damaging. As Brier judges them individually and they are relatively close to 0 their scores are quite low, and less damaging to the model. Countries like Chad and Iran are however punished more harshly by Brier than by PR. A few peace years are harmful, but many conflict years are predicted as low as 70%, earning them poor Brier scores as well. These would probably be well over the majority of peace years for the whole model, and so PR would not punish them as much as Brier.

Countries such as Djibouti and Tanzania, who have relatively well predicted peace periods are not as destructive by Brier because their predictions are no longer compared to those of other countries. China and Macedonia are also removed from the group, as their single units are not of as much consequence when the score is averaged over many years. The results is that the outliers by Brier appears more reasonable, perhaps apart from the fact that Uzbekistan is no longer on the list. The choice is of course not as straight forward as this, as there are pros and cons with either measure. Since PR is based on the rankings of probabilities, the model is not forced to predict extreme probabilities for conflict years in order to be rewarded, only higher than the vast majority of peace years. A model that gives all peace years conflict probabilities of 0 to 20% and all conflict years 40-60% will be judged as excellent by PR, but poor Brier score.

Probabilities in the mid range are a problem for both measures if we think of our predictions as time series rather than single years. Take for an example a hypothetical country with predictions that give a fifty-fifty chance for conflict over two years. This will yield a very poor score by Brier, and if the peace year is predicted minutely higher than the conflict year then PR will also be poor. The predictions were however completely correct: The chance of a conflict in the period was on average 50%, and conflicts occurred in 50% of the period. The examples of Angola and the DRC illustrate how the cross-sectional time-series structure of the data is ignored by the measures used here. The average predicted conflict level is much closer to the observed in Angola and the DRC than Sierra Leone, yet by Brier score their results are classified as similar. The two also have a much greater effect on PR AUC than Sierra Leone, even though they intuitively appear to be much better at correctly identifying cases. The measures of predictive power used here discriminate against the middle spectrum of probabilities.

Treating each year as an individual unit means that we regard missing a conflict onset by a single year a complete failure. It also means we are saying that probabilities above a certain threshold are predictions of definite conflict while those below are of definite peace. This is not how probabilities work. The many redraws of conflict outcomes does not correct this problem, it only helps explore the scenarios where the less probable outcome occurs. A probability is something that must be tested by repeated experiments, not by examining the same case again and again expecting something different to happen. In our case this can be done by examining how probabilities match reality over time, with each year as an experiment.

If the goal of forecasting is to uncover the underlying risks in countries rather than attempting perfect onset forecasts, it would seem that the measures being used to optimize forecasting models are less than ideal. Averaging the predicted probabilities and observed values allows predictions to be in the middle of the spectrum without being unduly punished. Comparing average predicted probability with proportion of years with conflict occurrence is not a completely realistic way of evaluating results, but it is better than completely ignoring the fact that units are countries over a period, not individual country years.

There are fewer inexplicable countries returned as destructive outliers with Brier score compared to PR AUC. It is however clear that model complexity makes it difficult to trace how each country affects the overall results. Brier does make it easier, as each country's individual Brier score can be calculated and compared to the full model. By comparing the country's score with its effect on the full model one can calculate to what extent the country also has indirect effects, which can then be traced further. It is however clear that starting with a fully developed model has lead to problems, especially due to the random country

effects used. To accurately track effects a research design should start with a completely basic design, before gradually adding complexity in the form of variables, interactions, neighborhoods and country effects.

6.2 Conclusion

I have shown that there are indeed some countries that have greater influence on model predictive power. Importantly, these are not the same as those found by a conventional test of outliers that estimates differences in β-coefficients. While there is a reasonable overlap between the results it is clear that effect on coefficients $\neq$ effect on model predictive power, which means that testing for influential units when using predictive power requires its own methods. I have presented different ways of performing such a test using different means. The simple test of differences in other units predicted probabilities is a simple way of measuring the degree to which a unit affects the rest of the set. The expanded tests reveal how different units have different effects on model predictive power depending on the measure used.

The countries that have the most detrimental effect on predictive power differ in many aspects, yet there are clear patterns that show where and when the model fails. The outliers with a negative effect on predictive power measured by PR AUC had less conflict than average, but the outliers with a negative effect on Brier score had considerably more than average. Despite this, the two groups have common features. The countries whose negative effect was due to their own poor predictions followed two patterns of conflict history. The first is a change in conflict state between largely peaceful and largely conflict stricken, or vice versa, that coincides with the partitioning of data between estimation and evaluation periods. The second pattern is a recent conflict history that is either plagued by several transitions between peace and conflict.

The first pattern is cause by a problem with cross-validation using cross-sectional time-series data in general. K-fold cross-validation using random partitioning is highly problematic, and the best alternative is a temporal cut-off point that has to be arbitrarily determined by the scientist. The timing of this point and the length of the gap between estimation and evaluation affects the outcome of the cross-validation greatly. Any country that happens to have a change between peace and conflict near to the estimation-evaluation divide will have several years in its evaluation data that is inaccurate as the model takes time to adapt to a change in status. A way of compensating for this could be to cross-validate using multiple variations of estimation, evaluation and gap periods and averaging the results. Using my data as an example, the end of the estimation period could vary between 2000 and 2005, and the gap could vary between 0 and 10 years, with a minimum

evaluation period of 5 years. A script that accomplishes this can easily be written, and computationally the task is done in a matter of seconds. Averaging over many periods in this fashion would reduce the problems that arise from having to use only the latest years of the dataset for evaluation.

The second conflict pattern arises from the conflict operationalization used in the UCDP/PRIO armed conflict dataset. The strict 25 battle deaths per year definition does not pick up on the nuances of conflict situations. I have shown examples of conflicts that are coded as ended and then restarted simply because the middle year had 4 fewer battle deaths. The conflict is, by qualitative standards, ongoing for the whole period, but the purely quantitative definition of conflict sees the tiny fluctuations in casualties as peace. This creates data that is impossible to predict accurately, requiring huge leaps back and forth in predicted probabilities. The model is behaving as it should, but it is being given an impossible task. This leads to an unfair punishment that misleads us regarding its true accuracy. The problem can be addressed by using a composite conflict indicator that takes into account other factors as well as battle deaths.

The problem can also be seen as a result of the our evaluation methods not taking into account the time-series aspect of the data. All the measures of predictive power used here judge years by how well they are predicted on an individual basis, but this is not necessarily the optimal solution if our goal is to predict risk of conflict in an upcoming time period.

When it comes to choice of predictive power measure there are some differences that are noteworthy. The main difference comes from how they treat the results, with PR pooling them all together and Brier judging each individually and then averaging. This makes Brier less affected by single years, especially with longer evaluation periods.

My overall impression is that PR is better as a means of evaluating model separation, as it evaluates the model responses more as a whole. Brier does have its advantages when looking for outliers. It is much easier to determine which countries have the worst predictions using Brier score, as it judges the results on their own merits. The Brier outlier test thus returns countries with the worst predictions rather than those that are necessarily the worst for the models overall ability to separate conflict and peace. It is also easier to unravel indirect effects using the Brier score, as the effect of the countries own predictions can easily be calculated.

A problem I have faced has been the complexity of the research design. While the model I have applied my test of outliers to is representative of what theory prescribes should be included, and is at the cutting edge of the forecasting field, it has lead to problems identifying causal mechanisms. Further research on the subject would be better served by starting with a simpler model and gradually adding complexity. By gradually adding variables, neighborhoods and random effects to a baseline model, one would be

better able to see how countries create problems.

A further issue is that when examining unit effects on coefficients I have used the control model without reestimating the multilevel model and subsequently the country random effects that are extracted from it. It is reasonable to assume that these effects are affected by dropping countries. Any change in coefficients will also lead to a change in the country specific intercepts which are used as random effects here. The exact changes in all coefficients and random effects could be found by modifying the simulator to extract these. This would provide the true effect of each country on the coefficients rather than the approximation I have used.

To conclude, the potential of a test of influence on predictive power is great. Much can be learned about the variables used and effect out single units by examining their effect on summary statistics. Using different measures of predictive power will reveal outliers with different characteristics, each useful in their own way. While the work presented here has much room for improvement I have also come to two major conclusions that I believe will be robust. The first is that a threshold based conflict definition can be problematic for evaluating forecasting models. The second is that how data is split between estimation and evaluation can have severe consequences for evaluation results. I therefore recommend the implementation of a more flexible conflict definition, and the use of more than one evaluation set.

Bibliography

Adams, D.

2002. *The Salmon of Doubt: Hitchhiking the Universe One Last Time*. Harmony Books.

Bakan, D.

1966. The test of significance in psychological research. *Psychological Bulletin*, 66:423–437.

Beck, N.

2001. Time-series-cross-section data: What have we learned in the past few years? *Annual review of political science*, 4(1):271–293.

Blattman, C. and E. Miguel

2010. Civil war. *Journal of Economic Literature*, Pp. 3–57.

Bolt, J. and J. L. van Zanden

2013. The first update of the maddison project; re-estimating growth before 1820. *Maddison-Project Working Paper WP-4, University of Groningen, January*, 5.

Buhaug, H.

2010. Dude, where's my conflict? lsg, relative strength, and the location of civil war. *Conflict Management and Peace Science*.

Buhaug, H. and S. Gates

2002. The geography of civil war. *Journal of Peace Research*, 39(4):417–433.

Buhaug, H., S. Gates, and P. Lujala

2009. Geography, rebel capability, and the duration of civil conflict. *Journal of Conflict Resolution*, 53(4):544–569.

Buhaug, H. and K. S. Gleditsch

2008. Contagion or confusion? why conflicts cluster in space. *International Studies Quarterly*, (2):215–233.

Buhaug, H. and J. K. Rød

2006. Local determinants of african civil wars, 1970–2001. *Political Geography*, 25(3):315 – 335.

Cederman, L.-E., N. B. Weidmann, and K. S. Gleditsch

2011. Horizontal inequalities and ethnonationalist civil war: A global comparison. *American Political Science Review*, 105(03):478–495.

Cederman, L.-E., A. Wimmer, and B. Min

2010. Why do ethnic groups rebel? new data and analysis. *World Politics*, 62(01):87–119.

Chateau, J., R. Dellink, E. Lanzi, and B. Magné

2012. Long-term economic growth and environmental pressure: reference scenarios for future global projections. In *draft presented at the OECD EPOC/WPCID meeting in September*.

Collier, P.

1999. On the economic consequences of civil war. *Oxford economic papers*, 51(1):168–183.

Collier, P.

2004. Development and conflict. *Centre for the Study of African*.

Collier, P. et al.

2003. *Breaking the conflict trap: Civil war and development policy*. World Bank Publications.

Collier, P. and A. Hoeffler

1998. On the economic causes of civil war. *Oxford Economic Papers*, 50:563–573.

Collier, P. and A. Hoeffler

2004. Greed and grievance in civil war. *Oxford Economic Papers*, 56(4):563–595.

Collier, P., A. Hoeffler, and M. Söderbom

2008. Post-conflict risks. *Journal of Peace Research*, 45(4):461–478.

Collier, P. and N. Sambanis

2005. *Understanding Civil War: Africa*, volume 1. World Bank Publications.

Dahl, D. B.

2014. *xtable: Export tables to LaTeX or HTML*. R package version 1.7-4.

Davis, J. and M. Goadrich

2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, Pp. 233–240, New York, NY, USA. ACM.

D'Orazio, V., J. E. Yonamine, and P. A. Schrodt

2011. Predicting intra-state conflict onset: An event data approach using euclidean and levenshtein distance measures. In *69th Annual Meeting of the Midwest Political Science Association, Chicago, Il, USA*.

Fearon, J. D.

2011. Governance and civil war onset.

Fearon, J. D. and D. D. Laitin

2003. Ethnicity, insurgency, and civil war. *American Political Science Review*, 97(1):75–90.

Feenstra, R. C., R. Inklaar, and M. Timmer

2013. The next generation of the penn world table. Technical report, National Bureau of Economic Research.

Gates, S., H. Hegre, M. P. Jones, and H. Strand

2006. Institutional inconsistency and political instability: Polity duration, 1800–2000. *American Journal of Political Science*, 50(4):893–908.

Gates, S., H. Hegre, H. M. Nygård, and H. Strand

2012. Development consequences of armed conflict. *World Development*, 40(9):1713–1722.

Gill, J.

1999. The insignificance of null hypothesis significance testing. *Political Research Quarterly*, 52:647–674.

Gleditsch, K. S.

2007. Transnational dimensions of civil war. *Journal of Peace Research*, 44(3):293–309.

Gleditsch, N. P., P. Wallensteen, M. Eriksson, M. Sollenberg, and H. Strand

2002. Armed conflict 194-2001: A new dataset. *Journal of Peace Research*, 39(5):615–637.

Gneiting, T., F. Balabdaoui, and A. E. Raftery

2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Goldstone, J. A., R. H. Bates, D. L. Epstein, T. R. Gurr, M. B. Lustik, M. G. Marshall, J. Ulfelder, and M. Woodward
2010. A global model for forecasting political instability. *American Journal of Political Science*, 54(1):190–208.

Greene, W. H.
2003. *Econometric analysis*, 5 edition. Upper Saddle River, NJ: Pearson Education.

Gurr, T. R.
1970. Why men rebel.

Gurr, T. R.
1993. Why minorities rebel: A global analysis of communal mobilization and conflict since 1945. *International Political Science Review*, 14(2):161–201.

Gurr, T. R.
2000. *Peoples versus states: Minorities at risk in the new century.* US Institute of Peace Press.

Gurr, T. R., B. Harff, M. Levy, G. D. Dabelko, P. T. Surko, and A. N. Unger
1999. State failure task force report: Phase ii findings. *Environmental Change and Security Project Report*, (5):50–72.

Hastie, T., R. Tibshirani, and J. Friedman
2009. *The Elements of Statistical Learning - Data mining, Inference, and prediction.* New York: Springer.

Hegre, H., H. Buhaug, K. C. Calvin, J. Nordkvelle, S. T. Waldhoff, and E. Gilmore
forthcoming. Forecasting armed intrastate conflict along the shared socioeconomic pathway. .

Hegre, H., T. Ellingsen, S. Gates, and N. P. Gleditsch
2001. Toward a democratic civil peace? democracy, political change and civil war, 1816-1992. *American Political Science Review*, 95:33–48.

Hegre, H., J. Karlsen, H. M. Nygård, H. Strand, and H. Urdal
2013. Predicting armed conflict, 2010-2050. *International Studies Quarterly*, 57:250–270.

Hegre, H. and H. M. Nygård
forthcoming. Ac2. .

Hegre, H. and N. Sambanis
2006. Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution*, 50:508–535.

Hempel, C. G. and P. Oppenheim

1948. Studies in the logic of explanation. *Philosophy of science*, Pp. 135–175.

Hlavac, M.

2014. *stargazer: LaTeX/HTML code and ASCII text for well-formatted regression and summary statistics tables*. Harvard University, Cambridge, USA. R package version 5.1.

Kennedy

forthcoming. Making useful conflict predictions: Methods for addressing skewed classes and implementing cost-sensitive learning in the study of state failure. *Journal of Peace Research*.

King, G.

1989. *Unifying political methodology: The likelihood theory of statistical inference*. University of Michigan Press.

Kuhn, M. and K. Johnson

2013. *Applied Predictive Modeling*. Springer.

Lambdin, C.

2012. Significance tests as sorcery: Science is empirical - significance tests are not. *Theory and Psychology*, 22:67–90.

Long, J. S.

. Regression models for categorical and limited dependent variables. *Advanced Quantitative Techniques in the Social Sciences Series*, 7.

Lujala, P.

2010. The spoils of nature: Armed civil conflict and rebel access to natural resources. *Journal of Peace Research*, 47(1):15–28.

Marshall, M. G. and K. Jaggers

2002. Polity iv project: Political regime characteristics and transitions, 1800-2002.

Menard, S.

2010. *Logistic Regression – From Introductory to Advanced Concepts and Applications*. Thousand Oaks, CA: SAGE.

Murdoch, J. C. and T. Sandler

2002. Economic growth, civil wars, and spatial spillovers. *Journal of conflict resolution*, 46(1):91–110.

Murdoch, J. C. and T. Sandler

2004. Civil wars and economic growth: Spatial dispersion. *American Journal of Political Science*, 48(1):138–151.

O'Brien, S.

2002. Anticipating the good, the bad, and the ugly: An early warning approach to conflict and instability analysis. *Journal of Conflict Resolution*, 46:791–811.

Pfetsch, F. R.

2015. Conflict barometer 2014. Technical report, Heidelberg Institute for International Conflict Research.

Przeworski, A.

2000. *Democracy and development: political institutions and well-being in the world, 1950-1990*, volume 3. Cambridge University Press.

Raftery, A. E.

1995. Bayesian model selection in social research. *Sociological Methodology*, 25:111–163.

Salehyan, I. and K. S. Gleditsch

2006. Refugees and the spread of civil war. *International Organization*, 60(02):335–366.

Sambanis, N.

2001. Do ethnic and nonethnic civil wars have the same causes? a theoretical and empirical inquiry (part 1). *Journal of Conflict Resolution*, 45(3):259–282.

Sambanis, N.

2004. What is civil war? conceptual and empirical complexities of an operational definition. *Journal of Conflict Resolution*, 48(6):814–858.

Schrodt, P. A.

2014. Seven deadly sins of contemporary quantitative political analysis. *Journal of Peace Research*, 51:287–300.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer

2005. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):7881.

Sundberg, R., M. Lindgren, and A. Padskocimaite

2010. Ucdp geo-referenced event dataset (ged) codebook. *Uppsala, Sweden: Uppsala University*.

Themnér, L. and P. Wallensteen

2014. Armed conflicts, 1946–2013. *Journal of Peace Research*, 51(4):541–554.

Tuszynski, J.

2014. *caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc.* R package version 1.17.1.

United Nations

2013. *World Population Prospects: The 2012 Revisioncomprehensive Dataset in Excel.* United Nations Publications.

Valentini, G. and M. Re

2014. *PerfMeas: PerfMeas: Performance Measures for ranking and classification tasks.* R package version 1.2.1.

Vreeland, J. R.

2008. The effect of political regime on civil war unpacking anocracy. *Journal of Conflict Resolution*, 52(3):401–425.

Ward, M. D., B. D. Greenhill, and K. M. Bakke

2010. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375.

Weidmann, N. B. and M. D. Ward

2010. Predicting conflict in space and time. *Journal of Conflict Resolution*, 54(6):883–901.

World Bank Group

2013. World development indicators.

Ziliak, S. T. and D. N. McCloskey

2008. *The Cult of Statistical Significance.* Ann Arbor: The University of Michigan Press.

Appendix A

Tables

A.1 Variables

Variable name	Variable description
c1	Minor conflict at t-1
c2	Major conflict at t-1
ltsc0	Logged years since last conflict. This is also lagged by one year to avoid an effect at the first year after a conflict.
nc	Conflict in a neighboring country at t-1.
ncc1	$c1$ interacted with nc.
ncc2	$c2$ interacted with nc.
ltsnc	Logged years since the last neighboring conflict. Like $ltsc0$ this variable is also lagged by one year.
ncts0	$ltsc0$ interacted with $nc0$. $nc0$ is not included itself, but is coded as 1 when there are no neighboring conflicts.
lpop	Logged total population in thousands.
lpop_c1	$lpop$ interacted with $c1$
lpop_c2	$lpop$ interacted with $c2$
lGDPcap	Logged GDP PPP per capita in thousands of dollars.
lGDPcap_c1	$lGDPcap$ interacted with $c1$.
lGDPcap_c2	$lGDPcap$ interacted with $c2$.
lGDPcap_ltsc0	$lGDPcap$ interacted with $ltsc0$.
nb_lGDPcap	The average $lGDPcap$ value of all neighboring countries. For countries with no neighbors the value is set to their own $lGDPcap$ value.
ltimeindep	Logged years since the country became an independent nation.
dec50	Dummy coded as 1 for all country-years from the 1950's, 0 otherwise.
dec60	Dummy coded as 1 for all country-years from the 1960's, 0 otherwise.
dec70	Dummy coded as 1 for all country-years from the 1970's, 0 otherwise.
dec80	Dummy coded as 1 for all country-years from the 1980's, 0 otherwise.
dec90	Dummy coded as 1 for all country-years from the 1990's, 0 otherwise.
polity2	The country's Polity IV score.
polity2sq	$polity2$ squared.
polity2_c1	$polity2$ interacted with $c1$.
polity2sq_c1	$polity2sq$ interacted with $c1$.
polity2_c2	$polity2$ interacted with $c2$.
polity2sq_c2	$polity2sq$ interacted with $c2$.
nb_TSRC_5	A dummy variable coded as 1 if a neighboring country has experienced a regime change in the last 5 years.
random_1	The random country effect for minor conflicts.
random_2	The random country effect for major conflicts.

Table A.1: List of variables included in the model.

A.2 Country effects on coefficients and predicted probabilities

	Minor conflict	Major conflict
c1	2.271 (2.114, 2.55)	3.100, (2.807, 3.382)
c2	1.798, (1.553, 2.103)	4.837, (4.566, 5.447)
ltsc0	-0.058, (-0.1, -0.028)	0.021, (-0.07, 0.11)
nc	0.433, (0.299, 0.612)	1.067, (0.836, 1.454)
ncc1	-0.301, (-0.498, -0.153)	-0.700, (-1.057, -0.465)
ncc2	0.153, (-0.053, 0.361)	-0.098, (-0.554, 0.164)
ltsnc	0.007, (-0.006, 0.021)	0.021, (-0.02, 0.074)
ncts0	0.041, (0.013, 0.11)	0.223, (0.083, 0.381)
lpop	0.200, (0.186, 0.222)	0.283, (0.235, 0.322)
lpop_c1	0.044, (0.008, 0.073)	-0.004, (-0.073, 0.076)
lpop_c2	-0.119, (-0.156, -0.089)	0.088, (0.027, 0.171)
lGDPcap	-0.107, (-0.153, -0.076)	-0.221, (-0.281, -0.145)
lGDPcap_c1	0.042, (0.008, 0.098)	0.029, (-0.065, 0.121)
lGDPcap_c2	0.271, (0.23, 0.327)	-0.016, (-0.124, 0.056)
lGDPcap_ltsc0	-0.020, (-0.022, -0.017)	-0.056, (-0.068, -0.047)
nb_lGDPcap	-0.442, (-0.488, -0.421)	-0.104, (-0.186, -0.029)
ltimeindep	0.211, (0.199, 0.225)	0.131, (0.062, 0.177)
dec50	-0.872, (-0.978, -0.793)	0.413, (0.239, 0.596)
dec60	-0.157, (-0.237, -0.097)	0.839, (0.611, 0.953)
dec70	0.252, (0.172, 0.303)	1.090, (0.939, 1.214)
dec80	0.382, (0.299, 0.46)	1.275, (1.193, 1.403)
dec90	0.245, (0.202, 0.285)	0.655, (0.541, 0.768)
polity2	-0.006, (-0.012, -0.003)	-0.018, (-0.037, -0.004)
polity2sq	-0.013, (-0.015, -0.012)	-0.021, (-0.024, -0.018)
polity2_c1	0.040, (0.036, 0.049)	0.011, (-0.006, 0.03)
polity2sq_c1	0.020, (0.019, 0.022)	0.018, (0.015, 0.022)
polity2_c2	-0.027, (-0.041, -0.012)	-0.001, (-0.019, 0.021)
polity2sq_c2	0.015, (0.011, 0.018)	0.017, (0.011, 0.023)
nb_TSRC_5	0.067, (0.032, 0.12)	-0.176, (-0.257, -0.082)
random_1	1.048, (1.015, 1.079)	0.475, (0.427, 0.561)
random_2	0.269, (0.249, 0.299)	1.171, (1.084, 1.31)
intercept	-1.288, (-1.606, -1.051)	-5.803, (-6.206, -5.41)

Table A.2: Coefficients for the control case, with the minimum and maximum values from drops.

Table A.3: Differences in weighted coefficients. The absolute difference in coefficients is divided by the coefficient, then multiplied by the maximum possible effect of the variable (The coefficient multiplied by the maximum score of the variable in the data set.)

Rank	**Difference**	**Country**
1	2.19	Slovenia
2	3.10	Qatar
3	3.91	Norway
4	4.00	Bahrain
5	4.92	Luxembourg
6	5.40	New Zealand
7	5.52	Lithuania
8	6.02	Australia
9	6.13	Singapore
10	6.40	Bhutan
11	7.24	Switzerland
12	7.62	Slovakia
13	7.87	Denmark
14	8.33	German Federal Republic
15	9.09	Finland
16	10.60	Turkmenistan
17	10.78	Mauritius
18	11.48	Poland
19	13.23	Zambia
20	13.54	Czech Republic
21	15.25	Ukraine
22	16.00	Greece
23	17.50	Japan
24	17.94	Canada
25	19.08	Estonia
26	20.50	Gambia
27	20.67	United Arab Emirates
28	24.00	Netherlands
29	26.82	Portugal
30	26.94	Kyrgyz Republic
31	27.35	Italy/Sardinia
32	27.96	Guyana

Continued on next page

Table A.3 – continued from previous page

Rank	Difference	Country
33	28.80	Hungary
34	29.24	Rumania
35	30.27	Fiji
36	30.50	Trinidad and Tobago
37	30.61	Equatorial Guinea
38	31.32	Ecuador
39	33.75	Albania
40	35.64	Solomon Islands
41	35.84	Lesotho
42	36.59	Malawi
43	38.36	Kuwait
44	43.96	Cyprus
45	45.10	Surinam
46	46.88	Latvia
47	47.73	Belarus (Byelorussia)
48	48.54	Austria
49	50.14	Comoros
50	51.23	Cape Verde
51	54.39	Ireland
52	54.74	Sweden
53	56.58	Taiwan
54	58.28	Saudi Arabia
55	58.85	Jamaica
56	60.59	Uruguay
57	61.24	Vietnam, Democratic Republic of
58	64.43	Botswana
59	67.46	Togo
60	71.50	Eritrea
61	75.21	Belgium
62	78.94	Chile
63	79.49	Dominican Republic
64	81.33	Namibia
65	85.36	Swaziland
66	92.18	Guinea-Bissau

Continued on next page

Table A.3 – continued from previous page

Rank	Difference	Country
67	93.46	Gabon
68	96.91	Bolivia
69	97.71	Jordan
70	99.15	Mali
71	104.51	Egypt
72	106.54	Brazil
73	108.22	Armenia
74	109.02	Costa Rica
75	112.83	Kazakhstan
76	116.42	Mauritania
77	116.43	Honduras
78	116.92	Mongolia
79	119.52	Korea, People's Republic of
80	119.88	Guatemala
81	120.09	Tunisia
82	122.59	Mexico
83	130.43	Spain
84	130.56	Moldova
85	134.16	Tanzania/Tanganyika
86	137.14	Uzbekistan
87	138.99	Ghana
88	139.28	Panama
89	139.51	Macedonia (Former Yugoslav Republic of)
90	158.14	Central African Republic
91	161.87	Madagascar
92	162.27	Burkina Faso (Upper Volta)
93	164.68	Senegal
94	170.69	Paraguay
95	173.98	Cuba
96	176.71	Nigeria
97	185.02	Ethiopia
98	186.96	Djibouti
99	190.33	Haiti
100	202.27	Bulgaria

Continued on next page

Table A.3 – continued from previous page

Rank	Difference	Country
101	203.40	Oman
102	205.39	Zimbabwe (Rhodesia)
103	211.92	Cote D?Ivoire
104	212.31	Georgia
105	221.81	United Kingdom
106	222.33	Libya
107	224.52	Venezuela
108	226.25	Argentina
109	239.86	Guinea
110	247.76	Cameroon
111	254.23	Benin
112	256.37	Bangladesh
113	262.62	Korea, Republic of
114	263.01	Liberia
115	266.98	Mozambique
116	268.51	Turkey (Ottoman Empire)
117	268.75	Sierra Leone
118	274.50	Thailand
119	300.23	Croatia
120	308.10	Papua New Guinea
121	312.96	Yemen (Arab Republic of Yemen)
122	315.31	Bosnia-Herzegovina
123	341.30	Israel
124	343.72	Afghanistan
125	347.92	Tajikistan
126	357.91	South Africa
127	379.99	Rwanda
128	383.17	Niger
129	404.64	Morocco
130	429.15	Algeria
131	470.76	Philippines
132	472.86	Somalia
133	482.35	El Salvador
134	523.60	Pakistan

Continued on next page

Table A.3 – continued from previous page

Rank	Difference	Country
135	527.43	Angola
136	544.60	Syria
137	566.71	Sri Lanka (Ceylon)
138	606.57	Sudan
139	608.65	Burundi
140	628.05	Peru
141	695.39	Azerbaijan
142	728.51	Cambodia (Kampuchea)
143	733.87	Chad
144	781.50	Russia (Soviet Union)
145	809.34	Uganda
146	813.45	Iraq
147	837.31	Nepal
148	914.39	Malaysia
149	975.65	Kenya
150	1012.33	Iran (Persia)
151	1051.74	India
152	1256.57	Myanmar (Burma)
153	1366.15	China
154	1436.02	Nicaragua
155	1462.12	Indonesia
156	1501.44	Colombia
157	1577.27	Congo
158	1820.34	France
159	1889.56	Congo, Democratic Republic of (Zaire)
160	2027.80	Lebanon
161	3645.44	Laos
162	5050.84	United States of America

A.3 Predicted probability differences

Table A.4: Differences in predicted probabilities by country

Rank	Country	Difference
1	Bosnia-Herzegovina	39.18
2	Mauritius	40.26
3	Gabon	41.25
4	Azerbaijan	41.71
5	Liberia	42.01
6	Slovenia	42.15
7	Netherlands	42.22
8	Madagascar	42.35
9	Niger	42.47
10	Fiji	42.49
11	Solomon Islands	42.49
12	Austria	42.71
13	Afghanistan	42.76
14	India	42.84
15	Pakistan	42.99
16	Lithuania	43.13
17	Burundi	43.40
18	Benin	43.43
19	Sweden	43.44
20	Laos	43.44
21	Bhutan	43.62
22	Sudan	43.67
23	DRC	43.83
24	Swaziland	43.84
25	Trinidad and Tobago	43.86
26	Singapore	43.88
27	Mauritania	44.02
28	Guinea	44.07
29	South Korea	44.15
30	Cape Verde	44.16
31	Uzbekistan	44.22
32	Italy	44.31

Continued on next page

Table A.4 – continued from previous page

Rank	Country	Difference
33	Guyana	44.39
34	Cameroon	44.53
35	Dominican Republic	44.58
36	Slovakia	44.60
37	Croatia	44.62
38	Canada	44.62
39	Cuba	44.65
40	Mali	44.66
41	China	44.67
42	Nicaragua	44.83
43	Estonia	44.94
44	Rwanda	44.95
45	Lebanon	45.16
46	CAR	45.17
47	Comoros	45.23
48	Papua New Guinea	45.51
49	Ukraine	45.52
50	Jamaica	45.71
51	Surinam	45.76
52	Iran	45.77
53	Venezuela	45.90
54	Ethiopia	45.90
55	Bolivia	45.93
56	Equatorial Guinea	46.21
57	Argentina	46.37
58	Haiti	46.41
59	Morocco	46.57
60	Thailand	46.74
61	Oman	46.81
62	Georgia	46.84
63	Belgium	46.85
64	Belarus	46.91
65	Japan	46.94
66	Rumania	46.99

Continued on next page

Table A.4 – continued from previous page

Rank	Country	Difference
67	Hungary	47.05
68	Syria	47.07
69	Switzerland	47.08
70	Egypt	47.08
71	New Zealand	47.22
72	Algeria	47.36
73	Czech Republic	47.41
74	Tanzania	47.52
75	Greece	47.59
76	Libya	47.69
77	Russia	47.77
78	Philippines	47.80
79	Ecuador	47.87
80	USA	47.88
81	Denmark	47.94
82	Jordan	47.96
83	Panama	47.97
84	North Korea	48.03
85	Cote D'Ivoire	48.10
86	Moldova	48.11
87	Somalia	48.13
88	Burkina Faso	48.29
89	Malaysia	48.33
90	Latvia	48.35
91	Tajikistan	48.36
92	Cyprus	48.49
93	Namibia	48.50
94	Senegal	48.51
95	Ghana	48.54
96	Kazakhstan	48.77
97	Ireland	48.86
98	Gambia	48.90
99	Botswana	48.93
100	Qatar	48.97

Continued on next page

Table A.4 – continued from previous page

Rank	Country	Difference
101	Mexico	49.00
102	Angola	49.26
103	Kuwait	49.28
104	United Kingdom	49.35
105	Albania	49.39
106	Vietnam	49.42
107	Nepal	49.51
108	Germany	49.51
109	Uganda	49.64
110	Israel	49.68
111	France	49.87
112	Nigeria	49.92
113	Poland	49.94
114	Kyrgyzstan	50.06
115	Turkmenistan	50.08
116	Malawi	50.09
117	Guinea-Bissau	50.13
118	Lesotho	50.26
119	Togo	50.31
120	Portugal	50.43
121	Cambodia	50.47
122	Tunisia	50.56
123	Guatemala	50.61
124	Indonesia	50.71
125	Sri Lanka	50.76
126	South Africa	51.04
127	Congo	51.06
128	Yemen	51.22
129	Mongolia	51.27
130	Eritrea	51.52
131	Mozambique	51.92
132	Honduras	51.94
133	Costa Rica	52.13
134	Chad	52.29

Continued on next page

Table A.4 – continued from previous page

Rank	Country	Difference
135	Macedonia	52.29
136	Paraguay	52.33
137	Spain	52.46
138	Uruguay	52.55
139	Brazil	53.75
140	Australia	54.42
141	Saudi Arabia	54.46
142	Colombia	54.53
143	Zimbabwe	54.63
144	Taiwan	54.89
145	Bulgaria	55.95
146	Djibouti	55.99
147	Peru	56.34
148	Chile	57.41
149	Zambia	57.44
150	Bangladesh	59.91
151	Iraq	60.01
152	El Salvador	60.65
153	Sierra Leone	63.00
154	Turkey	65.59
155	Finland	69.60
156	Kenya	70.32
157	UAE	70.80
158	Luxembourg	73.68
159	Bahrain	74.02
160	Armenia	78.87
161	Norway	86.93

A.4 PR results

A.4.1 PR AUC differences

Full period

Table A.5: All countries sorted by effect on PR AUC 2001-2013

Rank	PR AUC	Difference	Country
1	0.721	-0.047	Finland
2	0.723	-0.044	Bahrain
3	0.724	-0.043	Turkey (Ottoman Empire)
4	0.728	-0.039	Ethiopia
5	0.732	-0.036	Indonesia
6	0.732	-0.035	Luxembourg
7	0.734	-0.033	United Arab Emirates
8	0.736	-0.031	Sudan
9	0.737	-0.030	Kyrgyz Republic
10	0.738	-0.029	Yemen (Arab Republic of Yemen)
11	0.738	-0.029	Mexico
12	0.740	-0.028	Uganda
13	0.740	-0.028	Armenia
14	0.741	-0.027	Cambodia (Kampuchea)
15	0.741	-0.026	Czech Republic
16	0.742	-0.026	Philippines
17	0.742	-0.025	Italy/Sardinia
18	0.742	-0.025	Norway
19	0.743	-0.025	Algeria
20	0.744	-0.024	Lesotho
21	0.744	-0.023	United Kingdom
22	0.745	-0.022	Korea, People's Republic of
23	0.745	-0.022	Colombia
24	0.746	-0.021	Honduras
25	0.747	-0.021	Bolivia
26	0.748	-0.019	Argentina
27	0.750	-0.018	Greece
28	0.750	-0.018	Paraguay
29	0.750	-0.018	Rumania

Continued on next page

Table A.5 – continued from previous page

Rank	PR AUC	Difference	Country
30	0.750	-0.017	Switzerland
31	0.750	-0.017	Zambia
32	0.750	-0.017	El Salvador
33	0.751	-0.017	Moldova
34	0.751	-0.017	Burundi
35	0.751	-0.016	Ukraine
36	0.751	-0.016	Ghana
37	0.751	-0.016	Mongolia
38	0.752	-0.016	Comoros
39	0.752	-0.016	Burkina Faso (Upper Volta)
40	0.752	-0.015	Jamaica
41	0.752	-0.015	Cuba
42	0.752	-0.015	Estonia
43	0.753	-0.015	Guyana
44	0.754	-0.014	Vietnam, Democratic Republic of
45	0.754	-0.014	Rwanda
46	0.754	-0.013	Oman
47	0.754	-0.013	Senegal
48	0.754	-0.013	Mozambique
49	0.755	-0.013	Portugal
50	0.755	-0.013	Afghanistan
51	0.755	-0.012	Togo
52	0.755	-0.012	Mauritania
53	0.755	-0.012	Pakistan
54	0.755	-0.012	Mauritius
55	0.755	-0.012	India
56	0.756	-0.011	Solomon Islands
57	0.756	-0.011	Trinidad and Tobago
58	0.756	-0.011	Fiji
59	0.756	-0.011	Korea, Republic of
60	0.756	-0.011	Kuwait
61	0.756	-0.011	Australia
62	0.757	-0.011	Costa Rica
63	0.757	-0.011	Guinea-Bissau

Continued on next page

Table A.5 – continued from previous page

Rank	PR AUC	Difference	Country
64	0.757	-0.011	Belgium
65	0.757	-0.011	Surinam
66	0.757	-0.010	Tajikistan
67	0.757	-0.010	Bosnia-Herzegovina
68	0.758	-0.010	German Federal Republic
69	0.758	-0.010	Hungary
70	0.758	-0.010	Tunisia
71	0.758	-0.010	Israel
72	0.758	-0.010	Cameroon
73	0.758	-0.010	Russia (Soviet Union)
74	0.758	-0.010	Brazil
75	0.758	-0.009	Kazakhstan
76	0.758	-0.009	Peru
77	0.758	-0.009	Ecuador
78	0.759	-0.008	Albania
79	0.759	-0.008	Eritrea
80	0.760	-0.008	Guatemala
81	0.760	-0.007	Croatia
82	0.760	-0.007	Madagascar
83	0.760	-0.007	Singapore
84	0.760	-0.007	Thailand
85	0.761	-0.007	Netherlands
86	0.761	-0.007	United States of America
87	0.761	-0.007	Venezuela
88	0.762	-0.006	Cote D?Ivoire
89	0.762	-0.006	Latvia
90	0.762	-0.006	Nepal
91	0.762	-0.006	Malawi
92	0.762	-0.006	Namibia
93	0.762	-0.005	Saudi Arabia
94	0.762	-0.005	South Africa
95	0.762	-0.005	Liberia
96	0.763	-0.005	Lebanon
97	0.763	-0.005	Bangladesh

Continued on next page

Table A.5 – continued from previous page

Rank	PR AUC	Difference	Country
98	0.763	-0.005	Gambia
99	0.763	-0.004	Niger
100	0.764	-0.004	Haiti
101	0.764	-0.004	Egypt
102	0.764	-0.004	Benin
103	0.764	-0.003	Azerbaijan
104	0.764	-0.003	Somalia
105	0.764	-0.003	Belarus (Byelorussia)
106	0.765	-0.003	Bhutan
107	0.765	-0.002	Zimbabwe (Rhodesia)
108	0.765	-0.002	Congo
109	0.766	-0.002	Iraq
110	0.766	-0.002	Laos
111	0.766	-0.001	Botswana
112	0.766	-0.001	Nigeria
113	0.766	-0.001	Bulgaria
114	0.767	-0.001	Morocco
115	0.767	-0.001	Tanzania/Tanganyika
116	0.767	-0.001	Swaziland
117	0.767	-0.001	Papua New Guinea
118	0.767	-0.001	New Zealand
119	0.767	-0.001	Taiwan
120	0.767	-0.000	Denmark
121	0.767	-0.000	Central African Republic
122	0.767	-0.000	Libya
123	0.767	-0.000	Malaysia
124	0.768	0.000	Cyprus
125	0.768	0.000	Slovakia
126	0.768	0.001	Ireland
127	0.768	0.001	Mali
128	0.769	0.001	Guinea
129	0.769	0.001	Japan
130	0.769	0.001	Dominican Republic
131	0.769	0.001	Panama

Continued on next page

Table A.5 – continued from previous page

Rank	PR AUC	Difference	Country
132	0.769	0.001	Nicaragua
133	0.769	0.002	France
134	0.769	0.002	Sri Lanka (Ceylon)
135	0.770	0.002	Cape Verde
136	0.770	0.002	Turkmenistan
137	0.770	0.002	Chile
138	0.770	0.003	Kenya
139	0.771	0.003	Sweden
140	0.771	0.003	Austria
141	0.771	0.004	Georgia
142	0.771	0.004	Equatorial Guinea
143	0.771	0.004	Poland
144	0.773	0.005	Jordan
145	0.773	0.005	Uruguay
146	0.773	0.005	Gabon
147	0.773	0.006	Iran (Persia)
148	0.773	0.006	Sierra Leone
149	0.773	0.006	Lithuania
150	0.774	0.007	Canada
151	0.775	0.007	Chad
152	0.775	0.007	China
153	0.775	0.008	Uzbekistan
154	0.775	0.008	Slovenia
155	0.776	0.008	Syria
156	0.777	0.010	Congo, Democratic Republic of (Zaire)
157	0.778	0.010	Macedonia (Former Yugoslav Republic of)
158	0.779	0.011	Qatar
159	0.779	0.011	Angola
160	0.783	0.016	Spain
161	0.791	0.023	Djibouti

Limited period

Table A.6: All countries sorted by effect on PR AUC 2006-2013

Rank	PR AUC	Difference	Country
1	0.697	-0.084	Turkey (Ottoman Empire)
2	0.721	-0.060	Sudan
3	0.721	-0.060	Mexico
4	0.721	-0.060	Cambodia (Kampuchea)
5	0.721	-0.060	Kyrgyz Republic
6	0.722	-0.060	Uganda
7	0.724	-0.058	Ethiopia
8	0.728	-0.053	Yemen (Arab Republic of Yemen)
9	0.729	-0.052	United Kingdom
10	0.730	-0.052	Zambia
11	0.730	-0.051	Philippines
12	0.731	-0.050	Finland
13	0.734	-0.047	Luxembourg
14	0.734	-0.047	Italy/Sardinia
15	0.735	-0.046	Lesotho
16	0.737	-0.044	Bahrain
17	0.737	-0.044	El Salvador
18	0.737	-0.044	Afghanistan
19	0.738	-0.043	Czech Republic
20	0.738	-0.043	Rwanda
21	0.739	-0.042	United Arab Emirates
22	0.739	-0.042	Honduras
23	0.741	-0.040	Thailand
24	0.741	-0.040	Korea, People's Republic of
25	0.741	-0.040	Algeria
26	0.742	-0.039	Vietnam, Democratic Republic of
27	0.742	-0.039	Togo
28	0.743	-0.038	Switzerland
29	0.744	-0.037	Norway
30	0.745	-0.037	Greece
31	0.745	-0.036	Zimbabwe (Rhodesia)
32	0.745	-0.036	Surinam
33	0.745	-0.036	German Federal Republic
34	0.746	-0.035	Guinea-Bissau

Continued on next page

Table A.6 – continued from previous page

Rank	PR AUC	Difference	Country
35	0.746	-0.035	Ukraine
36	0.747	-0.035	Paraguay
37	0.747	-0.034	Burundi
38	0.748	-0.033	Tunisia
39	0.748	-0.033	Cuba
40	0.748	-0.033	Moldova
41	0.748	-0.033	Belgium
42	0.748	-0.033	Indonesia
43	0.749	-0.032	Comoros
44	0.749	-0.032	Bangladesh
45	0.749	-0.032	Mauritius
46	0.749	-0.032	Iraq
47	0.750	-0.031	Latvia
48	0.750	-0.031	Portugal
49	0.750	-0.031	Denmark
50	0.750	-0.031	Bolivia
51	0.751	-0.030	Guatemala
52	0.751	-0.030	Australia
53	0.751	-0.030	Somalia
54	0.751	-0.030	Hungary
55	0.752	-0.030	Saudi Arabia
56	0.752	-0.029	Jamaica
57	0.752	-0.029	Rumania
58	0.752	-0.029	India
59	0.752	-0.029	Armenia
60	0.752	-0.029	Brazil
61	0.753	-0.028	Belarus (Byelorussia)
62	0.753	-0.028	Lebanon
63	0.754	-0.027	Trinidad and Tobago
64	0.754	-0.027	Fiji
65	0.754	-0.027	Mongolia
66	0.754	-0.027	Sri Lanka (Ceylon)
67	0.755	-0.026	Kuwait
68	0.756	-0.025	Peru

Continued on next page

Table A.6 – continued from previous page

Rank	PR AUC	Difference	Country
69	0.756	-0.025	Mozambique
70	0.757	-0.024	Kazakhstan
71	0.757	-0.024	Eritrea
72	0.758	-0.024	Colombia
73	0.758	-0.023	Liberia
74	0.758	-0.023	Bulgaria
75	0.758	-0.023	Ghana
76	0.758	-0.023	Central African Republic
77	0.758	-0.023	United States of America
78	0.759	-0.022	Madagascar
79	0.759	-0.022	Croatia
80	0.759	-0.022	Solomon Islands
81	0.760	-0.021	Mauritania
82	0.760	-0.021	Burkina Faso (Upper Volta)
83	0.760	-0.021	Tajikistan
84	0.760	-0.021	Nepal
85	0.760	-0.021	Estonia
86	0.761	-0.020	Egypt
87	0.762	-0.019	Oman
88	0.762	-0.019	Malawi
89	0.762	-0.019	Albania
90	0.762	-0.019	Sierra Leone
91	0.762	-0.019	Korea, Republic of
92	0.762	-0.019	Senegal
93	0.763	-0.018	Guinea
94	0.763	-0.018	Gambia
95	0.763	-0.018	Cote D?Ivoire
96	0.763	-0.018	Chile
97	0.763	-0.018	Taiwan
98	0.763	-0.018	Bosnia-Herzegovina
99	0.763	-0.018	Costa Rica
100	0.765	-0.016	Israel
101	0.765	-0.016	Argentina
102	0.765	-0.016	Namibia

Continued on next page

Table A.6 – continued from previous page

Rank	PR AUC	Difference	Country
103	0.766	-0.015	Slovakia
104	0.766	-0.015	Benin
105	0.766	-0.015	Singapore
106	0.766	-0.015	Netherlands
107	0.767	-0.014	Ecuador
108	0.767	-0.014	Guyana
109	0.767	-0.014	New Zealand
110	0.767	-0.014	Ireland
111	0.768	-0.013	South Africa
112	0.769	-0.012	Kenya
113	0.769	-0.012	Poland
114	0.769	-0.012	Japan
115	0.769	-0.012	Russia (Soviet Union)
116	0.770	-0.011	Morocco
117	0.771	-0.010	Azerbaijan
118	0.772	-0.009	Syria
119	0.772	-0.009	Congo
120	0.772	-0.009	Haiti
121	0.772	-0.009	Sweden
122	0.773	-0.008	Swaziland
123	0.773	-0.008	France
124	0.773	-0.008	Dominican Republic
125	0.773	-0.008	Malaysia
126	0.774	-0.008	Uruguay
127	0.774	-0.007	Cameroon
128	0.774	-0.007	Gabon
129	0.774	-0.007	Georgia
130	0.774	-0.007	Iran (Persia)
131	0.775	-0.006	Equatorial Guinea
132	0.775	-0.006	Nicaragua
133	0.775	-0.006	Venezuela
134	0.775	-0.006	Chad
135	0.776	-0.005	Qatar
136	0.776	-0.005	Slovenia

Continued on next page

Table A.6 – continued from previous page

Rank	PR AUC	Difference	Country
137	0.776	-0.005	Uzbekistan
138	0.777	-0.005	Macedonia (Former Yugoslav Republic of)
139	0.777	-0.004	Panama
140	0.778	-0.004	Pakistan
141	0.779	-0.002	Niger
142	0.780	-0.002	Papua New Guinea
143	0.780	-0.001	Botswana
144	0.780	-0.001	Nigeria
145	0.780	-0.001	Jordan
146	0.780	-0.001	Turkmenistan
147	0.782	0.001	Cape Verde
148	0.782	0.001	Laos
149	0.782	0.001	China
150	0.782	0.001	Libya
151	0.782	0.001	Bhutan
152	0.782	0.001	Lithuania
153	0.784	0.003	Canada
154	0.784	0.003	Mali
155	0.785	0.004	Cyprus
156	0.785	0.004	Congo, Democratic Republic of (Zaire)
157	0.786	0.005	Spain
158	0.786	0.005	Austria
159	0.788	0.007	Djibouti
160	0.789	0.008	Tanzania/Tanganyika
161	0.799	0.018	Angola

A.4.2 Descriptive statistics by group

Countries are divided into three groups depending on how their presence affects model precision. The *destructive* group is comprised of the countries with the greatest negative impact on model precision, more precisely any country that is in the top ten list of either evaluation period. The *reinforcing* group are those on the other end of the scale, being those that lead to the greatest reduction in precision when removed. The others are the remaining countries

	Destructive	Reinforcing	Others
N	16	15	131

Table A.7: PR outlier group sizes.

	No conflict	Minor conflict	Major conflict
Destructive	0.88	0.07	0.05
Reinforcing	0.68	0.20	0.11
Others	0.86	0.10	0.04

Table A.8: The proportion of dyads in different conflict states by group.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
			ltimeindep			
Destructive	2.16	2.70	3.03	3.42	4.19	5.64
Reinforcing	2.20	2.97	3.42	3.84	4.78	5.64
Others	2.12	2.96	3.27	3.82	5.01	5.66
			ltsc0			
Destructive	0.06	1.41	2.11	2.26	2.95	4.73
Reinforcing	0.29	0.71	2.17	2.03	2.82	4.73
Others	0.00	1.74	2.40	2.55	3.23	5.64

Table A.9: Descriptive statistics for logged years since independence and logged years in a state of peace.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
			ncts0			
Destructive	0.00	0.00	0.74	1.34	2.22	4.65
Reinforcing	0.03	0.34	1.06	1.35	2.50	4.59
Others	0.00	0.42	1.40	1.63	2.55	5.64
			ltsnc			
Destructive	1.08	1.54	2.29	2.18	2.67	3.21
Reinforcing	0.78	1.62	1.98	2.00	2.39	3.21
Others	0.85	1.63	2.13	2.13	2.64	3.21
			lpop			
Destructive	6.14	7.62	9.05	8.99	10.10	13.77
Reinforcing	5.92	8.49	9.24	9.25	10.63	11.87
Others	5.89	7.99	8.75	8.84	9.64	13.47
			lGDPcap			
Destructive	6.42	7.40	8.50	8.47	9.59	10.73
Reinforcing	6.29	7.16	7.45	8.30	9.77	11.08
Others	6.19	7.33	8.30	8.23	9.05	10.36
			nb_lGDPcap			
Destructive	6.56	7.28	8.83	8.44	9.56	10.15
Reinforcing	6.79	7.24	8.65	8.53	9.69	10.02
Others	6.54	7.40	8.20	8.24	8.94	9.99
			polity2			
Destructive	-10.00	-5.08	-2.58	0.39	8.97	10.00
Reinforcing	-9.37	-4.01	-2.73	-0.58	2.84	10.00
Others	-10.00	-4.27	-0.27	0.64	5.84	10.00
			polity2sq			
Destructive	22.83	39.53	67.17	66.13	100.00	100.00
Reinforcing	9.00	30.84	44.58	52.23	76.17	100.00
Others	0.00	36.00	46.91	52.35	64.66	100.00
			nb_TSRC_5			
Destructive	0.00	0.33	0.54	0.50	0.75	0.83
Reinforcing	0.00	0.18	0.57	0.44	0.70	0.84
Others	0.00	0.27	0.52	0.47	0.71	0.86

Table A.10: Descriptive statistics for ncts0, ltsnc, lpop,lGDPcap, nb_lGDPcap, polity2, polity2sq and nb_TSRC_5.

A.5 Brier results

A.5.1 Brier score differences

Tables of all countries and the effect on model Brier score when dropping them from the dataset. As a lower score is better with Brier, the effects here are opposite of those on PR. A positive effect when dropped equals a negative impact, meaning the country is reinforcing, and vice versa for negative effects.

Full period

Table A.11: All countries sorted by effect on Brier score 2001-2013

Rank	Brier score	Difference	Country
1	0.067	0.006	Turkey (Ottoman Empire)
2	0.066	0.005	Norway
3	0.066	0.005	Finland
4	0.066	0.005	Bahrain
5	0.066	0.004	Armenia
6	0.066	0.004	United Arab Emirates
7	0.065	0.004	Luxembourg
8	0.065	0.003	Mexico
9	0.064	0.003	Cambodia (Kampuchea)
10	0.064	0.003	United Kingdom
11	0.064	0.003	Kyrgyz Republic
12	0.064	0.003	Italy/Sardinia
13	0.064	0.003	Fiji
14	0.064	0.003	Togo
15	0.064	0.003	Indonesia
16	0.064	0.003	Argentina
17	0.064	0.003	Korea, People's Republic of
18	0.064	0.003	Guatemala
19	0.064	0.003	Bolivia
20	0.064	0.003	El Salvador
21	0.064	0.002	Mozambique
22	0.064	0.002	Latvia
23	0.064	0.002	Yemen (Arab Republic of Yemen)
24	0.064	0.002	Tunisia

Continued on next page

Table A.11 – continued from previous page

Rank	Brier score	Difference	Country
25	0.064	0.002	Rumania
26	0.064	0.002	Burkina Faso (Upper Volta)
27	0.064	0.002	Zambia
28	0.064	0.002	Honduras
29	0.064	0.002	Vietnam, Democratic Republic of
30	0.064	0.002	Comoros
31	0.064	0.002	Czech Republic
32	0.064	0.002	Portugal
33	0.063	0.002	Surinam
34	0.063	0.002	New Zealand
35	0.063	0.002	Australia
36	0.063	0.002	Trinidad and Tobago
37	0.063	0.002	German Federal Republic
38	0.063	0.002	Ukraine
39	0.063	0.002	Oman
40	0.063	0.002	Malawi
41	0.063	0.002	Guinea-Bissau
42	0.063	0.002	Switzerland
43	0.063	0.002	Brazil
44	0.063	0.002	Philippines
45	0.063	0.002	Korea, Republic of
46	0.063	0.002	Belarus (Byelorussia)
47	0.063	0.002	Albania
48	0.063	0.002	Madagascar
49	0.063	0.002	Jamaica
50	0.063	0.002	Cuba
51	0.063	0.002	Belgium
52	0.063	0.002	Greece
53	0.063	0.002	Lebanon
54	0.063	0.002	Bosnia-Herzegovina
55	0.063	0.001	Uganda
56	0.063	0.001	Tajikistan
57	0.063	0.001	Netherlands
58	0.063	0.001	Ghana

Continued on next page

Table A.11 – continued from previous page

Rank	Brier score	Difference	Country
59	0.063	0.001	Singapore
60	0.063	0.001	Gambia
61	0.063	0.001	Ecuador
62	0.063	0.001	Costa Rica
63	0.063	0.001	Afghanistan
64	0.063	0.001	Ethiopia
65	0.063	0.001	Slovakia
66	0.063	0.001	Mauritania
67	0.062	0.001	Gabon
68	0.062	0.001	Zimbabwe (Rhodesia)
69	0.062	0.001	Solomon Islands
70	0.062	0.001	Taiwan
71	0.062	0.001	Paraguay
72	0.062	0.001	Mongolia
73	0.062	0.001	Mauritius
74	0.062	0.001	Algeria
75	0.062	0.001	Peru
76	0.062	0.001	Morocco
77	0.062	0.001	Croatia
78	0.062	0.001	Lesotho
79	0.062	0.001	Namibia
80	0.062	0.001	Saudi Arabia
81	0.062	0.001	Burundi
82	0.062	0.001	Haiti
83	0.062	0.001	Congo
84	0.062	0.001	Guyana
85	0.062	0.001	Uruguay
86	0.062	0.001	Moldova
87	0.062	0.001	Colombia
88	0.062	0.001	Nicaragua
89	0.062	0.001	Estonia
90	0.062	0.001	United States of America
91	0.062	0.001	South Africa
92	0.062	0.001	Swaziland

Continued on next page

Table A.11 – continued from previous page

Rank	Brier score	Difference	Country
93	0.062	0.001	Kuwait
94	0.062	0.001	Eritrea
95	0.062	0.001	Azerbaijan
96	0.062	0.000	Denmark
97	0.062	0.000	Cameroon
98	0.062	0.000	Senegal
99	0.062	0.000	Sri Lanka (Ceylon)
100	0.062	0.000	Sudan
101	0.062	0.000	Kazakhstan
102	0.062	0.000	Venezuela
103	0.062	0.000	Japan
104	0.062	0.000	Benin
105	0.062	0.000	Turkmenistan
106	0.062	0.000	Kenya
107	0.062	0.000	Bangladesh
108	0.062	0.000	Chile
109	0.062	0.000	Russia (Soviet Union)
110	0.062	0.000	Botswana
111	0.062	0.000	Libya
112	0.062	0.000	Equatorial Guinea
113	0.062	0.000	Hungary
114	0.062	0.000	Cape Verde
115	0.061	0.000	Egypt
116	0.061	0.000	France
117	0.061	0.000	Laos
118	0.061	0.000	Cote D?Ivoire
119	0.061	-0.000	Cyprus
120	0.061	-0.000	Liberia
121	0.061	-0.000	Bhutan
122	0.061	-0.000	China
123	0.061	-0.000	Guinea
124	0.061	-0.000	Rwanda
125	0.061	-0.000	Papua New Guinea
126	0.061	-0.000	Qatar

Continued on next page

Table A.11 – continued from previous page

Rank	Brier score	Difference	Country
127	0.061	-0.000	Georgia
128	0.061	-0.000	Sweden
129	0.061	-0.000	Panama
130	0.061	-0.000	Nepal
131	0.061	-0.000	Spain
132	0.061	-0.000	Dominican Republic
133	0.061	-0.000	Macedonia (Former Yugoslav Republic of)
134	0.061	-0.000	Tanzania/Tanganyika
135	0.061	-0.000	Nigeria
136	0.061	-0.000	Canada
137	0.061	-0.000	Israel
138	0.061	-0.000	India
139	0.061	-0.000	Lithuania
140	0.061	-0.000	Austria
141	0.061	-0.000	Bulgaria
142	0.061	-0.000	Jordan
143	0.061	-0.001	Uzbekistan
144	0.061	-0.001	Niger
145	0.061	-0.001	Malaysia
146	0.061	-0.001	Iraq
147	0.061	-0.001	Ireland
148	0.061	-0.001	Slovenia
149	0.060	-0.001	Mali
150	0.060	-0.001	Thailand
151	0.060	-0.001	Djibouti
152	0.060	-0.001	Syria
153	0.060	-0.001	Chad
154	0.060	-0.001	Sierra Leone
155	0.060	-0.001	Poland
156	0.060	-0.001	Pakistan
157	0.060	-0.002	Iran (Persia)
158	0.060	-0.002	Central African Republic
159	0.059	-0.002	Somalia
160	0.059	-0.002	Angola

Continued on next page

Table A.11 – continued from previous page

Rank	Brier score	Difference	Country
161	0.059	-0.003	Congo, Democratic Republic of (Zaire)

Limited period

Table A.12: All countries sorted by effect on Brier score 2006-2013

Rank	Brier score	Difference	Country
1	0.071	0.007	Turkey (Ottoman Empire)
2	0.070	0.006	Finland
3	0.069	0.006	Norway
4	0.069	0.006	Armenia
5	0.069	0.005	United Arab Emirates
6	0.069	0.005	Bahrain
7	0.069	0.005	United Kingdom
8	0.068	0.005	Kyrgyz Republic
9	0.068	0.004	Mexico
10	0.068	0.004	Korea, People's Republic of
11	0.068	0.004	Luxembourg
12	0.068	0.004	Tunisia
13	0.068	0.004	Cambodia (Kampuchea)
14	0.068	0.004	Vietnam, Democratic Republic of
15	0.068	0.004	Italy/Sardinia
16	0.068	0.004	Argentina
17	0.068	0.004	Fiji
18	0.067	0.004	Bolivia
19	0.067	0.004	Togo
20	0.067	0.003	Latvia
21	0.067	0.003	New Zealand
22	0.067	0.003	Rumania
23	0.067	0.003	Comoros
24	0.067	0.003	Guatemala
25	0.067	0.003	Indonesia
26	0.067	0.003	Portugal

Continued on next page

Table A.12 – continued from previous page

Rank	Brier score	Difference	Country
27	0.067	0.003	German Federal Republic
28	0.067	0.003	Honduras
29	0.067	0.003	Burkina Faso (Upper Volta)
30	0.067	0.003	Trinidad and Tobago
31	0.067	0.003	Afghanistan
32	0.067	0.003	Zambia
33	0.067	0.003	Surinam
34	0.067	0.003	El Salvador
35	0.067	0.003	Jamaica
36	0.066	0.003	Lebanon
37	0.066	0.003	Ukraine
38	0.066	0.003	Brazil
39	0.066	0.003	Sri Lanka (Ceylon)
40	0.066	0.003	Belarus (Byelorussia)
41	0.066	0.003	Belgium
42	0.066	0.002	Uganda
43	0.066	0.002	Madagascar
44	0.066	0.002	Malawi
45	0.066	0.002	Mozambique
46	0.066	0.002	Cuba
47	0.066	0.002	Czech Republic
48	0.066	0.002	Tajikistan
49	0.066	0.002	Gambia
50	0.066	0.002	Australia
51	0.066	0.002	Guinea-Bissau
52	0.066	0.002	Oman
53	0.066	0.002	Switzerland
54	0.066	0.002	Yemen (Arab Republic of Yemen)
55	0.066	0.002	Gabon
56	0.066	0.002	Korea, Republic of
57	0.066	0.002	Slovakia
58	0.066	0.002	Philippines
59	0.066	0.002	Senegal
60	0.066	0.002	Greece

Continued on next page

Table A.12 – continued from previous page

Rank	Brier score	Difference	Country
61	0.066	0.002	Bosnia-Herzegovina
62	0.066	0.002	Costa Rica
63	0.066	0.002	Solomon Islands
64	0.066	0.002	Zimbabwe (Rhodesia)
65	0.066	0.002	Ecuador
66	0.066	0.002	United States of America
67	0.066	0.002	Albania
68	0.065	0.002	Burundi
69	0.065	0.002	Algeria
70	0.065	0.002	Netherlands
71	0.065	0.002	Congo
72	0.065	0.002	Haiti
73	0.065	0.002	Eritrea
74	0.065	0.002	Denmark
75	0.065	0.002	Liberia
76	0.065	0.002	Azerbaijan
77	0.065	0.001	Ghana
78	0.065	0.001	Croatia
79	0.065	0.001	Mauritania
80	0.065	0.001	Colombia
81	0.065	0.001	Mauritius
82	0.065	0.001	Morocco
83	0.065	0.001	Ethiopia
84	0.065	0.001	Guinea
85	0.065	0.001	Cote D?Ivoire
86	0.065	0.001	Uzbekistan
87	0.065	0.001	Libya
88	0.065	0.001	Paraguay
89	0.065	0.001	Namibia
90	0.065	0.001	Mongolia
91	0.065	0.001	Saudi Arabia
92	0.065	0.001	Sudan
93	0.065	0.001	Taiwan
94	0.065	0.001	Kuwait

Continued on next page

Table A.12 – continued from previous page

Rank	Brier score	Difference	Country
95	0.065	0.001	Estonia
96	0.065	0.001	Nicaragua
97	0.065	0.001	Rwanda
98	0.064	0.001	Moldova
99	0.064	0.001	Singapore
100	0.064	0.001	Cameroon
101	0.064	0.001	Swaziland
102	0.064	0.001	Kazakhstan
103	0.064	0.001	Equatorial Guinea
104	0.064	0.001	Egypt
105	0.064	0.000	Guyana
106	0.064	0.000	Uruguay
107	0.064	0.000	Peru
108	0.064	0.000	South Africa
109	0.064	0.000	Benin
110	0.064	0.000	Bangladesh
111	0.064	0.000	Lesotho
112	0.064	0.000	Hungary
113	0.064	0.000	Cape Verde
114	0.064	0.000	Japan
115	0.064	0.000	Nepal
116	0.064	0.000	Jordan
117	0.064	0.000	India
118	0.064	-0.000	Turkmenistan
119	0.064	-0.000	Thailand
120	0.064	-0.000	China
121	0.064	-0.000	Iraq
122	0.064	-0.000	Cyprus
123	0.064	-0.000	Sweden
124	0.064	-0.000	Russia (Soviet Union)
125	0.064	-0.000	Georgia
126	0.064	-0.000	Botswana
127	0.064	-0.000	Bhutan
128	0.064	-0.000	Macedonia (Former Yugoslav Republic of)

Continued on next page

Table A.12 – continued from previous page

Rank	Brier score	Difference	Country
129	0.064	-0.000	Panama
130	0.064	-0.000	Ireland
131	0.064	-0.000	Laos
132	0.063	-0.000	Chad
133	0.063	-0.000	Nigeria
134	0.063	-0.000	Lithuania
135	0.063	-0.000	Chile
136	0.063	-0.001	Venezuela
137	0.063	-0.001	Qatar
138	0.063	-0.001	France
139	0.063	-0.001	Israel
140	0.063	-0.001	Bulgaria
141	0.063	-0.001	Kenya
142	0.063	-0.001	Papua New Guinea
143	0.063	-0.001	Malaysia
144	0.063	-0.001	Spain
145	0.063	-0.001	Dominican Republic
146	0.063	-0.001	Djibouti
147	0.063	-0.001	Tanzania/Tanganyika
148	0.063	-0.001	Sierra Leone
149	0.063	-0.001	Austria
150	0.063	-0.001	Canada
151	0.063	-0.001	Iran (Persia)
152	0.063	-0.001	Slovenia
153	0.062	-0.001	Central African Republic
154	0.062	-0.001	Niger
155	0.062	-0.002	Syria
156	0.062	-0.002	Mali
157	0.062	-0.002	Congo, Democratic Republic of (Zaire)
158	0.062	-0.002	Somalia
159	0.062	-0.002	Poland
160	0.061	-0.003	Pakistan
161	0.060	-0.003	Angola

A.5.2 Descriptive statistics by group

	Destructive	Reinforcing	Others
N	13	12	137

Table A.13: Brier outlier group sizes.

	No conflict	Minor conflict	Major conflict
Destructive	0.71	0.19	0.09
Reinforcing	0.87	0.11	0.02
Others	0.85	0.10	0.05

Table A.14: The proportion of dyads in different conflict states by group.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
			ltimeindep			
Destructive	2.16	3.03	3.03	3.27	3.33	5.64
Reinforcing	2.20	2.80	3.70	3.90	5.08	5.64
Others	2.12	2.93	3.33	3.83	5.01	5.66
			ltsc0			
Destructive	0.06	0.95	1.74	1.48	2.01	3.37
Reinforcing	0.84	2.16	2.74	2.95	3.48	5.11
Others	0.00	1.64	2.42	2.52	3.21	5.64
			ncts0			
Destructive	0.00	0.01	0.15	0.58	1.05	2.03
Reinforcing	0.00	0.94	2.15	1.92	2.64	4.59
Others	0.00	0.43	1.38	1.64	2.63	5.64

Table A.15: Descriptive statistics by group.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
			nc			
Destructive	0.13	0.38	0.86	0.68	0.93	1.00
Reinforcing	0.00	0.11	0.36	0.40	0.57	1.00
Others	0.00	0.03	0.38	0.43	0.80	1.00
			ltsnc			
Destructive	1.08	1.61	2.13	2.00	2.30	3.14
Reinforcing	0.78	1.68	2.13	2.01	2.22	3.21
Others	0.85	1.60	2.13	2.15	2.66	3.21
			lpop			
Destructive	7.60	8.62	8.99	9.23	10.38	11.34
Reinforcing	5.92	7.93	8.49	8.72	9.97	11.11
Others	5.89	7.85	8.87	8.88	9.69	13.77
			lGDPcap			
Destructive	6.42	6.81	6.90	7.53	8.06	9.94
Reinforcing	6.66	7.87	9.34	9.02	9.96	11.08
Others	6.19	7.30	8.34	8.26	9.14	10.73
			nb_lGDPcap			
Destructive	6.61	7.17	7.38	7.73	8.03	9.91
Reinforcing	7.25	8.53	9.33	9.03	9.73	10.02
Others	6.54	7.38	8.28	8.27	9.01	10.15
			polity2			
Destructive	-6.44	-4.63	-2.35	-2.03	-1.48	10.00
Reinforcing	-9.37	-4.25	1.74	1.50	10.00	10.00
Others	-10.00	-4.27	-0.27	0.65	6.02	10.00
			polity2sq			
Destructive	24.51	34.11	40.70	44.95	49.70	100.00
Reinforcing	9.00	32.24	72.64	65.51	100.00	100.00
Others	0.00	36.00	47.42	53.50	69.23	100.00
			nb_TSRC_5			
Destructive	0.38	0.59	0.73	0.68	0.81	0.86
Reinforcing	0.00	0.21	0.39	0.40	0.61	0.84
Others	0.00	0.25	0.52	0.46	0.70	0.84

Table A.16: Descriptive statistics for ncts0, ltsnc, lpop,lGDPcap, nb_lGDPcap, polity2, polity2sq and nb_TSRC_5.

Appendix B

Figures

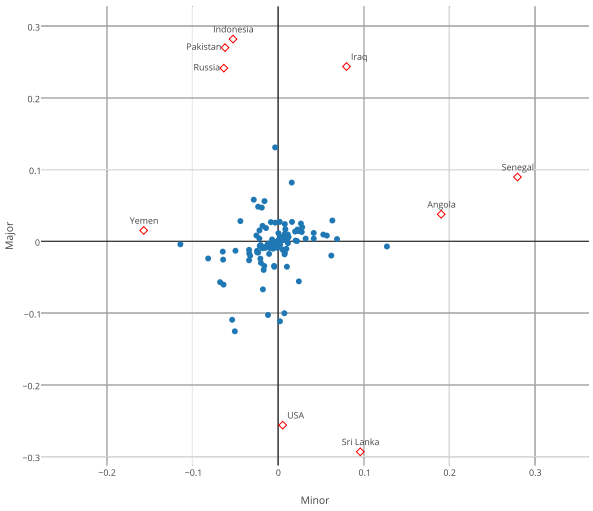
B.1 Coefficient effects

This appendix contains all figures referenced in the main text.

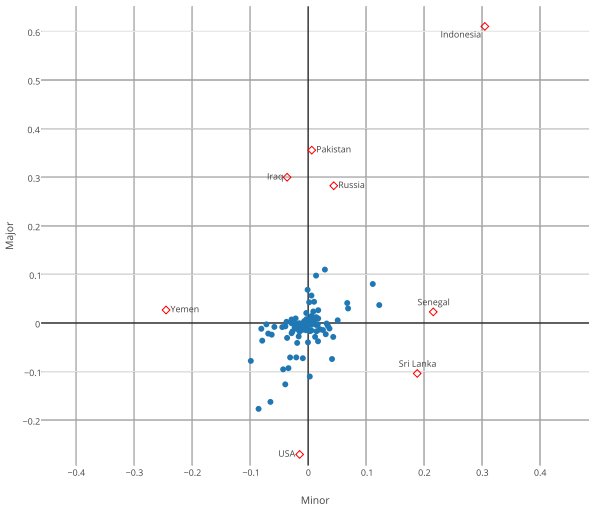
Figures B.1 and B.2 shows the centered effect of dropping each country on coefficients. The countries with the most extreme effects are highlighted and named.

Figures B.3 and B.4 shows the centered effect of dropping each country on coefficients with the destructive countries by PR highlighted.

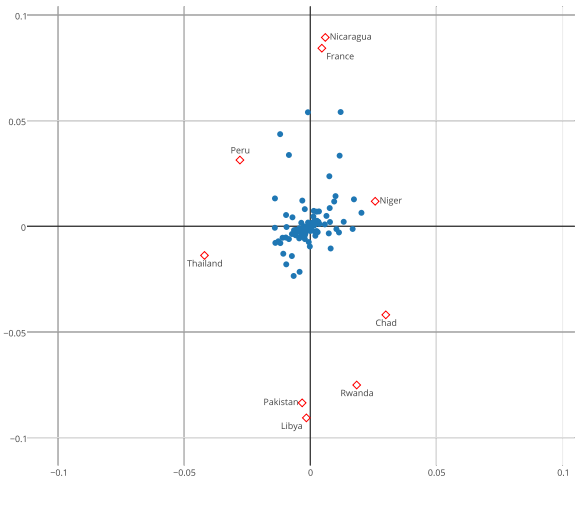
Figure B.5 shows distribution of countries' average polity scores by group.



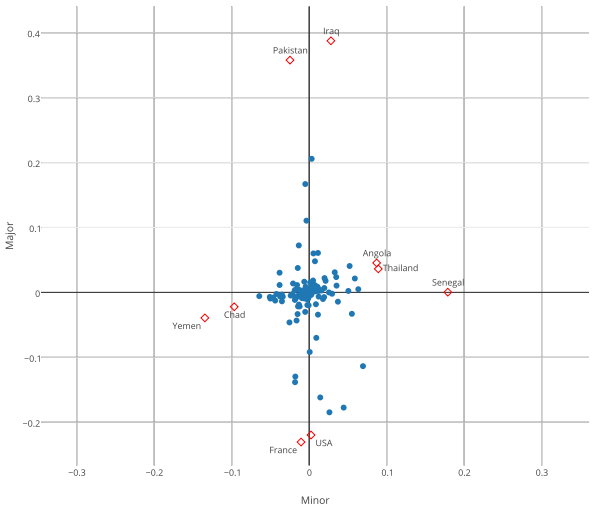
(a) c1



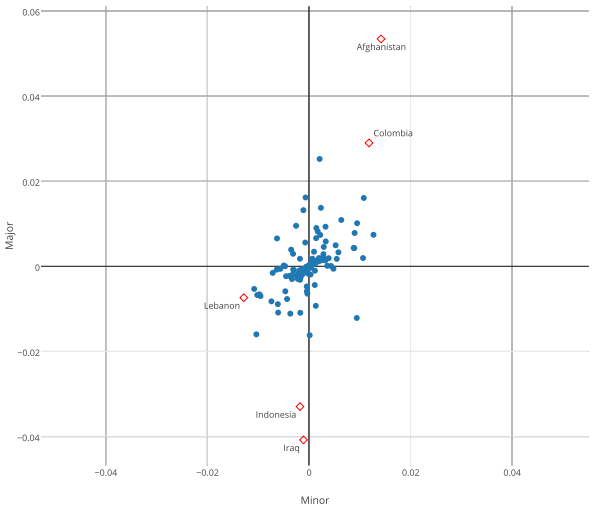
(b) c2



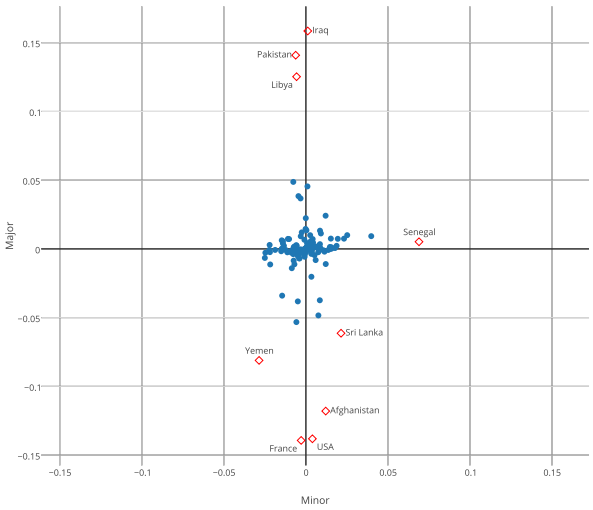
(c) ltsc0



(d) nc

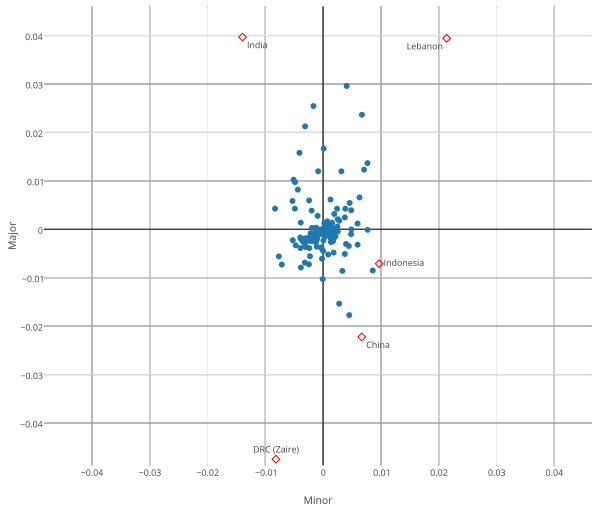


(e) ltsnc

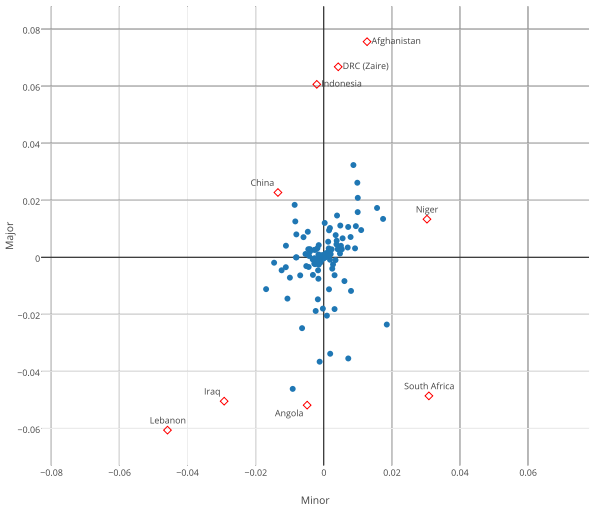


(f) ncts0

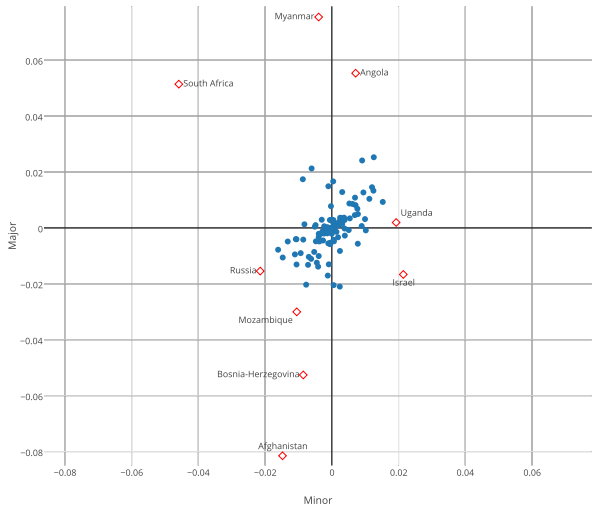
Figure B.1: Differences in coefficients resulting from country drops. X-axes are coefficients for minor outcomes and Y-axes are for major.



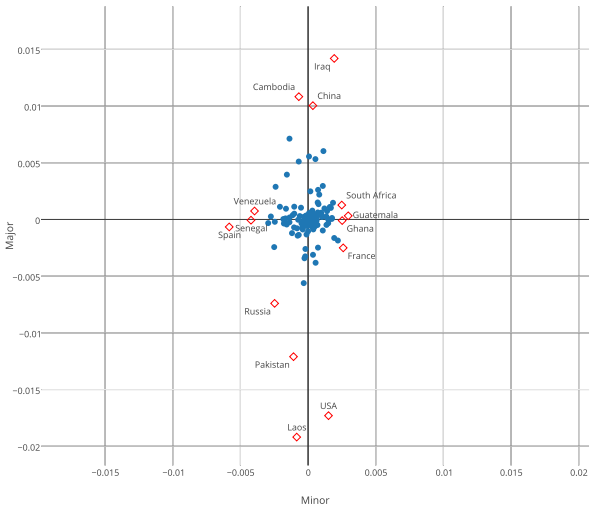
(a) lpop



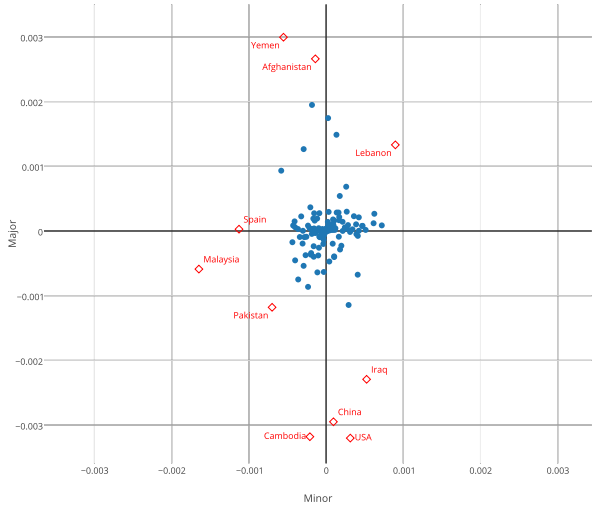
(b) lgdpcap



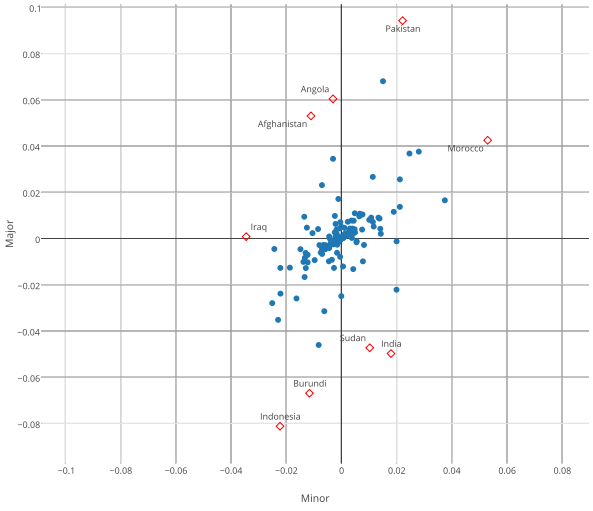
(c) nb_lgdpcap



(d) polity2

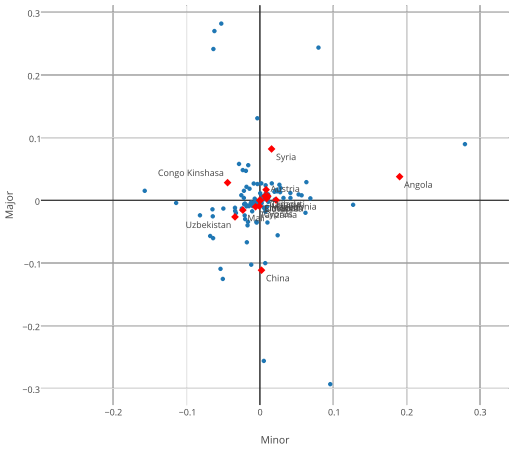


(e) polity22

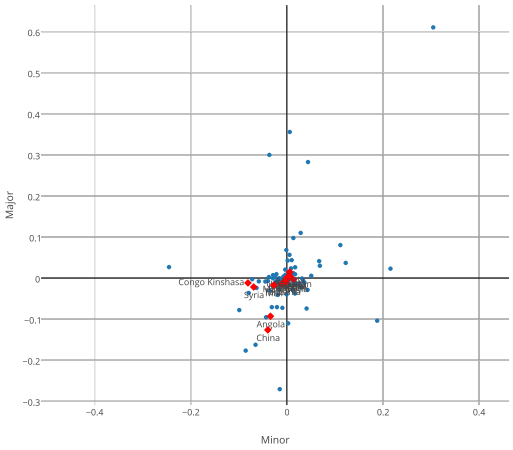


(f) nb_TSRC_5

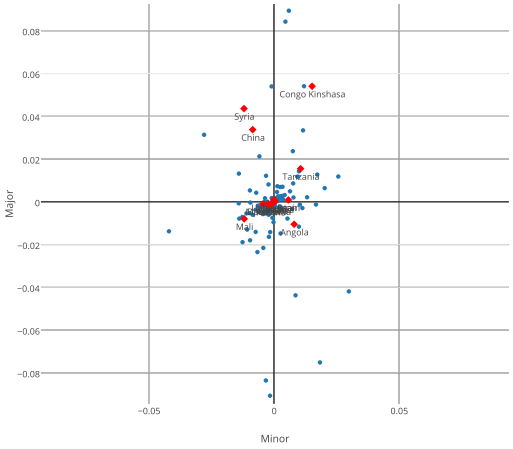
Figure B.2: Differences in coefficients resulting from country drops. X-axes are coefficients for minor outcomes and Y-axes are for major.



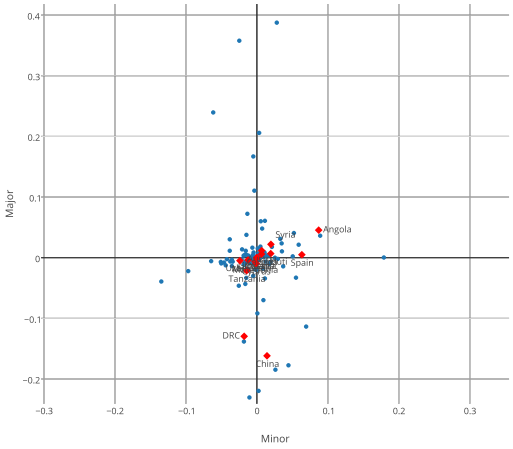
(a) c1



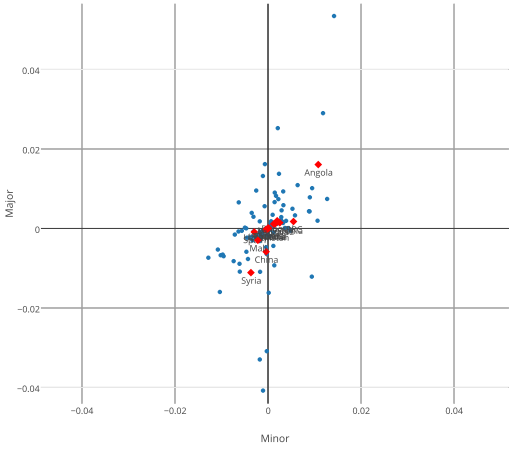
(b) c2



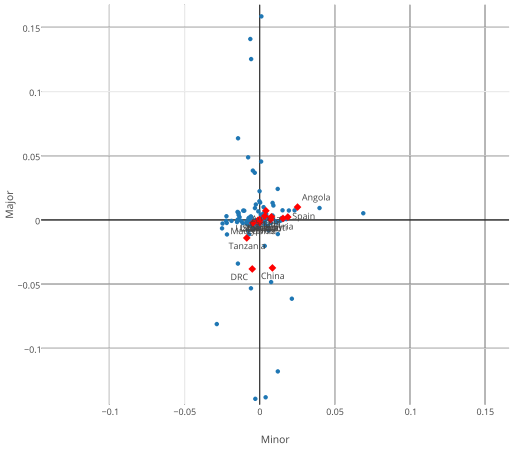
(c) ltsc0



(d) nc

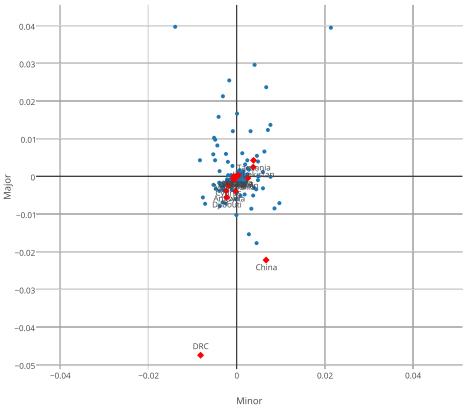


(e) ltsnc

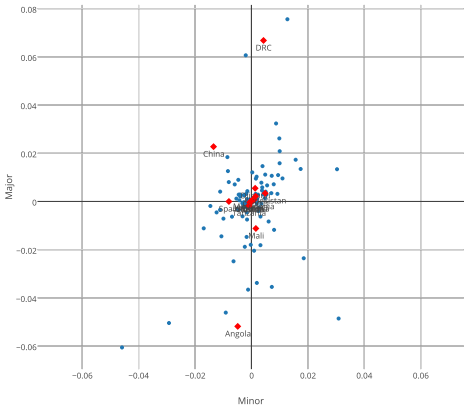


(f) ncts0

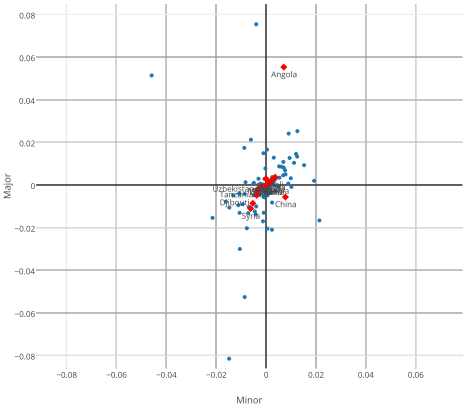
Figure B.3: Differences in coefficients resulting from country drops. X-axes are coefficients for minor outcomes and Y-axes are for major.



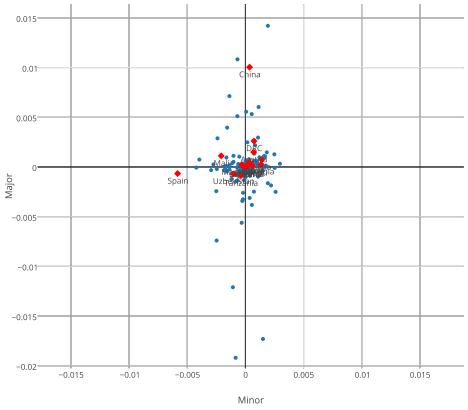
(a) lpop



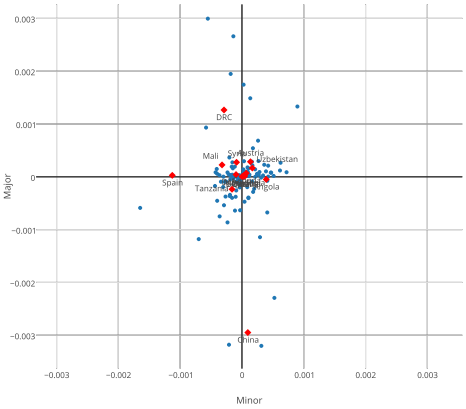
(b) lgdpcap



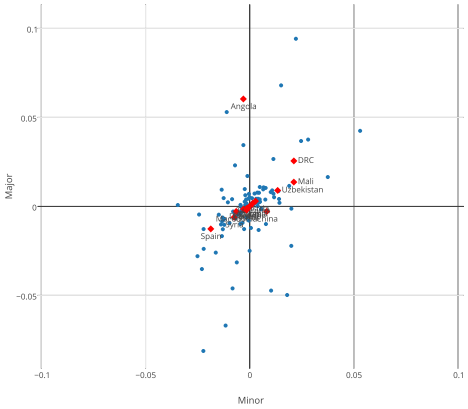
(c) nb_lgdpcap



(d) polity2

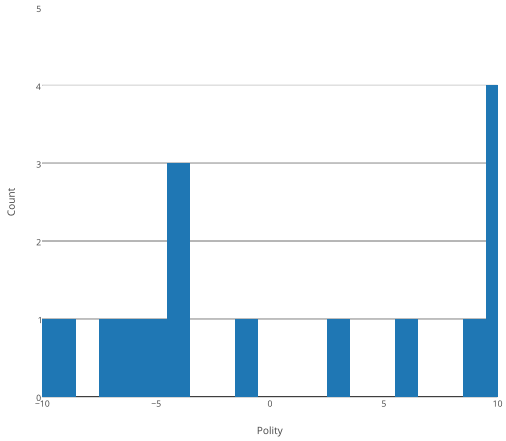


(e) polity22

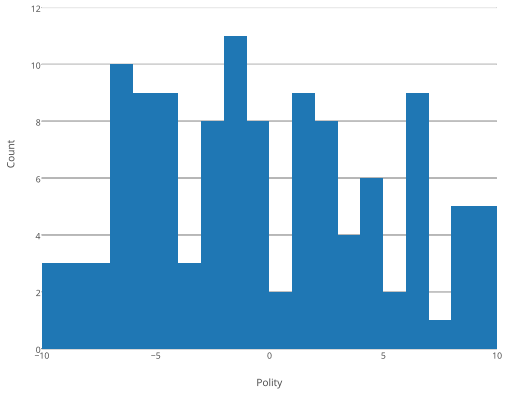


(f) nb_TSRC_5

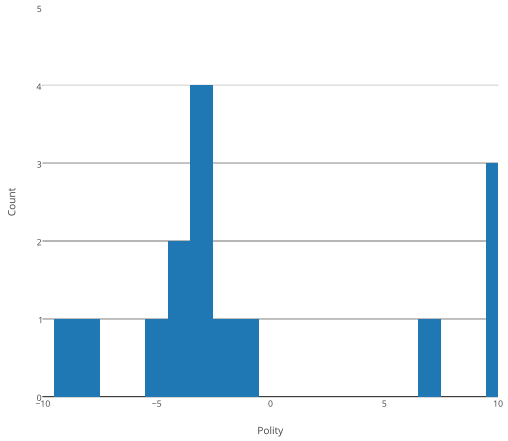
Figure B.4: Differences in coefficients resulting from country drops.
X-axes are coefficients for minor outcomes and Y-axes are for major.



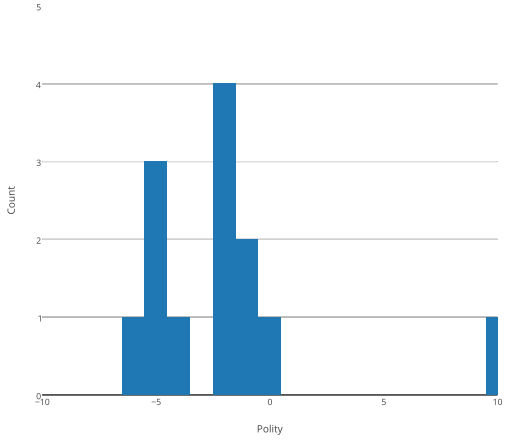
(a) PR outliers, destructive group



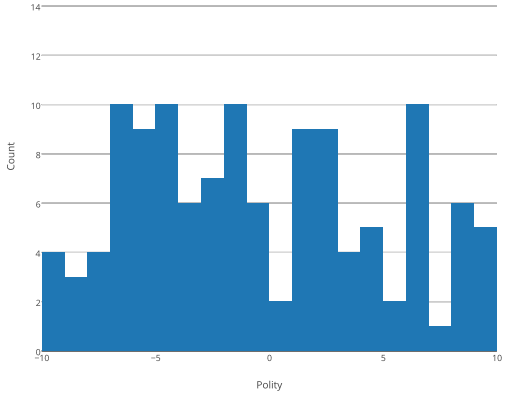
(b) PR outliers, main group



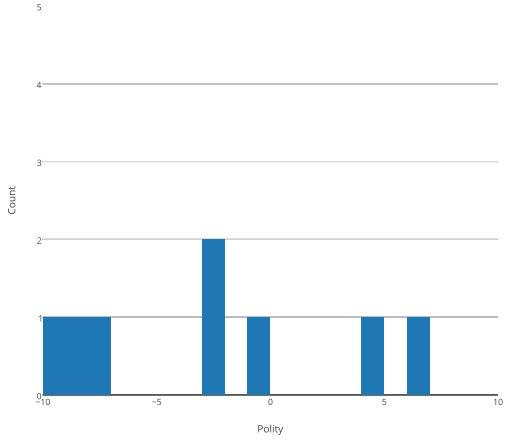
(c) PR outliers, reinforcing group



(d) Brier outliers, destructive group



(e) Brier outliers, main group



(f) Brier outliers, reinforcing group

Figure B.5: Histograms of the distribution of polity scores for the different groups of units.