Learning Conflict Duration: Insights from Predictive Modelling

Gokhan Ciflikli

A thesis submitted to the International Relations Department of the London School of Economics for the degree of Doctor of Philosophy, London, September 2018

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent.

I warrant that this authorization does not, to the best of my belief, infringe the rights of any party.

I declare that my thesis consists of 44,776 words.

Abstract

Why do some conflicts last longer than others? Previous work on conflict duration posits information asymmetries and credible commitment problems can cause protracted civil wars. The bifurcated nature of conflict studies, based on the notion that civil and interstate wars are qualitatively different, has so far prevented studies from including both types of conflict in the same dataset. Thus, empirical evidence is lacking as to whether the explanations apply to both types of conflict, or they are indeed separate phenomena.

This dissertation expands on the Cunningham and Lemke (2013) study on combining civil and interstate wars by including a large number of predictors taken from the rich civil war literature. The proposed framework unpacks the bargaining failure framework into three components governing power projection over distance, which I argue to be the main determinant of duration: material capability, politics, and geography. In doing so, I do not discriminate between the 'types' of war and provide a general theory of conflict duration.

I empirically test the general theory using a multi-method research design. First, I employ predictive modelling techniques such as machine learning, deep learning, and ensemble methods to demonstrate that the majority of predictive covariates of war duration are indeed common to both civil and interstate wars. Further, in most cases, the direction of the effect holds across types, suggesting that the underlying mechanism operates in a similar fashion. Second, I provide a shadow case study of the Sierra Leone Civil War to illustrate how capability shifts can occur on the ground that cannot be captured by observational data. Taken together, I contribute to the rationalist literature by providing a diverse set of empirical evidence showing that a unified model can explain the duration of both types of war.

Acknowledgements

My PhD experience in London had a rough start. I flew in the day before the course was scheduled to start, and promptly got my wallet stolen on my way to my new place. Luckily, there was someone waiting for me in the flat, so I could at least get in. Next day, without a working cell phone, I followed the path an offline Google Maps charted for me. Not being familiar with London, it took me about 1hr 45mins to walk from Peckham to Aldwych, where the LSE is located. It was a good ice breaker though, as my story quickly made the rounds in the department.

The rest of my PhD journey was a breeze in comparison.

During my first summer, I attended the ICPSR summer program at Ann Arbor, Michigan. Glad to have spent that time with my partners-in-crime: Calla Hummel, Chris Schwarz, and Andrew Brooks—you will all do well in your careers. A big shout-out to our awesome instructors: Daniel Martin Katz, Chris Zorn, and Jim Morrow.

The conference gang that I believe I was a late comer to: Luke Abbs, Corinne Bara, David Brenner, Kaisa Hinkkainen, and Rob Nagel. You made the acronym-soup conferences way more entertaining than they should be.

The Security, Conflict and Peace Studies Workshop at LSE, headed by David Rampton—talk about unorthodox teaching. It was a pleasure to be a part of the module while it lasted.

I got to spend a term at the Charles and Louise Travers Department of Political Science, UC Berkeley. Many thanks to Aila Matanock for agreeing to host me and her comments on my manuscript. Caroline Brandt and Andrew Reddie—you made the transition easier. It was great to run into an Uppsala colleague, Colin Walch, during our prospective exchanges.

I spent two months in Sierra Leone for fieldwork, and I probably have way too many people to thank for it. I can say I owe most of it to Kieran Mitton, who more or less set up all my initial contacts and prepped me for what's to come. Konstantin Born, it was invaluable having a reliable person on the ground. Finally, another Uppsala run-in—Sayra van den Berg. I think I like us more when we both reside in the same country, because both of us are terrible at keeping in touch.

We hosted the first Computational Social Science Hackathon at the LSE! The credit goes to the LSESU CSS Society board members: Kiran Phull, Pilar Elizalde, and Sebastian Mueller. I would like to thank LSE Annual Fund (AF) for their generous starter grant, along with LSE Social and Economic Data Science (SEDS) research unit (and its director Ken Benoit in particular) and GitHub for matching the AF grant to make it all possible.

Folks at the Alan Turing Institute: Kirstie Whitaker, Chanuki Seresinhe, and Merve Alanyali—thanks for hosting us. Also, my awesome Data Study Group (DSG) mates working on the U.K. Cabinet Office challenge—Jack Blundell, Haziq Jamil, and Kaisa—we did well!

For the more technical parts of this project, I am grateful to those who took their time to read and comment on the earlier versions of this manuscript: Karsten Donnay, David Muchlinski, and Julian Wucherpfennig. Konstantinos Travlos, my favourite Greek (high praise), thank you too for your sustained feedback over the years.

I wouldn't be writing on this topic if I hadn't studied Peace and Conflict Research at Uppsala University. The faculty would be too numerous to count, so I will not discriminate and thank all. During my two-year Master's studies and another two years of employment at the Uppsala Conflict Data Program (UCDP), I have made some close friends: Christopher Shay, Allard Duursma, David Larsson*, and Sayra. I have high hopes for you all. No, I will not explain what that asterisk means, David.

A later addition to that mix was Mihai Catalin Croicu, a welcome development in Turco-Romanian relations at large. Last but not least, my favourite person, Kristine Eck—tough times, tough people.

I have spent considerable time at UCL attending the Conflict Analysis Lab (COALA). The usual suspects during the time I was most frequently attending were Hannah Smidt, Sarah Leo, Altaf Ali and of course Nils Metternich, our convener. I have learned a lot during our discussions and they definitely made their mark in my work, including this one.

I have taught various courses at LSE. Departments of Government, Management, and Mathematics—thank you for giving me the opportunity to try out different subjects and teaching styles. Special thanks to Paul Mitchell, who trusted me to teach IR201 at the LSE Summer School for three years in a row. I owe the International Relations Department for all the financial support they have made available to me in the last four years—summer school, PhD exchange, fieldwork, and numerous conferences. I especially would like to acknowledge the Michael Leifer Estate, which provided my scholarship. Last but not least, I am grateful to Peter Trubowitz for agreeing to supervise me and for his continuous support and encouragement throughout my PhD.

Thanks to everyone who develop open-source applications—chiefly `R` and LaTeX—, make their code/data publicly available, and take their time to help others on Twitter, Stack Overflow, and GitHub. Special nod to Max Kuhn and Yihui Xie, as without `caret` and `bookdown`, this project would have taken so much more time.

Moving away from academia, I would like to thank Jimmy Slick's and Rafel Delalande, two true East End artists in their respective crafts. Peter Henderson, no, you are not an artist at all. However, I did not know where else to include you, so here it goes. I appreciate your London camaraderie, with special regards to consuming Chinese food and attending classical music concerts.

Pilar and Moritz—I would not dare demoting you to mere flatmates—it was so kind of you to share your home with me. Finally, I would be remiss not to mention Evelyn Pauls, whose companionship was dear to me throughout our PhD journeys.

Finally, I would like to acknowledge my mother. Raising a boy as a single mum in a traditionally patriarchal society while working two jobs, she was greatly concerned that I was lacking a healthy male role model in my life. I was given (many) books to read and sent to therapy for support. As I tried telling you this back then, your attempts were quite unnecessary—you are my role model. I hope I am strong enough to replicate your successes in life in my own way. Love you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The Puzzle

Why do conflict scholars study civil and interstate wars separately? In the literature at large, theories of conflict onset, duration, and termination differ by conflict type. For example, decades of formal work on bargaining[1] is built on state actor interactions and firmly situated in international relations. The divide in conflict scholarship is not confined to the theoretical realm; even when a theory crosses over from one realm to the other, as the work on civil war bargaining did in mid 2000s, the scope of empirical analyses continues to be limited to only one type of conflict. As a result, our understanding of conflict processes become conflict type-dependent: we do not entertain holistic applications to conflict research.

More specifically, even though there are empirical studies on both civil and interstate war duration, the great temporal variation found in war is exclusively studied in civil war settings. We ask why some wars last longer than others, but what we mean is why some *civil* wars last longer than others (Fearon, 2004). Simple descriptive statistics of war duration provide ample justification for this

[1] See Powell (2002) for an overview.

decision: on average, civil wars do tend to last longer than interstate wars.[2]

Yet, this scholarly divide prevents us from realising the true explanatory and predictive powers of our existing models. Why do civil wars last longer than interstate wars in general? Civil wars differ from their interstate counterparts in many aspects. First, there is an inherent asymmetry between the actors in civil wars in terms of international recognition and coercive power infrastructure (Driscoll, 2012; Clayton, 2013). Both factors are relatively more balanced in interstate war dyads. Belligerents in civil wars are geographically more constrained compared to state actors (Fearon and Laitin, 2003), whereas state actors can fall back to their sovereign territories. Further, rebel factions face more severe commitment issues as they are expected to fully demobilise during peace talks, while defeated state actors retain some levels of fighting capability (Walter, 1999).

However, they also share a multitude of commonalities. Both state and non-state actors behave strategically to achieve political goals (Atran et al., 2007). All actors require material capabilities (e.g. manpower, resources) to wage war (Bennett and Stam, 1996; Wood, 2010). Similar political and leadership issues play a crucial role in all wars (Cunningham, 2006; Weisiger, 2016); logistics matter (Kane, 2012), geography matters (Buhaug and Gates, 2002; Buhaug et al., 2009). Still, because we do not study them together, we cannot ascertain the level of empirical support for these commonalities between civil and interstate wars.

Further complicating matters, in some cases, the line between a state and a non-state actor can be blurred. One example is actor capabilities, which display a great amount of variation. Some rebel organisations are considerably more capable than the state actors they fight (for example the NFLP in Liberia), either

[2]The mean duration of civil war episodes since 1946 ($n = 304$) is 5.64 years, with 18% of all civil wars lasting longer than a decade ($n = 56$), 7% lasting longer then 20 years ($n = 21$), and slightly less than 2% going over 40 years ($n = 5$). Interstate wars, in contrast, last 1.95 years on average (with 10% ($n = 6$) going longer than 5 years; maximum length 11 years) as well as being far in-between in quantity ($n = 62$). All figures are based on the combined dataset compiled by Cunningham and Lemke (2013).

locally (PKK in South-eastern Turkey) or even nationally (Houthis in Yemen since 2016). The annual income of certain rebel organisations rival that of the states. The FARC was estimated to be worth $200m to $3.5 billion at its peak,[3] such that the high estimate would make it wealthier than 33 countries in terms of Gross Domestic Product.[4]

Some governments are fragile and fractured (Somalia) to a degree that most material capability advantages reserved for state actors hardly apply. In contrast, certain rebel organisations display strong central command and act as quasi-states (EFLP against Ethiopia prior to Eritrean independence). Rebels are sometimes pictured as being 'stuck' in their country in comparison to two sovereign state actors that can fall back to the protection of their own borders. However, some countries suffering from civil war are so vast in size (e.g. Mali, Sudan) that the distance between the capital (government power base) and the conflict zone (rebel power base) could be further apart than the average distance between two warring state actors.[5]

In cases like above, where does one draw the line? Do limited interstate wars behave more like civil wars than large-scale interstate wars? Conversely, do civil wars featuring militarily-strong governments and highly-capable rebels have more in common with interstate wars than small-scale civil wars? We do not have answers to such queries because conflict scholars specialising in either type of war do not talk to each other; and even if they try, they lack combined datasets to empirically test their claims.[6]

To test whether there are indeed empirical commonalities pertaining to war longevity, I posit a unitary framework for modelling conflict duration using a model of limitations on capability projection. Building on the rationalist concept

[3]The Economist (2014) The FARC's Finances: Unfunny money. [online] Available at: https://www.economist.com/the-americas/2016/04/14/unfunny-money [Accessed 7 Jul. 2018].

[4]World Bank (2018). World Bank National Accounts Data. [online] Available at: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD [Accessed 7 Jul. 2018].

[5]59% of all interstate wars are contiguous affairs (Slantchev, 2004).

[6]Cunningham and Lemke (2013) being a sole exception to the rule.

of bargaining, I propose a conflict-type agnostic model of limitations on power projection consisting of three components: baseline material capabilities in conjunction with physical and non-physical limitations acting upon them.

Baseline material capabilities refer to human, economic, and military capital of conflict actors—e.g. population, Gross Domestic Product, troop size. These represent the existing capabilities on an ideal level; in many conflicts actors do not mobilise completely (Wagner, 2000). The latter two components are modelled as constraints on the sustainability of applied force; i.e. negative force multipliers. For example, political factors such as veto players or a divided executive and geographical constraints such as vast distances and rough terrain can be thought of in this way: they can diminish the existing capabilities of an actor.

Finally, I expose the proposed unitary model to empirical testing by employing algorithmic predictive modelling to find out whether common patterns exist in each of the three components across both types of war. Instead of using a Null-Hypothesis Significance Testing (NHST) framework that looks backwards (favouring in-sample explanation), I employ a forward-looking forecasting approach that puts premium on out-of-sample predictive accuracy. The misuse and misinterpretation of NHST have been frequently criticised in political science (King, 1986; Gill, 1999; Gerber et al., 2008). Using predictive heuristics is one of the recommended solutions for addressing the over-deterministic nature of the traditional models that rely on statistical significance (Ward et al., 2010).

1.2 The Answer Writ Short

I provide strong empirical support for the proposed general framework that conceptualises duration as a function of limitations on material capability. Building on the rich empirical findings of the quantitative civil war literature, I demonstrate that the majority of predictive covariates that explain civil war

longevity also predict interstate war duration. Furthermore, excluding a handful of exceptions, the findings show that the direction of the predictors also hold across conflict types; if a covariate has a prolonging effect in civil wars, it also makes interstate wars longer.

More specifically, many operationalisations of material capability are ranked as top predictors of conflict duration. States with higher CINC[7] scores and Gross Domestic Product per capita (GDP p.c.) tend to fight longer than those with lower capabilities. Increasing total population and troop sizes, as well as increased military expenditures, are also consistent predictors of prolonged conflicts. Natural resources in certain forms—hydrocarbons (crude oil and natural gas) and gemstones—affect conflict duration: access to oil has a shortening effect while mining gems has a prolonging effect.[8]

Several political factors also play a crucial role as determinants of war duration. Politics in this context is conceptualised as non-physical limitations on the head executive regarding the continued application of military force. Both democracy as a regime type and the amount of political constraints on the head executive (Henisz, 2017) are the most influential covariates of this component that are associated with shorter wars. On the other hand, conflicts associated with coup d'etat have a slight prolonging effect on duration.[9]

Thirdly, time-dependency is also a robust predictive factor for both types of conflict. This is a topic of contestation in the literature—whether conflict is duration dependent or not—given the contradictory findings (Vuchinich and Teachman, 1993; Bennett and Stam, 1996). This project shows that

[7]Composite Index of National Capability (CINC) of a state is measured as an index using six indicators of material capacity: total population, urban population, military expenditure, military personnel, iron and steel production, and energy consumption (Singer, 1972).

[8]For a structured comparison of the effect of various natural resources on conflict intensity, see Lujala (2009).

[9]Although there are numerous outliers in civil war cases in which the direction of the effect is reversed (leading to shorter wars), a finding more in line with Fearon (2004). Further, data limitations must be carefully considered in the case of coups—there are significantly more cases of civil wars in the data.

time-dependency exists in war, further, it behaves similarly regardless of conflict type. Modelled as cubic splines, time t displays a slight shortening influence on both types of war. However, both the squared t^2 and cubic t^3 time are associated with longer conflicts. Taken together, this suggests the time-dependency of conflict has a specific functional shape that is consistent amongst both types: initially and up to a point, increasing duration is positively associated with conflict termination. However, as wars get more and more protracted, they become less likely to end. This can explain why we observe fewer protracted interstate wars in comparison to civil wars—actors with high baseline capabilities and projection capacity might cluster in time period t and rarely progress beyond (terminated early). Conversely, actors with low initial capabilities and projection rates are more likely to achieve their aims in time t and continue fighting.

1.3 Structure of the Dissertation

The rest of the dissertation is structured as follows. In Chapter Two, I lay out the theoretical framework guiding the project. I posit a general theory of conflict duration, primarily drawing on robust empirical findings borne out of the rich civil war literature. If the underlying data generating mechanism is indeed similar for both types of conflict, we should expect important predictors of civil war duration to perform well in interstate wars as well. Building on the extant literature, I describe a model of constraints revolving around the limitations and difficulties of sustained use of force. Using three main components—material capabilities, political and societal constraints, and geographical factors—I highlight the ways in which baseline fighting capacity can be hindered through projection.

Chapter Three is devoted to the research design. The first section covers the major methodological decisions influencing this mixed-methods project. First, I motivate why an integrative multi-method design is a more suitable approach than the more common empirical triangulation. Next, I justify why algorithmic predictive

modelling with an emphasis on out-of-sample prediction accuracy is likewise more apt than the Null-Hypothesis Significance Testing (NHST) approach. Third, I briefly cover the three most prominent approaches to conflict forecasting and situate the project in algorithmic modelling. Finally, I explore how case selection after quantitative analysis using extreme and deviant cases can lead to discoveries beyond what numbers can achieve. In the latter part of the chapter, I explain the data procedures undertaken for the large-n component of the project. Finally, I briefly summarise the semi-structured interview process and the respondent selection strategy.

Chapter Four is the first of the three inter-linked empirical chapters. It consists of a quantitative assessment of the conflict duration literature. First, I replicate 16 studies using Binary-Time-Series-Cross-Section (BTSCS) data on conflict duration that include time-varying covariates recorded at yearly intervals. Next, I run various feature selection algorithms to identify which covariates are highly predictive. Then, I run logistic regression, elastic net, and random forest models using out-of-sample cross-validation to see whether feature selection and predictive modelling findings overlap. Finally, I conclude with a list of top predictors of armed conflict duration.

Chapter Five offers the first machine learning application to conflict duration using data that includes both civil and interstate wars. I build on the Cunningham and Lemke (2013) study, to which I add the most predictive variables identified in Chapter Four. For this purpose, I select six machine learning algorithms using a distance-metric that maximises model diversity. Next, I fit linear and meta-model ensembles of the aforementioned models. Finally, using the state-of-the-art Keras infrastructure running a TensorFlow back-end, I employ deep learning to capture the complex non-linear interactions between the covariates.

Chapter Six is a stand-alone mixed-methods chapter that provides in-depth analysis of the predictive modelling findings alongside the shadow case study

of the Sierra Leone Civil War. The first section of the chapter focuses on the most robust findings of Chapter Five. Using the Local Interpretations of Model-agnostic Explanations (LIME) framework, I unpack the 'black box' findings of the complex machine learning algorithms. Doing so enables me to provide directionality of the results, making them more interpretable. In the second part of the chapter, I provide five pathways that can shift the material capabilities of an actor beyond the large-n findings.

Finally, Chapter Seven provides the conclusion of the project. First, I summarise the empirical findings of the predictive modelling enterprise. Next, I discuss the forecasting performance of the models. As the field moves towards a more predictive direction, it is important to provide performance metrics to which future work can be benchmarked against. I conclude by offering direction for future research and the need to collect more inclusive data to avoid empirical bottlenecks that can be caused by theory and tradition.

1.4 Contribution

Overall, this project makes three explicit contributions to the conflict literature. First, on a theoretical level, I provide a general theory of conflict duration. This conflict type-agnostic framework captures the commonalities of power projection that apply to both state and non-state actors by shifting the focus on actor capabilities. The proposed framework expands the mainstream bargaining approach to war by making it less prone to certain theoretical blind spots such as perpetual conflict and total war. A model built on material capabilities and constraints on power projection helps bound the theoretical implications of failed bargaining.

Second, empirically, the contribution is two-fold. One, I provide a sensitivity analysis of duration studies literature by replicating 16 BTSCS studies and

identifying the most consistently accurate predictors of conflict duration. I find that factors usually included in models as controls—democracy, GDP per capita, population, ethnic fractionalisation—have immense predictive power in both types of conflict. Two, I enrich the only existing combined dataset (featuring both types of war) by adding 20 more covariates to the original Cunningham and Lemke (2013) study. Doing so will allow other researchers to identify further variables that are robust to conflict type, as well as paving the way for replicating and expanding on the findings of this project.

Taken together, a general theory of conflict duration tested on a combined dataset answers the question of whether there are commonalities in conflict that can be captured without categorising wars. A unitary approach to modelling war duration reveals that instead of binning conflicts into binary types and studying them separately, one can treat actor capabilities on a continuous scale and investigate them in unison. The implication is that actors with similar material and projection capabilities—regardless of whether they are state or non-state—also do behave similarly. This opens up new avenues for the unitary forecasting of conflict onset, duration, and termination.

Third, on methodological grounds, I show the utility of employing machine learning and predictive modelling in conflict research. The random forest algorithm, which has been shown to adapt well to conflict studies (Muchlinski et al., 2016), greatly outperforms the literature-standard logistic regression in out-of-sample predictive accuracy. One commonly mentioned drawback relating to 'black-box' algorithms like the random forest is that even if their results are highly accurate, they are not interpretable. However, using various explainers designed to address this issue, I unpack the random forest predictions into interpretable chunks of information. This nullifies one of the main drawbacks of using similar algorithms, and it should encourage practitioners to consider adding such techniques to their methods toolbox.

Chapter 2

Theoretical Framework

2.1 Introduction

Why do some wars last longer than others? Further, to what extent do structural factors—absolute baseline material capabilities and their relative projections—govern armed conflict duration? In an attempt to answer these questions, I develop a general theory of conflict duration. I focus on three parameters of longevity: material capabilities in conjunction with non-physical (e.g. politics) and physical (i.e. geography) constraints that act upon them. I conceptualise material capabilities as a 'resource pool' (i.e. military capital) that can be spent on power projection. In other words, I model conflict duration as a function of the actors' sustainability of force projection.

Doing so allows me to put forward a general theory of duration that is applicable to and testable in both interstate and civil war settings. Most features differentiating civil wars from interstate conflicts can be attributed to differences in material capabilities.[1] Further, it extends the applicability of bargaining theory

[1]However, there are several notable differences that cannot be explained away by differences in material capability. For example, state actors are usually recognised in the international system whereas the rebels are not (Svensson, 2007). Further, unilateral disarmament of the rebel factions as a prerequisite for peace talks puts the rebels into a precarious scenario while

to duration studies by going beyond the established information asymmetry and credible commitment problem frameworks: a duration model built around projected material capabilities alleviates some of the existing theoretical blind spots of the bargaining approach (i.e. perpetual conflict, total war) to war.

The great temporal variation found in war has not escaped the scholars of conflict. From a bargaining perspective,[2] the rationalist arguments for why some wars last longer mirror their explanations of war occurrence: information failure, credible commitment problems, and issue indivisibility (Fearon, 1995). The informational approach posits that rational actors may go to war as long as they have incentives to misrepresent their own strength, or when the actual power distribution is not common knowledge. This type of unilateral miscalculation, or mutual optimism, is cited as the main reason for why parties go to war (Fey and Ramsay, 2007; Slantchev and Tarar, 2011).

Fighting loses its informational value once the belligerents agree on the actual distribution of power (Slantchev, 2003), which then leads to a negotiated outcome. Assuming wars erupt because of information issues, and recurrent fighting rectifies information asymmetries by offering non-manipulable[3] evidence from the battlefield, the rationalist theory postulates that actors are better off settling as soon as possible after the actual power distribution is known by all parties involved (Powell, 2004). In other words, they should not make the same mistake twice: at least one party has already miscalculated its chances and decided to wage a costly war instead of settling (Powell, 2006).

Consequently, one would expect an empirical examination of war duration to highlight this learning mechanism at work. However, when applied to protracted conflicts, the information failure explanation gives a skewed reading of history:

defeated state actors (post-WWII) keep their army intact (Walter, 1997).

[2]The conceptualisation of conflict as a bargaining problem can be traced back to Schelling (1960). See Powell (2002) for a survey of formal approaches to bargaining.

[3]Slantchev (2003) maintains that learning while fighting can occur in two ways: (1) Strategically manipulable negotiation behaviour, and (2) non-manipulable battlefield outcomes.

After several years of fighting, both sides have gathered enough information regarding their opponents' resolve and capabilities, but the fighting rarely ends afterwards (Powell, 2006). This observation has led to the view that civil wars are typically driven by problems of credible commitment (Fearon, 2004). Commitment issues arise when parties prefer a settlement that is beneficial to both, but they cannot credibly commit to uphold the agreement as the powerful side will have future incentives to renege on the terms of the deal once it has been signed.

The scholarly investigation of commitment problems in war has led to a multitude of explanations in the literature. Fearon (1995) provides three: pre-emptive war, preventive war, and conflict over issues that affect future bargaining power.[4] According to Fearon, either the winning or the losing side can be the instigator of the commitment problem.[5] Similarly, Walter (2002) claims that commitment problems arise because of the treacherous demobilization process that follows negotiated settlements, and agrees with Fearon that both sides can initiate the commitment problem.

Later studies further examine commitment issues in an attempt to pinpoint which faction is more likely to create the commitment problem. Svensson (2007) argues that there is a rebel-sided commitment problem: As the government[6] is a recognized international actor, formal talks with the previously unrecognised rebel faction will prompt the rebels to renege on the deal once their standing in the international arena is improved. In contrast, Powell (2012) asserts that state consolidation is the most likely exogenous shock to power distribution in civil wars, which suggests that the faction who controls the government creates the commitment problem.

[4] As Fearon himself readily admits, he does not claim to be first to draw attention to such mechanisms. Indeed, the foundations of both arguments can be traced back to the classical works of Blainey (1973) and Waltz (1979).

[5] See Gartzke (1999) for a more elaborated review of Fearon's proposed arguments.

[6] At least, the majority of the time vis-a-vis a non-state actor.

The arguments put forth by Fearon and other leading rationalists have contributed greatly to our understanding of underlying conflict mechanisms. However, the rationalist literature on war still suffers from several critical shortcomings. First, on a theoretical level, although commitment issues explain how parties fail to locate a mutually-beneficial agreement ex-ante or during hostilities,[7] they do not tell us much about how conflicts terminate endogenously. With the exception of attrition as a military strategy (Bennett and Stam, 2009; Langlois and Langlois, 2009), solutions to commitment issues usually involve the introduction of some exogenous factor into the equation. For example, Walter (1997) posits third-party guarantees can alleviate such problems by acting as a commitment device. While she offers empirical support in favour of her theory, it does not explain how conflict parties can overcome commitment problems by themselves.

More problematic is the hidden implication of perpetual conflict. If commitment issues are indeed so salient, actors only have two options once a conflict is under way: termination via complete annihilation of the opponent or suffer a never-ending war.[8] This deduction is again at odds with the empirical track record: 55% of all interstate wars (Walter, 1997) and 40% of all civil wars (Hartzell and Hoddie, 2007) end in negotiated settlements. Theories based on commitment issues thus fail to inform us about how nearly half of all conflicting parties eventually find a way to credibly commit to peace, nor do they shed light on how conflict duration is affected by this process.

Moving from theory to empirics, even when an effort is made to integrate conflict termination in the theoretical framework, it has only been tested in civil war settings. More specifically, the scholarly literature focuses on why some civil wars last longer than others.[9] The question of why civil wars as a whole last longer

[7] For a study on whether war is still inefficient ex post, see Chiozza and Goemans (2004).

[8] Wagner (2000) addresses this theoretical drawback of relying on commitment problems by differentiating between 'absolute war' (war-in-theory) and 'real war' (limited wars that we usually observe in reality). With that said, the formation of the conceptual divide stretches back to Clausewitz's seminal work *On War* (Clausewitz, 1832).

[9] For a review of the quantitative literature on civil wars, see Sambanis (2002).

than interstate wars, however, appears to be understudied. Only most recently, such an attempt has been made by Cunningham and Lemke (2013).

A theory build around baseline material capabilities and their limited projections, by design, less prone to extreme logical conclusions such as perpetual conflict or total war. Power projection in this context can be conceptualised as *usage rate*, with the baseline material capabilities acting as the main *resource pool*. War-as-attrition approaches in many scientific fields mimic this logic: animal contestation in biology (Bishop and Cannings, 1978), firm competition in economics (Bulow and Klemperer, 1999), and World War II tank warfare in operations research (Peterson, 1967). The main commonality across all these studies is the winning strategy—attrition warfare emphasises the gradual wearing down of the opposition via sustained casualties.

Higher usage rates, unless coupled with high baseline material capabilities, indicate shorter conflicts on average. In contrast, low usage rates even with moderate material capabilities can be sustained for longer periods of time. Certain edge cases—e.g. the current (legal) status of war between North Korea and South Korea since the 1953 armistice—will still be predicted as protracted conflict due to the infinitesimal usage rates and massive capabilities on both sides.[10] However, a capability-spending model makes bounded predictions for all cases of armed conflict (both interstate and civil) without relying on exogenous factors. In the next three sections, I make a case for why we should tackle conflict from an unitary perspective; posit a general duration model of limitations by extending the bargaining approach; and outline empirical expectations of such a model.

[10] I would argue that the case of Koreas is a significant outlier; and further, it can be excluded from analysis by employing defensible scope conditions either in theory or application (empirics) without loss of generality.

2.2 Bifurcated Study of War

Existing models of conflict are bifurcated on the basis of theory-driven war 'types' (Cunningham and Lemke, 2013). Conflict scholars formulate exclusive theories and perform separate empirical tests depending on whether we study interstate or civil wars. Case in point, a common way of motivating a duration study is to provide descriptive statistics on conflict duration stratified by type (see Fearon, 2004). However, once it is established that civil wars (either on average or in the extreme) last significantly longer than interstate wars—a well-established empirical fact—the authors then proceed to limit their analyses to civil wars only.

By limiting our analyses to certain subset of wars, however, we lose the opportunity to develop general theories of conflict. This has implications in both domains—theoretical and empirical. On theoretical grounds, we develop frameworks aiming to explain the temporal variation found in civil wars. Empirically, we only test our theories within the type of war we study. In other words, we have no way of knowing whether they will hold across war types. Indeed, if there are vast differences between interstate and civil wars, we should not necessarily expect that the findings will hold. On the other hand, if we had empirical evidence showing determinants of long civil wars do not overlap with that of interstate wars, this would make a strong case for the justification of the separate study of wars.

Alas, we have yet to see such a non-finding. We have come a long way in terms of providing explanations—especially the rationalist strain of conflict scholars—to why some civil wars last longer than others. But we do not speculate much on what explains long interstate wars; further, whether the underlying mechanism is the same or some interstate wars last longer than others for reasons separate from their civil war counterparts. Can some civil wars be conceptualised as localised, small-scale interstate wars? Or, do civil wars taking place between two highly-capable (i.e. in terms of manpower, resources) parties behave similarly

to interstate wars rather than small-scale civil wars? It is difficult to answer such questions because we consider them categorically different phenomena.

Instead, civil wars became the dominant focus as they possess a higher risk of running into commitment problems for a multitude of reasons in the literature. For instance, Walter (2002) suggests that the duration of civil wars can be drastically shortened if credible and potent third-party intervention is guaranteed. Svensson (2007) provides empirical evidence that the commitment problem caused by the rebel groups can be alleviated when mediators are biased in favour of the government. Yet, as stated above, both studies concern themselves only with a subset of all wars, rendering their findings incommensurable to the literature at large.

In sum, in the past two decades, the rationalist literature on war longevity is built upon civil war, both theoretically and empirically.[11] Commitment problems have come to be associated with civil wars to a degree that they are seldom applied to interstate wars, which is problematic. We rarely discuss deploying peacekeepers or sending mediators to alleviate commitment problems in interstate wars as we would in similar civil war settings.

To give an example, the US War on Terror has exceeded the median duration of interstate wars by tenfold, and the power distribution between the US and the Afghan Taliban has been common knowledge to both parties for many years. They would be better off if they located a mutually beneficial agreement, which always exists given the costs of fighting. The obstacle then, one might argue, is the inability to credibly commit to upholding the terms of such a settlement.[12] Yet, the conflict between the US and Taliban has not been labelled as a commitment failure in the literature, even after when Taliban opened a short-lived 'diplomatic'

[11] A curious development, given that the initial theorisation that led to the formation of concepts such as information failure and commitment problem focused solely on state actors and interstate wars. For example, Fearon (1995) does not mention civil wars or non-state actors once in his seminal article.

[12] Also see Lake (2002); Lake (2003).

office in Qatar.[13]

Altogether, the absence of empirical validation[14] relating to war duration begs several critical questions: What is the true nature of the relationship between the longer duration of civil wars and commitment issues? How do parties eventually overcome commitment problems and terminate hostilities? Is it even feasible for interstate actors, such as the most powerful ones like the US, to commit to war for decades? Do civil wars last longer than interstate wars because domestic opponents are somewhat less credible than their international counterparts? Or is it that civil wars are characterized by commitment problems because structural factors constraining the longevity of interstate wars do not apply fully to civil conflict?

2.3 A Unitary Framework

To this end, I provide a unitary model of conflict duration. The main motivation behind this undertaking is to unpack—both theoretically and empirically—what constitutes the variation in civil war duration and make those factors the main parameters of the model. Put simply, I aim to identify the structural determinants of conflict duration. Doing so gets rid of the notion that wars have unique characteristics depending on whether there is zero, one, or two state actors involved in it (Cunningham and Lemke, 2013). Instead of categorising conflicts into two based on whether 'Side B' is a government actor or not, I take conflict as it is and let the actor parameters (capabilities) dictate the outcome (duration).

The rest of the chapter is structured as follows. First, I offer a model of

[13]Taliban and Afghan officials hold 'reconciliation' talks in Qatar," The Guardian, May 2, 2015. Accessed May 5, 2015, http://www.theguardian.com/world/2015/may/02/taliban-and-afghan-officials-hold-reconciliation-talksin-qatar

[14]For an experimental testing of the rationalist explanations for war in a laboratory setting, see Quek (2017).

constraints that acts as a unitary model of armed conflict duration inspired by the standard rationalist model of conflict bargaining. The unitary nature of the proposed framework is based on the idea that one should be able to draw on the expanded civil war literature to identify the most important variables of both types of conflict. The implication is that if the process is really unitary—the underlying mechanism is structurally similar for both interstate and civil wars—the established findings from one domain should transfer to the other. The general model is characterised by three main components: limitations on physical ability, commonly thought as material capabilities of an actor; limitations on the use of such power, as political constraints on the head executive; and the loss of strength gradient, the waxing and waning of military force as it is projected over distance. Finally, I conclude by making explicit the empirical expectations borne out of the proposed theoretical framework.

2.3.1 A Model of Limitations

In this section, I parameterise a general theory of conflict duration. As the idea is to build a framework that does not rely on categorical labels such as interstate or civil war to capture the temporal variation, the theoretical parameters are designed to proxy the underlying commonalities between the two types of war.

General theories are important in empirical domains that rely on cumulative progression. International relations as a discipline also heralded the imminent unification. David Lake provided the following conjecture on the interstice of international relations and internal conflict as early as 2003 (Lake, 2003, pp. 81):

> "We are approaching a single, unified theory of political violence of which interstate and intrastate war may be particular forms. I emphasize approaching because this general theory has not yet been fully worked out and may because the particular forms of violence and the relationships between them have not yet been defined.

Nonetheless, considerable progress has been made."

Fifteen years later, we are not any closer to a general theory of conflict. Even though considerable progress has been made since 2003—both in our understanding of conflict dynamics and conflict data collection efforts—these innovations have remained exclusive to the type of conflict under scrutiny. To this end, I utilise the classic rationalist framework of bargaining space (Fearon, 1995) and expand on it to provide a general model of conflict duration. Figure 2.1 illustrates the conventional bargaining space approach demonstrated in a conflict dyad.

For simplicity, assume a one-shot game in which two conflict parties, denoted here as A and B, have well-defined preferences over the division of an issue (for example, a disputed territory). Both actors prefer to control all the territory, as this would maximise their gain. Without loss of generality, projected on a single dimension and bounded in respect to $[0, 1]$, the ideal point for A is all the way towards 1; conversely, B's ideal point is located at the very far left at 0. The division of the issue is determined according to the outcome of the contest q—representing war—which could be actual or expected. If the actors choose to fight over in order to alter this division, they incur costs a and b, respectively. As such, their net gain (as opposed to settling) obtained by fighting becomes $q - a$ for A and $q + b$ for B.

Since fighting introduces additional costs that are otherwise not applicable if the sides could agree on a settlement, this opens up what is deemed a *bargaining space*. This is the theoretical space—stretching between $q - a$ and $q + b$ in Figure 2.1—where any division of the issue located within is preferred to actual fighting, given the costs. Note that this formulation is not susceptible to future capability shifts. Assume p represents the expected outcome of a war under a new distribution of capabilities. Even if one side becomes more powerful and could shift the division to p, the bargaining space would simply shift to $p - a$ and $p + b$.

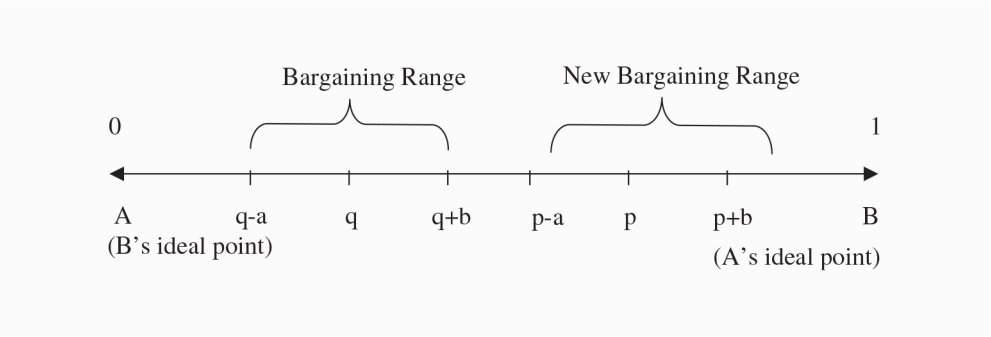

Figure 2.1: Bargaining space, as illustrated in Lake 2003. Settling is always preferable to fighting given the costs associated with war.

Thus, even though one side becomes more powerful and the old status quo (q) is no longer satisfactory, both parties still have an incentive to settle rather than wage war.

The rationalist bargaining space approach can be transformed into a duration framework by disaggregating the distribution of capabilities p into separate parameters. Indeed, many earlier interstate duration models have implicitly utilised this approach (Wittman, 1979; Morrow, 1985; Vuchinich and Teachman, 1993; Stam, 1996; Bennett and Stam, 1996).

The main assumptions of such frameworks are as follows. Once a conflict is under-way, rational utility-maximising leaders periodically make a decision to whether continue fighting or to settle. The conflict ends when no actor chooses to fight in a given period. The decision to terminate fighting is conceptualised as a function of expected benefits and costs. Different types of actors have different material—and even political—capacities that might affect their cost-benefit calculus. It follows that the duration function can be modelled using parameters that capture the actors' abilities to obtain war benefits and absorb accumulating costs of conflict.

Figure 2.2 offers a simple demonstration of this concept. Recall that in Figure 2.1 demonstrating single-shot bargaining, the status quo is denoted as q, and the new distribution of capabilities is p. Both points are displayed simultaneously to

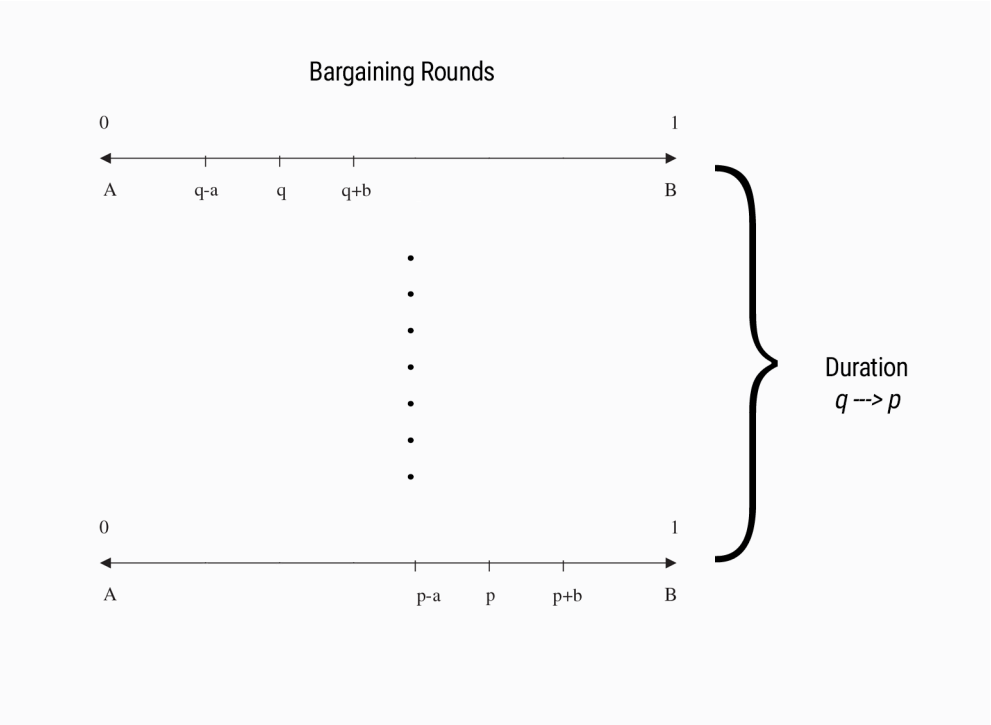

Figure 2.2: Conflict duration as a function of multiple-round bargaining.

drive home the implication that given the costs of fighting, there always exists a mutually-beneficial agreement—even in the case of power shifts such as $q \rightarrow p$. However, these shifts need not happen instantaneously. Instead, significant changes affecting the power distribution in a dyad take place over time.

Assume q represents the distribution of power at onset and p denotes the configuration at the time of conflict termination. In other words, belligerents start fighting based on the information (i.e. power parity) revealed by q, and cease fighting when the information is updated to p. Thought this way, the duration of a dyadic conflict becomes a function of the length of the iterated bargaining game between two players. For a conflict dyad i, such phenomena can be modelled as

$$duration = \Omega_i : f(q \rightarrow p).$$

Based on this formulation, the function Ω_i maps the two crucial states of war—onset and termination. Duration, then, becomes the length of time it takes for this state transformation. As the success of bargaining depends on credible power projection, the quantity (and the perturbations in such capacity) of force used—i.e. *applied power*—should act as a proxy for the underlying process.

In the next three subsections, I provide the parameters of this general duration model function Ω. The scope of the theoretical components is not constrained to those borne out of the bargaining literature. Instead, I cast a wide net to identify empirical regularities in the conflict literature at large. If these determinants of conflict duration are truly transcendent, one should observe their manifestations in both civil and interstate wars.

I identify three main components proxying the cost-benefit calculus of rational decision-makers: baseline material capabilities (population, troop size), non-physical (e.g. leader characteristics, issue salience) and physical (i.e. logistics, geography) limitations on power projection. Each component acts as an umbrella category that brings together a multitude of empirical findings from the civil war literature. Next, I provide a real life example of the War of the Triple Alliance as a stylised illustration of the proposed framework. Finally, I highlight the expected directionality of the empirical findings of such a model and how it can be tested using a predictive modelling framework.

Figure 2.3 demonstrates the generation of the force use as a function on material capability subject to limitations. The baseline material capabilities of an actor is denoted by γ. There are two possible constraints on this baseline; politics (non-physical) α and those relating to the nature of power projection (physical) β. Latter components can be thought of as negative force multipliers on the use of force. In the end, whatever force ends up being utilised to fuel the conflict takes the form $\gamma \cdot \alpha \cdot \beta$. I call this final product of force *applied power.* Doing so links the proposed general duration framework to the mainstream bargaining approach

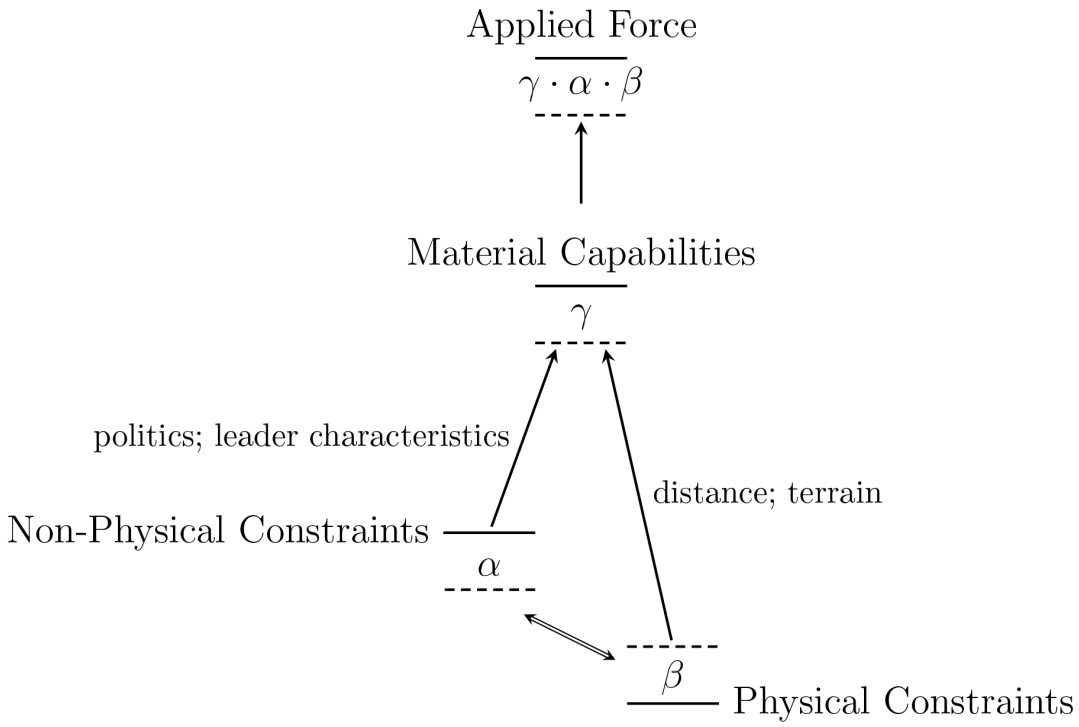

Figure 2.3: Illustration of the limitations on the amount of force application. Baseline material capabilities of an actor are subject to physical and non-physical constraints when projected away from the power base. All three components are dynamic and their values can change drastically over the course of the conflict.

adopted in the rationalist literature. Below, I unpack the main determinants of power projection as they relate to conflict duration Ω.

2.3.2 Material Capabilities

The first component of the general model encapsulates physical capacity to use force. This is the broadest category of the model, stretching from absolute and relative capabilities on one side of the spectrum to natural resources and other exploitables on the other.[15] This is not surprising; the study of power is often referred as the crux of international relations (Kennedy, 1987).

The fighting capabilities of the rebel organisations are found to affect civil war longevity. States are often unable to achieve a decisive victory against weaker rebel groups, as they often choose to engage in irregular warfare. In contrast, stronger groups are found to be more likely to receive concessions from the government (Cunningham et al., 2009; Lujala, 2009; Thomas, 2014), which results in relatively shorter conflicts.

Access to lootable natural resources (Stedman, 2001) as a means of increasing fighting capacity (Ross, 2004)—as well as providing a different set of incentives for the belligerents other than achieving military victory (Addison et al., 2002)—has been heavily studied since the earlier debates on greed vs. grievance (Collier and Hoeffler, 2002, 2004). Hydrocarbons[16] (Fearon, 2004), gemstones (Ross, 2004; Gilmore et al., 2005; Lujala, 2009), drug cultivation (Lujala, 2009), contraband (Fearon, 2004), primary commodity exports (Doyle and Sambanis, 2000), and smuggling (Conrad et al., 2018) are common independent and sometimes control variables included in civil war duration studies. Finally, natural resources may also aggravate existing commitment problems between the government and the rebel forces (Walter, 1997; Wagner, 2000).

[15] See Hendrix (2010) for a sensitivity analysis of various definitions and operationalisations of state capacity.

[16] That is, crude oil and natural gas.

Material capabilities can be enhanced through external interventions (Elbadawi and Nicholas, 2000; Balch-Lindsay and Enterline, 2000; Regan, 2002; Cunningham, 2010; Escribà-Folch, 2010). However, it is argued that its not the sanctions themselves, but the military force that usually accompanies them that creates the desired effect (Pape, 1997, 1998).

Moving the focus from non-state actor capabilities to that of state-actors, one of the most well-known power proxies in the literature is the Composite Index of National Capability (CINC) score (Singer, 1972). The CINC score of a state consists of six indicators of material capability: total population, urban population, military expenditure, military personnel, iron and steel production, and energy consumption. The six components are measured yearly in units relative to the system total, while the composite index score itself is the average of six components. In other words, they are indicators of relative material capabilities. However, the CINC score by itself or as a ratio has yet to show much significance as a reliable predictor in the literature (Maoz, 1983; Carroll and Kenkel, 2016).

In contrast, certain relative material capability indicators regularly do turn out to be statistically significant predictors of war duration. Population ratio (Vuchinich and Teachman, 1993; Hegre and Sambanis, 2006) and the balance of forces[17] (Bennett and Stam, 1996, 2009; Nilsson, 2012) are commonly cited as important covariates. Furthermore, in some cases absolute versions of the aforementioned variables—total population (Cunningham et al., 2009), geographic size (Buhaug et al., 2009), and total troop size (Bennett and Stam, 1996)—are found to be influential factors pertaining to conflict duration. Finally, military technology—such as the questionable effectiveness of conventional armies against irregular warfare (Lyall, 2009, 2010), the nullification of air-superiority in certain conflict settings (Kocher et al., 2011; Allen and Martinez Machain, 2017), and the efficacy of combined warfare (Caverley and Sechser, 2017)—also affect

[17] i.e. the ratio of the higher CINC score to the total CINC value of that dyad.

the longevity of violent conflict.

Military strategy and 'technologies of war' (Balcells and Kalyvas, 2014)—conventional, irregular, and symmetric non-conventional—also hold explanatory power in the literature. Bennett and Stam (1996) find that the interaction of strategy (maneouver, attrition, and punishment) and doctrine (offensive or defensive) is a strong predictor of interstate war duration, a finding that is also replicated by Nilsson (2012). Balcells and Kalyvas (2014) show that civil wars characterised by irregular (i.e. guerilla) fighting last longer than conventionally-fought civil wars.

However, I do not explicitly include strategy in the theoretical model as a main component. A general model is foremost focused on the core causes, not by-products. One of the commonly cited differences between the two types of war is that some civil wars are more likely to be fought in an irregular fashion,[18] even though there are cases of irregular interstate fighting as well; e.g. the Vietnam 'quagmire' (Krepinevich, 1986). Such, I expect the aforementioned material capability variables and their interactions to capture the empirical exposition otherwise explained away by technologies of war.

2.3.3 Non-Physical Constraints

The second component of the general model is the effect of non-physical constraints on the use and application of force. If material capability is conceptualised as force, non-physical constraints can be thought as a moderator. In other words, the latter can dampen or enhance the former. There is a wide range of factors that can be consolidated under this heading; the next three sub-sections briefly summarise some of the most commonly studied variables.

Politics

Regime type is one of the most-studied variables of conflict (Maoz and Abdolali,

[18] See Balcells and Kalyvas (2014) for a more through review of the subject.

1989; Maoz and Russett, 1993; Russett, 1994; Filson and Werner, 2004). Specifically, democracies are shown to be more pacific (Benoit, 1996; Weeks, 2008) than their autocratic counterparts when studied dyadically;[19] however the methodological validity of the so-called democratic peace findings is now challenged (Dafoe, 2011; Dafoe et al., 2013) or some of its explanatory power further unpacked into pre-existing socio-economic factors (Hegre, 2014). Recently, more nuanced parameters than categorical regime type are generated for the study of conflict. Political constraints is a CINC-like composite index that aggregates a multitude of political pressures on the head executive (Henisz, 2017). As the use of military force in a conflict is a top-down decision, we should expect regime type and political constraints to be important predictors of duration.

In bargaining, the number of veto players (Tsebelis and Yataganas, 2002) is shown to prolong civil wars by acting as a barrier to peaceful settlement (Cunningham, 2006). In the same vein, the number of actors in a conflict is also widely included in duration models (Cunningham et al., 2009). Internal cohesion of rebel groups (Elbadawi and Nicholas, 2000; Collier, 2000b; Bakke et al., 2012) as well as their fragmentation (Driscoll, 2012; Pearlman and Cunningham, 2012; Akcinaroglu, 2012; Fjelde and Nilsson, 2012; Brenner, 2015) can alter the number of conflict parties drastically. It must be noted that even though conflicts with more actors might run into coordination problems and thus influence war duration through executive decision-making, they can also work by affecting the material capability equation of the conflict—it is included here for theoretical coherence.

Societal Factors

The role of ethnicity is another important factor in conflict studies (Horowitz, 1985; Licklider, 1995; Kaufmann, 1996, 1998; Rose, 2000; Van Evera, 2001; de Rouen Jr and Sobek, 2004; Fearon, 2004; Kaufman, 2006; Cunningham et al., 2012). With

[19] Raknerud and Hegre (1997) show that, using non-dyadic modelling approaches, the tendency of democratic actors to join each other in wars is much pronounced than their avoidance of mutual fighting. Meaning, democracies are not necessarily less war-prone than autocracies.

that said, ethnic diversity or ethnic fractionalisation are not found to be significant predictors in various studies at the rebel-organisation level of analysis (Collier, 2000a; Fearon, 2004; Collier et al., 2004b; Cunningham, 2006; Brandt et al., 2008; Cunningham et al., 2009; Cunningham, 2010).[20] However, Wucherpfennig et al. (2012, p.111) empirically show that "ascriptive ethnicity and state-enacted exclusion along such categorical lines" indeed do lead to longer conflicts.

Leader Characteristics

Finally, leader characteristics can also greatly influence conflict dynamics (McGillivray and Smith, 2000, 2004; Chiozza and Goemans, 2004; Wolford, 2007; Gibler, 2008). Studies on leader tenure (Thyne, 2012), replacement (Tiernay, 2015; Weisiger, 2016), culpability (Prorok, 2018), and previous combat experience (Fuhrmann and Horowitz, 2014) show that leader characteristics and their priors (i.e. information) can have an effect on termination and duration dynamics.

2.3.4 Physical Constraints

The final component of the general model pertains to power projection, distance, and geography. If the non-physical component can be thought as a possible set of constraints on the use of force via 'soft' means (e.g. decision-making), this heading covers factors capturing the 'hard' constraints on existing material capabilities caused by its projection over distance.

On the linkage between proximity and power, the seminal work of Boulding (1962) is widely cited as the foremost of its kind. Figure 2.4 demonstrates the concept of Loss of Strength Gradient (LSG) graphically (Sakaguchi, 2011). At its core, it highlights the nature of the interaction between power and proximity: as nations project power further from their base, the projected power diminishes as a function of the distance. All states suffer from this loss-of-strength gradient, however

[20]Consult Saideman (2017) for a criticism of ethnic fractionalisation indices in quantitative research.

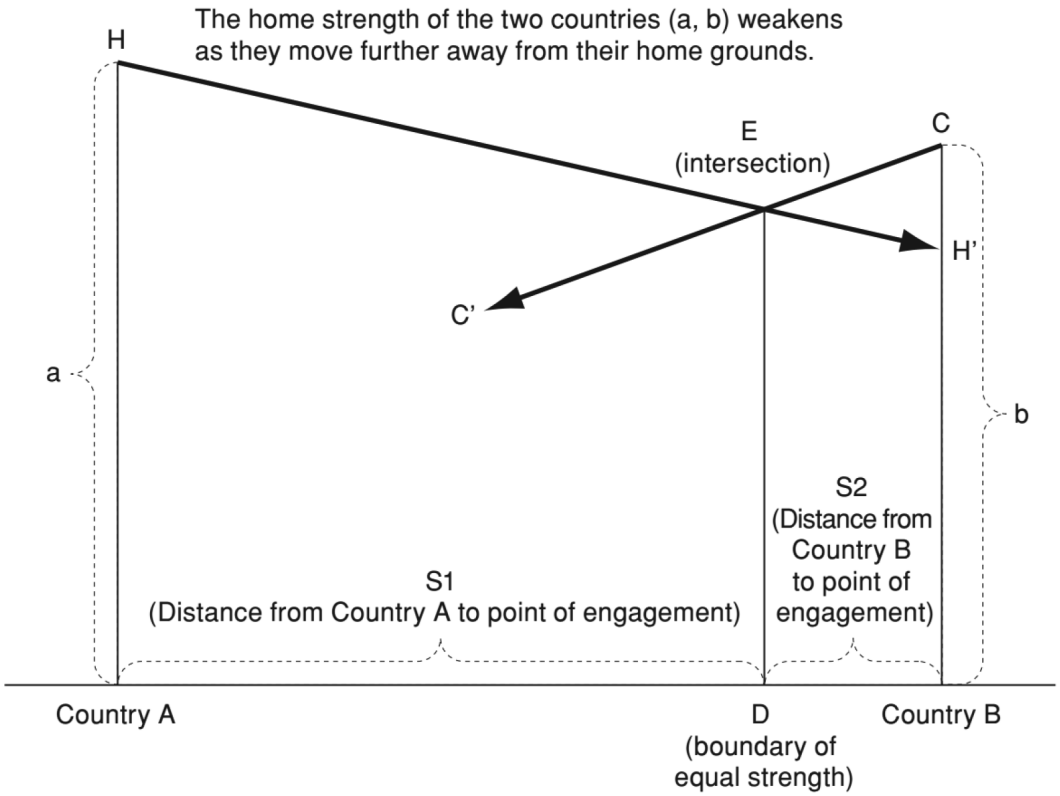

Figure 2.4: Boulding's loss of strength gradient concept, taken from Sakaguchi 2011

more capable actors can project further due to their higher baseline capabilities. Distance is cited as the most punishing penalty on power projection (Russett and Oneal, 2001). Further, government investment in power projection capabilities is found to help explain the historical polarity in international relations (Markowitz and Fariss, 2018).

Even though it was developed with state actors in mind, the LSG framework has been applied to civil wars as well. Buhaug (2010) finds that when the government possess high material capabilities, conflicts take place far away from the capital. Ruggeri et al. (2016) study where UN peacekeepers are deployed within the countries they have been sent. Finally, Tollefsen and Buhaug (2015) explain how various dimensions of inaccessibility influence the risk of localised conflict.

Furthermore, distance can be thought as a medium that can create or break power parity between actors. For instance, Gartzke and Braithwaite (2011) show that violent conflicts are more likely to occur at proximities where both states' capabilities are roughly equal to one another after applying a penalty for distance. Coupled with the above, we should expect distance and proximity indicators to hold predictive power on conflict duration as a modifier of material capability.

Various other impacts of geography on conflict are also well-studied. Terrain characteristics—such as dense forests and mountains—can act as another force multiplier for material capabilities (Fearon and Laitin, 2003; Buhaug et al., 2009). Research shows that conflicts last significantly longer when the rebel group operates in close proximity to remote international border areas, which may allow them to regroup outside the grasp of the government forces (Buhaug et al., 2009; Mukherjee, 2014). Finally, conflicts are known to cluster in space (Buhaug and Gleditsch, 2008).

2.3.4.1 The Paraguayan War 1864-1870: An Example

This calculus can be seen on display during the Paraguayan War 1864-70. Also known as the War of the Triple Alliance, named after the opposing bloc consisting of Brazil, Argentina, and Uruguay fighting against Paraguay, it is the most devastating war in the history of South America (Bethell, 1996). Several theories have been posited on the causes of the war; namely the colonial aftermath of the centuries long Portuguese-Spanish power struggle in Latin America (Whigham, 2002), the contested territories surrounding the fertile Platine basin that had already led to the Platine War (1851-52) in Uruguay (Box, 1967), and the conflicting interests of the regional hegemons (Brazil and Argentina) and the countries that they exercise influence over (Paraguay, Uruguay) (Centeno, 2002). Figure 2.5[21] shows the contested territories in the region and the configuration of the belligerents just before the onset of the war.

Further, the distribution of material capabilities across actors displays great variation. Brazil (10 million), Argentina (1.5 million), and Uruguay (200-350,000) were up against a total population of 300-400,000.[22] However, Paraguay actually enjoyed military superiority at the beginning of the war, and they were the initial aggressors (Hooker, 2008). With a standing army size estimated to be in the range of 28,000 and 57,000 plus about 25,000 reserves, virtually the whole male population of Paraguay mobilised for war (Bethell, 1996). The combined might, in terms of military troops that can be sent abroad,[23] of Brazil (17-20,000), Argentina (10-15,000) and Uruguay (5,000) was at best a match for Paraguay, but probably inferior. The Paraguayan army was also better-trained

[21] Vectorised map by Hoodinski, distributed under a CC BY-SA 3.0 license.

[22] Population estimates of Paraguay prior to the war is hotly contested in the literature. This mostly stems from the fact that initial reports put the total population of Paraguay about 1.3 million (Chartrain, 1972), making the post-war loss ratio closer to 90%. Later studies (Reber, 1988, 2002) corrected the pre-war population estimate to a more conservative 300-400,000; which is line with other studies claiming Paraguay lost about half of its population in a span of five years (Kleinpenning, 2002).

[23] For example, the Argentinians had an army size of 25-30,000; however, at least half of the army needed to stay put in the capital to ensure political consolidation of power during a period of "newly achieved internal unity and stability" (Bethell, 1996).

Figure 2.5: The region of Platine in 1864 showing the conflict parties of the War of the Triple Alliance and the location of contested territories

and better-equipped than its neighbours at the onset (Clodfelter, 2002).

After the initial Paraguayan offensive successes, however, the tide had quickly turned against them on the battlefield. More specifically, after the defeat of the Paraguayan navy by Brazil at the Battle of the Riachuelo taking place at the Paraná River on June 11, 1865, any threat to Argentina from Paraguay was neutralised. In the next three years, the Allies pushed towards interior Paraguayan territory, although opposition put up by Paraguayan soldiers greatly hindered their progress (Hooker, 2008). The Allied army finally entered Asunción in January 1869. Nevertheless, the end of the war did not come. Known as the 'Campaign of the Hills', the defeated Paraguayan president López retreated to the mountains where he led 9,000 resistance fighters against the occupying Allied forces (Esposito and Rava, 2015). The war finally ended when López was killed in the Battle of Cerro Corá on March 1, 1870.

The aftermath of the war was especially devastating for Paraguay. It is estimated that the immediate post-war population of Paraguay was around 150–160,000; of whom only 28,000 were adult males (Whigham and Potthast, 1999).[24] Still, even the estimated loss of 50-70% of their population puts Paraguay above Germany or Russia during World War II in terms of sheer magnitude, and it would take Paraguay 50 years to reach its pre-war population again (Clodfelter, 2002). Finally, Paraguay was also forced to cede 55,000 square meters of land to the victorious Allies, setting back their subsequent post-war reconstruction even more (Bethell, 1996).

Bartolomé Mitre, the Argentinian president and the supreme commander of the Triple Alliance forces, famously quipped that the Allies will be in Asunción in three months[25] (Rosa, 1968). However, it took four years for the Allies to

[24] It should be noted that at the time, infectious disease was as, if not more, deadly than the enemy itself (Clodfelter, 2002).

[25] "*…My fellow countrymen, I promise you: in 24 hours we shall be at the barracks. In two weeks, in Corrientes [the Argentinian province at the border that was attacked by the Paraguayan army]. And in three months in Asunción!*"

enter the Paraguayan capital and another year fighting against the guerilla warfare campaign put up by the loyalists. How come it took the Allies this long to subjugate Paraguay, a landlocked country—sharing vast borders with both Argentina and Brazil no less—slightly larger than Germany with a population of 400,000? The following account sheds some light on why the Triple Alliance—which was essentially reduced to Brazil after the first year of the war—had to fight a longer than anticipated war in which they enjoyed superiority in both material capability and military power for the vast majority of the conflict (Bethell, 1996, pp. 8):

> "…Brazilian governments faced enormous logistical problems, first organising, then transporting their troops thousands of kilometres either overland or by sea and up river, and finally supplying their troops. And breaking down Paraguay's excellent land and river defence was not an easy task. But it is also true that Brazilian commanders demonstrated a high degree of strategic and tactical ineptitude. On the other hand, the Paraguayan troops, indeed the Paraguayan people, remained loyal to Solano Lopez and fought with extraordinary tenacity and in the end, when national survival was at stake, heroically. This, and the Allied determination to pursue the war to the bitter end, also explains why the war was so bloody."

The limitations on power projection, as they were applicable to Brazil in the Paraguayan War, overlaps greatly with the existing theories on civil war duration.[26] The effect of distance (Buhaug, 2010), terrain (Fearon, 2004), and geography (Buhaug et al., 2009) in general; the advantages obtained by having larger populations (Hegre and Sambanis, 2006), military personnel/rebel combatants (Cunningham et al., 2009), and strong central command (Wucherpfennig et al., 2012); and leader characteristics (Uzonyi and

[26]See Hegre (2004) for a succinct introductory essay on the factors influencing civil war duration.

Wells, 2016) and technologies of war (Balcells and Kalyvas, 2014) are shown to be influential factors affecting civil war duration.

More specifically, the components of the proposed power projection model as a duration framework can be directly observed in this conflict. Even though the Triple Alliance had possessed impressive material capabilities on paper, they were never able to mobilise them to a great extent. The enormous combined landmass of Brazil and Argentina, coupled with the manpower available to Brazil, provided the Triple Alliance many advantages against the much smaller and less populous Paraguay. However, most of these advantageous features could not be translated into battlefield success. Even though they had high-capacity, the Alliance saw their power projection drastically dampened by political and non-political constraints.

First, the political situation in both Argentina and Brazil was not conducive to power projection. As alluded earlier, the Argentinian president was not popular at home, and he was forced to maintain a large contingent of the army garrisoning the capital against a potential putsch. The effect of this was two-fold: directly, it limited the amount of troops that could be sent to the front; indirectly, it constrained the deployment of such troops temporally (as the longer the soldiers stayed away, the greater the risk at home). On the other hand, the Paraguayan people supported their leader fervently. This effect was further exacerbated by the fact that, after the initial Paraguayan aggression, Paraguayans were now fighting for their survival and annihilation. Taken together, while the Alliance had their material capabilities reduced by political factors, Paraguay saw their lesser capabilities enhanced.

Second, the terrain nullified the material capability advantage of the Brazil. Distance acted as an equaliser; as the Alliance made progress further and further into Paraguay, there existed a power parity between the worn-down Paraguayan troops and the much larger Alliance contingent. The Paraguayans

were closer to their base, managed to set up elaborate defence networks, and their reinforcements—galvanised by Lopez and eager to follow him to the bitter end—were quick to replenish the soldiers in the ever-approaching front. Brazilian soldiers, in contrast, had to traverse a much longer distance, further away from their base of power. This was not helped by their non-existent logistical support, which was yet another factor limiting them from translating their potential capability to actual power on the ground.

In sum, the Triple Alliance acted as, in relative terms to Paraguay, a high-capacity yet low-projection power. They possessed high capacity because their latent power was considerable. However, due to political and non-political constraints on their power projection, they were not able to mobilise to their full extent. This, paired in a dyad with the low-capacity but high-projecting Paraguay, prolonged what should have been (given the drastic differences in capability) a swift contest otherwise.

2.4 Empirical Expectations

At its core, this project is a predictive enterprise build around important variables[27] and machine learning techniques. As such, I do not formulate the theoretical expectations using a Null Hypothesis Significance Testing (NHST) framework. The choice of empirical validation is given more exposition in the research design chapter.

Instead, I test the empirical validity of the proposed theoretical model based on its contribution to predictive accuracy and its overall ability to correctly forecast true positives and negatives. Given the predictors are selected based on their 'performance' in studies on civil war duration, the most important test is that whether established predictors of civil war duration are also important predictors

[27] *Variable importance* is a term in machine learning denoting a influential set of predictors that contribute positively to predictive accuracy. Roughly speaking, it can be thought of as a machine learning equivalent of 'statistical significance' in traditional statistics.

of interstate war duration. If this is indeed the case, it will act as strong evidence in favour of a common underlying mechanism governing conflict duration.

Unlike traditional statistical methods, most predictive modelling techniques do not establish directionality for their covariates. This makes the usual NHST formulation—e.g. x is positively correlated with y—an ill-fit for the purposes of this project. However, they are recent frameworks for extracting directionality from what are sometimes called 'black-boxes'. One such procedure—Local Interpretations of Model-agnostic Explanations (Ribeiro et al., 2016)—will be heavily utilised in the empirical chapters. Doing so will shed some light on how important predictors of conflict duration behave in a unitary model.

Further, directionality is paramount to the conduct of political science.Theoretical expectations guiding empirical research minimises the risk of identifying spurious correlations that might arise in the data. Even though I refrain from formulating alternative hypotheses against the null—a permanent feature of NHST studies—I nevertheless lay out the empirical expectations of the proposed general duration model in a predictive framework below.

First, from a probabilistic perspective, material capabilities γ are expected to enhance the fighting capabilities of an actor. This logic applies to both types of material capability; absolute and relative. Higher levels of absolute material capability—population, size, standing army size etc.—act as a larger reservoir of potential power. Relative capability, for example deployment numbers, are also indicative of such capability as the realisation of the potential power. It must be noted that both types of capability go hand in hand; further, their destructive effect is dependent on the ability of their opponent to take punishment. The dyadic nature of the model can be summarised as follows. A high-capacity actor projecting large amounts of force (high-potential, high projection) against a low-capacity actor can, on average, expect a shorter conflict. A high-capacity but low-projecting actor fighting the same low-capacity opponent would experience a

longer conflict, all other things being equal. If the high-capacity, low-projecting actor is against a high-capacity opponent, the expectation would be even longer. Plus, there will be further permutations depending on the projection ability of the opponent. This shows that while both types of capacity are important, they should be investigated in a dyadic setting as the process is one of interdependence.

Second, non-physical constraints α can either enhance or diminish the effect of material capabilities. In the former case, history provides a multitude of examples where the conventionally-weaker side has prevailed thanks to either popular support for their cause or the lack of it in their opponents. The US intervention in Vietnam is a well-known example of such a conflict. Similarly, in the case of the War of the Triple Alliance, Paraguay was able to keep on fighting a losing war longer than predicted by its opponents because the political power was consolidated and the people of Paraguay was desperate. Brazil and Argentina, on the other hand, neither had the popular support of their own populace nor were fighting a war of survival (as they were the invading party). Public support can also vary over time. Prior to Pearl Harbor, the US public opinion on their possible entry to the WWII was not in favour. However, after the Japanese surprise attack, the tide had turned which in turn allowed for the full mobilisation of the US population and its industry—and not wavered until the unconditional surrender of the Imperial Japan.

Similarly, the effect of physical constraints β on power projection is also dependent on baseline material capabilities. US projecting power all the way to Afghanistan in itself is a show of power. On the other hand, the fact that the military power of US inevitably decays as it is projected from its base, it gives the Afghan Taliban a fighting chance. Simply put, distance acts as an equaliser for power parity. This feature is also reflected in civil wars.[28] Often, in cases in which both the state and

[28]Of all the components considered, distance is one of the most defensible parameters in favour of a meaningful divide between interstate and civil wars as it conveys a different meaning depending on the setting. In interstate contexts, it is a display of power if one can project power over vast distances. On the other hand, in civil wars, distances are relatively constrained as they usually (but not always) take place in the same country. However, the difference between

the rebel forces are weak, the latter can put up a better fight if they can utilise the terrain to their advantage. Thus, we should expect physical constraints on power projection to prolong conflict in either case of war.

Taken together, the model of effective power projection as a determinant of duration can be thought of as the stability of a systemic reaction. Highly-capable conflict dyads will tend to be more chaotic systems than those which lack the capability. As a classic example, the duration of a hypothetical war between the Soviet Union and the US would be a volatile prediction given their immense capabilities. It can end in an instant, if the nuclear option is realised. It can be brief (but not instantaneous), if there was an initial escalation but the decision-makers decide to cooperate—perhaps in light of the first possibility. Or, if they choose to not directly engage each other but dabble in proxy wars, the conflict—depending on definition—can last decades. Plus, there are various other predictions that can be realised situated between these broad categorical outcomes.

Furthermore, there would be additional outcomes that are not even considered by political theorists or historians. For example, there could have been cases that allow for limited nuclear strikes rather than any nuclear option leading to Mutually-Assured Destruction (MAD). One of the parties can initiate a limited nuclear strike, and the other can sue for peace—probably in order to not to escalate the situation to MAD. With the aid of hindsight, we may not consider such possibilities; however from a modelling perspective, these would be additional outcomes that will not play out amidst low-capability actors.

In sum, the point being, higher destructive capacity $\gamma \cdot \alpha \cdot \beta$ leads to a large number of possibilities that are not available to low-capacity actors. This, in turn, makes the forecasting of such conflicts a more volatile affair. Probabilistically speaking,

the two can be traced back to material capabilities—state actors that can project power over distances can do so because of their high material capabilities, whereas state actors that are fighting rebels and are constrained by rough terrain do so because they lack the resources to nullify the effects of difficult terrain.

higher amounts of applied force should result in shorter conflicts. However, it still depends on the punishment-taking capacity of the opponent. If the opponent has vast reserves of manpower and material capacity themselves, they can replenish in time and prolong the duration of the conflict. This highlights the dyadic nature of the predictive model, which is a common way of studying conflict in the literature.

2.5 Conclusion

This chapter provides the foundations of a general model of war duration that aims to capture the common underlying dynamics of violent conflict. It is built on the notion that the findings borne out of theoretically-vast and empirically-rich literature on civil war dynamics should transcend to interstate war cases.

The model has three inter-related components: material capabilities limited by physical and non-physical constraints. The inter-connected nature of the theory makes it flexible enough so that there is enough dynamism in the conceptual framework to account for the temporal variation found in war. Indeed, the majority of the parameters that make up the three main components can vary within conflict, between conflict dyads, and from year to year.

The heading of material capabilities capture several absolute and relative metrics of physical force and potential destructiveness. Absolute capability refers to parameters that are not necessarily utilised to their full extent but nevertheless act as ceiling values (e.g. population, troop size). On the other hand, relative capability indicators convey magnitude in comparison to that of the opponent (population/troop ratio).

The non-physical constraints encapsulate domestic and/or international pressures on decision-makers, which often acts as a limitations on the usage of material capabilities. Higher number of actors—either as veto players or mere allies—can shift the distribution of capabilities on the ground. Regime type and political

constraints on the head executive are also closely intertwined with the application of military use of force.

Finally, difficulties associated with power projection via physical constraints can also penalise existing material capabilities. The crippling effect of distance on power, widely known as the loss-of-strength gradient, has a significant impact on conflict dynamics. It can create and break power parities, as well as clustering conflicts in space. Features of terrain that limit government reach are also shown to effect war longevity.

The contribution of a general model is two-fold. First, it enables comparative study of interstate and civil war that share similar characteristics. Some interstate wars are low-capacity conflicts fought between poor state actors. The Eritrean-Ethiopian War (1998-2000) was fought by two of the poorest countries in the world. In contrast, some rebel organisations have standing armies (Kachin rebels in Northern Myanmar), engage in taxation (FARC in Colombia), and have a higher GDP than thirty-something countries. At its peak, counter-terrorism specialists and security experts estimated the annual turnover of ISIS to be around $2 Billion.[29] These cases are historically studied separately. However, they might share more similarities with each other than they do with cases that their 'type' belongs to.

Second, it broadens our understanding of both types of conflict. Certain predictive variables are studied more thoroughly in one setting, or reveal themselves more readily in certain contexts. The US entanglement in Afghanistan is not usually thought as a credible commitment problem, whereas many similar civil war situations—the Israeli-Palestinian Conflict; Tamil Tigers in Sri Lanka to name a few—are. By putting forward a general model, we allow these otherwise disconnected findings to inform one another.

[29]Forbes (2014). The World's 10 Richest Terrorist Organizations. [online] Available at: https://www.forbes.com/sites/forbesinternational/2014/12/12/the-worlds-10-richest-terrorist-organizations/#5fda6de34f8a [Accessed 7 Jul. 2018].

The general theory provides us a set of possible predictive covariates of conflict duration that should explain war duration. Mostly studied in civil war settings, this rich set of variables should also hold exploratory and predictive power in forecasting models. This provides the empirical benchmark for which the proposed framework will be tested against; that is, whether the empirical results in the extant civil war literature can be generalised to include interstate wars as well.

Chapter 3

Research Design

This chapter lays out the overall design of the project. The heavy emphasis on empirics throughout the dissertation—theory-building, sensitivity analysis, replication studies, and predictive modelling using feature selection—necessitates a multitude of methodological choices to be made. For the same reason, the project utilises a large number of datasets, which requires adherence to common standards and data wrangling procedures. Both of these points are covered separately next.

The first part of the chapter deals with the general methodological approach undertaken throughout the project. Its contents provide justifications for the various choices of inference and validation. It starts off with a comparison of empirical triangulation (Webb et al., 1966; Jick, 1979; Tarrow, 1995) vs. integrative multi-method research (Seawright, 2016). Next, the dominant Null-Hypothesis Significance Testing (NHST) approach (Neyman and Pearson, 1933; Fisher, 1937, 1956) in the social sciences is compared to algorithmic predictive modelling (Marascuilo and McSweeney, 1977; Kuhn and Johnson, 2013). I posit several arguments in favour of the latter as being better suited for the needs of the project. Thirdly, I briefly summarise the three prevalent forecasting approaches in conflict research and where I situate the project.

Finally, I conclude the section with justifying the qualitative component—case selection after quantitative analysis—of the multi-method design.

The second part of the chapter pertains to data. Chapter Four provides a quantitative assessment of the conflict duration literature; and as such, utilises 16 replication studies. The selection process resulting in these studies and various data transformations to ensure overall compatibility and coherence are discussed here. Next, I justify the selection of Sierra Leone as a shadow case by explaining its conflict actor structures and how this can be leveraged to identify possible shortcomings of predictive modelling.

3.1 Methodology

This section consists of four inter-related debates on methodological choices. First, I motivate why integrative multi-methods research is a better methodological approach than empirical triangulation. Second, in the same vein, I make a case for predictive accuracy being a better indicator of empirical assessment than mainstream p-value significance testing. Third, I outline strengths and limitations of algorithmic conflict forecasting. Forth, I provide justifications for why the case of Sierra Leone Civil War is apt for the purposes of this research, and how it can be utilised to highlight the empirical blind spots of algorithmic forecasting.

3.1.1 Triangulation vs. Integrative Multi-Method Research

In social sciences, triangulation as an empirical strategy is the most common application of mixed-methods research (Seawright, 2016). The concept of triangulation is named after the geometrical concept of using two known points in some Euclidean space to situate an unknown point located in the said space (Mertens and Hesse-Biber, 2012). In the same vein, researchers may employ

multiple methods or empirical techniques to inquire about a question of interest. In doing so, they will be able to make causal inferences supported by two different strains of empirical methods. Such findings are thought to be superior to that of those that are borne out of a single empirical method (Olsen, 2004).

However, the added-value of triangulated research has been questioned (Seawright, 2016). Putting aside the criticism raised on post-modernist and post-structuralist grounds,[1] there are at least two major flows inherent in triangulation frameworks.

Assume a study that utilises both qualitative and quantitative methods. Further, the practitioner has been implementing both methods in order to answer the same research question. One possible outcome of this enterprise is that the empirical findings may not actually converge. In other words, what conclusion should (or can) be drawn when the results contradict each other? Even though this shortcoming has been elaborated at length in the research methodology literature (Robson, 2002), there is no clear answer that can be generalised.

Second, even if the findings do overlap, what inferences can be made? Seawright (2016) gives the following example[2] using the finding of mountainous terrain (as a logged percentage of state's territory) in Fearon and Laitin (2003). The logit coefficient of 0.219 is statistically significant at the conventional levels, meaning increasing coverage of mountainous terrain has a positive effect on the onset of civil war. Seawright (2016) then provides several anecdotes relating to the role of mountainous terrain in Colombia; i) naturally, different parts of the country had varying levels of elevation, ii) mountainous terrain did indeed had some positive effect on conflict onset at certain times (but not always), and conversely iii) many highly mountainous areas in Colombia have not seen conflict.

What conclusion can one triangulate combining the 0.219 logit coefficient with

[1] See Howe (1988) for the epistemological paradigm 'incompatibility thesis' on the mixing of qualitative and quantitative methods.

[2] Even though he gives the account on the subject of non-overlapping results, the same implications apply.

the stylised facts supplied above? It can be argued that the coefficient captures the average effect, and the sometimes-contradictory qualitative evidence is how such an average effect can manifest itself in real life. However, given the great epistemological differences underlying the two methods, the comparison can only be made in an abstract manner.

A better approach is to aim for integrating mixed-methods rather than triangulating. Integrative mixed-methods can be described as '…multi-method designs in which two or more methods are carefully combined to support a single, unified causal inference. With such a design, one method will produce the final inference, and the other is used to design, test, refine, or bolster the analysis producing the inference' (Seawright, 2016, pp. 8).

Given the aim of the project—analysing civil and interstate wars using combined data and a general theory—the more systematic method should be the main method. Thus, I employ quantitative methods to generate the main inferences from the data. However, for the reasons that will be explained in more detail in the following section, I do not employ traditional statistical methods. Instead, I employ algorithmic approaches drawn from various machine learning, ensemble, and deep learning methods. These approaches are better suited to uncovering non-linear effects and interactions, and such they are more appropriate for an exploratory general theory empirics.

In contrast, I rely on the qualitative component to highlight any potential shortcomings or empirical blind spots of the quantitative analysis. To achieve this, I first conduct the quantitative analysis and let the systematic findings guide me in my case selection process. Selecting a case study after quantitative analysis informs the practitioner of the possible shortcomings of the latter and acts as an empirical 'follow-up' (Seawright, 2016).

Indeed, an integrative mixed-methods design can evoke the concept of *complementarity.* For example, by tackling the same case from two different

paradigms, one can better understand how the data are constructed (and subsequently, re-constructed). Rich concepts that can be scrutinised greatly using qualitative methods are often reduced to inherently less-discriminatory numerical values in quantitative research. By combining the data coming in from both approaches, it is possible for a researcher to attain a holistic perspective on the issue under investigation.

Complementary mixed-methods can enrich and illuminate the empirical findings beyond the means of one singular approach. A quantitative scholar undertaking a cross-country study is unlikely to be an expert in every single case in the dataset. However, if she is inherently familiar with some of the cases and opts to gather further information (e.g. conducting surveys, fieldwork), this can be used to elaborate on the systematic findings borne out of quantitative studies. In the case of this research project, the qualitative findings based on my fieldwork in Sierra Leone informs the predictive modelling findings achieved by machine learning. To this end, in Chapter 6, I conclude the shadow case study with a side-by-side comparison of the data collected in the field and the algorithmic predictions of the Sierra Leone Civil War using the Local Interpretations of Model-agnostic Explanations (LIME) framework (Ribeiro et al., 2016). Doing so makes the quantitative findings more interpretable and provides possible pathways for how detail-rich narratives can manifest in systematic studies.

3.1.2 Null-Hypothesis Significance Testing (NHST) vs. Predictive Modelling

Null-Hypothesis Significance Testing (NHST) is arguably the most common procedure in quantitative social science (Nickerson, 2000). Succinctly summarised, the method (Fisher, 1937, 1955, 1956) allows the researcher to compute the probability of observing a result that is at least as extreme as a test statistic (t-value), under the assumption that the null hypothesis h_0 positing no effect is

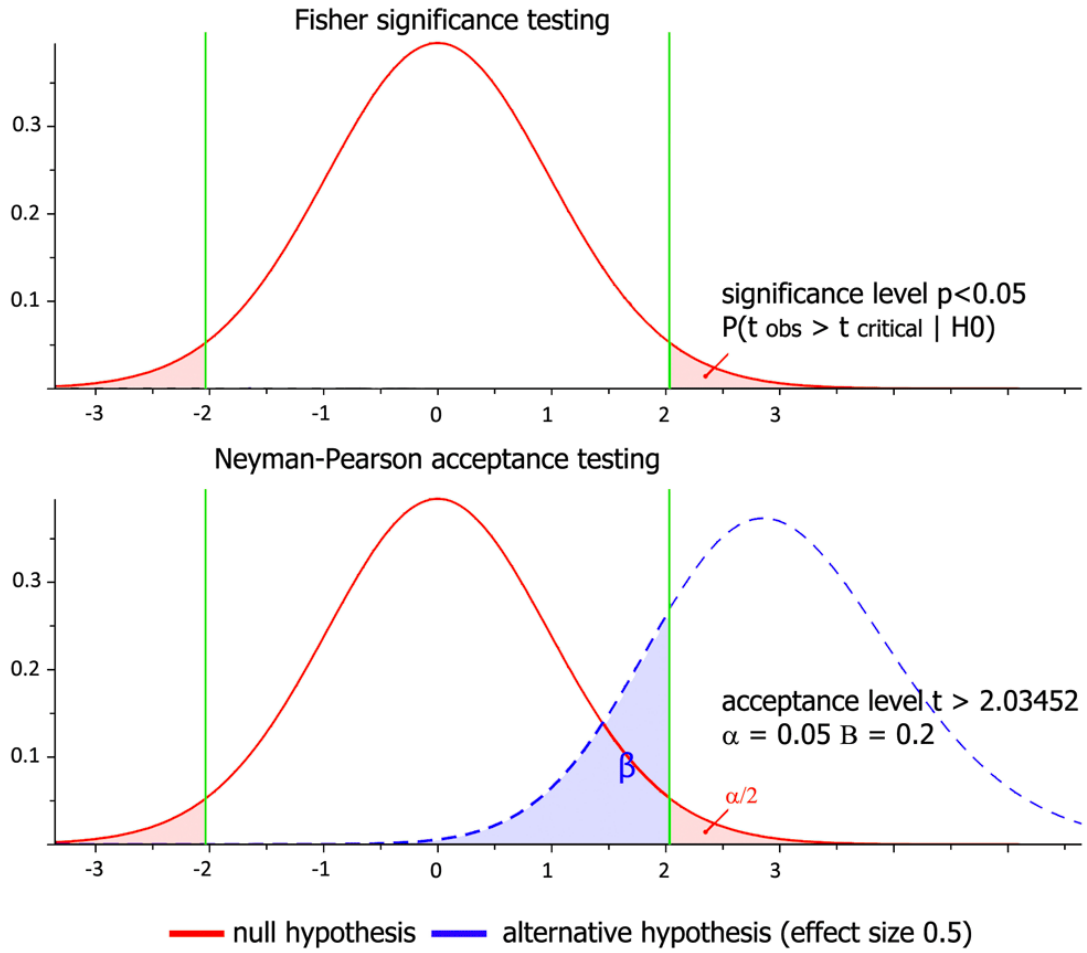

Figure 3.1: Significance and p-values, taken from Turkheimer et al. 2004

true. This p-value in turn denotes the conditional probability of achieving the observed or a larger outcome, making it a cumulative probability as opposed to a point estimate (Nickerson, 2000).

Figure 3.1 demonstrates two common approaches to NHST; The Fisher test of significance and the Newman-Pearson test of acceptance. In Fisher's formulation, the p-value[3] estimation equals the area under the null probability distribution curve starting from the observed test statistic and ending at the tail of the null distribution (Turkheimer et al., 2004). This formulation has led to the notion that the Fisher significance test operates via 'proof by contradiction' (Christensen, 2005).

[3]For an informative piece on the misinterpretation of p-values, see Cohen (1994).

1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies with the same P value provide the same evidence against the null hypothesis.*
6	*P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.*
7	*P = .05 and P ≤.05 mean the same thing.*
8	*P values are properly written as inequalities (eg, "P ≤.02" when P = .015)*
9	*P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
10	*With a P = .05 threshold for significance, the chance of a type I error will be 5%.*
11	*You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.*
12	*A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*

Figure 3.2: Twelve common p-value misconceptions by Goodman 2008

However, since its introduction NHST has been deemed as a controversial technique (Rozeboom, 1960; Pearce, 1992). The criticisms are manifold (Bakan, 1966; Branch, 2014) and such, beyond the scope of this section. Still, I provide some of the common criticisms associated with the NHST procedure below.

Most commonly, the concept of p-value is commonly misinterpreted or misused by social scientists (Cohen, 1994). Figure 3.2 shows the twelve common p-value misconceptions described by Goodman (2008). As most of the quantitative research has been done using the NHST framework up until now, the vast majority of our body of scientific knowledge stems from studies that exclusively focus on in-sample explanation. Such studies usually have low out-of-sample generalisability (Kukull and Ganguli, 2012); further, statistically significant findings are not automatically good predictors (Lo et al., 2015).

Next, many assumptions underlying the NHST procedure are not met regularly in published work (Lykken, 1991). This is a contributing factor to what is generally known as the replication (reproducibility) crisis in science (Moonesinghe et al., 2007; Begley and Ioannidis, 2015). Especially in studies with low statistical power, the p-value has a large variance across repeated samples, which makes it unreliable for the purposes of precise replication (Halsey et al., 2015).

In the same vein, the policy of enforcing stringent requirements of statistical significance in scientific journals exacerbates this problem. Commonly referred

as publication bias (Begg and Berlin, 1988), this type of path-dependency is especially harmful to science given its cumulative nature, as eventually findings will be skewed to a point where existing body of research will no longer be balanced (Song et al., 2010). Perhaps in light of this and other contributing factors, prominent political methodology journal *Political Analysis* has recently announced updated procedures relating to publications using statistical significance (Gill, 2018).

Other pitfalls can also arise; including the practitioners' mixing of Fisher and Newman-Pearson methods interchangeably (Tukey, 1960); lack of sufficiently large sample sizes (Biau et al., 2008); susceptibility to subjective nature of hypothesis definitions (Gigerenzer, 2004); the inability to account for prior beliefs and/or given data (Masson, 2011); and the arbitrarily low threshold ($p < .05$) of statistical significance (Benjamin et al., 2018).

Finally, the perils of p-value driven research are also studied in the specific context of political science (Ward and Bakke, 2005; Ward et al., 2010). In yet another example of methodological issues relating to Fearon and Laitin (2003), which was used in a New York Times Op-Ed written by the academic Jacqueline Stevens as a proof for political scientist being lousy forecasters.[4] One main reason as to why the Fearon and Laitin (2003) study fails at prediction is their emphasis on in-sample explanation, as opposed to out-of-sample prediction (Ward et al., 2010, pp. 479):

> "The poor predictive performance is not an indictment of Fearon and Laitin's contribution, nor is it evidence that prediction is too treacherous to attempt. Rather, it points to an opening for social scientists and to the benefits of embracing prediction as a concept. First, it establishes a framework for rigorous and ongoing

[4]The New York Times (2012). Political Scientists are Lousy Forecasters. [online] Available at: https://www.nytimes.com/2012/06/24/opinion/sunday/political-scientists-are-lousy-forecasters.html [Accessed 7 Jul. 2018]

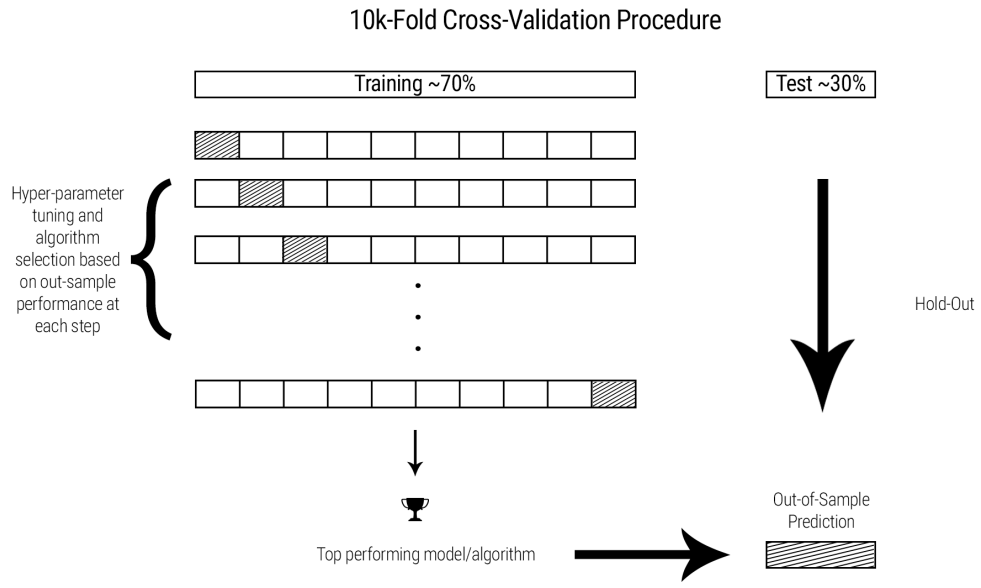

Figure 3.3: Cross-validation and data splitting procedures

cross-validation of our models. This cross-validation offers us the opportunity to test our theories, their scope, and their portability, which can provide valuable input in the theory-building process. Finally, generating predictions makes the implications of our research more accessible to the policy community and the general public. Specifically, it underscores the opportunity for developing better models."

A predictive modelling approach build on out-of-sample cross-validation does not share the above shortcomings characterising the NHST procedure. Figure 3.3 demonstrates the cross-validation procedure undertaken in for this project.

The general process of cross-validation can be described as follows. First, the available data—assumed to be sampled from the same population—is split into two groups; training and test.[5] The ratio of this split depends on multiple factors,

[5]Data permitting, it can also be split into three parts—training, validation, and test. Validation set is used for hyper-parameter tuning, so that the best tuned algorithm can be

including but not limited to, the overall sample size, class-imbalance, temporal dynamics of the data, and domain knowledge of the practitioner (Breiman, 1984; Kohavi et al., 1995; Varma and Simon, 2006). For the purposes of this project, the data are split into approx. 70%/30% to training and test samples, respectively.

The training set is referred to as the cross-validation set, in which an iterative process similar to the one described in Figure 3.3 occurs. This 70% split of the data is then further split into ten folds. During each iteration, one fold is held out for validation. That is, during hyper-parameter tuning and/or algorithm selection processes, the models use nine folds to train on and then predict the values of the unseen tenth fold. This hold-out fold changes with every iteration. Once all individual folds are tried, the top performing model based on its out-of-sample[6] prediction accuracy within the 10k-fold cross-validation is selected as the candidate model for the final forecast.

The test data (30% from the initial data split), also called the hold-out, is never exposed to the learning algorithm during the cross-validation process. This ensures the integrity of the validation test; the algorithm has never seen the test data and can only predict the outcomes if the patterns it learned during the cross-validation generalises to the test data. The cross-validation process thus optimises for external validity, and it is a superior approach to in-sample validation (which maximises internal fitness) for research that focuses on prediction. At this stage, the top performing model predicts the outcome values of the final hold-out, and this is what is reported as out-of-sample accuracy.

In sum, I opt for algorithmic predictive modelling using out-of-sample validation as the main quantitative component of the project. Doing so i) ensures

applied to a never-seen-before data (that is, test). However, if the sample size is small, it is generally advised to skip the validation set (Kuhn and Johnson, 2013). Also see Korjus et al. (2016).

[6]This should not be confused with the test data split that is held-out from the beginning. Given the folds are created within the 70% training data split, it is technically 'in-sample'; however, the selection procedure of 10k-fold still maximises external validity as seen in Figure 3.3. In this dissertation, the 10k-fold cross-validation is referred to as 'in-sample', whereas the predictions based on the true hold-out (test split) is called the 'out-sample'.

the replicability of the results; ii) provides a better fit for empirically testing generalisable theories; and iii) gets rid of many methodological concerns stemming from using NHST procedures (Cranmer and Desmarais, 2017). Finally, I add a limited case study, which is selected after the conclusion of the large-n study, to uncover the limitations of the covariates used in predictive modelling.

3.1.3 Forecasting in Conflict Research

Conflict research, similar to its preceding fields of political science and international relations, has historically favoured backward-looking causal explanations over forward-looking predictive power (Schneider et al., 2011). However, in the last decade, conflict forecasting has gained considerable momentum thanks to the innovations in variable measurement, the gathering of disaggregated data, and the introduction of more complex computational techniques to the practitioners (Schrodt et al., 2013).

There are three main approaches to conflict forecasting: expert-driven, game-theoretic, and algorithmic modelling (Schneider et al., 2011). Further, these three components can be combined in various ways to create ensemble forecasts (Montgomery et al., 2012). Expert predictions are probably the most well-known and the most visible type of political forecasting. However, they are not necessarily the most precise. In a highly-cited study, expert predictions on geopolitical events over the course of two decades failed to outperform random guesses on average (Tetlock, 2005). There are reasons why this may be the case: experts are incentivised to have strong positions, otherwise they run the risk of being 'dull'; and they rarely suffer any serious setbacks in case they fail (Chadefaux, 2017). Still, human predictions can aid computational forecasting, especially via an iterative process that selects for high marginal utility. For example, human forecasters in tournaments have beaten sophisticated algorithms and other 'superforecasters' (Tetlock and Gardner, 2016).

Game-theoretic approaches to conflict forecasting emphasise formal models to generate predictions. The need to formally specify models puts a high premium on theoretical completeness, making them more precise in comparison to informal models (Powell, 1996). Formal modelling focusing on prediction is a two-step enterprise. First, an expert identifies a set of relevant factors. Then, the model is constructed to capture the predicted interactions in the data. This type of setup has been successfully employed in several studies (Gurr and Lichbach, 1986; De Mesquita, 2010). On the other hand, game-theoretic expectations of human behaviour are usually at odds with observed empirical patterns because of their complexity (Axelrod, 1984; Kahneman and Egan, 2011). Agent-based approaches—models following a simple set of rules but come with a high computational cost—can overcome the shortcomings of game-theoretic models as they allow for the comparison of different scenarios and the evaluation of the counterfactuals (Cederman, 2002). However, agent-based models are complex in their own regard, making them difficult targets to draw causal inferences from (Chadefaux, 2017).

Finally, algorithmic modelling is a theory-led approach to conflict forecasting with a computational bend. It can utilise logistic regression and other members of the increasingly sophisticated generalised linear methods family (Rummel, 1969; Goldstone et al., 2010; Weidmann and Ward, 2010; Bell et al., 2013; Gleditsch and Ward, 2013; Hegre et al., 2013; Ward et al., 2013; Owsiak, 2015; Bagozzi, 2015; Hegre et al., 2016; Chiba and Gleditsch, 2017; Witmer et al., 2017) or Bayesian approaches (Brandt et al., 2011, 2014), but lately expanded to include machine learning and neural networks (Brandt and Freeman, 2006; Muchlinski et al., 2016; Bessler et al., 2016; Colaresi and Mahmood, 2017; Bagozzi and Koren, 2017). The introduction of highly disaggregated data, both for spatial and temporal domains, has allowed conflict researchers to identify and predict episodes of violence with increasing precision (Schrodt and Gerner, 2000; Chadefaux, 2014). In a similar vein, the automation of news report coding has led to more accurate

forecasts (Schrodt, 2009). However, automated text analysis still suffers from several shortcomings (Eck, 2012; Weidmann, 2015). First, complex sentence structures and implied meaning are difficult to capture (Croicu and Weidmann, 2015). Further, while temporal disaggregation in events can be achieved with high accuracy, spatial disaggregation is still lagging behind in comparison (Croicu and Hegre, 2018).

For the purposes of this project, I predominantly follow the algorithmic approach. However, I also incorporate components from the other two approaches. First, similar to formal modelling, I utilise a real life case—the Paraguayan War 1864-1870—to identify the relevant set of covariates for my theory. Next, I draw upon the rich civil war duration literature and evaluate many empirical operationalisations of the proposed theory. Finally, using a wide range of feature selection and predictive modelling algorithms, I pinpoint which covariates are the top predictors of conflict duration.

The aim of the predictive modelling enterprise, by definition, is to accurately estimate a future outcome given some contemporary covariates. Yet, there is no general consensus on whether conflict can be truly predicted; and if it indeed can, what type of events should be prioritised (or even feasible) (Cederman and Weidmann, 2017)? The oft-repeated quip, usually attributed to the Danish physicist Niels Bohr on the nature of quantum physics—"prediction is very difficult, especially about the future"—also accurately reflects the challenges of forecasting in other scientific fields beyond physics (Ellis, 1970). War can indeed be in the error term (Gartzke, 1999).

Further, even if it is not (i.e. war is predictable), some events are categorically different than others from a forecasting perspective. Popularised by Nassim Nicholas Taleb, the term 'black swan' refers to the unexpected, high-profile events of large magnitude that mostly fall outside of the empirical detection of scientific models given their astronomically low chances of occurrence ex-ante (Taleb, 2007).

The dissolution of the Soviet Union and the September 11 terror attacks in the U.S. are some of the common examples of black swan events. In contrast, Gleditsch (2017) argues that predicting 'white swan' events—that is, regularities in conflict—has greatly increased our understanding of various conflict processes over time.

I subscribe to this point of view as well, evidenced by the unified nature of my proposed theory and the inclusiveness of my empirical testing without stratifying conflict by type. Plus, duration forecasts—especially using a structural model—are more prone to displaying convergent properties than onset predictions, which are more likely to be the result of more idiosyncratic factors. In sum, I employ algorithmic predictive modelling to uncover the regularities pertaining to armed conflict duration that are applicable to both civil and interstate wars. However, every conflict prediction undertaking shares the same limitations and this project is not an exception. First, several strong assumptions are required for prediction: i) the covariates truly capture the phenomena they proxy and they are linked to the underlying data generating mechanism, ii) the linkages between the predictors and the outcome captured in the past will continue to hold in the future, and iii) exogenous factors ('the world') will largely stay the same. These assumptions are readily made; however they can be challenged—either singularly or as a group.

Some of the most important predictors of conflict are nigh-impossible to capture. Leader personalities, characteristics, and even 'moods' have an immense influence on conflict processes that are conceptualised as bargaining or signalling games. Yet, they are difficult to measure accurately (Chadefaux, 2017). Further, conflict settings are usually characterised by an interacting set of decision-makers who all have incentives to break rules and avoid pattern-detection, which makes prediction problematic (Cederman and Weidmann, 2017). As alluded to earlier, black swan events can alter the course of the history in ways beyond the adaptation capacity of a predictive model. In addition to unknown-unknowns, known phenomena

can also change the world as we know it (e.g. oil depletion, climate change, post-antibiotics). Thus, while I abide by the three rather strong assumptions, I am aware of the potential pitfalls surrounding conflict forecasting.

3.1.4 Case Selection after Quantitative Research

The literature on qualitative methodology is abound with case selection strategies (Seawright and Gerring, 2008). Still, given the possible range of options to choose from, the practitioner needs to justify why they opt for one technique over another. As the role of the case study in this particular project is discovery, certain case selection strategies make more sense than others.

For example, deliberate sampling with the intention of maximising variation found in the data (King et al., 1994) and completely random sampling (Fearon and Laitin, 2008) are two possible case selection strategies. However, both approaches are ill-suited if the goal of the case study research is to learn beyond what is already established by the large-n component (Seawright, 2016).

In contrast, both deviant and extreme-on-X case selection strategies (Seawright and Gerring, 2008) are shown to be most successful at i) identifying the sources of measurement error (King et al., 1994); ii) searching for omitted variables (Collier et al., 2004a); iii) testing causal paths (George and Bennett, 2005); and iv) establishing substantive boundaries of the set of cases sharing the same underlying causal mechanism (Collier and Mahoney, 1996).

Given the proposed general model focusing heavily on material capabilities, the case of the Sierra Leone Civil War ticks many boxes for being both a deviant and an extreme case. First, I motivate for which independent variables it is considered as an extreme case. Even though a case can be selected on the outcome (extreme-on-Y), it is shown to be problematic for causal inference (Seawright and Gerring, 2008; Seawright, 2016). Thus, I only consider which variables in the

Sierra Leone case on the right-hand side of the equation have extreme values with regards to the large-n averages.

Seawright and Gerring (2008) define extreme as "an observation that lies far away from the mean of a given distribution; that is to say, it is *unusual.*" More formally, the Extremity (E) value of the ith case can be defined using the sample mean ($\bar{X}$) and the standard deviation (s) for that variable as the following:

$$E_i = \left| \frac{X_i - \bar{X}}{s} \right|$$

which is equal to the absolute value of the Z-score (Stone, 1996) for the ith case. Given the proposed theoretical framework, a stylised typical case can be defined as

$$E(duration_i) = \beta_0 + \beta_1 Capability_i + \beta_2 Politics_i + \beta_3 Projection_i.$$

The case of Sierra Leone consists of several extreme values pertaining to material capability. For starters, both main conflict actors—the Government of Sierra Leone and the Revolutionary United Front (RUF)—possessed minimal capabilities at the onset. Sierra Leone has the 27th lowest CINC score and 6th lowest military expenditure in 1991.[7] The Sierra Leonean Army (SLA) was mostly ceremonial (Richards et al., 1998), and only had slightly more than 3,000 military personnel—in a country with a population of four million and half the size of England—when the war broke out. Similarly, the RUF is estimated to have 100 combatants[8] at the beginning of their insurgency. This dyadic lack of capabilities puts them under the symmetric non-conventional designation of technologies of rebellion (Kalyvas and Balcells, 2010), which accounts for only 13% of all civil

[7] Excluding countries with zero or missing values in military personnel or military expenditure variables (Singer, 1972).

[8] Uppsala Conflict Data Program (Date of retrieval: 30/06/18) UCDP Conflict Encyclopaedia: www.ucdp.uu.se, Uppsala University.

wars since 1944.[9]

If extremity implies unusualness, then deviantness signals anomalousness. Deviant cases, in reference to some accepted understanding of an issue—common sense, established theory or a proposed model—demonstrate a *surprising* value and they are therefore "closely linked to the investigation of theoretical anomalies" (Seawright and Gerring, 2008). More formally,

$$Deviantness(i) = abs[y_i - E(y_i|x_{1,i}, \ldots, x_{K,i})] = abs[y_i - b_0 + b_1 x_{1,i} + \cdots + b_K x_{K,i}]$$

Defined this way, cases on the regression line has a deviantness score of 0, while the upper bound of the measure is theoretically positive infinity. As a result, one should be interested in selecting from the set of cases with the highest overall estimated deviantness (Seawright and Gerring, 2008).

In certain components of the proposed theoretical framework, the case of Sierra Leone displays certain deviant qualities, especially in conjunction with other components. For example, there was a lack of material capabilities affecting both main warring parties as alluded above. However, both had access to rich natural resources; the Sierra Leone Civil War is commonly included in the 'blood diamond' conflicts (Le Billon, 2008). On the other hand, in many cases, it was difficult for the rebels to take control of the mines[10] and to maintain control once they had captured them[11]. Thus, it is not clear how much the rebels had benefited from conflict resources, or the full extent of the influence of conflict diamonds on the fighting capacity of the rebels.

In addition to extreme/deviant qualities, the Sierra Leone case has other advantages that makes it a suitable focus study. Most importantly, it had a

[9]The percentage rises to 26.09% for the post Cold War period (1990–2004).
[10]Big Daddy (senior RUF commander), personal interview, Makeni, 09/03/2017.
[11]Security advisor, personal interview, Freetown, 18/03/2017.

diverse set of conflict actors. Even though the war started between the SLA and the RUF, many domestic and international actors ended up joining the fight. Early on, the government had secured the services of the South African private military company known as the Executive Outcomes (EO). The EO was quite capable; commonly referred as well-trained and well-equipped (Harding, 1997). They were exclusively stationed in the mining areas, where they were credited to stop the rebel advances and made sure the diamond revenue was flowing toward the government forces (Fitzsimmons, 2013). Domestically, tribal hunters from the Mende ethnic group known as *Komojors* entered the frey, fighting against the both sides at certain times during the war. They continued to be important actors throughout the conflict (Zack-Williams, 1997), but they have been accused of major human rights violations as well (Ero, 2000). In 1997, the Economic Community of West African States Monitoring Group (ECOMOG) intervened in Sierra Leone, consisting of mostly Nigerian soldiers numbering around 16,000 possessing armoured vehicles and fighter jets (Adebajo, 2002). The UN peacekeeping force (UNAMSIL) entered Sierra Leone in 2001 with a force largest on its kind at the time (Olonisakin, 2008). Finally, the UK was involved in several crucial military engagements towards the end of the conflict (Keen, 2005).

Even though the combination of weak governments, rebel infighting, militia formations, and peacekeeping operations are not special to the case of Sierra Leone, the involvement of numerous actors that possess such a diverse set of material capabilities and political constraints—as well as distance-related projection problems in some cases—makes it an apt target for further scrutiny. In addition to idiosyncratic actor capabilities and political interests, many actors also saw their capabilities and political agenda transform during the course of the conflict. Both the between- and within-actor variations are quite valuable as a means of going beyond the typical case and learn more about the limitations of the large-n findings.

Finally, there are several other factors mainly relating to feasibility that makes Sierra Leone an ideal choice. Sierra Leone is one of the safest post-conflict settings to conduct fieldwork; the civil war ended fifteen years ago and there have been no serious relapses since then. Second, even though Creole (Krio) is de facto language amongst the local populace including in Freetown, owing to their British colonial history, one can get by conducting interviews in English. Third, the case of Sierra Leone has been widely studied by scholars, resulting in a research infrastructure that is immensely beneficial to first-time interviewers.

3.2 Data

The next two sub-sections provide empirical background and motivate various secondary methodological choices influencing the data. First, the selection and filtering procedure of conflict duration studies using binary-time-series-cross-section data is explained. Next, I describe the replication procedure. Finally, I provide background information on the case of Sierra Leone, including the interviewing strategy and the raw-data processing.

3.2.1 Replication Studies

Similar to the methodological approaches taken by Hegre and Sambanis (2006) and Hendrix (2010), I start out by quantitatively assessing the conflict literature to identify which covariates are consistently chosen as good predictors of conflict duration. This enterprise can also be seen as a sensitivity analysis on the determinants of war longevity. However, several guidelines need to be established to ensure apples are indeed compared to other apples.

First, I locate existing quantitative research on armed conflict duration without

discriminating between inter- and intra-state wars.[12] This search results in about 1,698 matches in total, containing 46 eligible studies. Table 3.1 provides basic descriptive statistics of these studies: number of observations, number of predictors, degrees of freedom, time coverage, conflict type, and the choice of statistical model. The most common data structure used in this batch of studies takes the shape of binary-time-series-cross-section (BTSCS). Within that subset, the vast majority of studies have yearly-data when they include time-varying covariates. I drop studies that use disaggregated data at the level of days (e.g. Weisiger, 2016) to maintain uniformity, as well as studies without replication data and do-files. In the end, I am left with 16 BTSCS studies—two on interstate, thirteen on civil war, and one combined study.

I follow the original model formulation of the authors; however two interventions are made. First, most duration studies using traditional statistical tools naturally employ various parametric (e.g. Weibull) and semi-parametric (Cox Proportional-Hazards) forms of survival analyses. Algorithmic predictive modelling does not do well with survival processes (Zupan et al., 2000), especially in the presence of time-varying covariates (Ripley and Ripley, 2001).

One common approach to circumvent this shortcoming is to convert survival analysis into a classification problem (Abbott, 1985). In contrast to survival analysis, classification algorithms are well-developed in the machine learning literature (Weiss and Kulikowski, 1991). Most simply, a duration study can be transformed into a logistic regression in which time is included as a covariate (see Cunningham, 2006, for an example). However, the inclusion of cubic splines is advised to properly capture the inherent time-dependency found in duration data (Beck et al., 1998; Carter and Signorino, 2010). Thus, all selected studies are turned into classification problems with their original model specifications left intact:

[12]LSE library search results with the keywords 'conflict | war + duration' in title, English language, peer-reviewed, journal article, political science | international relations | war | conflict fields.

$$y \sim x_1 + \cdots + x_k + t + t^2 + t^3.$$

Other common data transformations are also applied. Most machine learning algorithms perform better when the underlying data is either normalised or standardised (Witten et al., 2016). All predictors are thus centred and scaled. Algorithms also find missing values hard to deal with (Jerez et al., 2010). As such, variables with severe missingness (>= 25%) are dropped from the models. Less severe missingness is dealt with out-of-bag imputation (Stekhoven and Bühlmann, 2011).[13] All predictive modelling is done using `caret` and `caretEnsemble` packages (Kuhn, 2018; Deane-Mayer and Knowles, 2016) in `R` (R Core Team, 2018).

Finally, the replication enterprise is built on the fact that the predictors are made consistent across studies. Even though a certain percentage of the independent variables (including controls) such as GDP P.C., population size etc. are labelled consistently in all studies, certain alterations are made to others to ensure uniformity.

First, variable names are reduced to the core concept they are measuring. Meaning, transformations ('log', 'square')[14], descriptors (e.g. 'size', 'total', 'per'), and various other qualitative labels are discarded. The initial stemming process is done using the `quanteda` package (Benoit, 2018) and then manually checked for accuracy.

Second, the data section of the articles are consulted to identify how variables are operationalised. This is more salient when a variable is constructed specifically for the problem at hand. For example, Bennett and Stam (1996) operationalises 'balance of forces' as the ratio of the higher CINC value to the sum of the CINC

[13] Also see Honaker and King (2010) for a primer on dealing with missingness in time-series cross-section data.

[14] With the exception of cubic splines $t + t^2 + t^3$.

values of all participants. This is allowed to exist alongside the regular CINC variable as it provides additional information regarding capabilities; however, any other study that has CINC ratio as a variable has its label changed to 'balance of forces'. This ensures uniformity across studies and minimises the risk of including the same operationalisation under different labels.

3.2.2 Case Study Interviews

For the qualitative component of the project, I conducted 19 semi-structured interviews in Sierra Leone between January-March 2017. Kajornboon (2005) quotes the following explanation of the semi-structured interview technique (Corbetta, 2003, pp. 270):

> "The order in which the various topics are dealt with and the wording of the questions are left to the interviewer's discretion. Within each topic, the interviewer is free to conduct the conversation as he thinks fit, to ask the questions he deems appropriate in the words he considers best, to give explanation and ask for clarification if the answer is not clear, to prompt the respondent to elucidate further if necessary, and to establish his own style of conversation."

They also provide a suitable framework to the informants to say what they have to say in their own terms (Carruthers, 1990). Plus, the open-ended nature of the interviews allows for additional observations depending on what the correspondent chooses to disclose (or not), the vocabulary they use, and the linkages they make (Drever, 1995).

Further, these are in addition to the common advantages of conducting interviews as a means of obtaining data such as i) being well-suited to the exploration of attitudes, values, beliefs and motives (Richardson et al., 1965); ii) providing the opportunity to assess the validity of the respondent's answers via observing

non-verbal indicators, especially regarding sensitive issues (Gorden, 1975); iii) facilitating comparability by making sure all questions are answered by each respondent (Bailey, 1987); and iv) ensuring that the respondents cannot receive assistance from others while formulating their own responses (Bailey, 1987).

If it was permitted, I recorded the interviews (12/19) given the open-ended nature of the semi-structured interview process. However, in seven cases, the interviewees expressed concerns about being on record and directly quoted. The information obtained from these interviews still informed my thinking, however they are not explicitly expressed or attributed in the manuscript.

I targeted a wide range of interviewees from both sides of the conflict. I paid special interest to ex-combatants and military personnel who were active during the civil war, as well as individuals close to the powers-that-be. The former group ended up including several high-ranking officers in the army, a senior RUF commander, and a multitude of rebel rank-and-file soldiers. In contrast, the latter set of correspondents work in the national security apparatus, the law enforcement, or involved in the personal security of high-ranking civilian administrators.

Special attention was also paid to select individuals who were involved in the conflict from day one. For example, a senior RUF commander who fought the whole war turned out to be an invaluable source of information on how the targeting strategies of the RUF had changed as a result of shifting fighting capabilities and other information obtained in the battlefield. In other cases, some of the interviewees originally fought on the side of the rebels, but later integrated into the army at the later stages of the civil war. Interviewing these soldiers provided a rare glimpse into how certain developments such as foreign military interventions were perceived differently by the warring factions.

Non-military personnel were also targeted. I interviewed local academics, human rights lawyers, UN personnel, and a Western diplomat. Doing so made me aware of the concerns that were secondary to that of those expressed by the combatants.

Some of the ex-combatants that I interviewed were involved in major human rights violations themselves, making such inquiries an especially sensitive topic. This, combined with the semi-structured interview technique, meant that such violations were only expressed if the interviewee chose to disclose them willingly. This was alleviated greatly by interviewing civilians with the aforementioned qualifications.

Table 3.1: Quantitative studies on armed conflict duration ($n = 46$) as identified by the LSE Library keyword search

Study	Obs	Features	df	Start	End	Type	Model
Akcinaroglu, Radziszewski 2005	103	8	12.88	1946	1992	Interstate	CoxPH
Aliyev 2017	240	10	24.00	1991	2015	Civil War	CoxPH
Aydin, Regan 2011	1617	13	124.38	1945	1999	Civil War	CoxPH
Bagozzi 2016	2464	17	144.94	1945	2004	Civil War	CoxPH
Balcells, Kalyvas 2014	906	15	60.40	1944	2004	Civil War	Weibull
Balch-Lindsay, Enterline 2000	152	14	10.86	1820	1992	Civil War	CoxPH
Bennett & Stam 1996	169	22	7.68	1823	1990	Interstate	Weibull
Briffa 2014	44	5	8.80	1823	2003	Interstate	Logit
Buhaug, Gates & Lujala 2009	1412	12	117.67	1946	2003	Civil War	Weibull
Burgoon et al 2015	1378	16	86.12	1975	2000	Civil War	CoxPH
Caverley & Sechser 2017	615	22	27.95	1967	2003	Civil War	Weibull
Collier, Hoeffler & Soderbom 2004	732	20	36.60	1960	1999	Civil War	Exponential
Conrad et al 2018	586	20	29.30	1990	2009	Civil War	CoxPH
Cunningham & Lemke 2013	1586	12	132.17	1946	2008	Combined	CoxPH
Cunningham 2006	15932	10	1593.20	1946	2003	Civil War	Logit
Cunningham 2010	1223	15	81.53	1946	1998	Civil War	CoxPH
Cunningham, Gleditsch & Salehyan 2009	2426	19	127.68	1945	2003	Civil War	CoxPH
DeRouen, Sobek 2004	92	16	5.75	1944	1997	Civil War	Competing Risks
Escriba-Folch 2010	608	21	28.95	1960	1998	Civil War	Logit
Fearon 2004	128	6	21.33	1945	1999	Civil War	Weibull
Fukumoto 2015	2201	10	220.10	1946	2003	Civil War	Weibull
Hartzell 2009	105	12	8.75	1945	1999	Civil War	CoxPH
Kirschner 2010	68	15	4.53	1945	2004	Civil War	CoxPH
Koch 2009	588	11	53.45	1945	1992	Interstate	Weibull
Koch, Sullivan 2010	793	19	41.74	1960	2000	Major Power Intervention	Competing Risks
Krustev 2006	1450	9	161.11	1950	1992	Interstate	CoxPH
Langlois, Langlois 2009	55	9	6.11	1823	1990	Interstate	Weibull
Lyall 2010	307	9	34.11	1800	2006	Counterinsurgency	Weibull
Meernik, Brown 2007	871	13	67.00	1948	1995	US Interventions	CoxPH
Metternich 2011	1013	15	67.53	1946	2003	Civil War	CoxPH
Moore 2012	94	11	8.55	1946	2002	Civil War	CoxPH
Mukherjee 2014	116	16	7.25	1945	1999	Civil War	Weibull
Nilsson 2012	150	23	6.52	1823	1978	Interstate	Weibull
Ohmura 2017	2272	8	284.00	1946	2003	Civil War	CoxPH
Prorok 2016	21200	15	1413.33	1980	2011	Civil War	CoxPH
Regan 2002	13048	14	932.00	1944	1999	Civil War	Weibull
Regan, Aydin 2006	13243	12	1103.58	1945	1999	Civil War	Weibull
Shannon, Morey, Boehmke 2010	55048	14	3932.00	1950	2000	Interstate	Weibull
Shirkey 2012	34984	16	2186.50	1816	1997	Interstate	Weibull
Slantchev 2004	104	8	13.00	1816	1991	Interstate	Log-logistic
Stanley, Sawyer 2009	78	15	5.20	1816	1990	Interstate	Weibull
Thyne 2012	782	10	78.20	1975	2004	Civil War	Weibull
Thyne 2017	17319	7	2474.14	1950	2009	Civil War	CoxPH
Uzonyi & Wells 2016	2361	11	214.64	1945	2003	Civil War	CoxPH
Weisiger 2016	36322	12	3026.83	1823	2003	Interstate	CoxPH
Wucherpfennig et al 2012	1941	15	129.40	1946	2005	Civil War	CoxPH

Chapter 4

A Quantitative Assessment of Duration Studies

How sensitive are the empirical findings of the civil war duration studies? There is no equivalent study that can be compared to Hegre and Sambanis (2006), who focus on the common predictors of civil war onset drawn from a pool of 88 variables taken from the literature. Sensitivity analysis allows one to summarise the literature in a succinct way by separating the more robust findings from the more idiosyncratic ones. In order to find out which predictors are more persistent in the literature, I will begin by establishing a baseline of common determinants of armed conflict duration.

The value of a model lies in the quality of its predictions (Miller, 2014). As such, unlike the approach taken by Hegre and Sambanis (2006), who uses statistical significance as their criterion, I instead focus on predictive accuracy. As an initial step, I first employ various empirical strategies to identify which covariates are better at prediction than others in the literature at large. To do so, I replicate a representative sample of published armed conflict duration studies using binary-time-series-cross-section (BTSCS) data. This data format allows for

time-varying covariates (Beck, 2008) and is widely adopted in conflict research,[1] which makes it an apt choice for the task at hand.

However, some BTSCS studies could not be replicated due to several reasons: i) not employing traditional survival analysis, e.g. Briffa (2014), Fukumoto (2015); ii) replication data unavailability (Conybeare, 1992; Bennett and Stam III, 1998; Balch-Lindsay and Enterline, 2000; Goemans, 2000; Langlois and Langlois, 2009; Kirschner, 2010; Aydin and Regan, 2012; Mukherjee, 2014; Prorok, 2018); and iii); replication script unavailability (Vuchinich and Teachman, 1993; Aliyev, 2017). Further, non-BTSCS duration studies such as Fearon (2004), Slantchev (2004), and Moore (2012) are also excluded, as well as most non-yearly BTSCS studies—in which variable values vary on month/week/day intervals—such as Shannon et al. (2010), Lyall (2010), Shirkey (2012), and Weisiger (2016).

I argue that the unavailability of some studies should not affect the validity of results for two reasons. First, the empirical expectations of the theoretical model are formulated in variable importance terms. This is in contrast to mainstream hypothesis testing, in which the practitioner usually posits a directed correlation against a null effect. However, as I aim to test whether there are common predictors of conflict duration, there is no precise directionality embedded with the theory. Second, the number of possible predictors is high enough to allow for robust variables to come through. Meaning, it is unlikely that removing one study and including another will severely shift the results, given the wide range of operationalisations—16 studies comprised of 232 independent variables are disaggregated into three components—captured by the whole replication procedure. Moreover, if multiple studies using different datasets (which might feature different operationalisations of similar concepts) identify a common variable, it only strengthens the notion that the results are robust and not idiosyncratic in nature (Eck, 2005).

For the next step, I stratify all independent variables specified by the authors into

[1] Consult Beck (2001) for a review of BTSCS studies in political science.

two groups: predictive and not predictive.[2] This is a fitting metric for a non-NHST study; predictive accuracy in machine learning can be thought of as statistical significance in traditional statistics in terms of explanatory impact. However, it should be noted that statistically significant variables are not necessarily good predictors (Lo et al., 2015).

Variable importance is one application for filtering out predictive covariates from noisy predictors. Feature selection is another common procedure that can greatly reduce the dimension of a dataset and identify relevant predictors (Kuhn and Johnson, 2013). I employ four different approaches when it comes to feature selection: recursive feature elimination, genetic algorithm, simulated annealing, and variable importance after fitting elastic net and random forest models. The entire replication enterprise fits about 300,000 models. Overall, the covariates that are selected more than others across studies—including both interstate and civil wars—will inform the model specification of the next chapter, in which I fit various machine learning ensembles and utilise deep learning on a combined dataset constructed by Cunningham and Lemke (2013).

The rest of the chapter is structured as follows. First, I briefly explore the BTSCS studies on conflict duration. Next, I analyse the replication studies to identify common methodological pitfalls and select appropriate pre-processing procedures. Then, I employ three feature selection algorithms and fit predictive models. Finally, I present in-sample performance metrics and out-sample predictive accuracy of the replication studies as the main empirical contribution of this chapter.

[2]The precise definition of the difference between predictive and not predictive depends on the algorithm at hand and explained accordingly during model fitting.

Table 4.1: Example non-BTSCS data subset

country	casename	waryrs	lpopl1	ef	_d
COLOMBIA	FARC, ELN, etc	1963-	9.730026	0.656000	0
SIERRA LEONE	RUF, AFRC, etc.	1991-	8.327484	0.763997	0
TURKEY	PKK	1984-99	10.776390	0.298504	1
AFGHANISTAN	v. Taliban	1992-	9.706864	0.750797	0

4.1 BTSCS Studies on Conflict Duration

4.1.1 Brief Review

Modelling techniques specific to duration entered mainstream conflict literature in late 90s (Beck et al., 1998). Previously, scholars either fit Ordinary Least Squares (OLS) to capture duration as an outcome given its continuous nature (De Mesquita, 1978) or simply fit curves without any independent variables (Morrison and Schmittlein, 1980). These approaches were found to be statistically inappropriate in the former case (as duration is always positive by construction) and uninformative in the latter (Bennett and Stam, 1996). Thus, the introduction of appropriate duration models to political science immensely aided the study of war longevity.

These earlier models, however, mostly featured single spells per event; whole conflicts would occupy only one row regardless of whether they lasted a month or several decades owing to data limitations at the time. Table 4.1 demonstrates this type of data structure using seminal work by Fearon (2004). The variable `waryrs` denote the time-frame of the conflict, and covariates such as ethnic fractionalisation `ef` and the binary outcome `_d` are not allowed to vary within that time-frame. Such covariates would either contain onset, termination, or averaged-over-time values; in other words, they are time-invariant.

Scholars quickly recognised the serious shortcomings of this approach, and opted for more dynamic models that include time-varying covariates (Beck et al., 1998).

Table 4.2: Example BTSCS data subset

sidea	sideb	year	parallelconflict	lgdppcl	territorial	_d
China	Peoples Liberation Army	1946	0	5.451038	0	0
China	Peoples Liberation Army	1947	1	5.420535	0	0
China	Peoples Liberation Army	1948	0	5.513429	0	0
China	Peoples Liberation Army	1949	0	5.545178	0	1

Table 4.2 demonstrates the concept of BTSCS data using a splice of Buhaug et al. (2009) study on conflict geography.

In the example above, covariates `parallelconflict` and `lgdppcl`—log of GDP per capita—are time-varying, as well as the outcome variable `_d`. Territorial conflict dummy `territorial` however is not. The values are updated every time unit, in this case `year`. It should be noted that even when a variable is time-varying with regards to the whole dataset, it could still be a constant within a cluster. For example, if the degree to which a terrain is deemed 'densely forested' is measured by some function of forest coverage in a specified area, its value may not vary during the life span of certain conflicts.

With that said, the effect of the added dynamism obtained by inclusion of time-varying covariates cannot be overstated (Beck, 2008). When covariates are allowed to vary within clustered observations (e.g. conflicts or dyads), models have access to a larger amount of possible sources of information that they can use to explain the variation in outcome. Further, time-varying values provide a more robust empirical challenge for the proposed theories under scrutiny, as the assumption of permanence is less defensible than allowing for variation. Finally, time-varying covariates allow for comparing different strata of variables within and between themselves. Case in point, the time-varying nature of the common conflict duration predictors (i.e. population, GDP p.c., troop size) allows me to test the effect of the same covariates on both types of war. It opens up the possibility that certain interstate wars, based on the set of values of their covariates, could be more similar to certain civil wars than they are to other

interstate wars that are characterised by a vastly different set of covariates (and vice-versa). In essence, the testing of this general hypothesis is the main empirical goal of this project.

4.1.2 Replication Procedure

The replication procedure is as follows. First, I identify peer-reviewed studies on armed conflict duration by conducting a curated search.[3] After filtering for quantitative studies with replication materials, this resulted in 46 studies in total covering the publication period from 1996 to 2018.[4] These studies cover a wide range of topics studied in conflict research. To maintain conceptual and empirical consistency, I further filter the initial batch of studies.

Quantitatively, more than half of the studies did not meet at least one of the necessary criteria for inclusion: i) covariates are time-invariant/single spell conflicts (Fearon, 2004), ii) contain only minor additions to an already-included study e.g. Stanley and Sawyer (2009), and iii) being in a format that is difficult to streamline (i.e. using daily data) such as DeRouen Jr. and Sobek (2004), Krustev (2006), Meernik and Brown (2007), Sullivan (2008b), Sullivan (2008a), Koch (2009), and Metternich (2011). These studies are accordingly dropped and not replicated using predictive modelling.

Qualitatively, I select for studies that focus on explaining the duration of violent conflict. Some papers (Hartzell, 2009) study peace duration as opposed to conflict. Others, such as Briffa (2014), focus on drawing parallels from theories of animal contestation to test the empirical relationship between material capabilities (i.e. disaggregated CINC components) and war duration. I leave out such studies and only include ones that aim to explain conflict duration making

[3]Precise search parameters are specified in section 3.2.1 of the Research Design Chapter.

[4]Full list of considered studies can be found in Table 3.1 in section 3.2.1 of the Research Design chapter.

a novel empirical contribution.[5]

After eliminating such papers, I am left with 16 BTSCS studies on conflict duration. I argue that 16 is a large enough number for a sensitivity analysis such as this, drawing on the sample size of similar studies. For instance, Carroll and Kenkel (2016) uses 18 replication studies to determine how well their new measure—called *dispute outcome expectations*—fares against the standard CINC measure, which they intend to replace.

Except for Cunningham (2006) and Escribà-Folch (2010), who use logistic regression, the remaining 14 studies utilise either parametric (Weibull/Accelerated Failure Time) or the semi-parametric (Cox Proportional-Hazards) survival models. Machine learning in general does really well on regression and classification problems, but less so in survival analysis (Zupan et al., 2000), one big drawback being the lack of support for time-varying covariates (Cruz and Wishart, 2006). Logistic regression can be used on survival (time-to-event) data if the inherent time-dependency is adequately controlled for. A common way of doing so is to introduce cubic splines in the form of $t + t^2 + t^3$, where t denotes the time to event (Beck et al., 1998; Carter and Signorino, 2010). Thus, I transform the survival models into logistic regression with added splines to allow for classification algorithms to be utilised.

4.2 Exploratory Data Analysis

Before training predictive models, I briefly explore the included replication studies. This serves two purposes. First, the included papers vary greatly in terms of their selection of predictors. Doing a quantitative assessment of the literature will inform the reader about the foci of the papers under scrutiny. Second, akin to traditional statistical models, machine learning algorithms perform better

[5] i.e. introducing a new variable, either in the form of measurement or operationalisation, or alternatively, data merging (bringing together novel covariates for the first time).

Table 4.3: Descriptive statistics of the replication studies

Study	Obs	Features	_df_	Start	End	Class Bal.	Type	Model
Bagozzi 2016	2464	17	144.94	1945	2004	0.86	Civil War	CoxPH
Burgoon et al 2015	1378	16	86.12	1975	2000	0.96	Civil War	CoxPH
Buhaug, Gates & Lujala 2009	1412	12	117.67	1946	2003	0.86	Civil War	Weibull
Bennett & Stam 1996	169	22	7.68	1823	1990	0.54	Interstate	Weibull
Collier, Hoeffler & Soderbom 2004	732	20	36.60	1960	1999	0.93	Civil War	Exponential
Cunningham 2006	15932	10	1593.20	1946	2003	0.99	Civil War	Logit
Cunningham 2010	1223	15	81.53	1946	1998	0.86	Civil War	CoxPH
Conrad et al 2018	586	20	29.30	1990	2009	0.72	Civil War	CoxPH
Cunningham, Gleditsch & Salehyan 2009	2426	19	127.68	1945	2003	0.85	Civil War	CoxPH
Cunningham & Lemke 2013	1586	12	132.17	1946	2008	0.85	Combined	CoxPH
Caverley & Sechser 2017	615	22	27.95	1967	2003	0.83	Civil War	Weibull
Escriba-Folch 2010	608	21	28.95	1960	1998	0.93	Civil War	Logit
Nilsson 2012	150	23	6.52	1823	1978	0.51	Interstate	Weibull
Thyne 2012	782	10	78.20	1975	2004	0.88	Civil War	Weibull
Uzonyi & Wells 2016	2361	11	214.64	1945	2003	0.85	Civil War	CoxPH
Wucherpfennig et al 2012	1941	15	129.40	1946	2005	0.86	Civil War	CoxPH

when certain conditions are met regarding the underlying data. Diagnosing these features helps me choose what type of pre-processing is required before fitting the models.

4.2.1 Summary Statistics

Several trends are readily visible in table 4.3. First, only three studies out of 16 include interstate wars; Bennett and Stam (1996) and Nilsson (2012) exclusively, and Cunningham and Lemke (2013) in conjunction with civil wars. The following studies only consider civil war duration: Bagozzi (2016), Burgoon et al. (2015), Buhaug et al. (2009), Collier et al. (2004b), Cunningham (2006), Cunningham (2010), Conrad et al. (2018), Cunningham et al. (2009), Caverley and Sechser (2017), Escribà-Folch (2010), Thyne (2012), Uzonyi and Wells (2016), and Wucherpfennig et al. (2012).

Studies on interstate wars also go back in time significantly more (1823 is the starting year for both studies) than civil war datasets, which start their coverage post-1945. It is also worth noting that interstate war duration studies include more variables than their civil war counterparts even though they

have significantly fewer observations—the median degrees of freedom *(df)*[6] for interstate studies is 7.1 whereas civil war studies enjoy 86.9 degrees. Given these figures, these studies are at risk of being what is termed 'garbage bin' regressions by Achen (2005). Citing the famous Anscombe quartet (Anscombe, 1973) demonstration,[7] Achen (2005) further posits that regression model findings by themselves are useless without "either a formal model or detailed data analysis". Additional robustness and sensitivity tests on top the main models are thus highly encouraged to establish the validity of the findings in such cases (Ray, 2003, 2005).

Finally, in terms of outcome classes, which are coded as 0/1 (continuation/termination) consistently for each row (i.e. year) across all included replication studies, class balance reflects the summary statistics of duration across war types. The column `Class Bal.` in Table 4.3 denotes the prevalence of the dominant class; the percentage of the observations having the dominant class label (i.e. *no event*). Interstate war studies are well-balanced (0.525). This contrasts the severe class imbalance inherent in civil war duration studies: on average, 87.5% of the observations in a dataset are non-events. If not addressed, class-imbalance might lead to 'lazy' algorithms that exclusively predict the dominant class. For example, predicting all zeros (no event) all across the board would lead to an accuracy of around 85% for most of these studies. Hence, a more nuanced performance metric is required to gauge the true informative value of the models, such as the Receiver Operator Characteristic (ROC). Note that ROC is not immune to falling prey to lazy models; however it is more resistant to them in comparison to naive accuracy.

Class imbalance can also be rectified using zero-inflated models (Lambert, 1992); however, to preserve consistency across multiple machine learning algorithms,

[6]The ratio of observations to the total number of variables; the rule of thumb being 30 degrees of freedom to satisfy common statistical assumptions.

[7]Commonly used to illustrate the importance of data visualisation, Anscombe's quartet comprises of four datasets that have nearly identical descriptive statistics—mean, sample variance, correlation, linear regression line, and R^2—that appear vastly different when graphed.

class imbalance is dealt with sub-sampling. At the most basic level, the data can be *up-* or *down-* sampled.[8] The former uses bootstrapping to ensure the least frequent class has as many observations as the most frequent class; in contrast, the latter reduces the number of most frequent class observations to match the least frequent class. In cases of severe imbalance, however, down-sampling will lead to immense data loss: all civil war duration studies would lose about 70% of their observations. For this reason, I utilise up-sampling in both feature selection and model fit stages. Anecdotally, up-sampling can lead to slightly lower out-of-sample ROC values compared to down-sampling (Kuhn and Johnson, 2013), however, the magnitude of potential data loss is too great to overlook.

4.2.2 Further Diagnostics

Even though the ill-effects of multicollinearity are felt more in traditional statistical approaches that estimate coefficient sizes, correlated variables can also affect the predictive performance of classification algorithms (Toloşi and Lengauer, 2011). Collinear predictors can lead to over-fitting as they contain similar information about the outcome (Hill and Judge, 1987). Thus, I start out with a simple correlation analysis of all independent variables in each study.[9] Figure 8.1 in the appendix displays the correlation matrices of all 16 studies. Even though there is no general pattern of cluster blocks, there are multiple variables in most of the datasets that are strongly correlated with each other in one direction or the other. This can be appreciated more when we move away from the big picture and zoom in on a single paper.

A useful study is Cunningham and Lemke (2013), as they combine both types of armed conflict in their dataset. Figure 4.1 highlights some basic correlations. We see that the log of total population and total number of troops are positively

[8] There are also hybrid methods combining both approaches such as `SMOTE` and `ROSE`. However, as both of them generate a large number of bootstrapped observations (much more than regular up-sampling), I do not consider them here.

[9] Outcome and cubic splines are left out of the process.

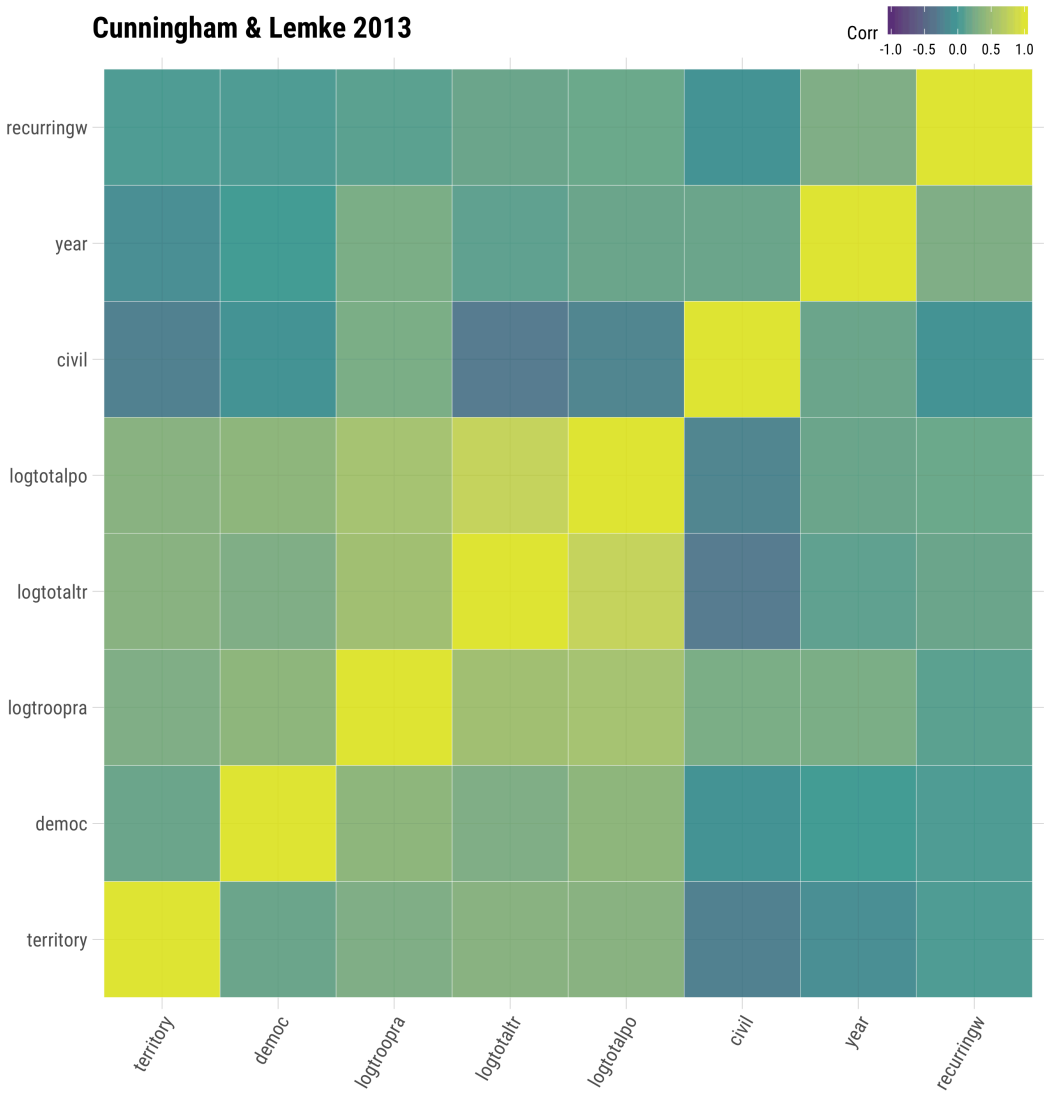

Figure 4.1: Correlation analysis of Cunningham and Lemke 2013

correlated with each other, while showing strong negative correlation with the civil war dummy. To prevent unwanted effects of collinearity, I pre-process the data by dropping the redundant correlated variables during model fitting.

Finally, I employ Principal Component Analysis (PCA) to determine how many components are needed for each dataset to explain 95% of the variation. PCA is a dimensionality reduction technique that relies on identifying orthogonal predictors (Wold et al., 1987). Figure 4.2 visualises the how a multivariate Gaussian distribution can be reduced to two components. If there is meaningful

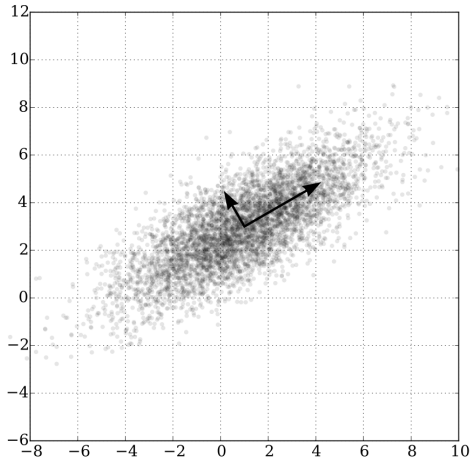

Figure 4.2: Example of a principal component analysis of a multivariate Gaussian distribution

variation across the studies in terms of how many principal components are required, this can help differentiate parsimonious models from over-specified ones.

Alas, the results fail to reveal any significant insights; for all 16 replicated studies only require two components to capture 95% of the variation.[10] One interpretation of this result is that virtually all variation in the outcome can be reduced to two dimensions, regardless of the original number of the independent variables. Although this makes comparison meaningless across replication studies, given the result holds across all, it could be signal of model over-specification: for example, Thyne (2012) has 10 predictors, while Nilsson (2012) has 23; however in both cases only two components are necessary. One takeaway is that all studies under scrutiny can be significantly reduced on the right-hand side of the equation: such a high reduction ratio obtained by PCA suggests the majority of the independent variables are linearly correlated with each other.

In sum, given the lack of variation across studies, coupled with the fact that PCA makes findings less interpretable, I choose not to pre-process the datasets using

[10]The results still hold when the threshold is increased to 99%.

PCA. Instead, I achieve dimensionality reduction using feature selection, which is explained in the next section.

4.3 Predictive Modelling with Feature Selection

Prior to model fitting, I utilise three wrapper algorithms for feature selection. In machine learning, feature selection—also known as variable selection—is the process of selecting a subset of relevant features (i.e. predictors) to construct models. Feature selection has many benefits (James et al., 2013), of which all are highly desired: i) model parsimony, so that they are easier to fit and explain; ii) greater generalisability, by reducing over-fitting; iii) avoiding the curse of dimensionality; and iv) computational efficiency.

I employ feature selection to reduce the models specified by the authors, not the whole dataset. To give an example, Bennett and Stam (1996) fits the following model to their data consisting of only 169 observations (cubic splines excluded):

Outcome ~ Strategy: OADM + Strategy: OADA + Strategy: OADP + Strategy: OPDA + Terrain + Terrain x Strategy + Balance of Forces + Total Military Personnel + Total Population + Population Ratio + Quality Ratio + Surprise + Salience + Repression + Democracy + Previous Disputes + Number of States

It is likely that their model is over-specified given the observations-to-predictors (n/p) ratio. Still, they only utilise 17 variables out of 37 included in the complete dataset. Meaning, there has already been a feature selection—the model specification. However, in light of the PCA results, expanding feature selection to the whole dataset (e.g. outside of the authors' specified model) is problematic for several reasons.

First, mere inclusion does not necessitate meaningful contribution; most datasets are built on others and contain multiple auxiliary variables (e.g. version,

backwards-compatibility codes etc.). Second, more common in older studies given how most statistical software worked back then, many constructed variables (i.e. natural logs and other transformations, interaction terms, normalisation etc.) are included next to their raw counterparts. Applying feature selection to such datasets will drop a vast majority of the features; however, it would be difficult to assess whether this can be traced to bad predictive performance per se or indicative of the uninformativeness caused by the redundant covariates. Third, the computational costs of increasing the number of variables—sometimes over hundred—in a dataset often exponential (Kuhn and Johnson, 2013) while not guaranteeing an improvement in predictive accuracy. Thus, I limit feature selection to the original model specifications and aim to further parsimonise the predictive models.

The following algorithms try to get at the best subset of covariates in terms of predictive accuracy. For all algorithms, both the internal and external performance measures are set to ROC maximisation. Selection is based on external ROC; as maximising internal ROC is prone to over-fitting. I use bootstrapped cross-validation (repeated 50 times) for both performance measures. I fit logistic regression models to leverage its computational efficiency. I do not fit a new model with the selected features as doing so will lead to selection bias (Friedman et al., 2001). Cubic splines are added to the model specifications to control for time dependency; however they are not reported if they are selected as predictive features. Next, I briefly introduce the algorithms and provide their respective pseudo-codes taken from Kuhn and Johnson (2013). Top predictors identified by the feature selection algorithms will not be covered here but in the upcoming findings section after model fitting.

Algorithm 1: Recursive feature elimination incorporating resampling

```
for Each Resampling Iteration do
    Partition data into training and test/hold-back set via resampling
    Tune/train the model on the training set using all predictors
    Predict the held-back samples
    Calculate variable importance or rankings
    for Each subset size S_i, i = 1 ... S do
        Keep the S_i most important variables
        [Optional] Pre-process the data
        Tune/train the model on the training set using S_i predictors
        Predict the held-back samples
        [Optional] Recalculate the rankings for each predictor
    end
end
Calculate the performance profile over the S_i using the held-back samples
Determine the appropriate number of predictors
Estimate the final list of predictors to keep in the final model
Fit the final model based on the optimal S_i using the original training set
```

Figure 4.3: Recursive feature elimination algorithm

4.3.1 Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a simple feature selection method that focuses on the size of the variable subsets. It is considered as a greedy algorithm as it only considers each subset once and never goes back again. Thus, it is susceptible to getting stuck in local maxima (Kuhn and Johnson, 2013). However, it is relatively fast to implement, making it an adequate starting choice as a benchmark.

Figure 4.3 provides the step-by-step guide of the RFE algorithm. Given the relatively lower number of covariates in the replication studies, I try all possible subset sizes; e.g. Nilsson (2012) has 23 predictors, so the RFE algorithm fits subset sizes of $1, 2, 3, \ldots, 22, 23$.

Table 4.4 shows the results of the RFE procedure across studies. The reduction factor achieved by the RFE is also included. Around 50% of the studies have a

Table 4.4: Recursive feature elimination results

Study	Covariates	Selected	Reduction %	ROC
Cunningham & Lemke 2013	10	8	20.0	0.819
Cunningham 2006	8	2	75.0	0.808
Bagozzi 2016	15	9	40.0	0.797
Nilsson 2012	21	21	0.0	0.742
Buhaug, Gates & Lujala 2009	10	10	0.0	0.741
Cunningham 2010	13	13	0.0	0.741
Wucherpfennig et al 2012	13	10	23.1	0.733
Bennett & Stam 1996	20	18	10.0	0.723
Cunningham, Gleditsch & Salehyan 2009	17	16	5.9	0.711
Uzonyi & Wells 2016	9	9	0.0	0.703
Caverley & Sechser 2017	20	18	10.0	0.684
Burgoon et al 2015	14	14	0.0	0.658
Thyne 2012	7	5	28.6	0.655
Conrad et al 2018	18	13	27.8	0.648
Collier, Hoeffler & Soderbom 2004	18	17	5.6	0.640
Escriba-Folch 2010	19	14	26.3	0.619

reduction rate less than or equal to 10%. The magnitude of the reduction ratio seems to go hand-in-hand with higher ROC values; the top three studies in terms of ROC (Cunningham and Lemke, 2013; Cunningham, 2006; and Bagozzi, 2016) have reduction rates of 20%, 75%, and 40%, respectively.

4.3.2 Genetic Algorithms

Genetic algorithms (GAs) simulate Darwinian forces of natural selection to locate optimal solutions to a function. Mimicking the original theory (Darwin, 1859), the underlying logic is that in an iterative process, less suited individuals (specific models) to the environment (prediction problem at hand) are less likely to survive (achieve high predictive accuracy) and thus, less likely to reproduce (selected for the next iteration of predictions).

More formally, initial sets of candidate solutions are created with corresponding fitness values. These are known as the *population*, whereas each solution is called an individual. These individuals with the highest fitness values are randomly

Table 4.5: Genetic algorithm results

Study	Covariates	Selected (Avg.)	Pop. Size	Elitism	ROC
Cunningham & Lemke 2013	10	7.5	4	1	0.818
Cunningham 2006	8	6.0	3	1	0.804
Bagozzi 2016	15	11.6	5	1	0.793
Nilsson 2012	21	15.2	7	2	0.747
Buhaug, Gates & Lujala 2009	10	8.5	4	1	0.735
Cunningham 2010	13	10.8	5	1	0.734
Bennett & Stam 1996	20	14.5	7	2	0.725
Wucherpfennig et al 2012	13	11.0	5	1	0.722
Cunningham, Gleditsch & Salehyan 2009	17	14.4	6	1	0.706
Uzonyi & Wells 2016	9	8.1	3	1	0.697
Burgoon et al 2015	14	10.5	5	1	0.665
Caverley & Sechser 2017	20	14.7	7	2	0.662
Conrad et al 2018	18	12.8	6	2	0.643
Thyne 2012	7	5.1	3	0	0.643
Escriba-Folch 2010	19	13.6	7	2	0.624
Collier, Hoeffler & Soderbom 2004	18	13.4	6	2	0.613

combined to procreate the next generation of solutions (Mitchell, 1998). During this process, the individual can undergo cross-over with a certain probability, as well as being subject to random mutations. This process is repeated many times, (theoretically) leading to better and better solutions.

The implementation of the GA is explained in Figure 4.4. For feature selection, the individuals are subsets of predictors that are encoded as binary based on whether they are included or not. The fitness values are the measure of model performance; in this case ROC. Similar to how hereditary characteristics become more pronounced as they are passed on generation after generation (assuming no or minimal cross-over), GAs can be aggressive during internal model fitting and are prone to over-fitting. Thus, to prevent this from happening, I set the number of generations to a relatively low number (20).[11] The cross-over probability is held at 0.8, population size is set to about one-third of the total number of covariates, and elitism (i.e. number of subsets to survive at each generation) is allowed with a probability of one-tenth.

GA findings are given in Table 4.5. Mimicking RFE findings, the same three

[11]Other numbers are also tried from 5 to 50; the selection of 20 generations is representative of the optimal accuracy/performance ratio.

Algorithm 2: A genetic algorithm for feature selection

```
Define the stopping criteria, number of children for each generation
  (GenSize), and probability of mutation (p_m)
Generate an initial random set of m binary chromosomes, each of length p
repeat
    for each chromosome do
        Tune and train a model and compute each chromosome's fitness
    end
    for reproduction k = 1 ... GenSize/2 do
        Select two chromosomes based on the fitness criterion
        Crossover: Randomly select a loci and exchange each
          chromosome's genes beyond the loci
        Mutation: Randomly change binary values of each gene in each
          new child chromosome with probability p_m
    end
until stopping criteria is met;
```

Figure 4.4: Genetic algorithm

studies come up on top in terms of ROC rates. Smaller population sizes (<6) tend to score a bit higher on average. As GA resamples internally and externally many times (20 generations with 50-times repeated bootstrapping), the selected covariate sizes are averages over resamples. Unlike the RFE rates, however, we do not find a lot of feature reduction taking place.

4.3.3 Simulated Annealing

The Simulated Annealing (SA) algorithm mimics the process of annealing in metallurgy. Annealing involves utilising the temperature to alter a material's physical properties. This is made possible by the changes in its internal structure: as cooling occurs, the new structure becomes fixed; which causes the metal to retain its newly-obtained properties. Figure 4.5 demonstrates the SA algorithm.

In simulated annealing, the temperature variable is utilised to simulate this heating process. It is initially set high, and then allowed to cool down with each passing iteration of the algorithm. When the temperature variable is high, the

Algorithm 3: Simulated annealing for feature selection. E is a measure of performance where small values are best and T is a temperature value that changes over iterations

Generate an initial random subset of predictors
for *iterations* $i = 1 \ldots\ t$ **do**
    Randomly perturb the current best predictor set
    [*Optional*] Pre-process the data
    Tune/train the model using this predictor set
    Calculate model performance (E_i)
    if $E_i < E_{best}$ **then**
        Accept current predictor set as best
        Set $E_{best} = E_i$
    else
        Calculate the probability of accepting the current predictor set

$$p_i^a = exp[(E_{best} - E_i)/T]$$

        Generate a random number U between $[0, 1]$
        if $p_i^a \leq U$ **then**
            Accept current predictor set as best
            Set $E_{best} = E_i$
        else
            Keep current best predictor set
        end
    end
end
Determine the predictor set associated with the smallest E_i across all iterations
Finalise the model with this predictor set

Figure 4.5: Simulated annealing algorithm

Table 4.6: Simulated annealing results

Study	Covariates	Selected	Reduction %	ROC
Cunningham & Lemke 2013	10	6	40.0	0.830
Bagozzi 2016	15	8	46.7	0.797
Cunningham 2006	8	4	50.0	0.794
Buhaug, Gates & Lujala 2009	10	8	20.0	0.728
Cunningham 2010	13	7	46.2	0.718
Cunningham, Gleditsch & Salehyan 2009	17	7	58.8	0.708
Nilsson 2012	21	11	47.6	0.706
Bennett & Stam 1996	20	7	65.0	0.705
Uzonyi & Wells 2016	9	6	33.3	0.695
Burgoon et al 2015	14	9	35.7	0.690
Wucherpfennig et al 2012	13	5	61.5	0.663
Thyne 2012	7	3	57.1	0.653
Collier, Hoeffler & Soderbom 2004	18	13	27.8	0.648
Caverley & Sechser 2017	20	9	55.0	0.645
Conrad et al 2018	18	11	38.9	0.599
Escriba-Folch 2010	19	8	57.9	0.548

algorithm can accept solutions that perform worse than the current solution. This is where SA diverges from RFE and GAs; as this gives the SA algorithm the flexibility to get out of local maxima located in earlier iterations. As the temperature cools down, it gets less and less likely to consider jumping out of such local maxima, allowing the algorithm to stabilise towards the end of its run. The process of gradual cooling tied directly to the flexibility of the algorithm is what makes SA remarkably effective at finding a close-enough optimum solution when dealing with problems containing multiple local maxima. In similar settings, greedy algorithms like RFE and GAs will be stuck with suboptimal solutions.

Table 4.6 displays the results of the SA feature selection. Again, the same studies occupy the top three spots in ROC calculations. However, we see that SA is highly successful at dimensionality reduction; the minimum reduction rate is 20%, the mean reduction rate is 46.34%, with seven studies being reduced by more than 50%.

4.4 Model Training

The feature selection algorithms are fit using logistic regression. Logistic regression is a commonly used classification algorithm that is both computationally-efficient and has no hyper-parameters that need tuning. However, as the focus of this chapter is find out which covariates are better at prediction than others, there is utility to be gained from switching to tune-able algorithms. Therefore, I select two suitable algorithms for the task: the elastic net, an extension of the generalised linear model that has a built-in feature selection tool; and random forest, an ensemble decision-tree algorithm that specialises in uncovering non-linear interactions between the predictors.

4.4.1 Elastic Net

The so-called *elastic net* is a generalised linear model that combines two common types of regularisation; $L1$ and $L2$. The $L1$ regularisation, commonly known as the LASSO—Least Absolute Shrinkage and Selector Operator, has the penalty form

$$|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|.$$

LASSO regression shrinks some coefficients to zero depending on their contribution; meaning it does feature selection (as multiplication by zero drops out the term).

In contrast, the $L2$ regularisation is called the Ridge regression. Ridge regression penalises large coefficients, which can have a disproportionate influence on the outcome. However, this makes them less interpretable than the LASSO. The elastic net combines the two adding a quadratic component to the penalty, which defaults to $L2$ when used by itself:

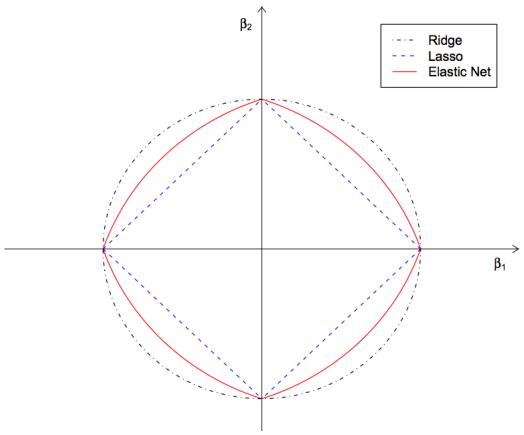

Figure 4.6: Elastic net vs. LASSO and ridge regression

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1)$$

Figure 4.6 demonstrates the relationship between elastic net and both types of regularisations. In `R`, elastic net can be fit using the `glmnet` package (Friedman et al., 2010). It has two hyper-parameters; *alpha* and *lambda*. The *alpha* parameter can take any value between $[0, 1]$ and denotes the type of penalisation: 0 for Ridge and 1 for LASSO. Any other value of *alpha* results in an elastic net, which is a hybrid of the two approaches. On the other hand, *lambda* is a continuous variable $(0, 1]$ and control the magnitude of the penalty. These can be supplied as value-pairs to the algorithm; 20 equally-spaced values of *lambda* are tried with *alpha* values $0, 0.2, 0.4, 0.6, 0.8$ and 1. Similar to the feature selection algorithms, repeated bootstrapped cross-validation (50) is used. In addition, as identified in the exploratory data analysis, the following pre-processing procedures are applied: near-zero variance and correlated variables are dropped; all remaining covariates are centred and scaled; and the least-occurring class label is up-sampled to match the frequency of the dominant class label. The aforementioned pre-processing steps help optimise the data for the machine learning algorithms as prescribed by Kuhn and Johnson (2013).

###Random Forest

Random forest is one of the most popular ensemble learners. It is also one of the most accurate machine learning algorithms (Caruana and Niculescu-Mizil, 2006; Fernández-Delgado et al., 2014). It grows many smaller and weak decision-trees into a strong, aggregated learner (Breiman, 2001). Random forests use the bagging technique, which is shown to reduce variance (Breiman, 1996). The randomness in the name refers to the process of randomisation that occurs when the algorithm selects variables to split on. Ensemble models tend to perform better when the underlying features are uncorrelated. The standard implementation of the bagging procedure produces highly-correlated trees and common features are shared widely between the trees. By randomising which covariates are available to each singular tree, the random forest algorithm grows less correlated trees. This type of randomisation also reduces the computation time required to train the forest.

The `ranger` package (Wright and Ziegler, 2017) in `R` ports a fast implementation of the original random forest algorithm (Breiman, 2001) written in C++. There are three hyper-parameters that can be tuned: `split rule`, `mtry`, and `min.node.size`. `split rule` determines the procedure used for splitting the trees; for classification, the allowed rules are `gini` and `extratrees`. `mtry` determines the number of covariates to possibly split at at each node. Finally, `min.node.size` sets the minimum allowed node size. Similar to the elastic net hyper-parameters, I supply various value-pairs consisting of `split rule` and `mtry` while holding the `min.node.size` at its default (one).

4.4.2 Variable Importance

Tables 4.7 and 4.8 display the best (selected) hyper-parameter tunings and their associated ROC values. Each selected model fit[12] on a replication study ranks the

[12]The algorithm fits many models using the value-pair combinations; *alpha* and *lambda* for the elastic net and `split rule` and `mtry` for the random forest. However, only the 'best' model—the one that has the highest external ROC—is selected for computing the variable importance.

Table 4.7: Elastic net selected hyper-parameters and ROC

Study	ROC	Reg. Type	Penalty	Top Predictors
Cunningham & Lemke 2013	0.812	Elastic Net	0.6316	
Bagozzi 2016	0.799	Elastic Net	0.0001	Territory + Ethnic Frac. + Democracy
Cunningham 2006	0.792	Elastic Net	0.0001	Coup d'etat
Bennett & Stam 1996	0.773	Elastic Net	0.0001	Strategy: OPDA
Cunningham 2010	0.750	Elastic Net	0.0001	Independent Intervention
Buhaug, Gates & Lujala 2009	0.741	Elastic Net	0.0001	Conflict at Border + Border x Distance + Democracy
Nilsson 2012	0.727	Elastic Net	0.0001	Strategy: OADP +Strategy: OPDA + Terrain
Wucherpfennig et al 2012	0.716	Elastic Net	0.0001	Territorial Control + Central Command + Democracy
Uzonyi & Wells 2016	0.710	Elastic Net	0.0001	Institutional Constraints
Cunningham, Gleditsch & Salehyan 2009	0.707	Ridge	1.0000	Coup d'etat + Fighting Capacity + Arms Procurement
Thyne 2012	0.677	Ridge	0.9474	Fight for Gov't + Lenient Veto + Coup d'etat
Burgoon et al 2015	0.675	Ridge	1.0000	UN PK + Strong Parity + Democracy
Caverley & Sechser 2017	0.662	LASSO	0.0001	Cold War + Natural Resources + Ground Mechanisation
Conrad et al 2018	0.656	Elastic Net	0.2632	Extortion + Contraband
Escriba-Folch 2010	0.632	LASSO	0.0527	Sons of the Soil
Collier, Hoeffler & Soderbom 2004	0.581	Elastic Net	0.0001	Primary Commodity Exports + Change in Commodity Price Index

covariates based on their contribution to predictive accuracy.

In Table 4.7, we see that the hybrid elastic net dominates the type of regression selection based on ROC maximisation. Ridge regression was selected three times, and the LASSO only two times. The most common penalty coefficient (*lambda*) is the smallest possible option (0.0001). Moving onto Table 4.8, on average, the gini impurity measure outperforms the extremely randomised trees in the `split rule` column for the random forest. The second hyper-parameter, `mtry` is kept to two for almost half of the studies.

Finally, the top three predictors excluding the time splines[13] are reported next to each study for both algorithms. Interestingly, even though Cunningham and Lemke (2013) has the highest ROC value in both model fits, only the time variables $t + t^2$ are selected as predictive. Thus, no variable from the original model specification is classified as a good predictor. This is striking, as the civil war

[13]The *war months* variable found in Cunningham (2006) is left in as it is part of the original model specification and not a later add-in.

Table 4.8: Random forest selected hyper-parameters and ROC

Study	ROC	Split Rule	mtry	Top Predictors
Cunningham & Lemke 2013	0.829	extratrees	2	
Cunningham 2006	0.801	gini	2	War Months + Population
Cunningham 2010	0.794	gini	2	Population + GDPPC + Ethnic Frac.
Bagozzi 2016	0.779	gini	2	Ethnic Frac. + GDPPC + Population
Escriba-Folch 2010	0.753	extratrees	10	Contraband + GDPPC + Population
Wucherpfennig et al 2012	0.751	gini	7	GDPPC + Population
Uzonyi & Wells 2016	0.748	gini	2	Institutional Constraints + Constraints x Tenure
Cunningham, Gleditsch & Salehyan 2009	0.746	gini	17	GDPPC + Population
Buhaug, Gates & Lujala 2009	0.743	gini	2	Distance to Capital + Border x Distance + Democracy
Nilsson 2012	0.734	extratrees	21	Balance of Forces
Burgoon et al 2015	0.723	extratrees	8	Media Reporting + UN PK + Amnesty
Bennett & Stam 1996	0.720	extratrees	2	Terrain + Territorial + Sum of Population
Caverley & Sechser 2017	0.693	extratrees	11	Distance to Capital + Democracy + GDPPC
Conrad et al 2018	0.650	extratrees	2	Democracy + Extortion + Contraband
Collier, Hoeffler & Soderbom 2004	0.613	extratrees	10	GDPPC + Ethnic Frac. + Missing Inequality
Thyne 2012	0.592	extratrees	8	Battle Deaths + Commitment Index

dummy is not found to be predictive of conflict duration in this combined dataset. Democracy, GDP per capita, and population variables are frequently selected as top predictors, as well as covariates proxying commitment issues (e.g. veto players, peacekeeping, commitment index).

4.4.3 Performance Metrics

I provide both in-sample and out-sample performance metrics to gauge model fit and predictive accuracy. First, I present the internal performance metrics: ROC, sensitivity, and specificity. The explanations of the confusion matrix statistics are provided in Figure 4.7 (Kuhn and Johnson, 2013). In-sample performance metrics tend to be more optimistic than their out-sample counterparts. However, there are several reasons why it is still a good idea to assess them in conjunction with the test data performance. As the cross-validation procedure generates multiple resamples during training, there is almost always more data available to assess model fit

		Reference	
Predicted		Event	No Event
Event		A	B
No Event		C	D

$$Sensitivity = \frac{A}{A + C}$$

$$Specificity = \frac{D}{B + D}$$

Figure 4.7: Confusion matrix statistics

at this stage. This comes in especially handy when analysing the true positive (sensitivity) and the true negative (specificity) rates and their spread—external performance metrics could be highly biased when they are not resampled enough times.

4.4.3.1 In-Sample Performance

Figure 4.8 shows the ROC, sensitivity, and specificity metrics of all replicated studies. Overall, across all resamples of elastic net and random forest models, the average values are: ROC: 0.721; Sensitivity: 0.784; Specificity: 0.479. In other words, the models do a much better job in identifying true positives in relation to the true negatives, which is predicted slightly worse than what random chance would dictate. If we stratify on model type, however, we see that the random forest algorithm is more extreme in its classification than the elastic net (Table 8.2 in the appendix). It does an extremely good job in detecting true positives (0.89), but really poorly in specificity (0.32). Elastic net performance, on the other hand, is more stable across all three metrics.

Model nuances and differences across studies are visualised in Figure 4.8. Elastic net metrics display more spread than their random forest counterparts, indicating higher variation. Interstate war duration studies do not display the

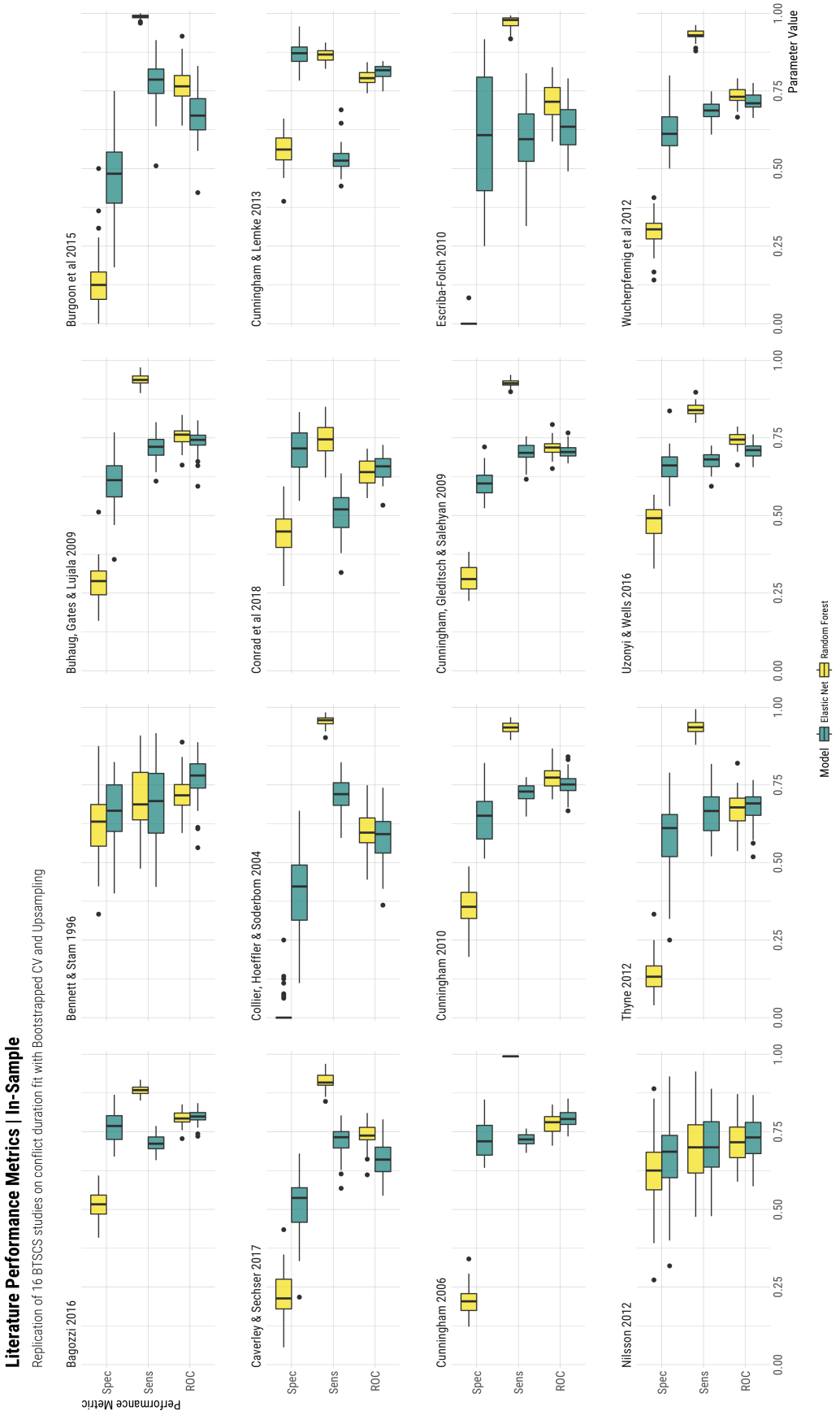

Figure 4.8: In-sample performance metrics

sensitivity/specificity trade-off plaguing the random forest.

4.4.3.2 Out-Sample Performance

Moving onto external performance, I use separation plots to visualise the classification performance. A separation plot "allows the analyst to evaluate model fit based upon the models' ability to consistently match high-probability predictions to actual occurrences of the event of interest, and low-probability predictions to non-occurrences of the event of interest" (Greenhill et al., 2011, pp. 991). It is touted as being insensitive to the arbitrary probability thresholds that are used to distinguish between true events and non-events. Figure 4.9 displays the separation plot of each study produced by the predictions of the best elastic net model. For reference, a perfect classifier will produce complete separation—red bars on one side (representing zeros) and white ones (denoting ones) on the other. A trace line is added to each plot to improve legibility.

It should be noted that the number of observations do affect the visual representation of the plots. Studies with lower n such as Bennett and Stam (1996) and Nilsson (2012) feature comparatively larger blocks than studies with larger n e.g. Burgoon et al. (2015). On the other hand, the aforementioned two studies both only analyse interstate wars and they have the lowest out-of-sample accuracy. Even though the difference in accuracy between the interstate war studies and the least-accurate civil war studies is not that large, there is still a categorical difference. While it is difficult to pinpoint why this is the case, the lower n of interstate war studies is a likely culprit.

Finally, even though we do not observe a clear cut separation in any of the studies, some are most discriminative than others (Buhaug et al., 2009; Caverley and Sechser, 2017; Escribà-Folch, 2010). The external accuracy of Buhaug et al. (2009) can be explained by the importance of geographical factors—many of which are highly predictive—in conflict duration forecasting: distance to capital and conflict

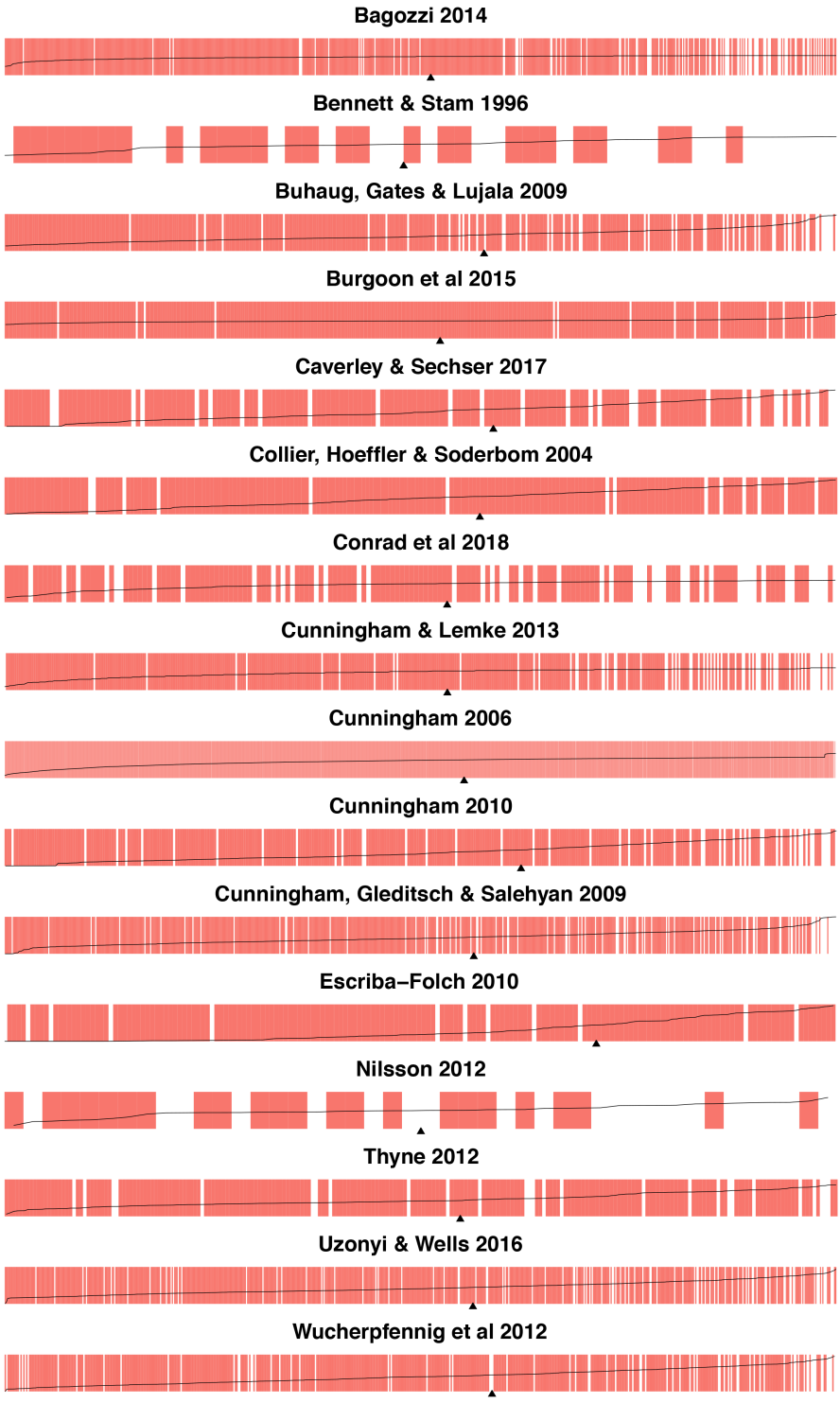

Figure 4.9: Out-sample (prediction) performance using separation plots

at the border are two of the top predictors of conflict duration. Caverley and Sechser (2017) focus on material and fighting capabilities as well as military tactics and logistics. Along with their new variable capturing army mechanisation, their model incorporates geography and regime type. Their analysis thus benefits from multiple sets of good predictors. Escribà-Folch (2010) highlights factors linked to economic sanctions and institutional constraints in his analysis. Again, a balanced mixture of covariates measuring natural resource exploitation, geography (terrain features), and political constraints pave the way for high out-of-sample accuracy.

The main takeaway of the out-of-sample predictions is that accurate duration forecasts require a complementary mix of covariates capturing different aspects of duration dynamics. As laid out in the theory chapter, duration can be conceptualised as a function of capability-spending consisting of baseline material capabilities and limitations (physical and non-physical) acting on them. Studies featuring a diverse set of variables that proxy for all three components—even in the form of control variables—are more likely to make accurate out-of-sample predictions compared to others that only focus on one aspect.

4.4.4 Top Predictors Across All Studies

I briefly summarise the findings borne out of five different algorithms: recursive feature selection, genetic algorithms, simulated annealing, and variable importance from elastic net and random forest model fits. Doing feature selection five times using a diverse set of algorithms resampled many times (293,600 model fits in total)[14] should offer a fair assessment of the predictive quality of the covariates under scrutiny.

Figure 4.10 visualises the top predictors (with a threshold of minimum ≥ 5 selections overall) across studies. Top predictor is defined as being selected as one

[14]RFE 50 resamples/study; SA 5000 internal and 5000 external resamples/study; GA 1000 internal and 1000 external resamples/study; Elastic Net 6000 resamples/study; and random forest 300 resamples/study.

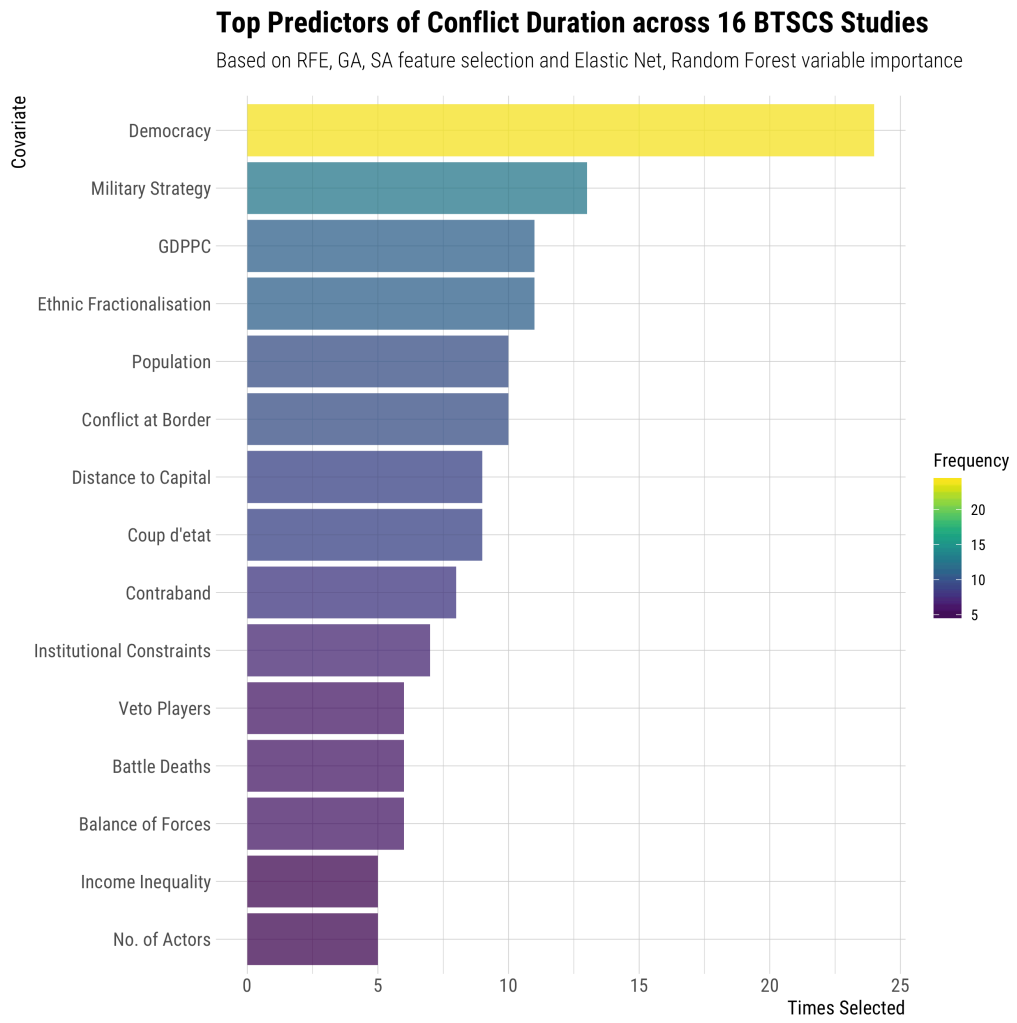

Figure 4.10: Top predictors of conflict duration

of the three top covariates *excluding* the cubic splines; for this reason, in some cases there are less than three selected covariates per study—more extreme result being the Cunningham and Lemke (2013) having no predictor getting selected except for the splines in several cases. In those cases, only the time splines are selected as predictive features and all other covariates are assigned an importance score of zero.

Polity, either in the form of a democracy dummy or a polity score, is by far the most selected feature (24) across studies. This is further complemented

by institutional constraints on the decision-maker, which is selected seven times. Structural determinants such as GDP per capita (11) and population size (10) are also deemed strong predictors. This is followed by factors capturing geography, such as having conflict at border (10) and distance to the capital (9). Balance of forces, operationalised as the ratio of composite index of national capability (CINC) of the belligerents (Singer et al., 1972), also makes the cut with six selections. The number of actors and income inequality share the last two spots with five selections each.

Moving away from physical capability, political constraints, and distance variables, the use of military strategy is the second most frequently selected covariate (13). However, it must be noted that this finding is likely to be driven by the same study (Bennett and Stam, 1996). Military coup is also designated as a consistent predictor of duration (9).

There is a caveat to assessing feature selection in this manner. Some covariates are more frequently included in datasets than others. Variables such as GDP per capita, population, regime type are ubiquitous. More specialised features like the *commitment index*, or covariates borne out of new data such as *ground mechanisation* are less likely to be included, let alone selected after a competitive filtering process.[15] Other than aggregating such distinct features into more general umbrella terms—for example, commitment index under bargaining or ground mechanisation under fighting capacity—this type of loss cannot be prevented. However, doing so would result in a loss of resolution and granularity that are present in the replication studies. Thus, I do not aggregate up distinct predictors and accept their loss, as only about less than 3% of all variables are truly idiosyncratic—their exclusion should not affect the systematic results in a significant way.[16]

[15]It should be noted that they can; the selection of institutional constraints as a top predictor being the case in point.

[16]Consult Table 8.1 in appendix for the complete list of included variables.

For the more common variables (predictors that are labelled more or less consistently), I take two precautions to reduce possible bias. First, as described in the research design, I aggregate features capturing related phenomena and report them under the top feature's label. This way, both *lenient* and *strong* veto players are coded as veto players. Second, I only consider the top three predictors in each study, using five different approaches. Each approach consists of many resamples using hold-out data, making sure each predictor is randomly selected enough times. More specifically, each best individual model fit from five different sources—three feature selection algorithms and two model fits—is a product of about 250 separate model fits aggregating into one. In order to be selected, a predictor needs to appear in the top three consistently across multiple approaches.

Expecting high predictive accuracy also undermines potential problems stemming from variable frequency. Even though nearly all replicated studies have either a GDP per capita or population size variable. However, while a more common variable—by definition—is more likely to be selected across studies given its higher frequency, it is not a given that they are good predictors. Setting a high threshold ensures such commonly occurring variables are only selected when they do exceptionally well in predicting the right outcome, such as being amongst the top three predictors in terms of accuracy. This instils a certain degree of robustness to the results. For instance, rebels having i) a legal political wing and ii) strong central command are frequently included variables across studies. However, neither of them are consistently selected as top predictors.

4.5 Conclusion

This is the first comprehensive quantitative assessment of the determinants of armed conflict duration. By leveraging both the internal and the external resamples and performance measures, I demonstrate which covariates do a

better job at predicting the true outcome. Normally, the predictors in a study are divided into two groups: independent variables and the controls. Even though both are just independent variables that are thought to be correlated with the outcome, being a control indicates that the authors are not interested in its substantive interpretation, and only include the term to account for an alternative explanation. In null hypothesis significance testing studies, these controls are merely mentioned in passing.

However, a closer look at 16 conflict duration studies using BTSCS data shows that features commonly delegated to the control status have immense predictive power. In line with the theoretical expectation of a nexus of physical capability, political constraints, and geographical factors governing war duration, I demonstrate that covariates that are used to operationalise these structural aspects are highly influential in forecasting.

The replication enterprise acts as a sensitivity analysis of the conflict duration literature. Under the assumption that the selected studies constitute a representative sample of the published literature, it identifies which predictors are more consistently better than others at forecasting. However, the study has its limitations.

First, as discussed recently, the consistency of variable labels across studies are manipulated for uniformity. There are other defensible ways of achieving consistency. For example, similar predictors can be aggregated up to a more general label. In this study, I treat different types of natural resources—e.g. hydrocarbons, gems, contraband—as unique predictors. Although existing research (Lujala, 2009) shows that they do behave differently, one could also combine them all under 'natural resources'. Doing so will make that variable highly predictive, as contraband alone is ranked ninth overall.

Second, even though the project is defensible on the grounds that it adequately represents yearly BTSCS studies, it does not necessarily mean

the representativeness can be extended to daily and monthly BTSCS studies. As data collection efforts improve, richer and more granular data become available to researchers. It is possible that while trends that take a while to manifest—changes in GDP P.C., military expenditures, troop size—are successfully captured in this study, finer trends might have been ignored.

Chapter 5

Machine Learning using Combined Data

In this chapter, I empirically the test the general duration model using a replication study (Cunningham and Lemke, 2013) containing both civil and interstate wars. In addition to the original model specification, I add a rich set of new covariates identified in Chapter 4 as the top predictors of conflict duration. I employ a diverse cast of machine learning algorithms, ensembles, and deep learning models. More specifically, I test the claim that whether operational differences between civil and intestate wars (conceptualised as a dummy indicator) can be successfully unpacked and explained away by predictors that capture baseline capabilities and the limitations acting on them. If the binary indicator of war type is not a consistent predictor of conflict duration, this serves as initial evidence suggesting that a similar underlying data generating process governing both types of armed conflict.

The bifurcated nature of conflict duration studies is reflected in the previous chapter: 15 out of 16 replicated studies only look at either interstate or civil wars. The sole exception is the Cunningham and Lemke (2013), a work aptly titled *Combining Civil and Interstate Wars.* The authors cite (pp. 610) two main

factors why conflict scholars traditionally separate the two types of warfare:

> "First, theoretical arguments about war within and between states were once quite distinct. Realism dominated international relations research when large-n statistical studies of war first became common. This approach viewed war as resulting from structural features such as the number of poles and the distribution of power in the international system. Comparativists studying internal conflict, by contrast, usually emphasized state-level features such as government institutions, state strength, and state-society relations, evaluating their theories almost exclusively against a handful of cases."

> "Second, data availability likely also played a role. The original Correlates of War data set included only interstate and extra-state wars, excluding civil wars. COW was the most commonly used data set for large-n analyses of conflict, and researchers interested in conducting these analyses were therefore limited to studying interstate and extra-state wars. By the time COW's intrastate war list became available in 1982, scholarly patterns likely had become fixed."

Both of these points have yet to be addressed in the conflict literature. Different types of war are still studied separately,[1] and data limitations have not been improved. To the author's knowledge, there has been no new study utilising a combined dataset at the time of writing.

To help alleviate these shortcomings, I replicate the study using the same guidelines established in Chapter 4. I take the original Cox Proportional-Hazards model with the specified covariates and transform it to a logistic regression.

[1] It should be noted that this is not an oversight on behalf of conflict scholars; they study civil and interstate wars separately as they believe them to be qualitatively different phenomena. I do not challenge this notion; rather, I focus on whether theoretical explanations of war longevity carry over across types given the similarities in how common measures of duration are operationalised in empirical applications.

I add cubic splines $t + t^2 + t^3$ where t is the duration of conflict to control for the inherent time-dependency. Contrary to Chapter Four, where I employ up-sampling to deal with severe class-imbalance, here I use down-sampling as the class-imbalance is relatively less severe. Down-sampling, even though it necessitates discarding some of the training data, is shown to result in better out-of-sample accuracy compared to up-sampling.[2] Finally, I pre-process the data by centring and scaling all the variables, taking the natural log of skewed numerical predictors, as well as dropping possible linear combinations that might exist in the data.

The rest of the chapter is structured as follows. First, I briefly outline the original model specification as published. Then, I add the covariates that I find to be highly predictive of conflict duration informed based on the results of the previous chapter. Next, I assemble a diverse set of shallow learning algorithms, including ensembles, and make out-of-sample predictions. I then switch to deep learning to generate additional insights from the data. Finally, I explain the predictions of the neural network by employing the Local Interpretations of Model-agnostic Explanations (LIME) framework.

5.1 Shallow Learning

First, in order to build on the findings of Chapter 4, I utilise several machine learning algorithms to predict conflict duration. Although a distinction between 'shallow' vs. 'deep' learning is rather artificial—representation (feature) vs. hierarchical learning being the established terms—there is utility in separating the two approaches.

In representation learning—that is, *regular* machine learning—, the focus is on the features (i.e. variables in a model). Furthermore, the models and the functional

[2] A comparison chart of four common sub-sampling methods—up, down, ROSE, and SMOTE—is supplied in the appendix.

forms are also explicitly specified by the practitioner. In fact, all of the components of this framework are designed by humans based on some heuristics.

'Deep' learning, on the other hand, does not put much emphasis on feature selection. Instead, it creates its own features during the vertical layering process—the moniker 'deep' highlights this aspect. In doing so, deep learning models are highly suited to capture complex, hierarchical interactions between the variables that would be quite difficult to otherwise capture using regular machine learning methods (Mhaskar and Poggio, 2016).

Thus, employing both learning approaches simultaneously helps me investigate the determinants of conflict duration in a more robust way. Shallow methods are valuable tools to identify which covariates have predictive power beyond traditional statistical significance. Deep learning techniques complement this by uncovering high-order interactions between features that also increase predictive accuracy.

5.1.1 Baseline Study

The original Cunningham and Lemke (2013) study features the following covariates:

Outcome ~ Civil War Dummy + Territory + Recurring War + Troop Ratio + Democracy + Total Troops + Total Population

Using a Cox Proportional-Hazards estimation, they find civil wars, wars featuring democracies,[3] and wars featuring larger populations tend to last longer, whereas skewed troop ratios have a statistically significant shortening effect on conflict duration.

[3]One caveat with this finding is that the authors use different operationalisations for interstate and civil wars. When they stratify their model based on war type, they find that democracies fight longer interstate wars but shorter civil wars.

The authors readily acknowledge that the statistical significance of the civil war dummy indicates that there are factors (that are thought to be common in both types of warfare) that are not captured by their model specification. Further, on page 621, they posit that:

> Had we better measures of the bargaining concepts motivating research on conflict duration, it is quite possible that the substantive and statistical significance of the civil war dummy would be considerably attenuated.

In a similar vein, although I do not focus on operationalising the concepts of bargaining per se,[4] I nevertheless take up their call and complement their dataset by adding covariates that capture the effects of absolute and relative material capabilities, political constraints on the head executive, and the geographic realities of power projection.

5.1.2 New Covariates

Based on the empirical findings of Chapter 4, the following variables are added to the original model specification to increase predictive accuracy of the original model. The selection of the new covariates is guided by how closely they resemble the top predictors and data availability. The latter can be a constraint in some cases; the closer the dataset in construction to the Cunningham and Lemke (2013) data, the higher the chances of compatibility. At this stage, I add covariates generously, as linear combinations and near-zero variances[5] will be dropped automatically during the data pre-processing prior to model fitting.

[4] It should be noted that many authors operationalise bargaining concepts using absolute or relative capabilities, so there is some overlap between the authors' conjecture and the goal of this project.

[5] Kuhn and Johnson (2013) recommend linear combinations and near-zero variance variables should be dropped as they are uninformative.

Material Capabilities

`cinc`: The quintessential national material capability indicator devised by Singer et al. (1972). It is a composite index consisting of six separate indicators: `milex`, `milper`, `pec`, `irst`, `upop` and `pop`.[6] While the aggregated `cinc` index is an indicator of relative capability (% of world resources possessed by a state; yearly global total adding up to one), the components themselves are proxies of absolute capability.[7] Thus, the indices are included separately in addition to the summary statistic; however the latter three variables are dropped as population is included in the original model. For interstate conflicts, it is the average; for civil wars, it is the score of the state actor.

`parallel`: The number of parallel ongoing conflicts in the same calendar year for that conflict dyad. From a capability perspective, the more dispersed the resources, the lesser the fighting capability. It is coded as a continuous variable to account for the possible magnitude of the resource dispersion as a function of increasing dyads.

`alldrugs, ALLGEMSP, hydroD`: A set of dummy variables indicating whether exploitable resources—drug cultivation, valuable gemstones, oil and gas—are present in the conflict zone (Buhaug et al., 2009). For interstate wars, conflict zone is defined as the whole country (where the conflict takes place).

`rebstrdum, figcapdum`: Dummy indicators for overall rebel fighting capacity taken from Buhaug et al. (2009). This provides power parity levels for the rebel organisations vis-a-vis the state. It is also used for imputing 'CINC' score categories for rebel factions. For example, if a rebel group is fighting a government with a CINC score of .02 and their dyadic relationship is coded '3' *(parity)* in the dataset, the average CINC score for that dyad is also .02.

[6]Military expenditure, military personnel, energy consumption, iron and steel production, urban population, and total population.

[7]Military expenditures are thousands of current year US Dollars, military personnel in thousands, and primary energy consumption in thousands of coal-ton equivalents.

`rgdppc`: Yearly GDP p.c. estimates of independent state compiled by (Gleditsch, 2002). As the dataset starts at 1950, measures for previous years are added using different sources. For consistency purposes, in interstate wars it is averaged, while for civil wars it takes the value of the state actor.

`major`: Binary indicator denoting whether a major power, as defined by the Correlates of War project,[8] is one of the conflict parties.

Non-Physical Constraints

`polconiii`: A composite index measuring the political constraints on the head executive (Henisz, 2017). This measure is positively correlated with the democracy (0.7) dummy included in the baseline model. However, it also captures additional constraints on the executive use of force that are not captured by the democracy dummy. It also acts as a proxy for institutional constraints. For interstate wars, it takes the average value for the dyad; for civil wars, the average is calculated by first imputing a score for the rebel side based on the Cunningham et al. (2013) data.

`coupx`: Binary variable representing whether the conflict resulted from a military faction seeking to overthrow the government (Cunningham, 2006).

Physical Constraints

All geographic indicators use Buhaug et al. (2009) for the majority (i.e. civil wars) of the observations. Values for interstate wars are manually coded. Missingness is dealt with imputation via k-nearest neighbours algorithm; variables with a large number of missing values (i.e. $> 25\%$) are not considered for inclusion as an added covariate.

`lndistx`: The geodesic distance in kilometres between the belligerents. For interstate wars, the distance between capitals as laid out by Mayer and Zignago

[8] Correlates of War Project. 2017. "State System Membership List, v2016." Online, http://correlatesofwar.org

(2011); for civil wars, the distance between the capital (government stronghold) and the conflict region obtained from Buhaug et al. (2009).

`confbord`: Dummy variable for having a conflict at the border. The rationale is that rebel groups can use an external state as a refuge and to conduct cross-border operations. For interstate wars, the variable captures whether the belligerents are contiguous states.

`borddist`: A multiplicative (interaction) term for border $\times$ distance, in order to moderate the effect of border.

`mt`: Percentage of the conflict zone that are covered with mountainous terrain. Indicators of rough terrain as an enabler of rebellion are commonly included in conflict studies.

`frst`: Same as above, but for forested areas.

Figure 5.1 visualises the correlation plot with the added covariates. Insignificant correlations are denoted with blank squares.

5.1.3 Algorithm Selection based on Maximum Dissimilarity

The choice of algorithm (or its family) can be crucial to the success of predictive modelling. All algorithms possess trade-offs that can be leveraged and exploited in some cases but not so much in others. Further, model diversity can also aid prediction accuracy on a more aggregate level. However, it needs to be intentional in design—algorithms resulting in similar predictions are less useful than divergent ones (Kuncheva and Whitaker, 2003), which can happen if they are picked randomly. One formal way of ensuring a diverse cast of models is to pick algorithms based on some distance metric. The Jaccard similarity coefficient (Jaccard, 1912) is a popular method commonly used in computer imaging to determine likeness and compare the similarity and diversity of sample sets.

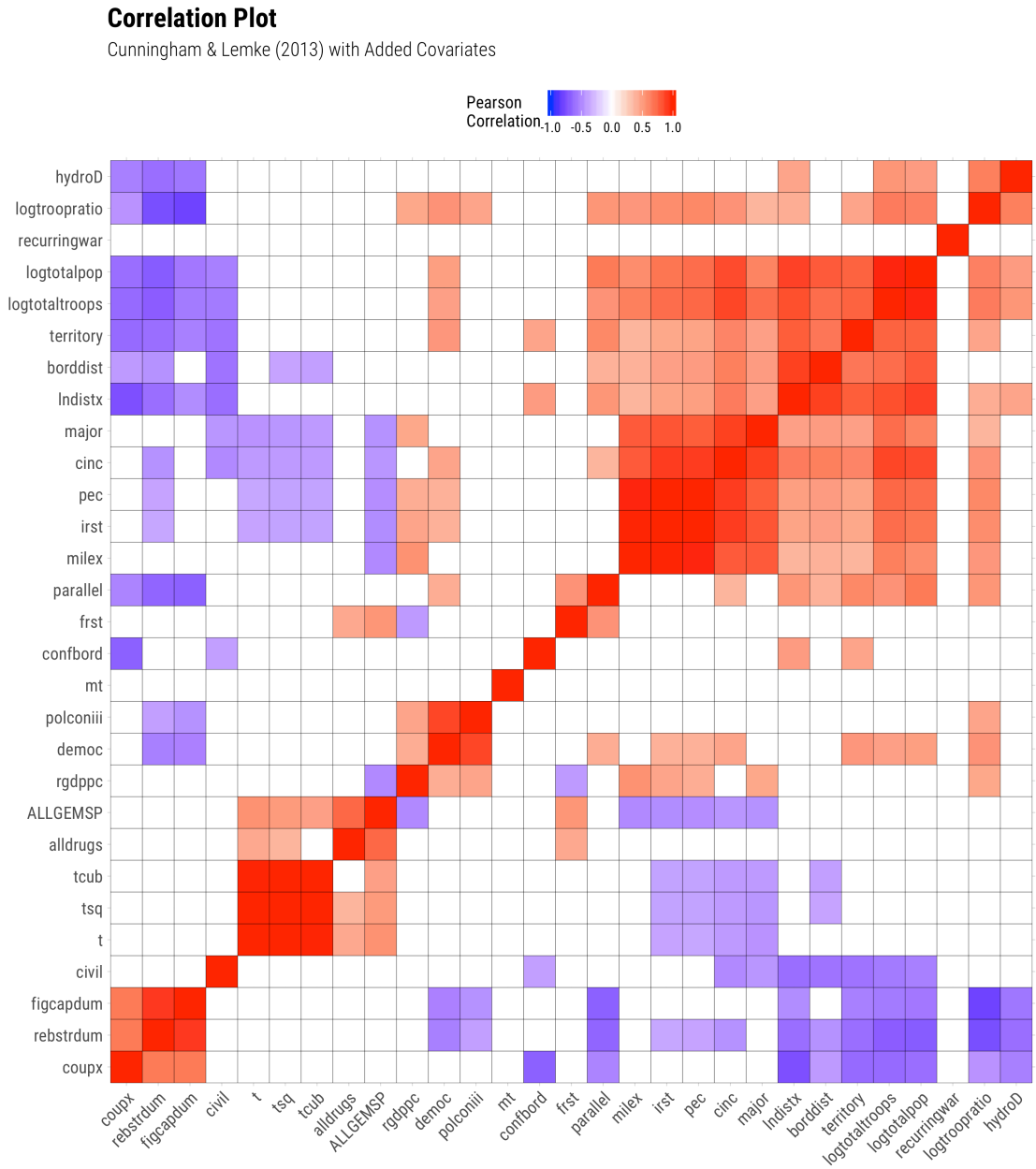

Figure 5.1: Correlation plot of Cunningham and Lemke 2013 with added covariates

The calculation of the dissimilarity distance metric of two samples, A and B, where a score of 0 indicates perfect overlap between the two samples, while 1 denotes no overlap (most dissimilar) can be shown as the following:

$$Dist(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cup B - A \cap B|}{|A \cup B|}.$$

At the time of writing, the `caret` package in `R` has 238 available models to train. Based on similarity tags featuring 57 identifiers, one can generate n number of most dissimilar algorithms *as a set* using their Jaccard distance from a specified input. I use the elastic net `glmnet` as the baseline model, as it is the closest in specification to logistic regression, the literature standard. I add six more algorithms to complement the elastic net in order to create a diverse ensemble without sacrificing too much computational burden.

Based on Jaccard distance, the following algorithms are selected: extreme gradient boosting, distance weighted discrimination with radial basis function kernel, support vector machines with radial basis function kernel, random forest, naive Bayes, and multilayer perceptron network with dropout. As the last algorithm is actually a neural network using the high-level Keras framework, which has its own section later in this chapter, I do not include it here. All in all, the formalised selection process has identified a quite diverse ensemble of algorithms including boosted and bagged trees, ensembles, discrimination models, and probabilistic classifiers. In the next paragraph, I provide a succinct summary of each classifier[9] with the exception of random forest `ranger`, which is already

[9]Most of the selected algorithm families are not widely employed in social science. For those interested in a technical yet accessible introduction, the *Elements of Statistical Learning* (Friedman et al., 2001) provides detailed explanations at the following chapters: Kernel smoothing methods in Chapter 6 (and the Naive Bayes classifier in subsection 6.6.3); gradient boosting and its variants in Chapter 10, with special attention to subsections 10.10.2 and 10.10.3; support vector machines and flexible discriminants in Chapter 12. Similarly, for those who are interested in the applied form with accompanying `R` code, *Applied Predictive Modeling* (Kuhn and Johnson, 2013) covers the following: Support vector machines and other non-linear regression models in Chapter 7 (also 7.3, 13.4); boosting and other rule-based models in Chapter 8 (also see 14.5); discriminant analysis and other linear classification models in Chapter 12 (also see 13.3);

covered in Chapter Four.

Extreme gradient boosting `xgbTree` is an efficient implementation of the gradient boosting framework (Friedman et al., 2000; Friedman, 2001). It is an ensemble learner consisting of many weak learners, and its regularised model formalisation helps reduce the risk of over-fitting (Chen and Guestrin, 2016). Support vector machine `svmRadial` is a discriminative classifier formally defined by a separating hyperplane (Cortes and Vapnik, 1995). Its support of kernel methods makes it a highly adaptable learner as the algorithm can change on-the-fly depending on the kernel function (Scholkopf and Smola, 2001). Distance weighted discrimination `dwdRadial` is based on the majorisation-minimisation principle to compute the entire solution path at a given fine grid of regularisation parameters (Marron et al., 2007). It was originally designed to solve the data piling issue found in support vector machine implementations (Wang and Zou, 2016). Finally, naive Bayes `nb` is a simple probabilistic classifier that is based on applying Bayes' theorem with a strong independence assumption between covariates (Rish, 2001). It is a popular choice for text classification and event models (McCallum and Nigam, 1998).

5.1.4 ROC, Sensitivity, and Specificity

As a first step, I start with providing in-sample performance metrics of the six selected algorithms to establish an empirical baseline. Even though in-sample metrics are more optimistic than their out-of-sample counterparts, it is nevertheless good practice to report both metrics so that comparisons can be made later. Figure 5.2 displays the ROC, sensitivity, and specificity measures across 50 resamples (10-k fold repeated five times). A comparative table of ROC performance metrics across four sub-sampling strategies is included in the appendix for reference; however down-sampling results in the highest ROC scores across all algorithms and thus reported here.

non-linear classification models including Naive Bayes in Chapter 13; and finally classification trees in Chapter 14.

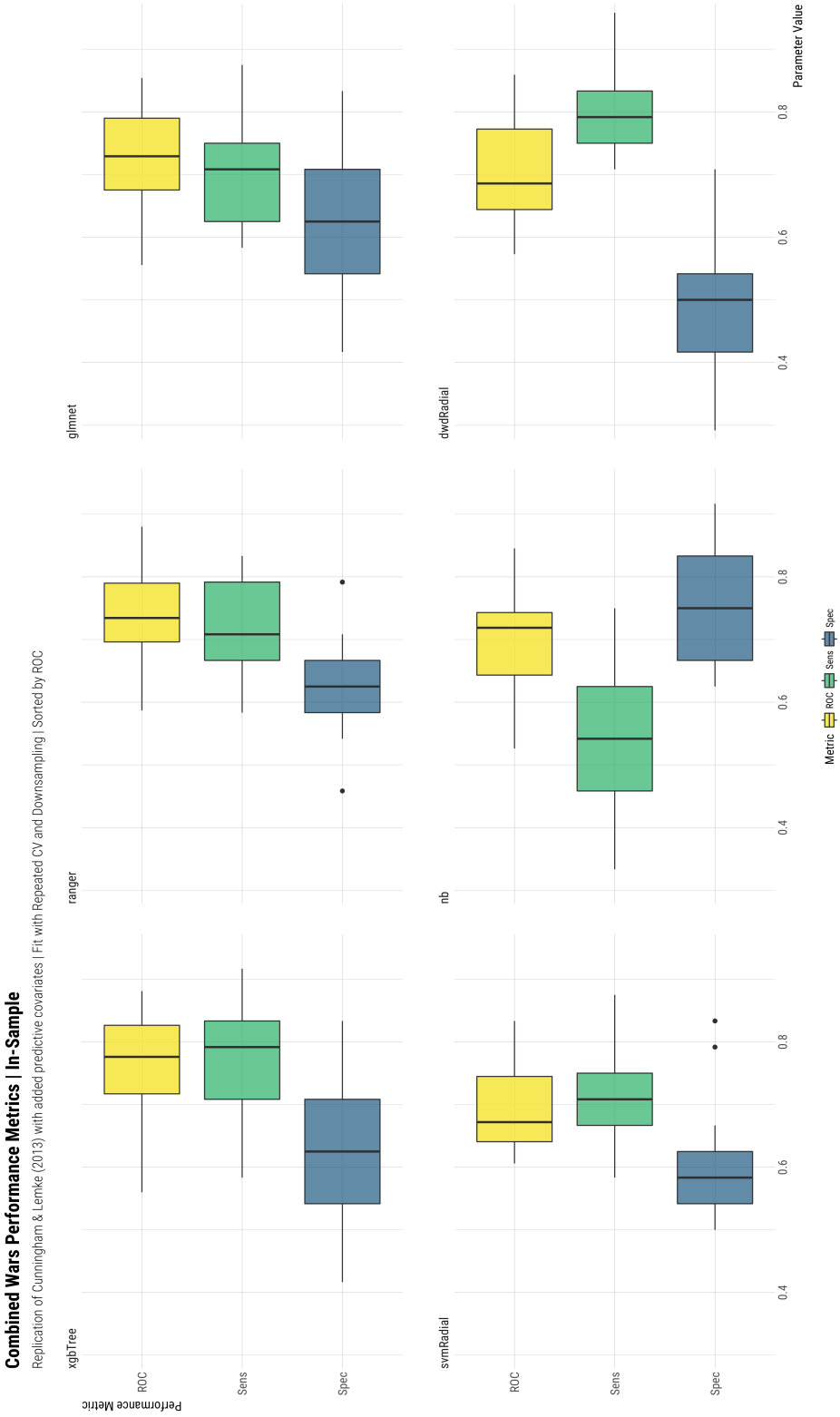

Figure 5.2: In-sample performance metrics

All six approaches yield similar ROC scores; meaning no single model performs significantly better than any other. This is not necessarily surprising; in fact it would be more concerning if the performance metrics display wild fluctuations based on algorithm selection. Still, the exercise acts as an algorithm sensitivity analysis—the findings are robust to the choice of algorithm; purely in terms of predictive accuracy, nothing fundamental is dependent on one algorithm. With that said, some algorithms do better than the others—extreme gradient boosting and random forest models have slightly higher ROC values, followed closely by the elastic net.

However, there is variation in detecting true positives and negatives across models. The naive Bayes model does only slightly better than random chance when it comes to detecting true positives, but it also produces the highest maximum true negative score. The discriminant analysis, on the other hand, suffers from the opposite trade-off: it is the best algorithm in terms of true positive detection, however it does very poorly in identifying true negatives. Such trade-offs are common in practice (Florkowski, 2008), and could be useful in ensemble settings if right algorithms can be leveraged for the correct cases.

5.2 Ensemble Models

A more direct approach to algorithm diversity is to create ensemble models. Ensembles models can be powerful, as they can harness the predictive power of multiple algorithms and outperform individual algorithms if appropriately constructed. This is especially the case when algorithms can complement each others' weaknesses, so that the final model (ideally) contains the 'best' parts of each algorithm included. However, as a trade-off, ensembles are more difficult to interpret, making them better choices for maximising predictive accuracy and when explanation is not the primary concern. For the purposes of this project, I only include ensemble models as a performance benchmark to which individual

algorithms can be measured against.

Although models can be ensembled in many different ways, two basic approaches are greedy and meta-model ensembles. Greedy ensembles are simple linear combinations consisting of individual model predictions. Meta-model ensembles, in contrast, can fit any algorithm on top of the existing predictions (Džeroski and Ženko, 2004). For example, a random forest classifier can be build using individual model predictions in which the predictions are the features. However, more complexity does not always lead to better predictive accuracy; in some cases, simple linear ensembles can perform better than more complex meta-model ensembles (Sollich and Krogh, 1996). Finally, making an ensemble prediction is also not guaranteed to outperform any one model's predictive accuracy, especially if one model is clearly better than the rest of the ensemble (Rokach, 2010).

5.2.1 Simple Linear Ensembles

The predictions of the previous six models are combined in a greedy ensemble. The model predictions are weighted based on their accuracy, and then a final linear model is fit to make new predictions on the hold-out (test) data. The ensemble uses identical cross-validation (repeated 10-k fold), sub-sampling (down), and pre-processing procedures that have been applied to the individual algorithms.

```
The following models were ensembled: glmnet, xgbTree, dwdRadial, svmRadial, ranger, nb
They were weighted:
-2.5535 1.9867 1.5765 2.6603 -2.7116 1.4449 0.3275
The resulting ROC is: 0.7546
The fit for each individual model on the ROC is:
    method       ROC      ROCSD
    glmnet 0.7327083 0.08021384
   xgbTree 0.7412500 0.07850042
 dwdRadial 0.7165972 0.08439435
 svmRadial 0.7119097 0.08244908
    ranger 0.7415625 0.06574621
```

```
nb 0.7121875 0.08059927
```

The code output provides summary statistics of the greedy ensemble, including the model weights and individual ROC values. Both the random forest and extreme gradient boosting algorithms have a ROC score of approx. 0.741, and the ensemble itself slightly improves on both with a ROC score of 0.7546.

5.2.2 Meta-Model Ensembles

More sophisticated ensembles beyond simple weighted linear combinations are also possible. Stochastic gradient boosting `gbm` is a refinement of the gradient boosting method in which at each iteration of the algorithm, a base learner is fit on a sub-sample of the training set drawn at random without replacement (Friedman, 2002). Figure 5.3 plots the relative contribution of each algorithm to the final `gbm` model fit.

The make-up of the meta-model ensemble is shown in Figure 5.3, and it is different from the simple linear ensemble. Even though the top algorithm is extreme gradient boosting, the next two top performing models are naive Bayes and the elastic net. Random forest, in contrast, is the second-to-last in terms of relative influence. The performance metrics of the meta-model ensemble are: ROC 0.818, Sensitivity 0.77, and Specificity 0.689—a vast improvement in ROC compared to the individual algorithms and the linear ensemble.

5.3 Predictive Accuracy

Even though we have achieved better ROC scores moving from individual algorithms to more complex ensembles, so far we have only evaluated in-sample performance. The true validation lies in test scores, as internal accuracy metrics

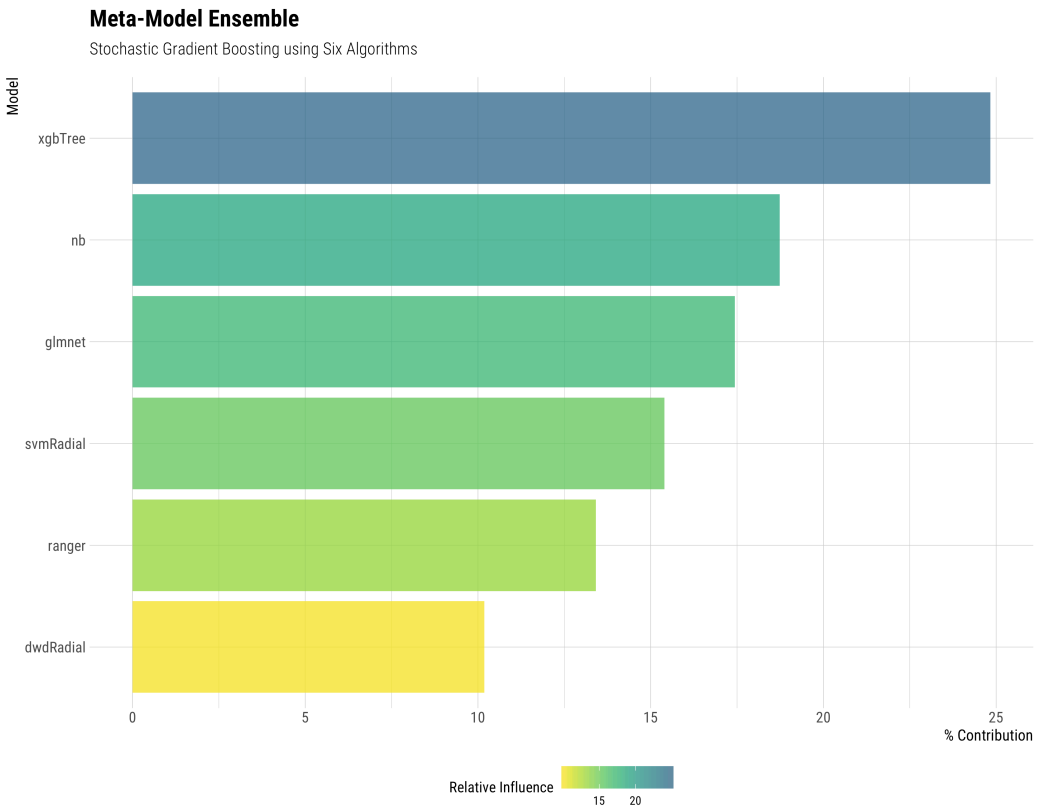

Figure 5.3: Meta-model ensemble relative influence graph

tend to be optimistic. Thus, I predict class probabilities for all observations in the held-out data using all six algorithms and two ensembles.

```
  Algorithm       ROC
1    ranger 0.8773902
2 svmRadial 0.7949076
3 dwdRadial 0.7889726
4    greedy 0.7858329
5   xgbTree 0.7814678
6    glmnet 0.7497164
7       gbm 0.7408645
8        nb 0.6632333
```

The ROC values of all models are displayed in the code output. The random forest algorithm separates itself from the pack with a ROC score of 0.877, while the naive Bayes predictions result in a similar distinction in the opposite direction (0.663). We also see that both ensembles, even though scoring higher in in-sample validation, are located in the middle of the pack. Furthermore, the simpler weighted linear ensemble (0.785) outperforms the more complicated meta-model ensemble (0.74).

Finally, we can extract variable importance from the elastic net, extreme gradient boosting, and random forest algorithms.[10] Figure 5.4 highlights the top predictors across all three models.

Variable importance plots are useful tools to illuminate how the predictive process underlying each algorithm unfolds. Recall that the elastic net has the lowest out-sample ROC value amongst the three. We see one possible reason why it under-performs: all predictive power is generated through the variable `coup` and the cubic time splines.[11] In other words, the determinants of conflict duration,

[10]Other algorithms featured in this chapter do not have an associated variable importance extraction technique.

[11]There are two caveats. First, the immense predictive power of coups, especially in the context of nullifying the effect of the civil war dummy, should be evaluated carefully. The main reason for this is that the variable itself is a very good predictor of civil wars. Second, the cubic splines, included in the algorithms for completeness (given their inclusion in the baseline logistic

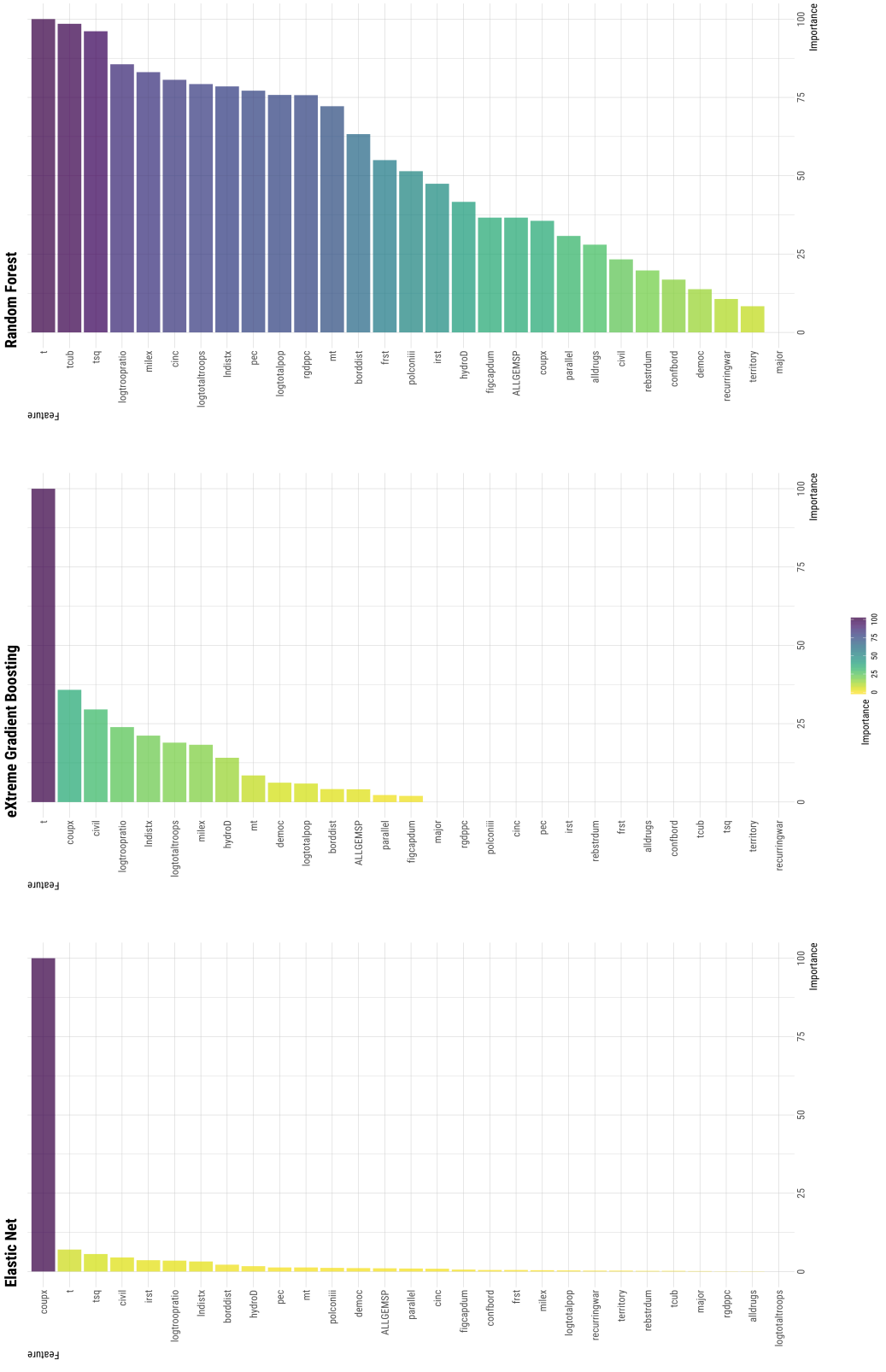

Figure 5.4: Variable importance after model fit

according to the model, is whether the conflict is a result of a military coup and the auto-regressive temporal aspect of the conflict (i.e. how long it has been going on). The elastic net fails to utilise the vast majority of the covariates, and its predictive accuracy suffers from it.

Extreme gradient boosting, in comparison, recruits a larger number of variables when it comes to prediction. Even though the top two covariates are coups and duration (first time spline), the following variables also contribute to predictive accuracy: civil war dummy, geographic factors such as distance and mountainous terrain, and several material capability predictors (e.g. population, troop ratio, troop size, military expenditure).

Finally, the random forest algorithm nearly fully utilises the available covariates. Unsurprisingly, the top three predictors are the cubic splines. The top three is then followed by a dozen covariates that proxy either material capabilities or aspects of geography. In contrast to the previous two models, out of 30 available variables, coups and the civil war dummy variables rank 20th and 23rd, respectively. This supports the theory that when the categorical differences between civil and interstate wars are captured by multiple capability variables, the dummy variable has minimal influence on duration prediction.

5.4 Deep Learning

The popularity of deep learning technologies in sciences—coined as a term in mid 80's, however theoretically neural nets have been around a couple of decades before that—has been on the rise since the computational power has risen up to match theory in the last decade. As immensely powerful learners, deep learning tools are especially adept at uncovering complex, layered interactions between individual predictors. Given the complexity inherent in social sciences in general and conflict

regression models) may not necessarily measure the same effect that is picked up by a logistic function.

research specifically, I complement the previous two approaches—shallow learning and ensemble models—with a neural net application.

More technically, deep learning (hierarchical learning) is a sub-field of machine learning that display the following characteristics (Deng et al., 2014): i) utilising a cascade of multiple layers of non-linear processing units for feature extraction and transformation; ii) each successive layer using the output from the previous layer as input; the model architecture can allow for no memory (each layer is a blank slate) or some memory retention (e.g. Recurrent Neural Networks; Long Short Term Memory models); iii) learning multiple levels of representations that correspond to different levels of abstraction, which form a *hierarchy* of concepts; and iv) can be supervised or unsupervised in nature.

5.4.1 Neural Nets with Keras

Keras is the most-popular high-level interface complementing the low-level TensorFlow back-end and provides a simple API written to reduce the cognitive load of the practitioner (Chollet and Allaire, 2018). TensorFlow is originally developed by Google engineers working at Google's Machine Intelligence Research organisation for the purposes of conducting machine learning and deep neural networks research (Abadi et al., 2016). Tensors are multi-dimensional data arrays. A single digit is a dimensionless (0-D) tensor. A vector of numbers is a 1-D tensor, a matrix is a 2-D tensor, an array of matrices is a 3-D tensor, and so forth.[12] Working with tensors rather than data frames allows TensorFlow to be used for fast prototyping, especially if the user has access to compatible Graphics Processing Units (GPU).

Figure 5.5 demonstrates an example neural network. The input layer is defined in such a way that its size is equal to the number of features in the data. Layer connections are almost always activated using an activation function, which

[12]Images can be represented as 4-D tensors, whereas videos can be captured in 5-D tensors.

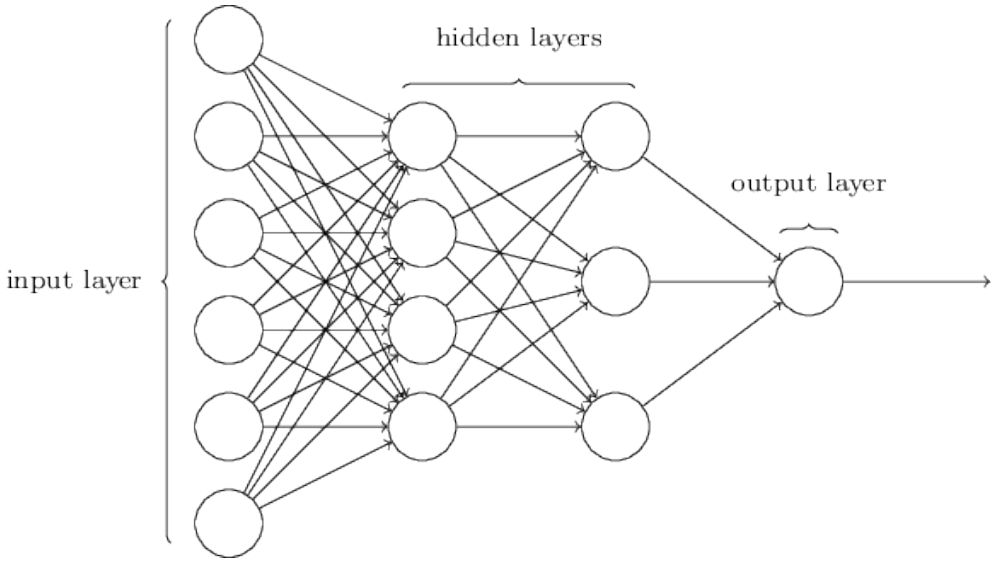

Figure 5.5: Example multilayer perceptron architecture

applies a transformation (such as relu; rectified linear unit) to the weights between layers. The intermediary layers—called *hidden* layers—are densely connected to both the input and the output layers. The practitioner specifies the number of units in these layers; larger numbers mean higher capacity (to learn) but they are also more-prone to over-fitting. The number of hidden units and layers depend on the problem at hand. Finally, the output layer contains the size of the expected output (e.g. one for binary classification tasks) and the activation function (e.g. sigmoid to obtain a value between $(0, 1)$ for the same task).

Neural networks further require functions in addition to activation and the optional normalisation: optimisers, loss, and metric. Optimiser functions such as the efficient ADAM (Kingma and Ba, 2014) are mainly derivatives of gradient descent algorithms used commonly in deep learning (Ruder, 2016). A loss function provides a target to minimise during model fit, for example binary cross-entropy for binary classification tasks. Finally, the metric is what the model aims to maximise, in this case, validation accuracy.

5.4.2 MultiLayer Perceptron (MLP) for Binary Classification

A MultiLayer Perceptron (MLP) is the 'vanilla' neural network, similar to what Ordinary Least Squares (OLS) is to linear regression. Neural network architectures are highly capable, and this high capacity can lead to over-fitting. Similar to shallow machine learning, there are several common counters that can minimise over-fitting. One approach is to include dropout layers (Srivastava et al., 2014). These layers randomly drop a user-specified fraction of input units each update during training and can also be set to keep mean and variance of inputs to their original values, ensuring self-normalisation (Klambauer et al., 2017). In addition to including dropout layers, I utilise regularisation (both $L1$ and $L2$, similar to the elastic net) and batch normalisation as suggested by Ioffe and Szegedy (2015). The latter procedure normalises the activations of the previous layer at each batch, applying a transformation that maintains the mean activation close to zero and the activation standard deviation close to one. Finally, I apply Gaussian noise to the dense layers, which is a natural choice for corruption processes for real-valued inputs (Choi et al., 2017).

Figure 5.6 visualises the training evaluation of the MLP model. To clarify jargon: twenty percent of the down-sampled *training* data is used for internal resampling at each epoch—this is referred to as (internal) validation below. The *held-out* validation is done using the untouched (no sub-sampling) *test* data—this is the out-of-sample (external) validation.

Ideally, models should run for just enough epochs until the validation accuracy stops improving while training accuracy continues to improve (i.e. divergence), as the difference between the two accuracy metrics represent over-fitting. The callback argument in Keras allows for early stopping when a user-specified monitored metric stops improving; however it was disabled in order to generate the plot so that whole 100 epoch performances can be seen. As epochs essentially represent different models (i.e. using different weights), one cannot average their

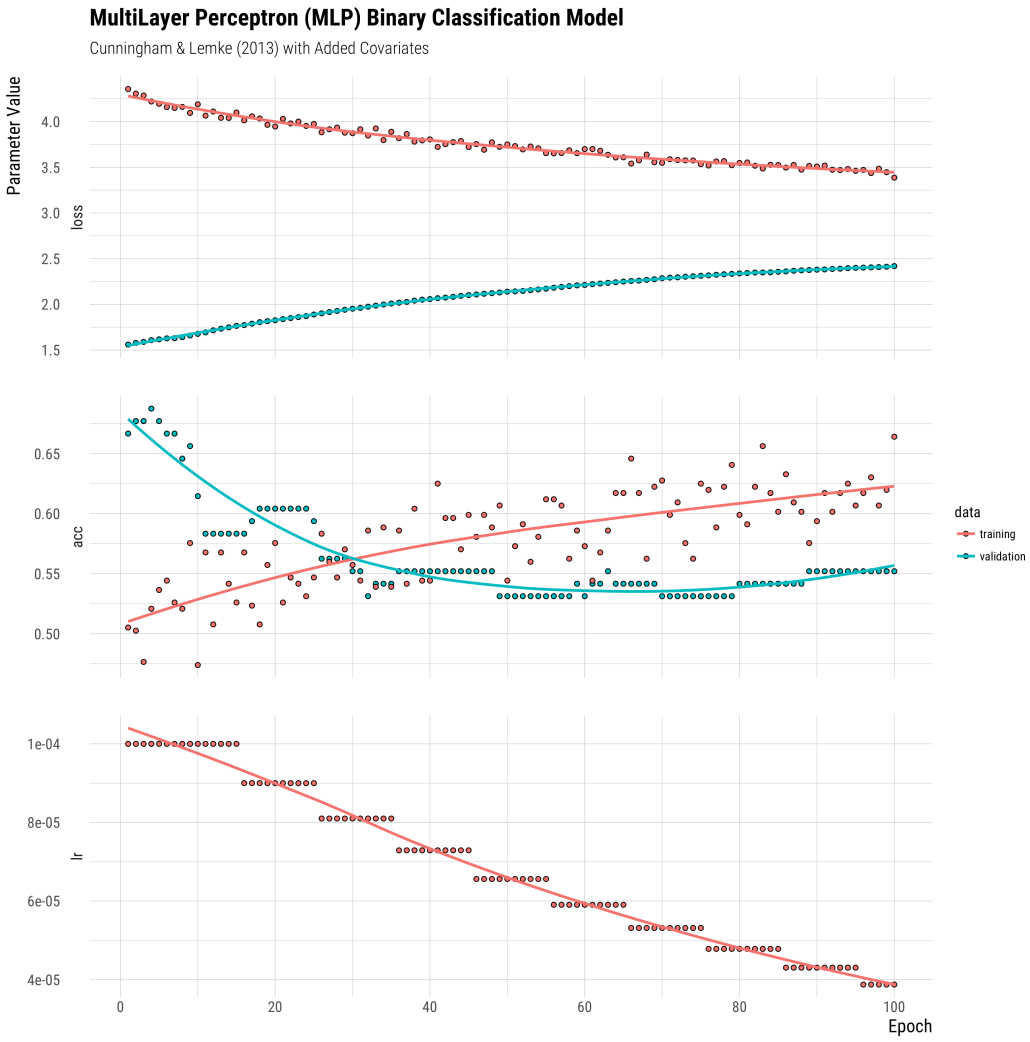

Figure 5.6: Multilayer perceptron training evaluation

Table 5.1: Multilayer perceptron external performance metrics

ROC	Sensitivity	Specificity	Accuracy	Precision	F1 Score
0.7321544	0.6764706	0.6570248	0.6604096	0.293617	0.4094955

accuracy over multiple runs. The top performing model in terms of validation accuracy is used for held-out validation.[13]

Another callback argument adjusts the learning rate when the model hits a plateau, which can be seen in the third row of the figure. Without such adjustments, the loss function would not be updated in-between the model runs, resulting in a flat-lining of the validation set. However, we see that incrementally lowering the learning rate (at a factor of 0.9) does not always lead to an improvement, evidenced by the fluctuations in validation accuracy over time.

5.4.3 Performance Metrics

By default, Keras provides accuracy (not ROC) as a performance metric, so other external measures need to be created separately. Table 5.1 displays calculated accuracy metrics. It has a ROC score of 0.732, on par with most of the shallow learning algorithms but far behind that of the random forest. The MLP is quite consistent with true positive and negative detection, showing no apparent trade-off between sensitivity (0.676) and specificity (0.657).

Precision, the fraction of relevant instances among the retrieved instances, is at 0.293. Finally, the F1 score—the harmonic average of the precision and recall; note that recall is equivalent to specificity—is about 0.41.[14]

```
Truth
```

[13]It should be noted that one-hit wonders—an epoch that has a significantly higher validation accuracy than its surroundings—are likely to be outliers. Top performing models located in peaks (e.g. performance drops both before and after that epoch) are more consistent models.

[14]F1 Score of 1 mean perfect precision and recall, similar to a ROC score of 1 indicating perfect sensitivity and specificity. Both metrics are bounded by $[0, 1]$.

```
Prediction Negative Positive
  Negative      318       33
  Positive      166       69
```

Finally, the confusion matrix provides a breakdown of the class predictions. Mirroring the summary statistics reported above, we see that the MLP model on average correctly predicts two-thirds of the held-out test observations for both categories ($n = 586$). Note that the class-imbalance does not seem to have a performance-reducing effect on class predictions.

5.4.4 Local Interpretations of Model-agnostic Explanations

Deep learning models are thought to be black boxes. However, recent developments have made significant progress in uncovering how deep learning predictions are made. One such method is the Local Interpretations of Model-agnostic Explanations (LIME) framework proposed by Ribeiro et al. (2016).

A LIME explanation is a local linear approximation of the model's behaviour as pictured in Figure 5.7. While the actual model is likely to be complex at the global level, it can be approximated within the proximity of a particular instance. Instead of trying to explain the model as a black box, the instance under scrutiny is perturbed. This results in an encompassing sparse linear model around that can be learned as an explanation. In the figure, the blue/pink background represents the model's decision function, which is non-linear. LIME explains the instance marked by a red cross. Next, the procedure samples instances around this area and weighs them (indicated by size) according to their distance to the area under inspection. Finally, a linear model (represented by the dashed line) is fit, which is used to approximate the model well in the local proximity, but not on a global scale.

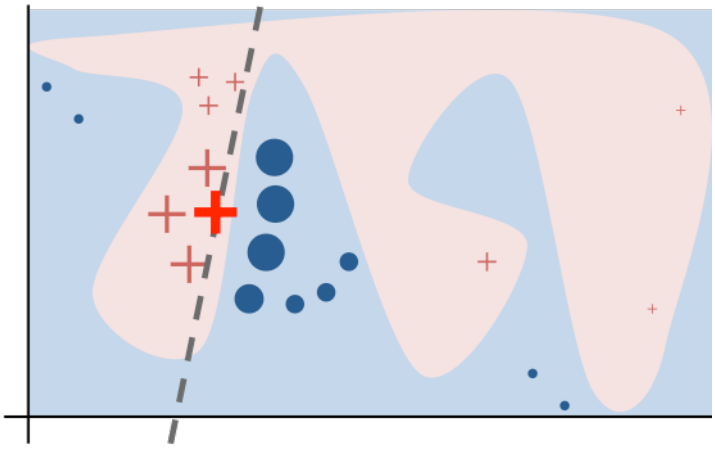

Figure 5.7: Illustration of a local interpretation according to the LIME framework

One of the features of LIME is the ability to generate explanation plots. After local interpretations are calculated, selected cases can be plotted to visualise i) which predictors contributed to that class prediction, and ii) the direction of their contribution (supporting or contradicting the prediction). As most of the covariates are scaled and centred, they are transformed to continuous values. LIME can bin these continuous variables into discrete quartile chunks. All explanation plots provide information in the following format: name of the conflict dyad, class prediction, class probability, and the R^2 value associated with the local linear approximation.

I first present ten positive cases—the observations that the MLP model predicted as 1—to scrutinise the selected predictors in detail. The most common predictors for explaining positive cases are coups and parallel conflict. However, their direction varies; coups tend to contradict whereas parallel conflicts support the positive predictions. We also see that time splines are picked up quite regularly, and they always support the positive forecasts. Distance is also identified as an important predictor, but its direction varies from case to case. Other important explanatory variables all relate to material capabilities: Composite index of material capability, primary energy consumption, iron and steel production, GDP p.c., and troop ratio.

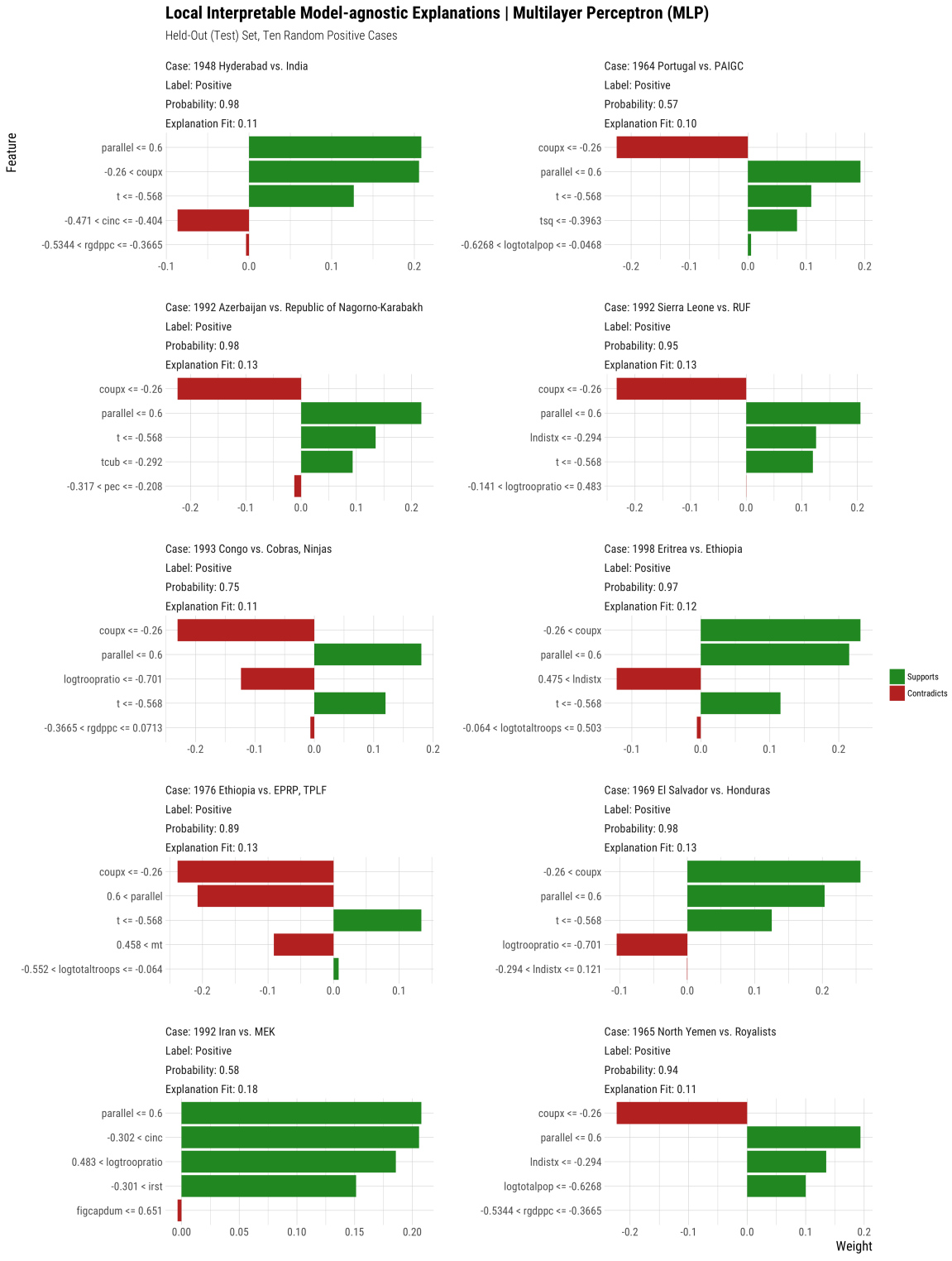

Figure 5.8: Explaining positive MLP predictions

Figure 5.9: Explaining negative MLP predictions

Moving on to the negative predictions, we find that most predictive covariates from the positive cases carry over to the negative cases as well. Coups and parallel conflict in conjunction with the cubic splines are the most common features. In contrast, more geographic features are included in negative predictions such as the interaction of border $\times$ distance, conflict at border, and mountainous terrain. Note that all geographical variables are associated with support; meaning these features are conducive to negative forecasts—that is, they predict prolonged conflict.

Finally, I pool the covariates and tabulate their frequencies similar to that of a variable importance chart. The number of parallel ongoing conflicts is selected 539 out of 586 times (92%) as an important predictor. Coups take second place with 439 selections (75% regularity). Time splines, controlling for the time-dependency, are picked 45%, 26%, and 24% of the time, respectively. This hints at the auto-regressive influence of time being strongest initially. Material capability indicators, such as GDP p.c., the composite index (cinc), and iron and steel production are the remaining covariates in the top ten. Geographic factors—distance, conflict at border, their interaction, and rough terrain—are also included, however to a lesser degree. Exploitable resources are very scarcely included as influential predictors. Political constraints also seem to have a minimal impact, and this time democracy does not make the cut at all.

parallel	coupx	t	logtroopratio	rgdppc
539	439	265	238	208
cinc	lndistx	irst	tsq	tcub
173	171	169	152	141
borddist	logtotalpop	logtotaltroops	figcapdum	mt
78	77	54	49	49
pec	frst	milex	polconiii	confbord
40	39	21	10	8
rebstrdum	alldrugs	ALLGEMSP	hydroD	
5	3	1	1	

Crucially, we see that the civil war dummy never makes into the top five in any of the cases in the test data. Coupled with the fact that many material capability and geographical variables are picked, this is further evidence that once the structural factors are controlled for, the effect of the civil war dummy is indeed attenuated as predicted by Cunningham and Lemke (2013).

Indeed, the repeated uninformativeness of the civil war dummy across a diverse set of machine learning algorithms provides strong evidence in favour of a similar data generating process underlying both types of conflict. At a basic level, in terms of the operational differences between interstate and civil wars, the dummy variable captures 'everything else' outside of what else has been explicitly specified in a model. Cunningham and Lemke (2013) find that the dummy indicator, even in the presence of several controls such as population and troop size, is still a statistically significant factor that explains why civil wars last longer than interstate conflicts. However, once relevant covariates—identified as the most consistently top predictors established in Chapter 4—are introduced, the dummy variable does not hold much predictive power. In other words, in terms of forecasting, the positive categorisation of being a civil war does not make a conflict more protracted.

5.5 Conclusion

This chapter aims to build on Cunningham and Lemke (2013)'s study on combined wars by enriching their existing model with highly predictive covariates identified via a quantitative assessment of the conflict duration literature. I argue that the explanatory power of the civil war dummy, proxying for the qualitative differences between civil wars and interstate conflicts, stems from lacking covariates that can capture the material capabilities of the belligerents.

First, I provide empirical evidence that the characteristics of civil war can be moderated and even completely nullified by introducing relevant capability indicators and limitations. I find that less successful models in terms of predictive accuracy rely on fewer predictors—mainly coups and the time dependency—and fail to generalise well to unseen data. In contrast, highly predictive models do well to diversify their predictors and channel the full potential of their features. With added features capturing various material capability and geographical factors pertaining to war, accurate classification rates increase across all algorithms.

Further, the most powerful predictive models do not identify the binary civil war indicator as an important predictor. From an empirical perspective, this is evidence in favour of the proposed theory—a unitary model can capture conflict duration in both types of war. When appropriately specified using relevant covariates capturing capability and the constraints acting on them, the effect of the civil war dummy is completely attenuated; it stops being an informative covariate.

Second, both absolute and relative material capabilities affect conflict duration. This is evidenced by the regularity of the selection of such variables: GDP p.c., primary energy consumption, iron and steel production, total population, and the total number of troops as absolute measures; and troop ratio, the composite index (as a fraction of world resources), and fighting capacity/parity for the relative

measures. Geography is also an important factor, especially the distance between the belligerents. As power projection is costly over distances, this findings fits in well with the general theory of the loss of strength gradient.

Third, on a more methodological note, 'shallow' learning algorithms out-perform their deep learning counterparts. More specifically, the random forest algorithm does significantly better when it comes to out-of-sample prediction. Random forests are known to perform well when underlying non-linear interactions hold predictive power, as they are well-equipped to capture such interactions. Given the superior performance of random forest over logistic regression in predicting conflict onset (Muchlinski et al., 2016), if the aim of the practitioner is to maximise predictive accuracy, more attention should be given to tree-based models. Further, the trade-off associated with using tree-based learners can be controlled to a degree, as the practitioner is able to dictate the depth of the trees. Shallow trees are easier to visualise, and arguably even more intuitive than the assumptions underlying logistic regression.

On the other hand, the neural network implemented using Keras seems to suffer when faced with class-imbalanced data and the low number of observations that come with down-sampling. However, they still provide value with their ability to go deeper than regular machine learning algorithms. Especially when paired with a framework like LIME, they can be useful in uncovering linkages that will not be picked up by more shallow learning methods. Yet, as evidenced by their pedestrian performance, conflict practitioners should not expect an improvement in predictive accuracy just by switching to deep learning technologies. The immense learning capacity of neural networks can lead to over-fitting much faster than any other algorithm, and the decision to switch should be backed up by either theoretical or methodological expectation (Cawley and Talbot, 2010).

Chapter 6

Analysis

In the previous two chapters, I have investigated the predictive determinants of conflict duration using algorithmic modelling. In this chapter, I move the discussion into a more in-depth analytical direction. Findings borne out of machine learning are useful, but require effort to be made interpretable. In addition, many numerical indicators—e.g. GDP p.c., population—cannot be changed in a short amount of time, making policy-recommendations that depend on them less useful. To this end, I provide two additional sets of empirics: i) establishing directionality of the predictive covariates of war duration, and ii) causes of capability shifts; what actors on the ground have to say about dynamic factors that can dampen or enhance material capabilities using a limited case study.

Establishing the direction of the predictive effects are important, as without direction, it is hard to understand the true relationship between the predictive variable and the outcome. It is one thing to identify a covariate—say, military coups—as a reliable predictor of war duration; however, if we cannot speculate on what type of effect military coups actually have on conflict longevity, our results are nevertheless less robust than what they could be otherwise. Thus, I provide an in-depth breakdown of each predictive covariate covering two points: whether

the effect is positive or negative, and whether the effect is consistent between civil and interstate wars. In doing so, I provide a firmer foundation in favour of a unitary model of conflict duration while simultaneously outlining some of its limitations.

Furthermore, there is a limit to how much we can explain using quantitative methodology. The choice of methodology comes pre-packaged with a certain set of assumptions about how the world operates. To this end, I employ a limited case study that focuses on the limitations of the quantitative component of the dissertation. More specially, I aim to identify possible pathways that the empirical operationalisations of the theory—material capabilities and non-physical (political and societal) constraints—can avoid detection by quantitative means.

The rest of the chapter is structured as follows. In the first part, I expand on Chapter 5 findings by establishing the directionality of the predictive covariates. Then, I analyse how the three theoretical components of the general theory influence conflict duration empirically. Second, I introduce a shadow case of Sierra Leone. Based on 19 semi-structured interviews, I provide several narratives of the interaction of material and political capabilities of the conflict actors—SLA, the Executive Outcomes, ECOMOG, UN Peacekeepers, and the UK Expeditionary Force. In sum, I bring together two sets of empirical evidence and analyse what we have learned so far about the predictive determinants of armed conflict duration.

6.1 Predictive Modelling

In this section, I investigate the directionality of the covariates found to be highly predictive in the conflict literature. More specifically, I look at whether the same features have similar effects on both types of warfare or the direction of the effect depends on conflict type. Unlike traditional statistical approaches, most machine learning algorithms do not provide readily-interpretable coefficients that denote

the direction of the observed effects. Instead, algorithms focus on which covariates contribute the most to predictive accuracy. As a result, an additional step is required to obtain directionality.

The procedure is as follows. I select the top performing algorithm across all categories—shallow learning, ensembles, and deep learning—from Chapter 5; which is the random forest. Then, using the best random forest hyper-parameter tunings to the data (selected for maximum external fitness; out-of-sample accuracy), I run the Local Interpretations of Model-Agnostic Explanations (LIME) procedure, similarly described in detail in Chapter 5. Put simply, the LIME framework computes variable importance and effect (coefficient) for black-box algorithms at the local level, the main idea being what is complex at the global level (the black box) is more interpretable at a lower (local) level where individual decisions are made for each case. Finally, I aggregate these effects with stratification (civil wars and interstate wars) and report their mean value and dispersion for both types of war.

Figure 6.1 demonstrates the directionality of the most important predictors according to the best random forest fit. The effects are local to the outcome (predicting '1' as the outcome, i.e. termination of conflict in that time-unit). The box-plots indicate the inter-quantile range, with the mean value denoted with a dash. Outliers are marked with dots. The zero line dividing the negative and the positive directions should not interpreted as p-value significance, meaning it can be crossed without losing importance. However, the line is nevertheless used for establishing the main direction of the covariate, based on where the median value lies. In some cases where the median value is nearly overlapping with the zero line, the width of the boxes can be more informative.

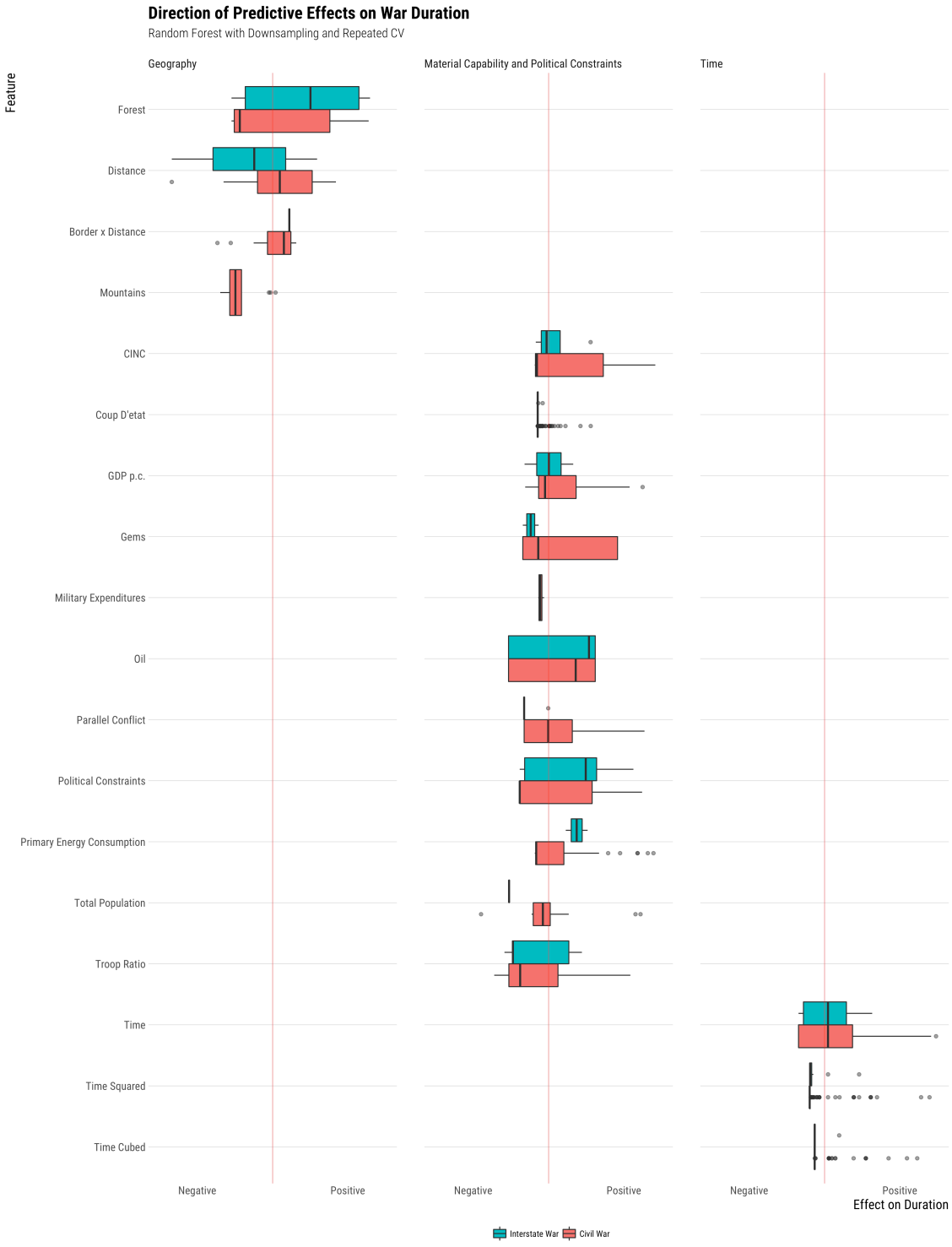

Figure 6.1: Random forest: directionality of the predictors

6.1.1 Material Capabilities and Non-Physical Constraints

For the purposes of this analysis, I discuss material and political variables in unison. The main reason for doing so is that some predictors are closely linked to both categories depending on context. For example, coups are sometimes considered as a constraint on material capabilities and other times as political proxies.

Material capacity and political constraints are also intertwined at the theoretical level. While geographical factors can affect power projection ex-ante—i.e. operational area being constrained by the loss of strength gradient—they are mostly constant throughout the conflict once it is under way. Political variables, either in the form of political constraints on the head executive or more generally regime type, can vary from year to year within a conflict.

Many absolute and relative material capability indicators are highly predictive of war duration. Unlike geographical factors, most of the capability indicators have consistent effects—war type does not usually change the direction of the effect. This is striking, as it provides support to the idea that same underlying processes might govern both types of conflict. At the very least, the fact that many capability indicators are selected as accurate predictors in both cases suggest same parameters are at play.

Starting with the covariates capturing absolute material capabilities, we see that they are conducive to longer wars. Higher Composite Index of National Capability (CINC), GDP per capita, military expenditures, and total population values are associated with protracted conflicts in both types of war. On the other hand, primary energy consumption has a shortening effect in interstate wars, while prolonging civil wars.

On relative capabilities, only the troop ratio is consistently chosen as an important

predictor. Surprisingly, higher (more skewed) ratios indicate longer conflicts. This effect is consistent for both war categories. Natural resources (oil, gems, drugs) and other capability-enhancing variables (e.g. contraband) are also included under this heading. However, only oil production and gemstones are accurate predictors, and they have opposite effects on duration. Having access to oil is associated with shorter durations, while dealing in gemstones prolong conflicts. These effects are again consistent for both types of war.

Finally, the three remaining important predictors are military coups, political constraints on the executive, and the existence of parallel conflicts. Conflicts stemming from coups have a prolonging effect on duration on average. However, there are many outlying cases of civil war falling into the positive (shortening effect) side. Parallel conflicts also extend war duration with the same caveat.

The effect of political constraints depends on war type. In interstate wars, it has a strong shortening effect—politically constrained leaders are associated with shorter international wars. On the other hand, similarly-constrained civil war leaders are good predictors of protracted wars.

Overall, most of the predictive capability indicators behave quite similarly in both types of conflict. Equally important is the fact that all capability indicators were selected as highly accurate predictors for both types of war. Taken together, these findings provide strong support to the notion that, as far as our modelling approaches and choices of operationalisations go, conflict is conflict—it can be modelled in a unitary fashion.

6.1.2 Physical Constraints

Geographical factors have the most diverse effects. Densely forested terrain has a positive effect on interstate war termination, however the effect changes is negative for civil wars. Mountainous terrain has a similarly prolonging effect on civil war

duration. There is no interstate war entry for mountainous terrain as it is not a good predictor of interstate war duration and as such, not picked up by the algorithm.

The effect of distance also varies according to war type. For interstate wars, longer distances have a negative effect. However, the sign is reversed for civil wars. It should be noted that the median value is close to zero, and there are many cases on either side of the line. Still, the changing sign of the distance predictor is not surprising. Given the wide range of distance values and the apparent predictive power the variable holds, future research should consider more refined operationalisations of distance, including cubic polynomials i.e. $distance + distance^2 + distance^3$. This would be in line with existing research that looks into power and proximity (Gartzke and Braithwaite, 2011). Finally, the interaction term consisting of border and distance has a positive effect on both types of war. The only caveat is that there are only a handful of interstate cases fulfilling the criteria, and it is possible that the result is driven by a dominant case.

Moving away from directionality, these four geographical factors[1] are consistently selected by the algorithm as accurate predictors. This supports the theoretical notion that such factors are important in both types of war. The large variances of the covariates, however, indicate that there is more than meets the eye concerning conflict geography. Difficult terrain types have a prolonging effect on civil wars, a finding that is in line with the literature. However, the effect of different types of terrain is not identical. Mountainous terrain has a more precise prolonging effect on civil war duration; while many civil wars—both long and short—were fought on densely forested terrain. One implication is that forest cover has an interactive effect depending on some other factor.

The effect of increasing distance—shortening civil wars and prolonging interstate wars—is contradictory to literature expectations. However, it is likely that the variable is capturing the essence of high absolute material capabilities and the

[1]With the exception of mountainous terrain for interstate conflict.

commitment to the fight as a proxy. When state actors project power over vast distances—U.S. in Iraq and Afghanistan—this implies that they have very high levels of operational (material) capabilities and they are committed to the fight (Gartzke and Braithwaite, 2011).

6.1.3 Time Effects

I discuss the influence of the effects of time next. Although the cubic splines are added to alleviate methodological concerns, they nevertheless provide insights on the underlying temporal dependency found in conflict processes. In addition, the stratified design makes it possible to assess whether the temporal effects manifest similarly in both types of warfare.

The findings suggest the time effects are uniform; they behave similarly in both civil and interstate wars. The first spline t, which indicates a linear functional form with regards to the outcome, is slightly positive (i.e. has a shortening effect, as the positive coefficient means contributing to a '1' prediction of termination). Both the second t^2 and third t^3 splines have a negative (prolonging) effect on duration, however both splines have numerous outliers in the positive direction. The median of t^3 is also slightly less than the median of t^2, indicating diminishing returns over time.

All three cubic splines are important predictors of duration. Taken together, time-dependency in armed conflict seems to manifest itself with a predictable functional form. Initially, some wars terminate soon after their onset, increasing the rate of termination and in effect, lead to shorter durations. However, once a critical point in time is reached, wars are less likely to terminate as more time passes.

The consistent time-dependency trend across war types may suggest explanatory factors such as commitment problems should apply similarly to both types of

conflict. This, however, assumes that the cubic splines capture what is not explicitly modelled; e.g. information asymmetries and/or commitment issues. As conducted, this project cannot assert that this is the case. Yet, it indicates that whatever effects the time variables proxy for via omission (in the model), they behave similarly in both types of conflict.

6.1.4 Variable Importance

Next, I analyse the variable importance more in-depth. First, parameter rankings relating to the tree structure are summed up to identify the most important variables. More important (i.e. better at splitting) predictors have desirable scores in several tree structure features. Second, a predictive performance plot based on both accuracy and Gini importance is shown to assess the relationship between the two measures. Both visualisations are useful for evaluating the robustness of the depicted metrics. Further, they can aid in making more qualitative inferences regarding the random forest model fit.

Figure 6.2 visualises the three main tree structure features of a random forest: the mean depth of the first split on a covariate (x-axis; smaller is better), the number of trees in which the root was split on that covariate (y-axis; larger is better), and the total number of nodes in the forest that split on that covariate (dot size; larger is better). The colour of the dots denotes whether the covariate—based on its overall summed up ranking for the three aforementioned features—is in top ten (blue) or not (black). Certain clusters are readily apparent from the plot. The cubic splines, capturing the effects of time, score highly on both axes and by far the most important variables. Distance and CINC is another top performing cluster. Further down, the total number of military personnel and mountainous terrain are located in close proximity to each other. Finally, troop ratio, primary energy consumption, and military expenditures form the remaining important predictors.

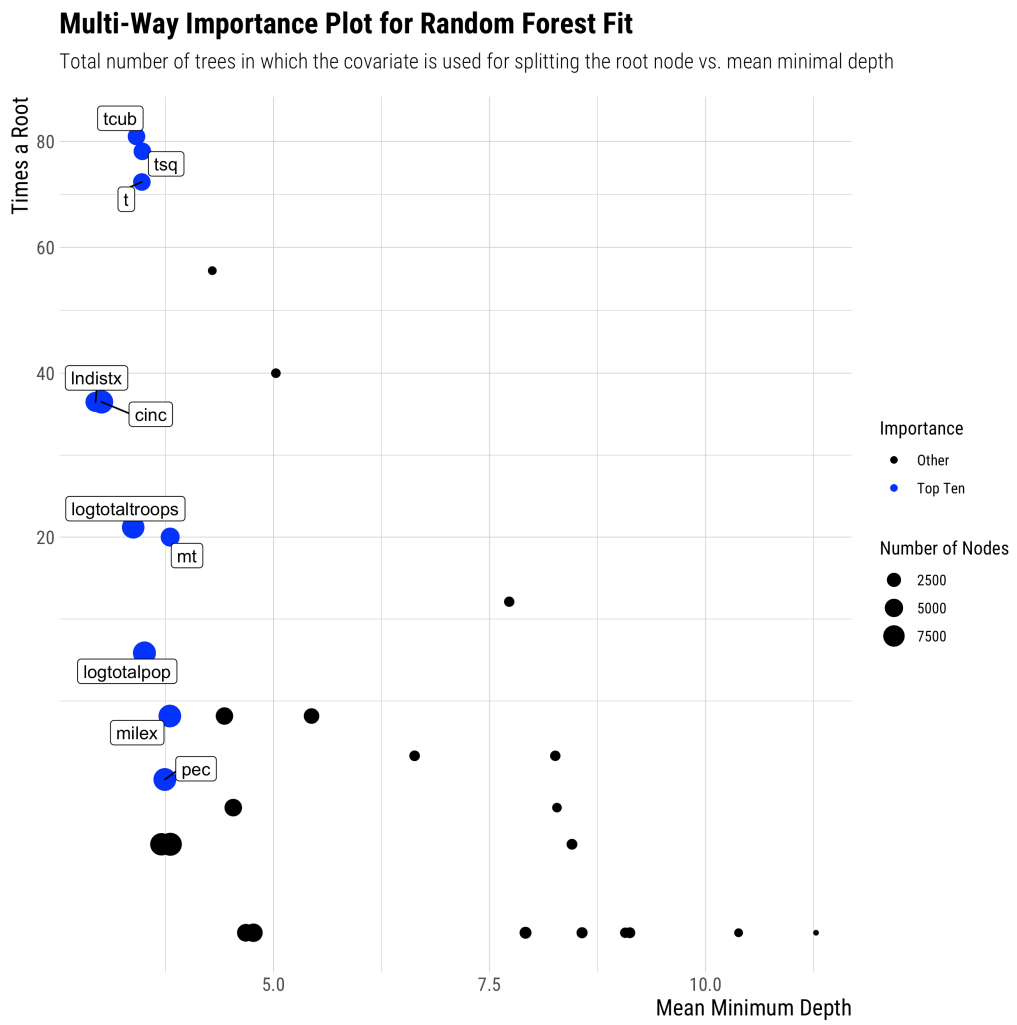

Figure 6.2: Random forest multi-way importance plot: tree structure metrics

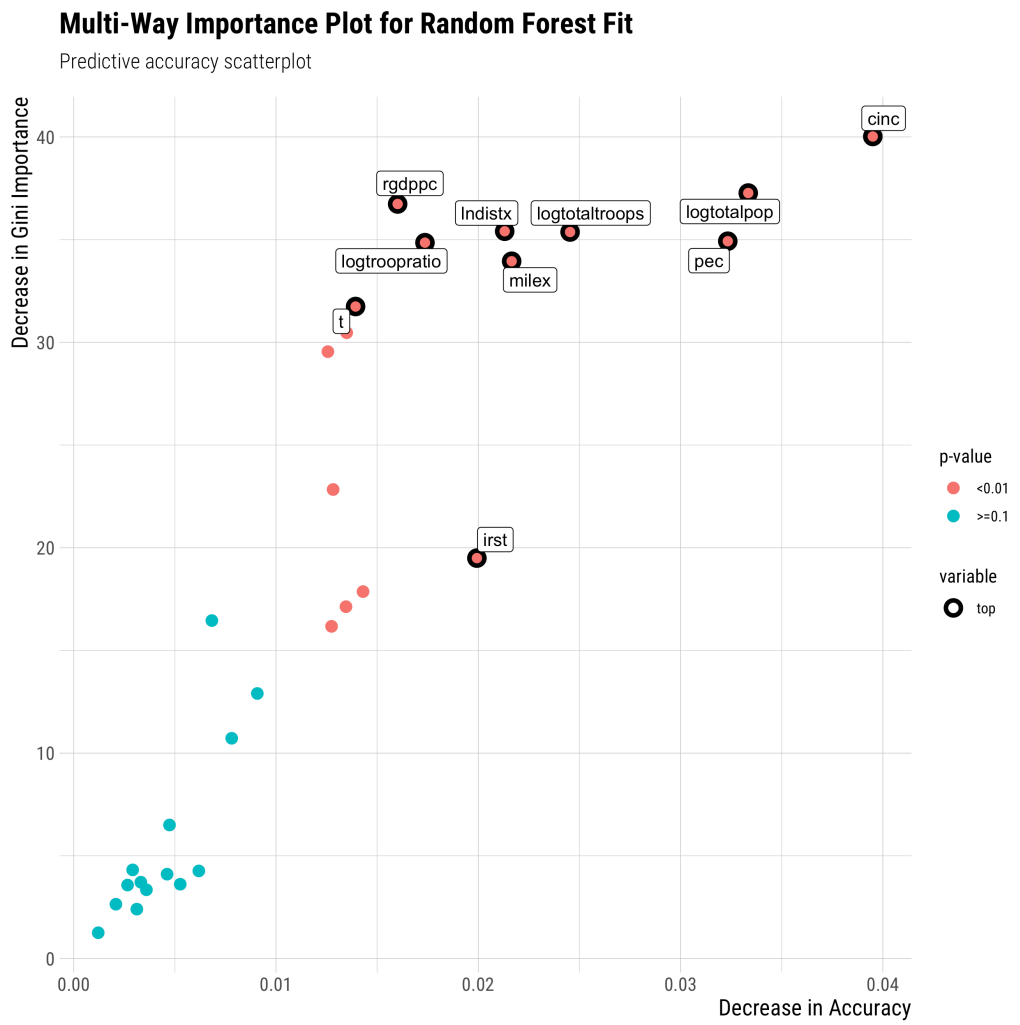

Figure 6.3: Random forest multi-way importance plot: predictive covariates

Figure 6.3 emphasises predictive accuracy of the covariates. The top predictor, based on the amount of performance loss caused in accuracy and Gini importance, is the CINC score. It is followed by population and primary energy consumption. Remarkably, the vast majority of the top ten variables pertain to material capability: number of military personnel, military expenditures, troop ratio, GDPPC, and iron and steel consumption. Only time t and the geographical factor distance are considered top variables without being a material capability indicator.

Taken together, multi-way importance plots provide several insights on the inner workings of the random forest predictions. Tree structure metrics identify which predictors are influential in terms of splitting the data. Time variables are indisputably the best at this task. Material capability indicators also do a good job at variable splitting, as well as geographical factors such as distance and terrain. No political covariate is identified as an important variable.

One takeaway from Figure 6.2 is that time and absolute material capabilities of an actor are highly discriminatory. The intervals of time are measured in years in this project. As such, higher order splines are better at splitting: t^3 is the most important variable, followed by t^2 and t, respectively. On the other hand, given the yearly intervals of t can be considered as an idiosyncrasy of this project, studies using more fine-grained time intervals (months, weeks, days) may not be able to replicate this result this strongly.

Moving away from time, it is also not surprising that absolute material capabilities (e.g. total troops, population, military expenditure) are more discriminatory than their relative counterparts (i.e. measured in ratios). In absolute terms, these indicators—even after log transformation is applied during the pre-processing stage—are more likely to be skewed in real life while calculated ratios are naturally smoother in their distribution.

In contrast, accuracy measures draw a different picture. Time loses its

prominence; only t is included in the top ten. This is strikingly different compared to their importance in data splitting: while they are excellent for splitting on (i.e. categorisation or binning), they are not necessarily accurate predictors. Material capability indicators plus distance, on the other hand, are the most accurate predictors of conflict longevity. Given that the civil war dummy is neither an important nor a predictive variable (placed 17^{th} out of 30), these findings support the notion that the proposed capability-projection model attenuates its influence.

6.1.5 Case Explanations

Finally, I visualise how the random forest case predictions can be unpacked on a case-by-case basis. Figures 6.4 and 6.5 display ten randomly selected case explanations drawn from civil and interstate wars, respectively. These showcase the similarities in the underlying trends governing conflict duration. Coupled with the preceding analysis of common predictive covariates, there is ample evidence supporting the notion that armed conflict is governed by similar characteristics regardless of type.

A quick glance at both figures provides additional support for a unitary predictive model that does not discriminate based on war type. First, in many cases, the top five most influential variables for both civil and interstate wars overlap. Second, there is a healthy amount of both exogenous and within-variable variation.[2]

Exogenous variation here refers to directionality that is dependent on the outcome—i.e. the case labels 'No' and 'Yes' in the figures. In Figure 6.4, the effect of coup d'etat (determined as <= -0.285 after applied transformations) differ in cases of Cambodia vs. Khmer Rouge in 1973 (predicted No with a probability of 0.53) and Philippines vs. CPP in 1978 (Yes with a 0.69 probability).

[2]It should be noted that both types of variation are not uniform; however this is expected as no effect is perfectly consistent in large-n studies.

Figure 6.4: LIME Random forest: ten randomly selected case explanations (civil wars)

Figure 6.5: LIME random forest: ten randomly selected case explanations (interstate wars)

Within-variable variation is the change of the effect direction based on differing values of the variable. In Figure 6.5, access to oil (variable *hydroD*) has a negative effect on Israel vs. Egypt in 1970 (label 'No') when the value of the covariate <=0.417. On the other hand, when *hydroD* is >0.417, such as the case of China vs. Vietnam in 1981 (label 'No'), this time it has a positive effect.

6.1.6 Theoretical Implications

Based on the empirical evidence presented in this chapter, several theoretical implications can be made. First, higher amounts of material capability are associated with longer conflicts. High CINC values, GDP per capita, military expenditures, and population all have a prolonging effect on conflict. This supports the notion that actors (or actor dyads) with higher material capabilities are able to continue bearing the costs associated with protracted fighting. An alternative reading is that highly-capable actors are more likely to select into prolonged conflicts. This could explain why increasing values of troop ratio, which suggests imbalance (as higher ratios are more skewed in favour of the stronger party), are also indicative of longer wars.

Oil and political constraints on the executive exert the strongest influence on shorter wars. Given that the mining of valuable gems are associated with longer conflicts, there are multiple possible explanations for the divergence in the results. First, conflict type can account for some of the variation: oil-wars are predominantly interstate affairs while gem mining is more prevalent in civil wars (de Soysa et al., 2009; Lujala, 2009; Hendrix, 2017).

Unlike duration analysis (Buhaug et al., 2009), the type of terrain matters in predictive modelling. Conflict in densely forested areas is associated with longer civil wars but shorter interstate wars. One explanation is that interstate actors have access to military technology that nullifies the effect of rough terrain. This can also explain why mountainous terrain is not a predictive factor for interstate

conflict but a strong predictor of longer civil wars. Non-state actors, in contrast, are affected by difficult terrain for a multitude of reasons: i) they may lack the capability to traverse rough terrain, ii) conversely, the government forces may not be able to project force into such terrain, and iii) there could be strategic incentives for the rebels to stay in relatively inaccessible areas where the government forces cannot reach them easily.

Finally, the changing influence of distance between the power bases of the warring parties depending on conflict type is telling. Interstate actors fight longer wars when they project power further away from their base. Conversely, civil wars are relatively shorter affairs when the distance between the capital and the conflict zone is not large. Both findings are well-established in the literature (Gartzke and Braithwaite, 2011; Buhaug et al., 2009). Similar to the effect of forested terrain, variables that have diverging influences have the most explanatory power when it comes to unpacking the civil war dummy.[3]

6.1.7 Conclusion

I provide further empirical support in favour of the proposed theory: not only the top predictors of conflict are consistent across war types, the majority of them also overlap in direction. In other words, the influence of the top predictors of war duration is mostly persistent: whatever prolongs civil wars also increases interstate war duration, and vice-versa. This is a strong indication that both types of conflict are governed by a similar underlying process.

Further, knowing the directionality of the predictors help unpack what is usually captured by the civil war dummy indicator. Recall that the rest of the original model specification of Cunningham and Lemke (2013) consists of variables capturing territory, recurring war, troop ratio, democracy, total troop size, and population size. The addition of the literature covariates identified by the

[3]The civil war dummy is further unpacked using multi-way importance plots in the appendix.

algorithmic replication process makes the categorical dummy redundant.

One key takeaway is that a diverse set of predictive covariates are required to attenuate the predictive power of the binary civil war indicator. The original model specification mostly focuses on absolute and relative material capabilities. Given the complexity of conflict processes, it takes the addition of additional capability indicators (e.g. natural resources), alongside with geographical constraints, to make the civil war dummy uninformative from a forecasting perspective.

With that said, some predictors have divergent effects on duration based on war type. Geographical constraints on power projection are sensitive to conflict type. The exploratory (and algorithmic) nature of this project does not organically lend itself to finely articulated theory; however, further research can be more discriminatory in this regard. Particularly, the influence of distance and various terrain features should be explored further to pinpoint the conditions under which the divergence occurs.

6.2 Shadow Case Study

I complement the quantitative findings with a shadow case study of Sierra Leone. I conducted 19 semi-structured interviews with retired and active officers from the Sierra Leonean Army (SLA), ex-combatants (both mid-level commanders and rank-and-file) from the Revolutionary United Front (RUF), defence ministry officials, UN personnel, national security advisers, local academics, and a Western diplomat between January-March 2017. The majority of the interviews took place in Freetown, the capital. The rest were conducted in Makeni, known for being the headquarters of the SLA 4th Brigade and where 500 UNAMSIL peacekeepers were famously disarmed and taken hostage by the rebel forces back in May 2000.

6.2.1 Conflict Parties

The Sierra Leone Civil War (1991-2002) featured a rich set of actors, both domestic and international. Even though the majority of fighting can be linked to the Government of Sierra Leone vs. the RUF dyad, both sides had multiple actors intervening on their behalf. The government forces were supported, at different times (and sometimes in an overlapping fashion) by the many international interveners—the Executive Outcomes, the South African mercenary group; the ECOMOG, the Nigerian-led West African Task Force; UNAMSIL peacekeepers; and the UK Expeditionary Force. The RUF, on the other hand, joined forces with the Komojors—influential hunters from the Mende tribe—as well as receiving support from NPFL[4] in Liberia. In the next section, I briefly introduce the conflict parties before covering the dynamics of capability shifts on the ground.

6.2.1.1 Domestic Powers

SLA

Throughout most of its history, The Sierra Leonean Army (SLA) was a ceremonial power. Founded in 1961, the SLA was modelled after the former British Royal West African Frontier Force. When the civil war broke out in 1991, it had around 3,000 personnel. To bolster its forces against the rebels, President Momoh expanded the army ranks to include "mainly drifters, rural and urban unemployed, a fair number of hooligans, drug addicts, and thieves" (Clapham, 1998). This trend continued when Captain Strasser, who took control of the government following a coup d'etat in 1992, recruited young criminals, school drop-outs, and semi-literate youths. The size of the SLA rose up to nearly 14,000 as a result. At the same time, the SLA was constantly under-armed and under-paid. Many soldiers came to the realisation that they could benefit from the war by joining

[4]National Patriotic Front of Liberia, the main rebel group led militarily by Charles Taylor during the First Liberian Civil War (1989-1996).

the rebels looting the civilians in the countryside. This led to the notion of *sobels*; soldiers by day, rebels by night (Feldman and Arrous, 2013). For civilians, the line separating the government forces and the rebels was quite thin.

RUF

The Revolutionary United Front (RUF) was the main rebel organisation during the civil war. The 'stated' goals of the RUF were to overthrow the All People's Congress (APC) regime that governed Sierra Leone, to 'liberate' the peasantry, destroy corruption, ensure an equitable distribution of the wealth of Sierra Leone's natural resources and institute a multi-party democracy (Day, 2015). However, while the RUF had succeeded in mobilising relevant grievances, it did not have a genuine guiding ideology. Their main motivation was to defeat the government with the exclusive goal of replacing them. The lucrative diamond mining, smuggled through Liberia, was the main source of income for the otherwise under-funded rebels. The RUF had humble beginnings, with 100 or so fighters at the onset of the conflict. In 2000, it had around 15,000 combatants.[5]

Komojors and the CDF

The Komojors were a group of traditional hunters from the Mende ethnic group found predominantly in the southern and eastern parts of Sierra Leone. The Komojors first joined forces with the government to fill in as the main security forces at the wake of the ouster of the South African mercenaries (the Executive Outcomes). This integrated security force was called the Civil Defence Forces (CDF). Most estimates put their total number somewhere between 10,000 and 30,000 in 1997 (Hoffman, 2007). Some parts of the CDF eventually merged with the RUF when the latter took control of Freetown in 1998.

[5]Uppsala Conflict Data Program, Additional Information on RUF. Retrieved from http://ucdp.uu.se/additionalinfo?id=532&entityType=0

6.2.1.2 International Powers

ECOMOG

The Economic Community of West African States Monitoring Group (ECOMOG) is a West African multi-lateral armed force established by the Economic Community of West African States (ECOWAS). During the civil war, two main factions were present in ECOMOG: the larger Nigerian contingent, and the Ghanian strike force ('the bombardiers'). Nigeria provided at least ninety percent of ECOMOG troops (12,000 out of 13,000) and its funding during the military intervention in Sierra Leone. It should be noted that the Nigerians were also under similar obligations to the parallel ECOMOG mission in neighbouring Liberia against Charles Taylor and his the National Patriotic Front of Liberia (NPFL), stretching their material capabilities and political capital across two conflicts.

Executive Outcomes

The paramilitary mercenary group from South Africa known as Executive Outcomes (EO) arrived in Sierra Leone in March 1995. They costed around $1.5 million per month.[6] They were given three objectives: i) return the diamond and mineral mines back to government control; ii) locate and destroy the RUF's headquarters, and iii) operate a successful propaganda program that would encourage local Sierra Leoneans to support the government instead of the rebels. The military force of EO consisted of about 500 military advisers and 3,000 highly-trained and well-equipped soldiers with extensive combat experience, backed by tactical air support and transport (Singer, 2011). As a military force, EO was extremely capable and conducted a highly successful counter-insurgency against the RUF during their tenure (Howe, 1998).

[6]The New York Times (1997). Pocketing The Wages Of War. [online] Available at: https://www.nytimes.com/1997/02/16/weekinreview/pocketing-the-wages-of-war.html [Accessed 7 Jul. 2018]

UN Peacekeepers

The United Nations Mission to Sierra Leone (UNAMSIL) began arriving in Sierra Leone in December 1999. The main objective of UNAMSIL was to assist with the disarmament process and enforce the terms established under the Lome Peace Agreement (Olonisakin, 2008). Unlike other previous neutral UN peacekeeping forces, UNAMSIL brought in serious military power. At that time, the maximum number of troops to be deployed was set at 6,000. However, a follow-up UN resolution authorised the deployment of 11,000 combatants after a few months. In March 2001 that number was increased to 17,500 troops, making it at the time the largest UN force in existence (Johnstone, 2006). They were mainly deployed in the RUF-held diamond mining areas.

Despite these impressive numbers, UNAMSIL was frequently rebuffed and humiliated by the much smaller RUF. They were regularly being subjected to attacks, obstruction and disarmament. In an infamous incident in May 2001, over 500 UNAMSIL peacekeepers were captured by the RUF and held hostage. Using the weapons and armoured personnel carriers of the captured UNAMSIL troops, the rebels then advanced towards the capital. For over a year, the UNAMSIL force avoided intervening in RUF-controlled mining districts to prevent another humiliation. Only after Operation Palliser and Operation Khukri by the intervening British, the situation had stabilised and UNAMSIL regained control in Sierra Leone.

UK

In May 2000, British Paratroopers were deployed in Operation Palliser to evacuate foreign nationals and establish order in the capital. Their intervention stabilised the situation, and they were the catalyst for a ceasefire that helped end the war for good (Penfold, 2013). The British forces, commanded by Brigadier David Richards, expanded their initial mandate[7], which was originally limited to

[7]Western diplomat, personal interview, Freetown 01/03/2017.

evacuating commonwealth citizens out of Sierra Leone only. At the time, the RUF was still in control of considerable territory. The 1,200 strong ground force was further supported by air and naval power. One decisive British action was the raid against the West Side Boys, a RUF splinter group that took several British soldiers hostage outside of Freetown. The nature of the successful rescue and the lob-sided casualty figures—one British soldier against 25 WSB combatant deaths plus 18 captured including their leader—ended any lingering threat of further obstruction to peace.

6.2.2 Dynamics of Capability Shifts on the Ground

As declared in the research design, the intention of mixing methods is not to triangulate quantitative and qualitative results. Instead, the goal of the shadow case study is to bring into focus the more dynamic characteristics of capability that cannot be easily captured in observational studies. Five inter-related themes emerge out of the interviews in regards to how conventional understanding of capability can be more complex on the ground: having access to specialist equipment, re-arming the enemy through incompetence, battle discipline/training, operational mandate and military doctrine, and widely-shared beliefs in the supernatural.

Specialist Equipment

One of the first points that is brought up by both the former fighters and active soldiers is the enhanced capability provided by specialist military equipment.[8] Note that specialist equipment in this specific context refers to night-vision goggles and smoke bomb coverage provided by Chinook helicopters, and not high-calibre weapons and ammunition, artillery, or other air strike capabilities that is available to the military at large. Abu Bakarr Jaward, who fought for the RUF for three

[8]RSLAF Major, personal interview, Makeni 09/03/2017; Abu Bakarr Jaward, personal interview, Makeni, 09/03/2017.

years before integrating into the Sierra Leonean Army and later received training from the British military advisors, explains the added capability gained by using such equipment:

> "Normally, you cannot just attack at day time like that. [Both] in conventional or guerilla warfare…It is difficult, you have to be tactical. But now with the specialist equipment, in the night time, you can see even the small ants on the ground. So you use your night-vision googles and you go very close to the enemy, and get rid of them. And sometimes with the specialist equipment, like the smoke bomb, you can put your enemy to sleep for some hours and you get close and disarm them. You finish your mission and pull out."

The importance of such specialist equipment seems to be the way they widen the operational capabilities available to one side alone, leading to an asymmetry. First-aid and other medical support on the ground is yet another factor that can influence capability projection, especially over time.[9]

Another influential factor associated with having access to specialised equipment pertains to target selection and troop movement. The Guinean contingent of the ECOMOG forces were known as the 'bombardiers'. They were known to be very robust with their mortars and artillery, which was their first choice of engagement. The RUF commanders had to strategise around this fact when they were fighting them[10]. On the other hand, they were also specifically targeted for their military hardware.[11]

Re-arming the Enemy

One of the consistent narratives in the Sierra Leone Civil War is the successful re-armament of the RUF through defeated enemy forces. At the onset of the

[9]RSLAF Major, personal interview, Freetown 03/03/2017.
[10]Big Daddy, personal interview, Makeni 09/03/2017.
[11]Abu Bakarr Jaward, personal interview, Makeni 09/03/2017.

insurgency, the size of the RUF was believed to be around 100 fighters in total. The security establishment at the time was not sure whether this local uprising would take hold beyond its immediate proximity, and subsequently not much attention was given to them. However, even though they came really close to be military defeated by the SLA in 1992, the RUF persevered—mostly owing to the Strasser regime not challenging them in the provinces—and became stronger and stronger by obtaining arms and equipment from the intervening forces.

A senior RUF commander, going by the nom-de-guerre 'Big Daddy' who first fought in the Liberian civil war and then joined RUF as an experienced commander in 1991, on his target selection strategy:[12]

> "During the time of the ECOMOG, I enjoyed fighting. Because everything I wanted, I got it from them…They have sophisticated weapons…So every time I attacked the ECOMOG, I got good weapons and ammunition.
>
> I: So they had a lot of supplies and good equipment, but the RUF would just take it from them?
>
> Yes, yes, easy. Simple to collect it from them. The first weapon I got from them, four anti-aircraft guns, mounted on pick-ups."

The South African mercenary force, the Executive Outcomes, were highly-trained and possessed advanced weaponry. However, as they were mostly protecting the mining sites, they were not on the offensive. Even though the RUF had engaged them from time to time to gain control of the mining areas, they were pushed back.[13] A security official who was in Kono during the war posits the RUF decision to repeatedly attack the EO was ill-founded, given how well-trenched the EO was at the mining sites and how difficult it would be to hold the mines once they were captured.[14]. The general perception in the security establishment was that the

[12]Big Daddy, personal interview, Makeni 09/03/2017.
[13]RSLAF Brigadier General, personal interview, Makeni 09/03/2017.
[14]Security official, personal interview, Freetown 18/03/2017.

EO did not suffer any casualties caused by the RUF during their deployment in Sierra Leone,[15] and none of their material capabilities went into the hands of the RUF.

The UNAMSIL force also ended up greatly enhancing the fighting capabilities of the rebels:[16]

> "The arms they brought in all ended up at the hands of the rebels. Tanks and all those equipment…That is also a problem for us, for the military. In a way, the [aim of the] intervention was to meant to really pacify, to calm down; instead, they feared the rebels…They have given them their armoury."

In sum, even though the military capabilities of the ECOMOG and UNAMSIL forces vastly out-matched that of the rebels, the RUF was able to capitalise on their victories and absorb the fighting capabilities of their enemies. More disciplined forces, such as the professional mercenaries of EO and the UK expeditionary force, were able to maintain control of their equipment and engaged their opponents with an advantage in weapons technology.

Battle Discipline

Military discipline and training also cited as an important determinant of force application. This heading captures reconnaissance, intelligence gathering, and unit cohesion. The British forces made an immediate impact in Freetown:[17]

> "The manner they land, you know that this is a fighting force. Because you see, when they land, when real fighting soldiers land, you come close and they tell you 'Move back! Move! Move!' They hold position, they push you. So you know this guy means something.
> I: But you could approach the UN Peacekeepers?

[15] Solomon B. Caulker, personal interview, Freetown 14/02/2017.
[16] Al-Shek Kamara, personal interview, 14/02/2017.
[17] Abu Bakarr Jaward, personal interview, Makeni 09/03/2017.

Easily. They will not reject you, they will not say 'No, no, go back'."

Battle discipline can also manifest in the apparent commitment to fight. Several interviewees brought up the point that Sierra Leoneans did not think highly of their fellow African soldiers in terms of professionalism.[18] ECOMOG forces were also perceived as corrupt,[19] and they committed a multitude of human rights violations themselves.[20] Sheka Forna, son of Dr. Mohamed Sorie Forna—former minister of finance and deputy prime minister of Sierra Leone who was executed by the regime in the 70s—motivates what he sees as differences in mental approaches to battle:[21]

> "I am not convinced by the efficacy of the African armies. A lot of individuals across Africa would join because it is a means of earning a living, as a profession. I do not think they are particularly effective fighting forces…[On part of the Nigerian soldiers] How commited were they to the fight? A British, American, European soldier would be more committed to battle than someone who joined the army out of expediency. In the UK, these are people who willingly join and know that they may be sent to the front line. I would hazard very few Africans join the army in the anticipation that they may be put in the front lines."

The EO mercenaries are also reported as having high battle discipline by the rebels:[22]

> "I fought against the EO. I found it somewhat difficult. They were trained by the South African army, they had training in bush [war]. They have the same experience…It was hard to ambush them."

[18]Security official, personal interview, Freetown 18/03/2017.
[19]Solomon Caulker, personal interview, Freetown 14/02/2017.
[20]Al-Shek Kamara, personal interview, Freetown 14/02/2017.
[21]Sheka Forna, personal interview, Freetown 08/03/2017.
[22]Big Daddy, personal interview, Makeni 09/03/2017.

In addition to the overall fighting capabilities of the EO as a unit, individually they were perceived as a serious threat as well:[23]

> "One of their commanders, he and I, we met [on the battlefield] one-to-one in Koidu (Kono) once. For every step I take and every step he take, I knew that he was [dangerous].
> I: So you were well-matched?
> Yes. We fought, we wanted to capture but they resisted and resisted and resisted."

Mandate and Doctrine

Another point that has been brought up is the capacity to use force. RUF had a very good understanding of the existing limitations and constraints (or there lack of) on their enemies and capitalised on this.[24] Foday Sankoy, the leader of the RUF, was previously a corporal in the SLA. He would monitor the radio channel frequencies that the military was using to coordinate their attacks, and relay the enemy movements to his commanders so that they can either move out of the said area or prepare an ambush.[25] Similarly, many rebels had family connections in Freetown, who provided them with information regarding the ECOMOG presence in the capital, prior to the January 6 massacre.[26]

RUF command was also aware of the UNAMSIL mandate, even before the peacekeeping force set foot in the country.[27] Even rank-and-file soldiers were aware of the fact that they were in Sierra Leone under Chapter 6 (and not Chapter 7) Omar Lebbie, former RUF fighter and now a corporal in the SLA, explains:[28]

[23] ibid.
[24] Al-Shek Kamara, personal interview, Freetown 14/02/2017.
[25] Big Daddy, personal interview, Makeni 09/03/2017.
[26] Oswald Hanciles, personal interview, Freetown 09/02/2017.
[27] Omar Lebbie, personal interview, Freetown 13/03/2017; Abu Bakarr Jaward, personal interview, Makeni 09/03/2017.
[28] Omar Lebbie, personal interview, Freetown 13/03/2017

> "They are two chapters. Chapter 6, pure peacekeeping. Chapter 7–peace enforcement. When you know that peacekeepers are coming…[peacekeepers] they never open fire until order came from the above. That's how the RUF took them in Makeni. The peacekeepers saw the RUF coming, but their mandate was not to open fire."

This is why the Makeni incident—a ragtag band of RUF fighters utterly disarming and abducting 500 Kenyan Peacekeepers—went down the way it did. Overnight, the UN lost credibility[29], perceived as 'pushovers' to the rebels,[30] and "even made ECOMOG look good".[31] This also set the table for UK to take credit,[32] which was "a masterclass in psychology"[33]. On the other hand, the British had their hands forced into action:[34]:

> "It was a blessing in disguise that they [West Side Boys] captured British troops. To us, it was a blessing. Because that was what caused the justification for the British intervention."

However, even it may be the case that UK was forced into action or experiencing mission creep—it was mentioned out that the military commander on the ground used his initiative for the operation—[35]this was one of the outcomes they had considered in their contingency plans. A Western diplomat with close ties to the UK High Commission in Sierra Leone shares that at the time, the UK government was not "risk-averse" in regards to the military intervention, and they were prepared for "acceptable losses".[36] This belief was also mirrored in the UK:[37]

> "There is a different dynamic between the US and the UK. US seems to be much more concerned with casualties than Britain is. Had the

[29]Francis Stevens George, personal interview, Freetown 07/03/2017.
[30]Solomon Caulker, personal interview, Freetown 14/02/2017.
[31]Security Officer, personal interview, Freetown 15/03/2017.
[32]Dr. Ibrahim Bangura, personal interview, Freetown 04/02/2017.
[33]Dr. Henry Mbawa, personal interview, Freetown 10/02/2017.
[34]Al-Shek Kamara, personal interview, Freetown 14/02/2017.
[35]Western diplomat, personal interview, Freetown 01/03/2017.
[36]Western diplomat, personal interview, Freetown 01/03/2017.
[37]Sheka Forna, personal interview, Freetown 08/03/2017.

> result been the same, as in Britain was engaged in Sierra Leone and had had brought peace to the country, if that had been at the cost of 15 [British] military personnel, I think that would be acceptable…Britain seems to be prepared to accept the unfortunate death of individuals as a part of engaging militarily. Had there been a tangible result, the British public would have accepted that."

Finally, the multi-national character of the ECOMOG force contributed to the lack of clarity in their military doctrine. Ret. Colonel Simeon Sheriff highlights what he sees as doctrinal differences stemming from culture:

> "Culture is very important here [Africa]. For instance, when I went to the UK, we also went to Holland, Belgium…It was just like we were moving in one country. Here in Africa, when you move from one country to another, you see differences. A lot of differences. That is the aspect of doctrine. What, in effect, doctrine is really about *what do we believe in* and what type of equipment we think we should buy [and training]."

Supernatural Abilities

Finally, material capabilities also suffered shifts caused by a more intangible factor. Komojors, ethnic hunters from the Mende tribe dominating the south, are believed by some to possess supernatural powers. One of these powers include being bulletproof:[38]

> "They told us boys that, 'If I wash your body and I hit you and your body, when I shoot you–you won't die.' This is how many of them died."

and[39]

[38]Abu Bakarr Jaward, personal interview, Makeni 09/03/2017.
[39]Big Daddy, personal interview, Makeni 09/03/2017.

> "They have traditions…They have talismans hanging on them. They say they were protections against the bullets. So when you confront your enemy, the bullet will not kill you. They deceived many, many men. So many people."

A more grounded explanation was given: The rebels were terrible marksmen, often shooting at targets at a distance of 150-400 meters, resulting in a very low accuracy which contributed to the perception of invincibility.[40]

Another commonly-shared Komojor ability was teleportation. The Komojors knew the terrain quite well owing to their life-long profession as hunters, and they were apt at utilising short-cuts and trails unbeknownst to the rebels. This gave the illusion that they can appear and disappear at will.[41]

Finally, the Komojors could 'sense' who was a rebel. Al-Shek Kamara, Assistant Inspector General of Police in Freetown who dealt with Komojor-related cases during the war, provides an example:[42]

> "[the Komojors] were in Freetown, when the attack was imminent [January 6 massacre]. They confessed to possess some powers, spiritual powers, that they could detect rebels…They got a hold of one man, they said 'We can see this man is a rebel', and shot him. I knew he was an innocent man. He had never been to the provinces. But this [type of behaviour] was accepted, because the Komojors [are believed to] possess powers."

Komojors used to eliminate individuals that they have prior beefs[43] with by 'identifying' them as rebels so that they can be killed and their properties can be looted.[44]

[40] ibid.
[41] RUF ex-combatant, personal interview, Freetown, 16/02/2017.
[42] Al-Shek Kamara, personal interview, Freetown 14/02/2017.
[43] Dr. Ibrahim Bangura, personal interview, Freetown 04/02/2017.
[44] Oswald Hanciles, personal interview, Freetown 09/02/2017.

Even though the Komojors eventually increased their numbers—estimates given as 15,000—and were provided weapons (AK-47s, RPGs)[45] by the government after their official transition into the Civil Defence Force (CDF),[46] the belief in their supernatural powers did affect the strategic target selection of the RUF and the ECOMOG forces (Komojors at times fought against both sides in the war) when they lacked high levels of material capability. As a result, they were able to hold onto territory beyond their actual capabilities. On the other hand, they were instances where their professed powers also made them high-priority targets. More secular RUF commanders would round them up upon capture and mass-execute them in front of their troops to prove they are indeed mortal and susceptible to gunshots like any other enemy.[47]

6.2.3 Insights from Integrative Mixed-Methods

As alluded in the research design chapter, certain empirical set-ups of multi-method studies can lead to holistic insights when data generated from two different paradigms are combined. In this section, I look at the determinants of conflict duration in the Sierra Leone Civil War with evidence from two approaches: case study insights and the LIME procedure. The utility of doing so is two-fold. First, the underlying processes are traced and clearly linked to the theoretical model of conflict duration; second, classification predictions of the machine learning algorithms are interpreted from a duration perspective to provide insights from learning conflict duration in a predictive modelling framework.

The proposed theory posits that duration is a function of the length of the iterated bargaining game, which begins with a power distribution q and terminates when the status reaches p. This is so, because the ability to project power is the crux

[45]Solomon Caulker, personal interview, Freetown 14/02/2017.
[46]Bangaly Monorma Bah, personal interview, Freetown 02/03/2017.
[47]Big Daddy, personal interview, Makeni 09/03/2017.

of the model. Further, the directional expectations inform us that capacity and projection go hand-in-hand; their interaction (both within actor and dyadically) provides insights into the duration function.

The case of the Sierra Leone Civil War is one of low-capacity and low-projection. Both the government forces and the RUF had limited means to wage war. Material capabilities were scarce on both sides in absolute terms; however the SLA had the numerical advantage in terms of manpower.[48] When the conflict was initially under-way, the SLA outnumbered the RUF 30:1. In fact, the RUF was at the brink of defeat by the end of the second year of the insurgency.

Several factors were crucial as to how the RUF did not succumbed there and then. First, coup leader Strasser—the youngest head of state at the age of 25—was happy to stay in Freetown. From hi perspective, there was no political incentive to chase after the rebels when they have no credible means of threatening the capital. Further, the conflict provided plausible cover for the government to take part in the lucrative diamond trade. Many high-level government officials were involved in the mining business in the east. These factors, both top-down, limited any further military action against the rebels when they were down.

Second, eventually the RUF was able to *gain* material capability beyond what was otherwise not available to them to attain. The intervention forces of ECOMOG and later UNAMSIL were equipped with equipment that were undisputedly superior to that of the RUF's. However, owing to their knowledge of terrain and the discrepancy between the rules of engagement between the RUF and the intervention forces, the RUF was able to commandeer arms, supplies, and even armoured personnel carriers. Doing so greatly enhanced the fighting capability of the rebels, especially vis-a-vis the government forces. Once the rebels were strong enough, they stormed the capital and sacked in 1996—an otherwise unthinkable development previously.

[48]It must be noted that, even though the SLA was more sizeable than the RUF, both sides lack sufficient training and discipline for the majority of the conflict.

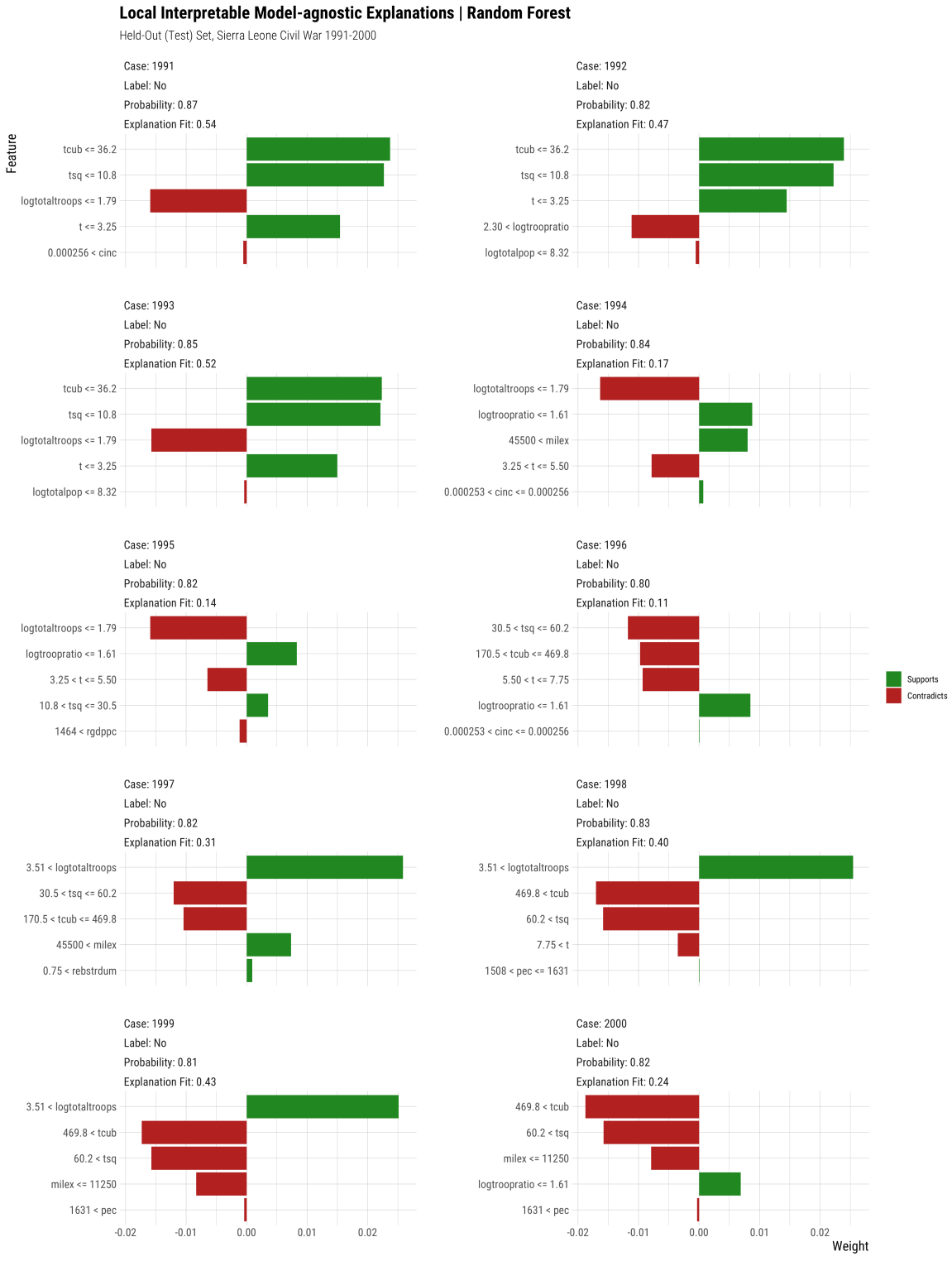

Figure 6.6: LIME Random forest: Sierra Leone Civil War Predictions 1991-2000

Moving onto predictive modelling, so far studies using time-series data had been transformed into classification problems as machine learning development is lacking when it comes to duration analysis. Before concluding the empirical analysis, I offer additional insights by reverse-engineering the classification findings and interpret them from a duration point of view.

In line with the concept of *complementarity*, Figure 6.6 displays the LIME classification predictions for all years of the Sierra Leone Civil War. First takeaway from the graph is that the algorithm always predicts continuing war—getting it right for the first nine years but resulting in a false negative in the tenth. Additionally, all predictions are backed up with about 80% probability, so the covariates strongly favoured the forecasts.

Excluding the effect of time (cubic splines) given that they are not theorised as a model component, the most important covariates utilised in the prediction of the Sierra Leone Civil War are the logged total number of troops, troop ratio, and absolute material capability indicators (CINC, military expenditures, GDP p.c., Primary Energy Consumption).

More specifically, we see that when the logged total number of troops is ≤ 1.79, they contradict the predictions of continuing conflict. In contrast, when it is ≥ 3.51 in the latter half of the conflict, they now support the predictions of prolonged war. Further, the logged troop ratio parameter supports the null predictions when it is ≤ 1.61, however it contradicts when its value is ≤ 2.3 in 1992. Finally, the infinitesimal CINC value of $\leq .000256$ places the government Sierra Leone towards the very bottom of absolute military capability rankings.

Taken together,[49] the interaction of the very low capacity of the Sierra Leonean government and their relatively low projection (except for the first couple of years when they had the numerical advantage) levels results in shorter duration

[49]As increasing logged values of total troops and troop ratio represent high projection and high relative capability, respectively.

predictions. This is especially true for the first three years of the conflict (1991-1993), and indeed, the RUF barely recovered in this period thanks to the lack of political willingness on Chairman Strasser's part.

6.2.4 Conclusion

Even though material capabilities and political constraints are shown to have predictive power, the in-depth study of the Sierra Leonean case of military interventions reveal multiple pathways that can lead to capability shifts that cannot be captured by observational studies. Furthermore, they provide a novel insight. Material capabilities, once brought to the ground, can end up enhancing the fighting capacity of the enemy if captured. On a theoretical level, the implication is that there is more to be modelled beyond absolute and relative capabilities of the actors themselves, but also the level of capacity that is available on the ground which can be captured and utilised by others. Such conditional interactive effects are difficult to capture using observational data, given the currently achievable data-granularity levels in conflict research.

To sum up, unsecured material capability can be appropriated by the enemy, making them a more robust fighting force. The RUF started their rebellion with minimal material capability in 1991, but after years of leeching Nigerian and Guinean military hardware, they were able to storm Freetown in 1997. Similarly, the UN peacekeeping force, at the time the largest of its kind in the world, ended up bolstering the RUF fighting capacity even more as a result of their constrained rules of engagement. Only when the British conducted several successful military operations the UN could regain control of the military situation.

Further, the added value of greater weapons technology is not limited to conventional artillery support or air strike capabilities. Certain types of military hardware can enhance the overall capability of a fighting force more than others. Specialist equipment that make otherwise-impossible maneouvers—night-vision

and smoke screens—allows for a wider range of tactical options to be taken against the enemy.

Second, training and mandate can make or break an intervention, regardless of absolute material capability. The Nigerian-lead ECOMOG forces numbered at 16,000. They were backed up by jet fighters, attack helicopters, and armoured personnel carriers. The Guinean contingent supported mortar and artillery support. However, the Nigerian soldiers were not fully committed to the fight, and sidelined their mission in order to pursue economic activities. The UN peacekeepers were embarrassed by the rebels when they were abducted in broad daylight as a result of their restrictive rules of engagement. The British, on the other hand, fielded the smallest amount of troops in comparison to the other international actors. Still, their battle discipline and military doctrine—combined with the timing of their intervention—allowed them to discredit their opposition and prevail militarily.

Finally, widely-shared regional beliefs can affect the performance of various international actors and how they choose to mobilise their military capabilities. Certain traditions and beliefs were shared between the local Sierra Leoneans and the intervening Nigerian forces. The latter, in some cases, chose not to engage such enemies. In contrast, the actions and the strategy of both the UN peacekeepers (predominantly Kenyan) and the British forces were not influenced by such factors. One implication of this phenomenon is that local approaches such as 'African solutions to African challenges' (Duursma, 2015) can have unintended side effects and there could be a trade-off between regional solutions and extra-regional interventions.

6.3 Conclusion

This chapter brings together various empirical findings to identify and explore the structural determinants of conflict duration. The results are leveraged and explained in a self-contained manner, as empirical triangulation is not the underlying goal of this chapter. Instead, I summarise the main findings under three headings: the examination of the proposed theoretical framework, the observation of the commonalities between the two types of war, and the synergistic insights borne out of utilising predictive modelling and case studies together.

First, there is ample evidence that the proposed theory of conflict duration encapsulates the most predictive set of variables that are out there. Various operationalisations of material capability, political constraints, and geographical factors are frequently selected as informative covariates in explaining war duration. While the empirical counterpart is not as parsimonious as the theoretical framework, it is evident that more than one operationalisation of each aspect is required to achieve high predictive accuracy. As the quality of the data gets better, it is possible that the parsimony of the theoretical approach can be implemented empirically as well.

Second, the predictive modelling findings suggest that the vast majority of the predictive variables behave the same way in both civil and interstate war settings. The implication of this insight is that armed conflict can be modelled using a unitary framework that does not discriminate between the types of warfare. So far, the literature on conflict, both in regards to onset, termination, and duration, are bifurcated by conflict type. These results should further the agenda of combining interstate and civil war studies under a common theoretical and empirical umbrella. Future research has several avenues. Most importantly, more joint data collection is required. One of the most challenging parts of this research was putting together a compatible dataset that includes variables for both types of

war. Another path forward is to understand where the divergences occur. Some covariates have indeed opposite effects depending on war type. The source of these incompatibilities should be unpacked to better understand the underlying data-generating mechanism.

Third, bringing together predictive modelling and case study findings together generates additional insights that would not be as revealing if only one method was utilised. While the quantitative aspects pertaining to actor capability is captured using observational data, the perception of such capability is captured via semi-structured interviews. Narratives borne out of such consultations highlight several qualitative nuances. Some actors—ECOMOG, UNAMSIL—who would appear materially 'capable' on paper were thought to be not so. In contrast, ethnic bands of hunters known as Komojors were widely respected and successful, as the common belief (which was crucially shared by both sides in the war) was they had supernatural powers. The Komojors were able to exploit this perception of themselves and over-achieved in the battlefield. Finally, the UK, even though having contributed the least amount of troops internationally, managed the get the lion's share of the credit for ending the eleven-year civil war. The British, in addition to their successful PR narrative, accomplished this feat via military precision and discipline, which helped them achieve the largest yield with the minimal capacity required for the task (Ucko, 2016).

Chapter 7

Conclusion

The divided study of conflict based on war type is currently at a state that is ripe for consolidation. Decades of theorising on interstate war situated in the vast international relations literature, combined with the rich and diverse empirical findings borne out of civil war studies in the last twenty years, can be a powerful combination together. However, we have yet to fully harness the potential of this union.

This project sets out to challenge the widely-adopted notion of studying civil and interstate wars separately by demonstrating that a unitary model can successfully capture the most important predictors of armed conflict duration. To do so, I posit a simple general model, built by aggregating the most consistent quantitative predictors of war longevity, comprised of three main components: material capability, and the physical and non-physical constraints acting on it. Conceptualising the sustained effort to continue fighting as a function of successful power projection in a world of limited resources, I argue for a dynamic model of limitations that can help explain the temporal variation found in war. This does not mean there are no qualitative differences between civil and interstate wars; rather, it is an extension of the notion (Lake, 2003) that the extant models in the literature should be applicable to both types of conflict.

7.1 What Have We Learned?

In support of the general theory, the empirical findings shed light on various determinants of war longevity exerting influence in both types of conflict. More importantly, the vast majority of predictors behave similarly in both conflict settings; whatever prolongs one type of conflict also increases the duration of other. On the civil war front, this is perhaps not terribly surprising; however, the project still makes an empirical contribution to the literature: similar to that of Hegre and Sambanis (2006), but instead using algorithmic predictive modelling on BTSCS data by doing a sensitivity analysis of common civil war duration predictors. On the interstate war front, this project is the first of its kind to leverage covariates identified in civil war literature and use them explicitly to predict interstate war duration. The following three paragraphs provide more specifics as to how certain covariates influence the duration of political violence.

Several material capability indicators come out as top predictors of armed conflict duration. Standard measures of state capability—Composite Index of National Capability (CINC) and Gross Domestic Product per capita (GDP p.c.)—are shown to increase conflict duration. Further, similar effects on longevity are also caused by having large civilian populations, high number of military troops, and increased military spending. Taken together, the empirical findings demonstrate that actors with higher material capabilities are associated with longer wars. This is in line with the proposed theory that treats capability as a resource pool that can be 'spent' on power projection.

Further, natural resources—specifically hydrocarbons (crude oil and natural gas) and valuable gems—influence conflict longevity: while the presence of oil is linked to a shorter wars, gems are associated with longer conflicts. This divergence of the effect direction is also found by Lujala (2009) on conflict onset.[1] It should be

[1]However, other than being more intense in terms of battle-related deaths, she also notes that oil conflicts are generally longer.

also noted that even though the effect of oil reported here is based on its median value, there are numerous cases associated with shorter and longer wars.

Moving on to the second theoretical component, politics operationalised as non-physical limitations of the continued application of military force, encapsulates several predictive covariates that are common to both types. Political constraints on the head executive, either as a composite measure by itself or via categorical regime type (democracy) are the most influential covariates in this class that lead to shorter wars. This demonstrates the dampening effect of power projection through non-physical means and mirrors the extant findings in the literature (Stam, 1996; Bennett and Stam, 1996), even though the advantages of democracy are shown to be declining over time (Bennett and Stam III, 1998). Further, it suggests that a similar mechanism underlying the canonical example of U.S. withdrawal from Vietnam is applicable to conflict at large.

In contrast, conflicts associated with military coups are slightly prolonged in duration. In interstate settings, this can be attributed to either diversionary war (Miller, 1999), rally-around-the-flag effect (Mueller et al., 1973; Baker and Oneal, 2001), or to a counterbalanced military (Belkin and Schofer, 2005). With that said, again there is a large amount of outliers found in civil war cases, in which the effect of coups are in fact shortening. For example, Thyne (2017) shows that coups can act as shocks to otherwise protracted bargaining situations, aiding their termination.

Regarding geographical factors, conflicts taking place near international borders are identified as drivers of longer conflicts for both types of war. One common explanation is the transnational dimensions of war (Gleditsch, 2007) and contagion (Buhaug and Gleditsch, 2008). The finding suggests that cross-border operations do have a dampening effect on military force projection regardless of actor type. There are multiple possible explanations on why this is the case (also see Forsberg,

2016): actors could be constrained by international pressure, or they might select international targets that they can rapidly defeat. This effect is more precise in interstate settings, as certain civil wars are prolonged as a result of having cross-border sanctuaries (Salehyan, 2009). However, the overall effect is consistent for both conflict types.

Finally, this study provides further evidence that conflict is time-dependent (Vuchinich and Teachman, 1993). The effects of time, measured in the form of t, t^2, and t^3 can be described as follows. The linear effect of time, that is t, is slightly shortening on average. Meanwhile, both the squared and the cubed transformations of time, which allows for slope changes, are associated with longer conflict durations. In terms of robustness, the linear time effect is the most commonly observed out of the three; however, this could be driven by the much-smaller set of observations containing the most protracted conflicts.

The empirical analyses conducted in this project also reveal certain limitations. Even though the existence of common predictors of armed conflict duration across war types is a novel finding in conflict forecasting;[2] these results, in the end, should still be seen as exploratory in nature. The process of replicating a large number of existing studies always comes with a margin of error. Even though a diverse set of algorithms and feature selection methods are utilised using bootstrapping procedures, only a representative set of candidate studies are used. These sixteen studies were selected in order to enforce uniformity across observations, variables, and units of analysis. It is hard to conjecture how the results would look like when we expand the scope of the replications to beyond yearly BTSCS studies. Some variables may perform better when captured in finer time intervals (i.e. months, days).

Further, covariates pertaining to the geographical constraints on power projection still come out as conflict type-dependent. These findings could be driven by

[2] Cunningham and Lemke (2013) test their hypotheses using a Null-Hypothesis Significance Testing framework which focuses on in-sample explanation rather then predictive performance.

data issues: they might be flagging up incompatibilities regarding variable measurement and operationalisation across conflict types, or they could be manifestations of omitted variable bias. For example, the effect of rough terrain might be attenuated if army mechanisation or air force capabilities are accounted for. On the other hand, the divergent results could be theory-related. Distance has a different meaning across war types. For interstate wars, the ability to project force over vast distances imply immense military capabilities (Gartzke and Braithwaite, 2011). In civil war settings, increased distance between the capital and the conflict zone could hinder the effectiveness of the government response (Buhaug and Gates, 2002). In the former case, the willingness to project force over such distances could proxy for the salience of the issue for the instigating actor (Rummel, 1979). In the latter cases, the government actors might feel less pressure to quench civil strife in the periphery, especially when the conflict-ridden areas are not of primary interest along economic, ethnic, or strategic lines (Fearon and Laitin, 2011). With that said, foreign interventions can also suffer from the same and become quagmires (Taliaferro, 1998). Thus, one avenue for future research would be focusing on unpacking the conditions under which the divergence occurs.

7.2 Predictive Performance of Machine Learning

Forecasting in conflict research has been on the rise (Schneider et al., 2011). However, many forecasting applications continue to utilise traditional statistical tools adopted from the literature-standard Null-Hypothesis Significance Testing framework. Only very recently there have been attempts to broaden the range of analytical tools available to conflict researchers (Colaresi and Mahmood, 2017).

A diverse set of classification algorithms have been implemented in this project. The literature standard logistic regression was also included to act as a baseline to compare against various shallow, deep, and ensemble learning algorithms.

The complexity of machine learning algorithms should not be thought of as certain advantage; in many fields traditional statistical approaches outperform machine learning algorithms (Makridakis et al., 2018)—more complex does not automatically mean more accurate.

However, in the case of conflict duration forecasting, this study shows that random forest and extreme gradient boosting algorithms do greatly outperform (by 5-7%) logistic regression in out-of-sample predictive accuracy. These two algorithms are commonly named as top performing in many fields in science (Moisen et al., 2006; Ogutu et al., 2011; Freeman et al., 2015). Coupled with other recent studies showing that machine learning algorithms do better than traditional statistical approaches (Muchlinski et al., 2016; however see Neunhoeffer and Sternberg, 2018), this should serve as a reminder to conflict researchers that they should not limit their choices solely to generalised logistic regression umbrella of models (e.g. zero-inflated, negative binomial) and be more open-minded about introducing machine learning techniques to their methods toolbox.

In contrast, deep learning approaches using the state-of-the-art Keras front-end do not result in better predictive accuracy compared to the baseline logit. Similar to 'shallow' machine learning techniques discussed above, more complexity does not necessarily mean better performance. Instead, models with higher complexity are more prone to over-fitting given their enormous learning capacity. Neural networks are commonly cited as the most 'capable' machine learning method available (Gevrey et al., 2003). Thus, it is not surprising that they are overtaken by most machine learning algorithms.

One possible domain-related explanation for the underwhelming performance of the MultiLayer Perceptron model is that the existing BTSCS data is not 'rich' enough to leverage the full potential of the neural network. This study uses the Cunningham and Lemke (2013) dataset with added predictive covariates identified in the quantitative civil war duration literature. With about slightly less than

2,000 observations and 30 covariates, there may be not enough 'dynamism' in the data for the neural network to unpack. Instead, the neural network—regardless of many cautions taken to prevent it via altering the model architecture—resorts to over-fitting very rapidly; i.e. it learns the specific noise of the data rather than the true underlying pattern. This is not a criticism of the authors of the original datasets; rather, it is a representation of the overall state of conflict data collection efforts.

Another possible explanation is the severe class-imbalance inherent in conflict research (Cederman and Weidmann, 2017). Shallow machine learning methods are more apt at recovering the true signal in the presence of class-imbalance (Muchlinski et al., 2016). Studies using BTSCS data, by construction, will be ridden with class-imbalance problems than most other types of data formats. It is conceivable that ensemble learners such as random forest will continue to be a better fit for such applications compared to neural networks in the near future. On the other hand, given its superior performance in other fields, deep learning can be more conducive to nascent image-as-data approaches to conflict studies—for example, see Alanyali et al. (2016).

Finally, statistical models can only be as good as the underlying data. Further, the necessary step of variable operationalisation can be another source of hindrance. A shadow case study focusing on the warring factions in the Sierra Leone Civil War highlights several possible shortcomings of large-n research. Access to specialist equipment, re-arming the enemy by losing assets that are not otherwise available to them, battle discipline, nuances of mandates and military doctrine, and shared local beliefs can all cause capability shifts on the ground. These variables are unlikely to be captured fully using observational data ex-ante. However, researchers can use this type of value-added information to be more wary of their research designs and perhaps consider multi-method approaches that are aimed at minimising possible empirical blind spots.

7.3 Implications for Future Research

The effects of the bifurcated nature of conflict studies can be seen at both the theoretical and empirical levels. Nearly two decades of theory-building using the civil war template has increased our understanding of conflict dynamics greatly. However, it has also led to a several theoretical blind spots, as its practitioners only sought to explain the temporal variation found in civil wars. Similar frameworks are discussed in both civil and interstate war literatures, but explicit linkages to one another prove elusive. This lack of theoretical coherence pertaining to the study of armed conflict hinders the accumulation of knowledge, especially the coveted positivist end-goal of discovering the true underlying data generating processes.

Similarly, restricted empirical testing of such type-specific theories are by nature under-powered. Limiting oneself to only a certain subset of political violence—even though it is mostly in favour of the dominant class of events that is civil war—is akin to discarding observations from a dataset meeting a certain criteria (e.g. missingness). However, the empirical impact of excluding interstate wars (as opposed to the methodological debate on dropping missing observations in datasets) is rarely discussed. Combined datasets including observations for both types of conflict over time enriches the empirical scope and can lead to general insights applicable to all, as demonstrated by this project.

Specifically, conflict researchers should focus on the standardisation of variable operationalisations across conflict types. Some concepts are easier to proxy than others. For example, for a rebel group with explicit ethnic links as identified in the Ethnic Power Relations (EPR) Dataset (Vogt et al., 2015), the geographical area dominated by that ethnic group and its population can be the rebel equivalent of country size and population, respectively. In other cases, however, there may not be a clear cut answer (democracy) or the logic can be dependent on conflict type (i.e. the differing meaning of distance). Perhaps, there are other meaningful

operationalisations of democratic behaviour and proximity rather than regime type and the distance between conflict parties. We should strive to come up with such operationalisations for all relevant predictors of conflict that are applicable to both contexts.

More generally, joint data collection efforts should be encouraged and must be pursued as a part of a larger agenda of advancing predictive modelling in conflict research. There are two available avenues: expanding existing repositories and applying for grants for new data-gathering projects. There are pros and cons for each option. Existing conflict databases such as UCDP/PRIO and the Correlates of War project, if they can be expanded in scope, provide the easiest way of establishing a common empirical ground for war studies. On the other hand, popular datasets are usually products of many decades of established rules and tradition, making them resistant to structural changes given the relatively high entry costs.

Conversely, new data gathering projects can prioritise the inclusive scope if planned specifically for the task. With the wide adoption of open-source frameworks amongst conflict researchers, we could be approaching the 'ripe' moment for taking such an initiative. However, start-up costs associated with undertakings with this calibre of ambition necessitate, probably multiple, large grants. Given the relative scarcity of research grants in social sciences, securing a large enough starter fund might prove challenging. On the other hand, automated approaches to data collection can be a crowd-sourced alternative. Conflict researchers should be in the driving seat of gathering such data, as their domain knowledge and expertise can cut down the costs associated with ambitious data collection efforts by optimising the most time-intensive parts (i.e. identification, operationalisation) of the process.

Chapter 8

Appendix

The appendix provides additional robustness checks and sensitivity analyses pertaining to the empirical chapters. For Chapter Four, I expand on the exploration of the replication studies using BTSCS data. In addition, I demonstrate the process employed to achieve consistency among variables across multiple studies. Next, I present additional in-sample performance metrics based on elastic net and random forest model fits.

The addentum for Chapters 5 and 6 focus on the Cunningham and Lemke (2013) study, both using the original model specification and with added predictive covariates. First, I compare the performance of the sub-sampling process for the random forest model fit. Next, I provide deeper insights into the random forest algorithm by: i) unpacking the determinants of armed conflict duration using the `randomForestExplainer` package (Paluszynska and Biecek, 2017); and similarly, ii) investigating what goes into the civil war dummy variable. Finally, for completeness, I include logistic regression and survival analysis (employing Cox Proportional-Hazards) analogues of the algorithmic modelling enterprise using the NHST framework using the `survminer` package (Kassambara and Kosinski, 2018).

8.1 Chapter 4

8.1.1 Replication Studies

Figure 8.1 visualises the correlation analysis of all 16 replicated studies using BTSCS data.

Table 8.1 contains original model specifications of all 16 replicated studies. For consistency across multiple studies, some variables are renamed. These new names can be found under the 'New Label' column.

Table 8.1: Variables from replication studies (with relabels where applicable)

Replication Study	Variable	New Label
Bagozzi 2016	Malaria prevalence	
	Rebel strength	
	Malaria x Rebel strength	
	War on core territory	Sons of the soil
	ELF index	Ethnic fractionalisation
	Ethnic conflict	
	Democracy	
	Ln GDP per capita	GDDPC
	Two or more dyads	Number of actors
	Territorial control	
	Ln population	Population
	Percentage tropics	
	Africa	
Burgoon et al 2015	Media reporting	
	Human rights violations	
	UN peacekeeping	Peacekeeping
	Media reporting x UN peacekeeping	
	Territorial control	
	Rebel strong/parity	Rebel strength
	Legal political wing	
	Ethnic conflict	
	GDP per capita (log)	GDDPC
	Population (log)	Population
	Democracy	
Buhaug, Gates & Lujala 2009	Distance to capital (ln)	
	Conflict at border	

Table 8.1: Variables from replication studies (with relabels where applicable) *(continued)*

Replication Study	Variable	New Label
	Border distance	
	Rebel fighting capacity at least moderate	Rebel strength
	Gemstones in conflict zone	Gems
	Petroleum in conflict zone	Hydrocarbons
	Drugs in conflict zone	Drugs
	Mountains in conflict zone (%)	Terrain
	Forest in conflict zone (%)	Terrain
	Democracy score at onset	Democracy
	GDP capita at onset (ln)	GDPPC
Bennet & Stem 1996	Strategy: OADM	Military strategy
	Strategy: OADA	Military strategy
	Strategy: OADP	Military strategy
	Strategy: OPDA	Military strategy
	Terrain	
	Terrain x Strategy	
	Balance of forces	
	Total military personnel	Military personnel
	Total population	Population
	Population ratio	
	Quality ratio	
	Surprise	
	Salience	
	Repression	
	Democracy	
	Previous disputes	
	Number of states	Number of actors
	Year	
Collier, Hoeffler, Soderbom 2004	Income inequality	
	Missing inequality	Income inequality
	Per capita income	GDPPC
	Ethnic fractionalization	
	Ethnic fractionalization square	
	ln population	Population
	1970s	
	1980s	
	1990s	
	3rd and 4th years of war	
	5th and 6th years of war	
	7th year of war and beyond	

Table 8.1: Variables from replication studies (with relabels where applicable) *(continued)*

Replication Study	Variable	New Label
	Change in commodity price index (CPI)	
	Primary commodity exports/GDP (sxp)	
	CPI x sxp	
Cunningham 2006	Strict veto players	Veto players
	Lenient veto players	Veto players
	Coup	Coup d'etat
	Log population	Population
	Ethnic fractionalization	
	War months	
Cunningham 2010	Clearly independent interventions	Military intervention
	Quasi-independent interventions	Military intervention
	Non-independent interventions	Military intervention
	Any intervention	Military intervention
	Lootable resources	Natural resources
	Logged battle-deaths	Battle deaths
	Democracy	
	Log population	Population
	Incompatibility	
	Log GDPpc	GDPPC
	ELF	Ethnic fractionalisation
	Proportion of neighboring democracies	
	Cold war dummy	Cold war
Conrad et al 2018	Extortion	
	Smuggling	Contraband
	Extortion x smuggling	
	Territorial control	
	Mobilization capacity	Rebel capability
	Arms capacity	Rebel capability
	Coup	Coup d'etat
	International intervention	Military intervention
	Ethnic conflict	
	Ln(GDP per capita)	GDPPC
	Democracy	
	Ln(Population)	Population
Cunningham, Gleditsch, Salehyan 2009	Territorial control	
	Strong central command	
	High mobilization capacity	Rebel capability
	High arms-procurement capacity	Rebel capability
	High fighting capacity	Rebel capability

Table 8.1: Variables from replication studies (with relabels where applicable) *(continued)*

Replication Study	Variable	New Label
	Legal political wing	
	War on core territory	Sons of the soil
	Coup d'etat	
	ELF index	Ethnic fractionalisation
	Ethnic conflict	
	Ln GDP per capita	GDPPC
	Democracy	
	Two or more dyads	Number of actors
	Ln population	Population
Cunningham & Lemke 2013	Civil war	
	Peacekeeping	
	Territorial war	Territorial conflict
	Recurring war	
	Troop ratio	
	Democracy	
	Total troops (logged)	Military personnel
	Population (logged)	Population
Caverley & Sechser 2017	Ground mechanization	
	Aircraft mechanization	
	Combined arms	
	Distance to capital	
	Conflict at border	
	Border distance	
	Rebel fighting capacity	Rebel capability
	Rebels' relative strength	Rebel strength
	Natural resources	
	Rough terrain	Terrain
	Incumbent democracy	Democracy
	GDP per capita	GDPPC
	External support: rebels	
	External support: government	
	Sons of the soil	
	Insurgency	
	Post-cold war years	Cold war
Escriba-Folch 2010	Mountains	Terrain
	Forests	Terrain
	Log population	Population
	Log GDP per capita	GDPPC
	Mineral exporting	Gems

Table 8.1: Variables from replication studies (with relabels where applicable) *(continued)*

Replication Study	Variable	New Label
	Oil exporting	Hydrocarbons
	Oil production	Hydrocarbons
	Diamond production	Gems
	Ethnic fractionalization	
	Ethnic fractionalization Square	
	Contraband	
	Number of borders	
	Army size (log)	Military personnel
	Deaths/year	Battle deaths
	Ethnic war	Ethnic conflict
	Sons of soil war	Sons of the soil
	Military intervention	
	Economic sanctions	
	Sanction duration	
	Threat	
	Imposed sanction	
Nilsson 2012	Strategy: OADM	Military strategy
	Strategy: OADA	Military strategy
	Strategy: OADP	Military strategy
	Strategy: OPDA	Military strategy
	Terrain	
	Terrain x Strategy	
	Balance of forces	
	Military personnel	
	Total population	Population
	Population ratio	
	Quality ratio	
	Surprise	
	Salience	
	Repression	
	Demoracy	
	Previous disputes	
	Number of states	Number of actors
	Offense-defense	
	Balance	
Thyne 2012	Institutional constraints	
	Political constraints	
	Political polarization	
	Parliamentary	Political constraints

Table 8.1: Variables from replication studies (with relabels where applicable) *(continued)*

Replication Study	Variable	New Label
	Exec's Longevity	
	Exec party's longevity	
	Opposition vetoes	Veto players
	Battle deaths (ln)	Battle deaths
	GDP/capita (ln)	GDPPC
	Fight for gov	
	Coups	Coup d'etat
	% Forest (ln)	Terrain
Uzonyi & Wells 2016	Ln(Tenure)	
	Institutional constraints	
	Ln(Tenure) x Institutional constraints	
	Strong central command	
	Legal political wing	
	Multiple actors	Number of actors
Wucherpfennig et al 2012	Ethnic linkage	Ethnic conflict
	Ethnic linkage with included group	Ethnic conflict
	Ethnic linkage with excluded group	Ethnic conflict
	Territorial conflict	
	Strong central command	
	Legal political wing	
	Territorial control	
	Democracy	
	ln GDP p.c.	GDPPC
	ln Population	Population
	Natural resources	
	Sons of the soil	
	Ethnic linkage x Territorial control	
	Veto players	

Figure 8.1: Correlation analyses of all replicated BTSCS studies

Table 8.2: Elastic net and random forest in-sample performance metric averages

model	metric	mean	sd
Elastic Net	ROC	0.7132	0.0801
Elastic Net	Sens	0.6785	0.0952
Elastic Net	Spec	0.6318	0.1469
Random Forest	ROC	0.7290	0.0674
Random Forest	Sens	0.8900	0.1013
Random Forest	Spec	0.3255	0.2055

8.1.2 Performace Metrics

Table 8.2 displays the in-sample performance metric averages of the elastic net and the random forest algorithm used in model fitting.

8.2 Chapters 5 & 6

8.2.1 Random Forest Performance based on Sub-sampling

Figure 8.2 demonstrates head-to-head in-sample and out-sample performance of the random forest model fits. Four sub-sampling techniques are implemented: up-sampling, down-sampling, ROSE, and SMOTE.

The `ROSE` package (Lunardon et al., 2014) provides functions to address binary classification problems in the presence of class-imbalance. Artificially balanced samples are generated using a smoothed bootstrap approach which in turn aids both the accurate evaluation of the classifier in the presence of a rare class and the phases of estimation.

The Synthetic Minority Over-sampling Technique (SMOTE) oversamples the rare class by employing bootstrapping and k-nearest neighbour approaches to synthetically generate additional observations for that class. The algorithm is included by the `DMwR` package (Torgo, 2010).

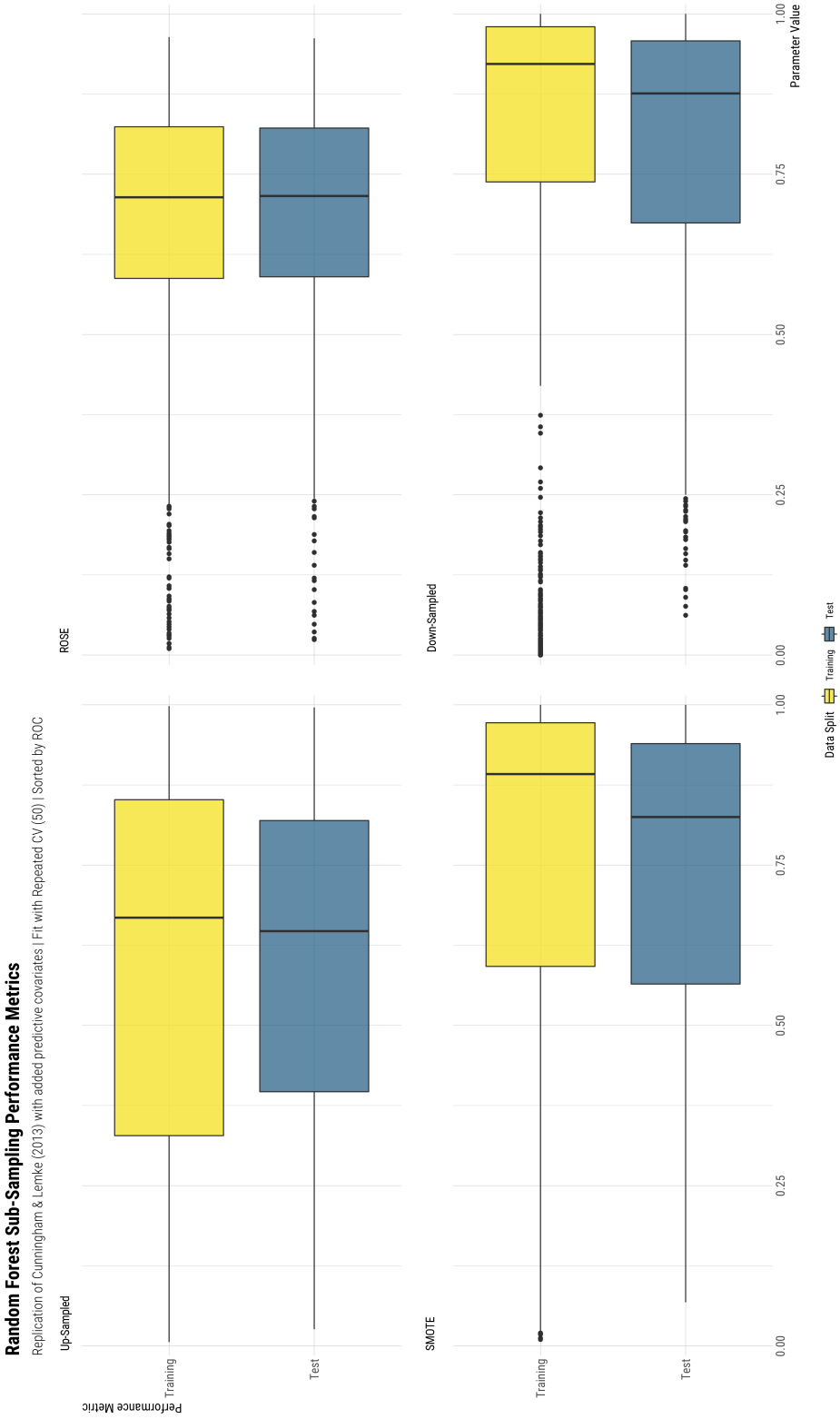

Figure 8.2: Random forest sub-sampling performance

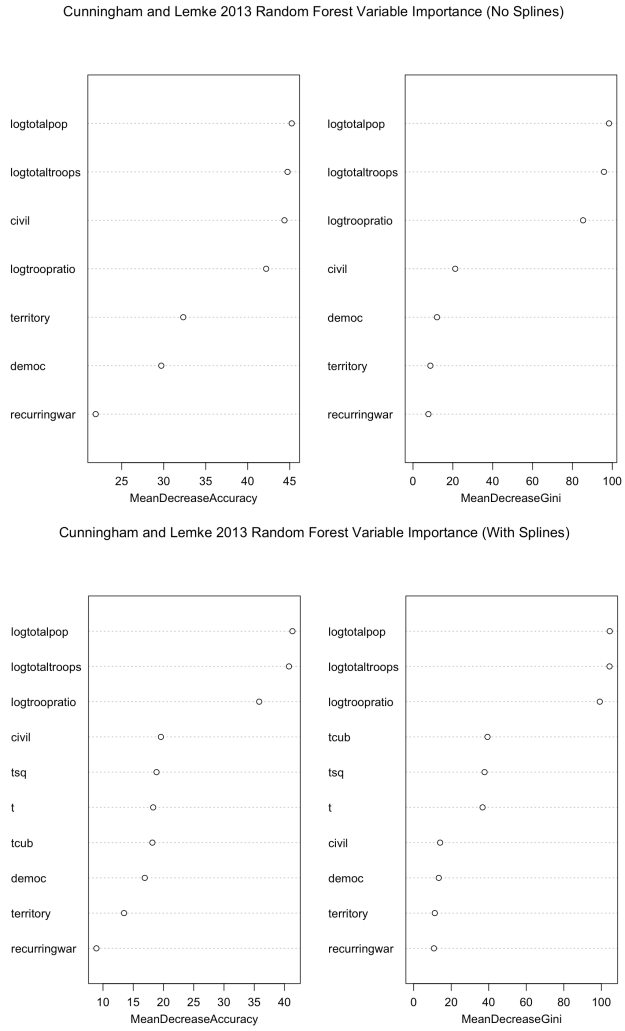

Figure 8.3: Variable importance for Cunningham and Lemke 2013 without cubic splines

8.2.2 Random Forest: Original Model Specification

Figure 8.3 shows the random forest variable importance plots for the Cunningham and Lemke (2013) study using the original model specification and with added cubic splines, respectively.

8.2.3 Random Forest Explained: Duration

Figure 8.4 plots the bilateral relations between several importance measures.

Figure 8.5 shows the bilateral relations between the rankings of variables according to selected importance measures.

Figure 8.6 visualises the distribution of minimal depth amongst the trees. The mean of the distribution is shown by a vertical bar with a value label. The scale of the x-axis ranges from 0 to the maximum number of trees in which any covariate was used for splitting. Minimal depth for a predictor in a tree equals to the depth of the node which splits on that predictor and is the closest to the root of the tree. If this value is low, this suggests a large number of observations were divided into groups on the basis of this covariate.

Figure 8.7 reports the 30 top interactions calculated by the mean of conditional minimal depth. The horizontal line displays the minimum value of the selected statistic amongst interactions for which it was derived. The interactions are considered in the following way: root variables first, and then all possible values for the second variable.

Table 8.3 shows various importance measures of the random forest fit.

Figure 8.8 plots the predictions of the random forest depending on values of components of the interaction between distance and CINC with the values of remaining predictors are sampled from their empirical distribution.

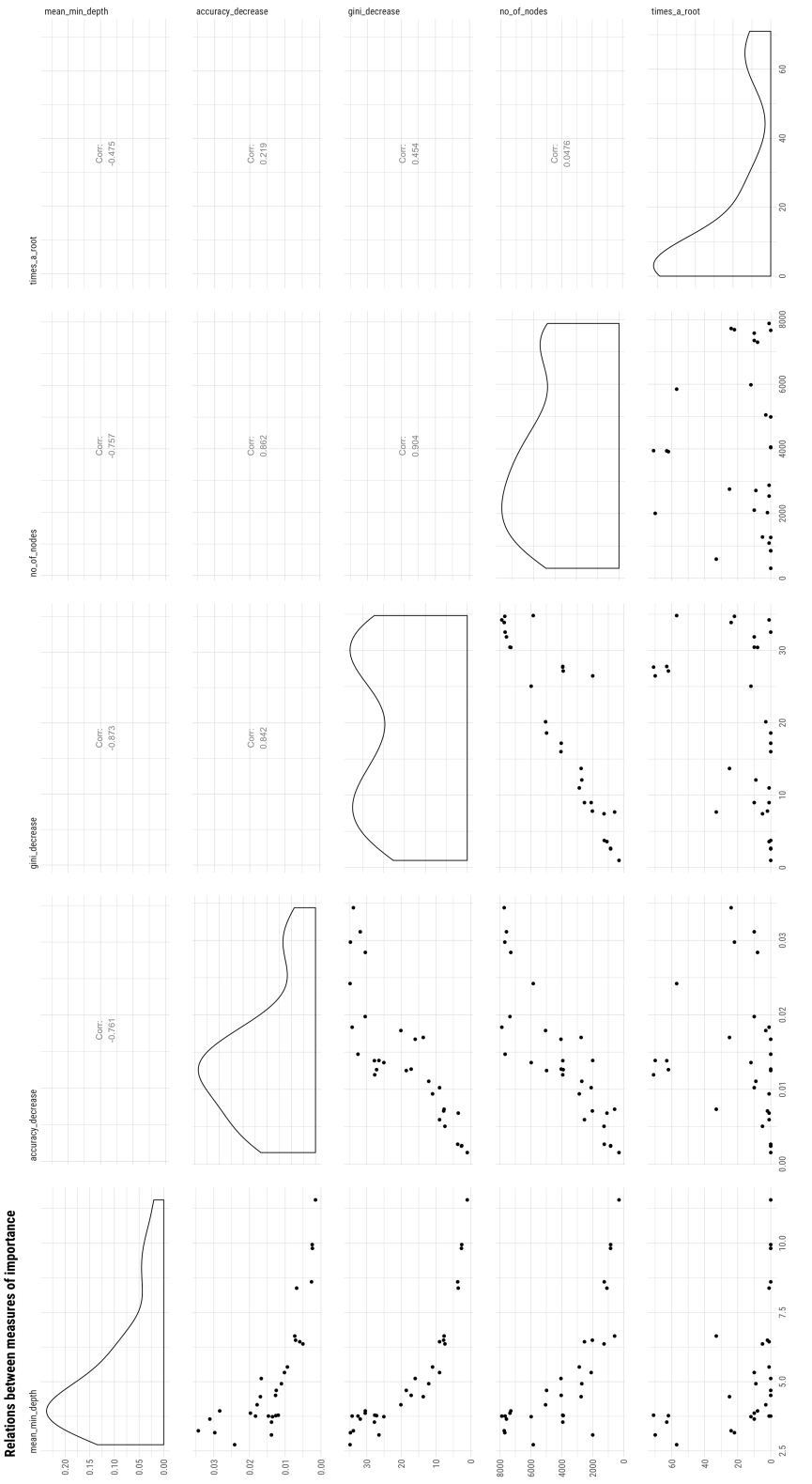

Figure 8.4: Relations between measures of importance

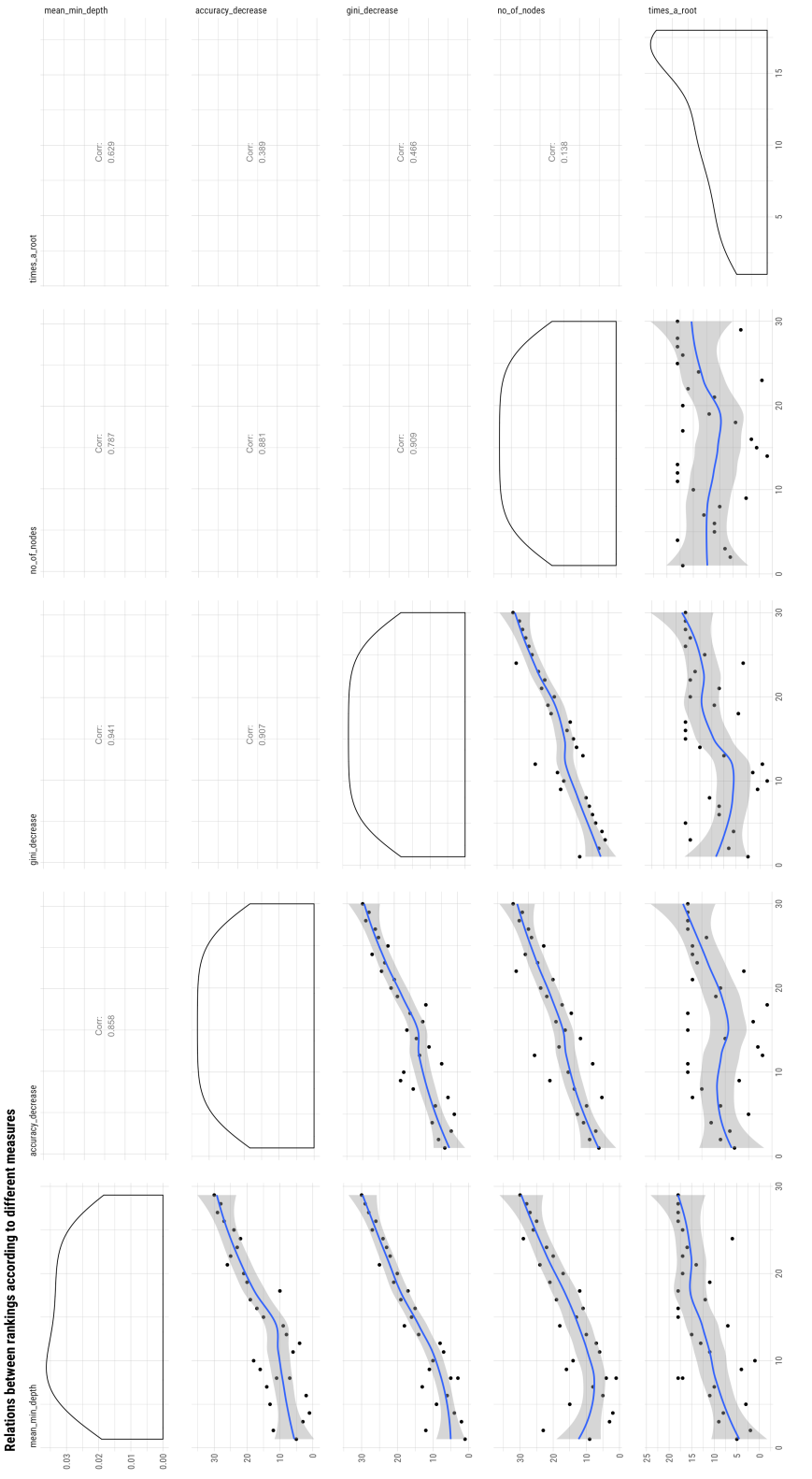

Figure 8.5: Relations between rankings according to different measures

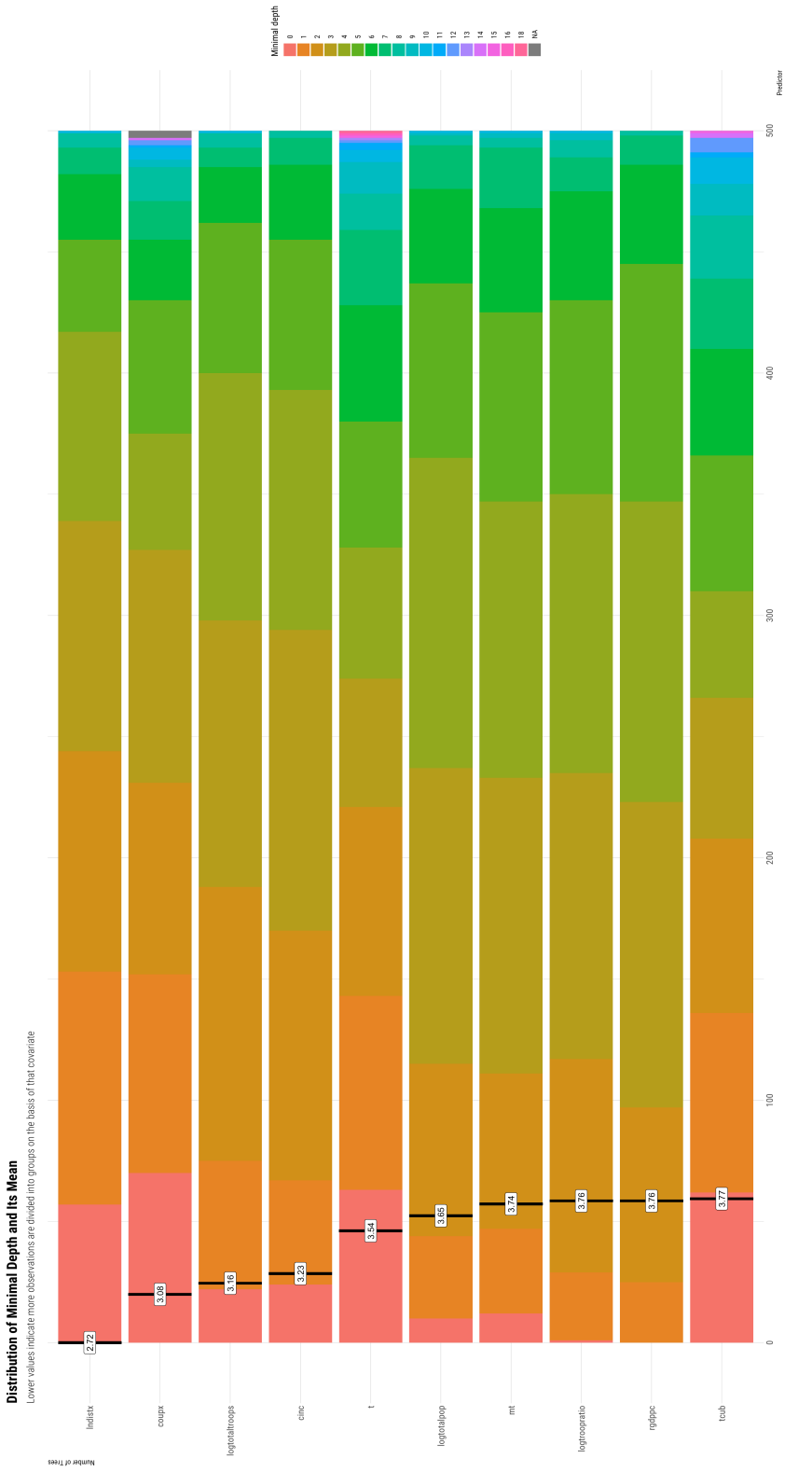

Figure 8.6: Distribution of minimum depth and its mean

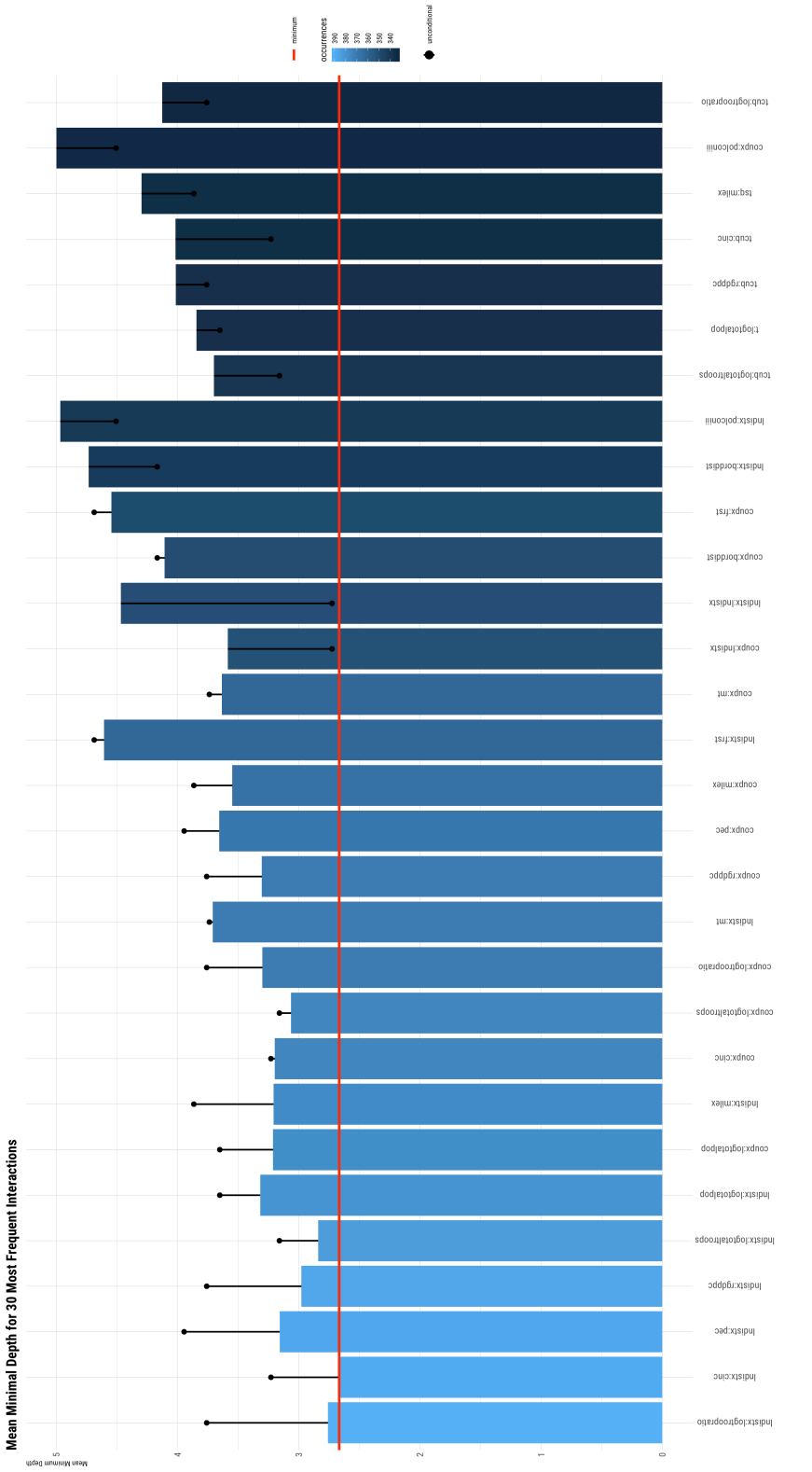

Figure 8.7: Mean minimal depth for 30 most frequent interactions

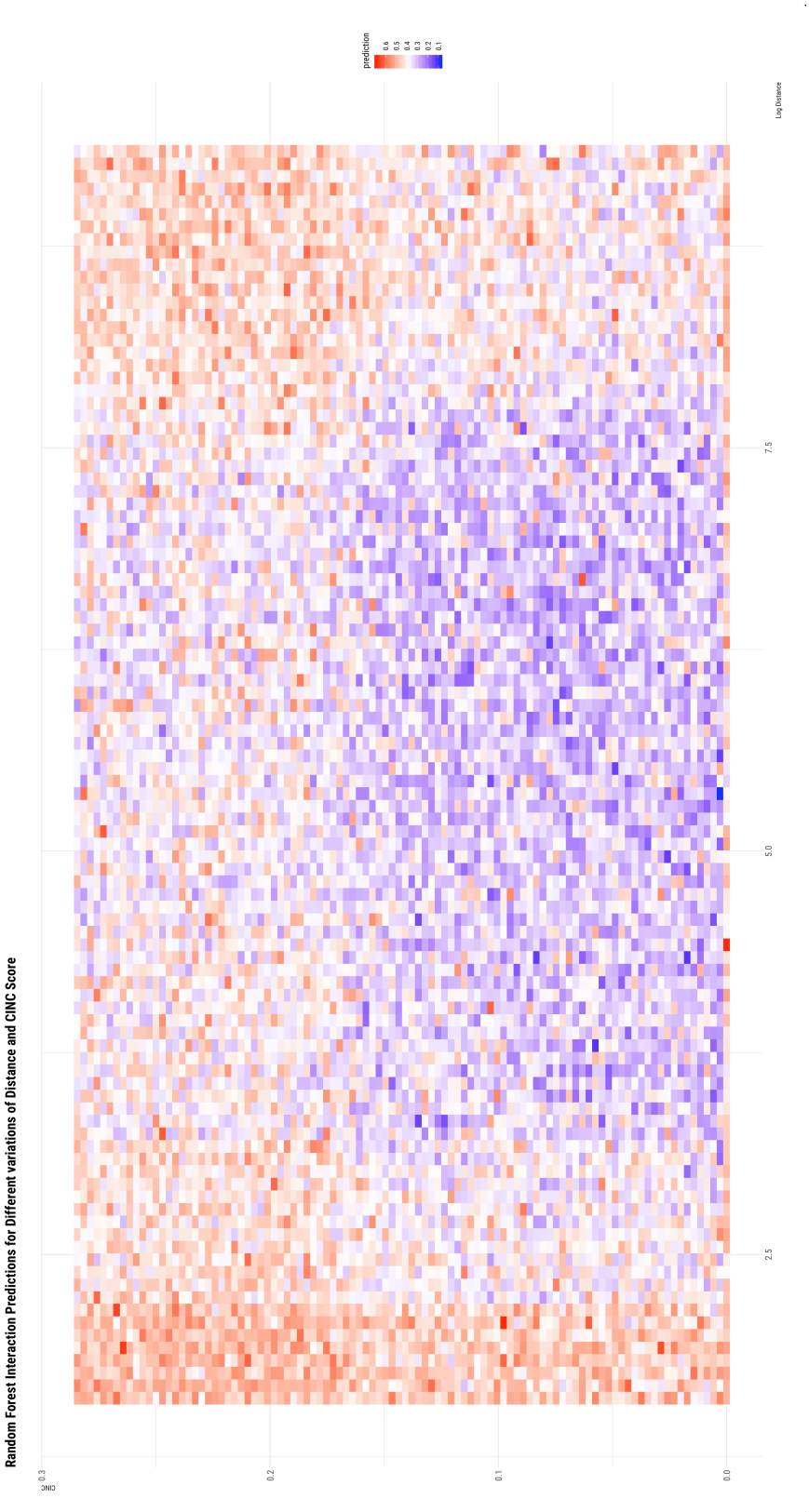

Figure 8.8: Interactive predictions for different values of distance and CINC

Table 8.3: Random forest importance measures sorted by accuracy

Variable	Mean Min. Depth	No. of Nodes	Accuracy Decrease	Gini Decrease	No. of Trees	Times a Root	p-value
cinc	3.010	9161	0.040	40.023	500	36	0
logtotalpop	3.506	8957	0.033	37.270	500	10	0
pec	3.742	8762	0.032	34.925	500	3	0
logtotaltroops	3.376	8473	0.025	35.371	500	21	0
milex	3.800	8650	0.022	33.949	500	6	0
lndistx	2.940	6339	0.021	35.409	500	36	0
irst	4.768	5149	0.020	19.496	500	0	0
logtroopratio	3.702	8517	0.017	34.853	500	1	0
rgdppc	3.806	9087	0.016	36.737	500	1	0
polconiii	4.534	4541	0.014	17.865	500	2	0
t	3.476	4425	0.014	31.745	500	72	0
tcub	3.414	4420	0.013	30.470	500	81	0
borddist	4.432	4469	0.013	17.130	500	6	0
mt	3.804	5467	0.013	22.836	500	20	0
frst	4.680	4639	0.013	16.173	500	0	0
tsq	3.482	4418	0.013	29.547	500	78	0
civil	5.027	935	0.009	12.905	479	40	1
parallel	5.440	3294	0.008	10.724	500	6	1
coupx	4.292	764	0.007	16.455	463	56	1
ALLGEMSP	7.728	1110	0.006	4.266	457	14	1
democ	8.454	1240	0.005	3.620	461	1	1
confbord	6.632	1188	0.005	6.503	472	4	1
figcapdum	8.261	1129	0.005	4.108	462	4	1
territorial	9.071	1144	0.004	3.348	452	0	1
hydroD	8.570	1328	0.003	3.717	470	0	1
alldrugs	10.382	771	0.003	2.405	396	0	1
recurringwar	7.916	1611	0.003	4.316	486	0	1
territory	9.124	1256	0.003	3.578	457	0	1
rebstrdum	8.280	932	0.002	2.645	437	2	1
major	11.278	495	0.001	1.256	319	0	1

8.2.4 Random Forest Explained: Civil War Dummy

Figure 8.9 combines two multi-way importance plots. The first figure follows the same guidelines described in Figure 6.2. The second plot displays two importance measures that were derived from the role a covariate plays in prediction. The p-value is based on a binomial distribution of the number of nodes that were split on that covariate assuming that the covariate selection process was random.

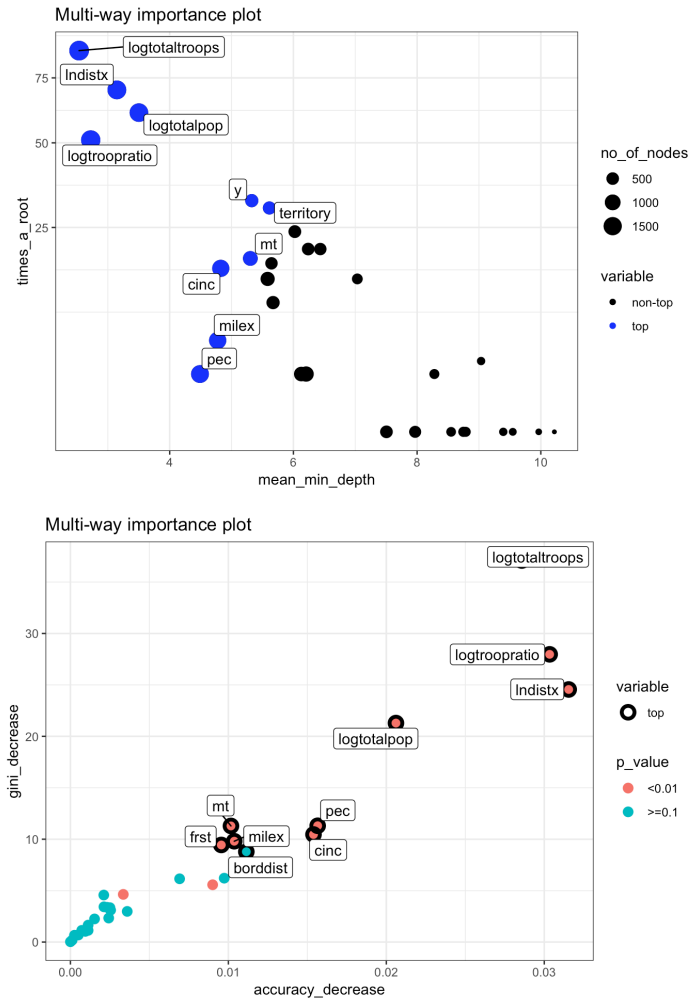

Figure 8.9: Multi-way importance plot for civil war dummy

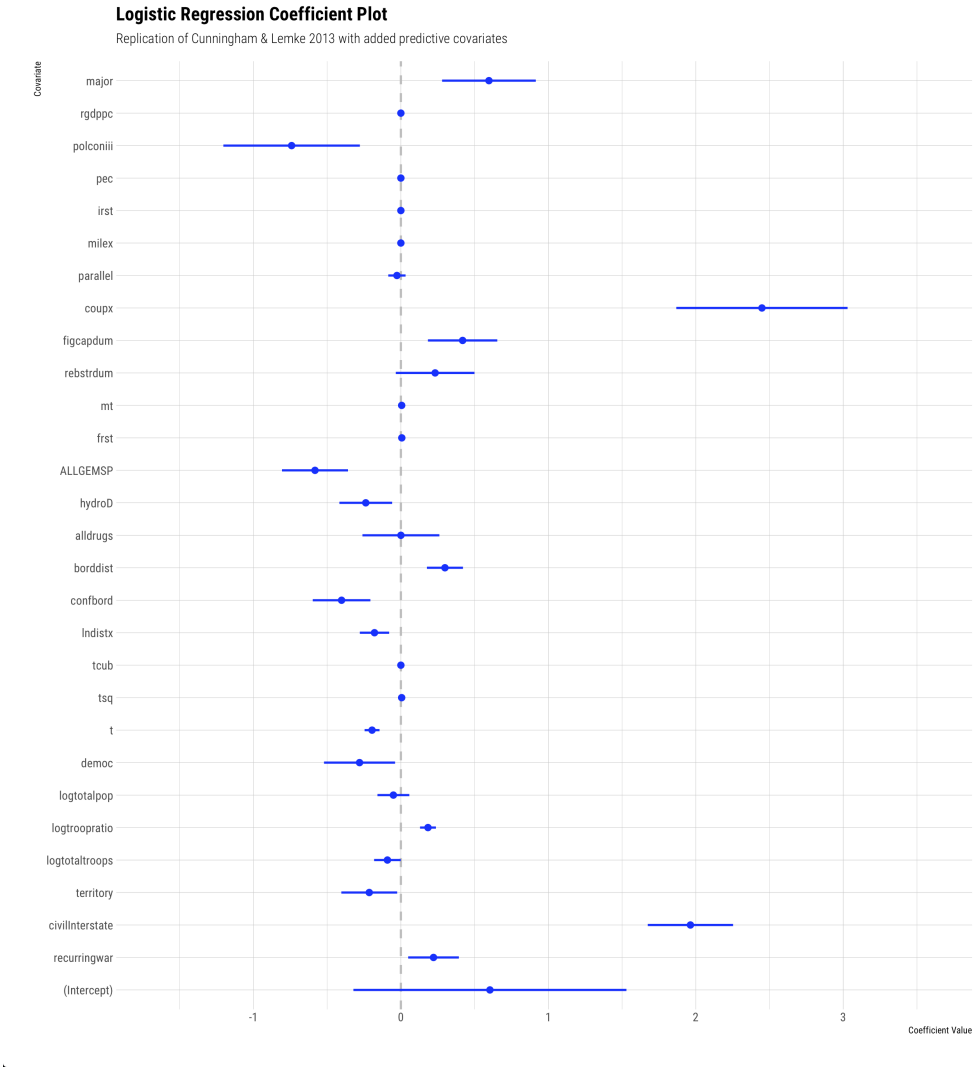

Figure 8.10: Logistic regression coefficient plot

8.2.5 NHST Replication: Logistic Regression

Figure 8.10 plots the coefficient estimates of the logistic regression analogue of the Cunningham and Lemke (2013) study with added covariates.

Figure 8.11 visualises the coefficient effects based on the logistic regression model fit.

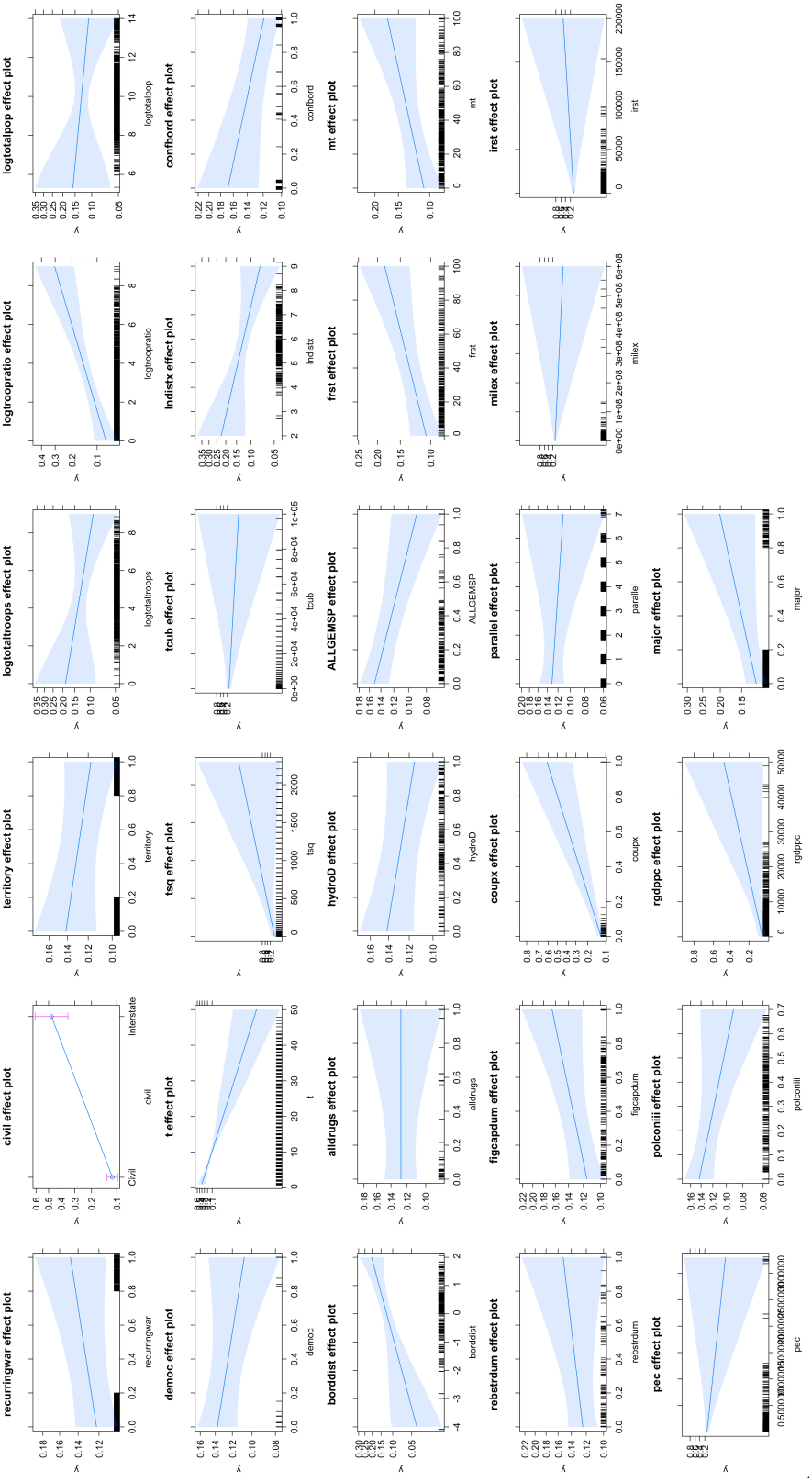

Figure 8.11: Logistic regression variable effects

8.2.6 NHST Replication: Survival Analysis

Figure 8.12 plots survival curves stratified by conflict type with 95% confidence intervals. The p-value denotes the log-rank test score. Additionally, the cumulative number of events and censors are shown in both absolute and relative terms.

Figure 8.13 displays the Cox Proportional-Hazard coefficient estimates of the Cunningham and Lemke (2013) study using the original model specification.

Figure 8.14 replicates Figure 8.13 with added predictive covariates.

Figure 8.15 shows the results of the scaled Schoenfeld residuals for each covariate in Figure 8.14.

Figure 8.16 visualises influential observations and outliers by plotting the estimated chages in the coefficient when each observation is removed in turn.

Figure 8.17 plots deviance residuals that are normalised Martingale residuals.

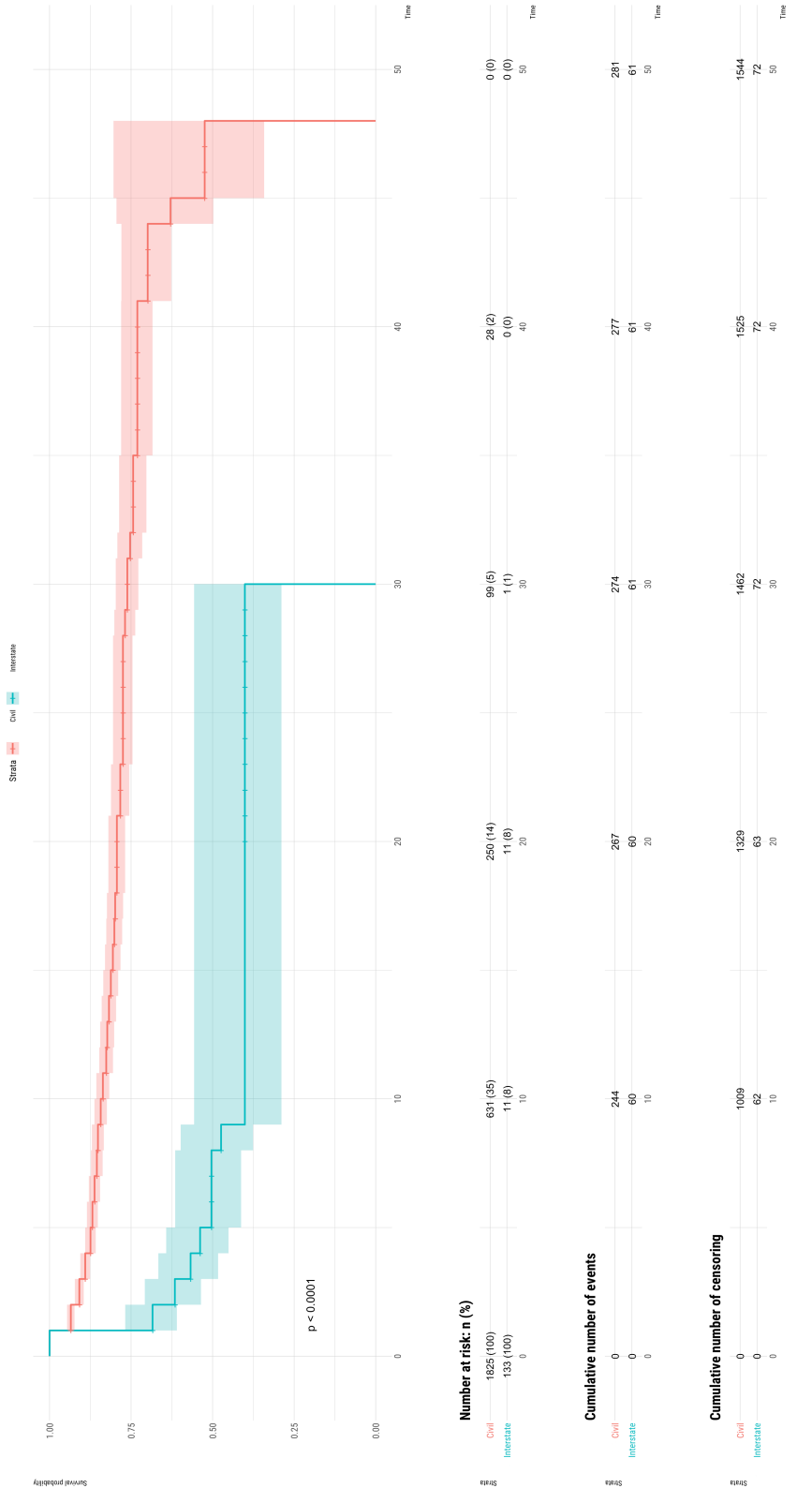

Figure 8.12: Cox-PH fit stratified by war type

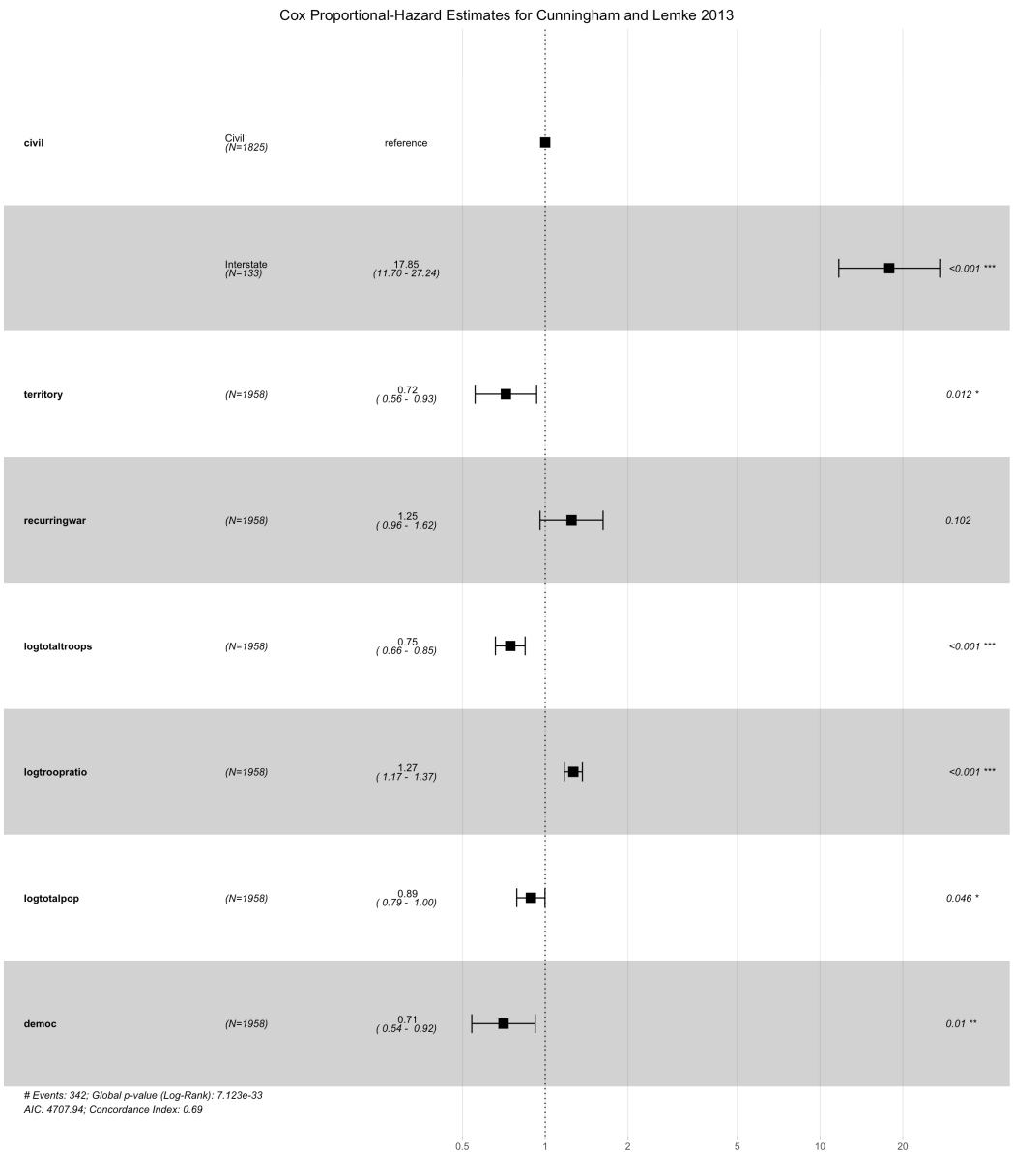

Figure 8.13: Cox-PH estimates for Cunningham and Lemke 2013

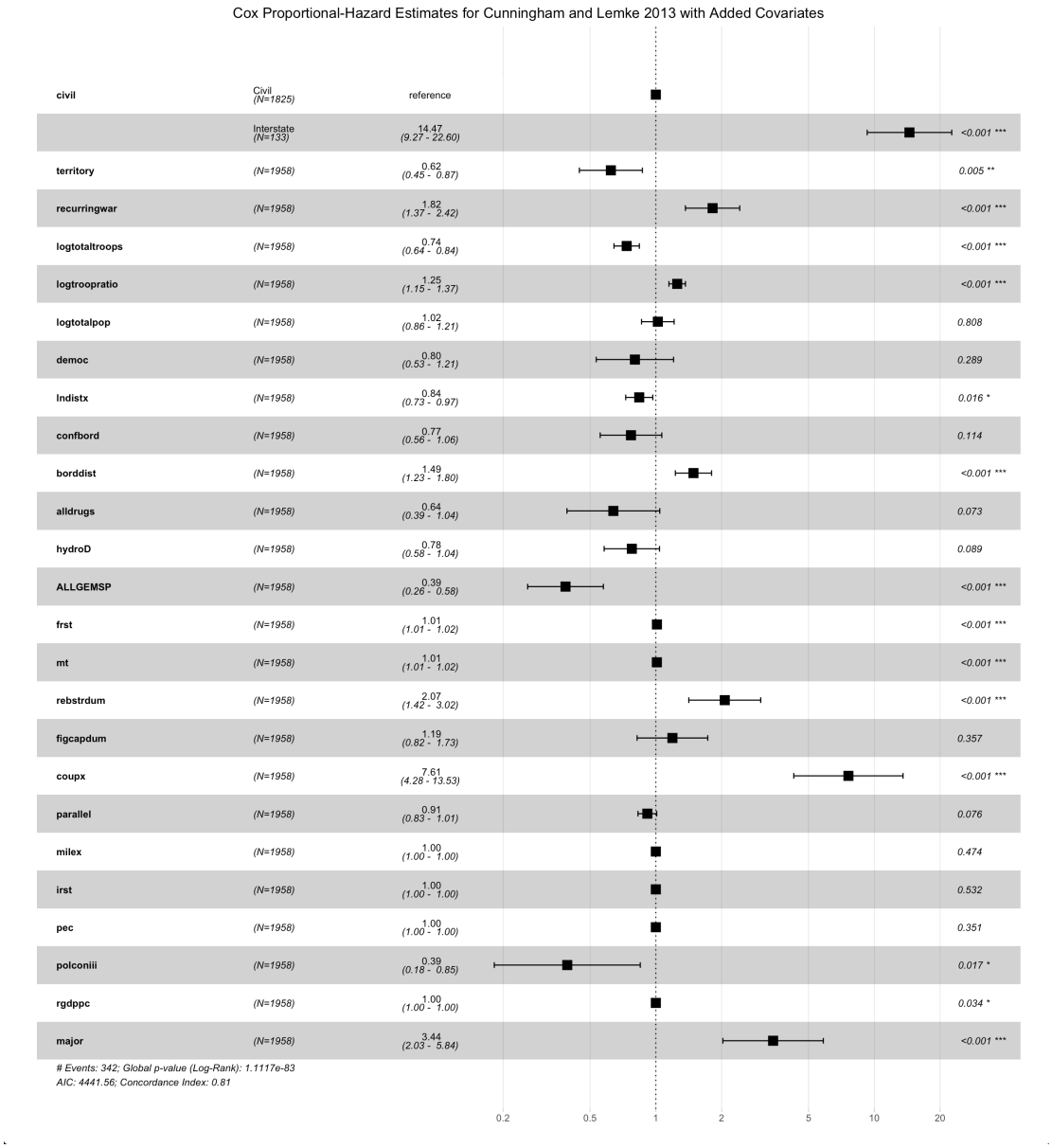

Figure 8.14: Cox-PH estimates for Cunningham and Lemke 2013 with added covariates

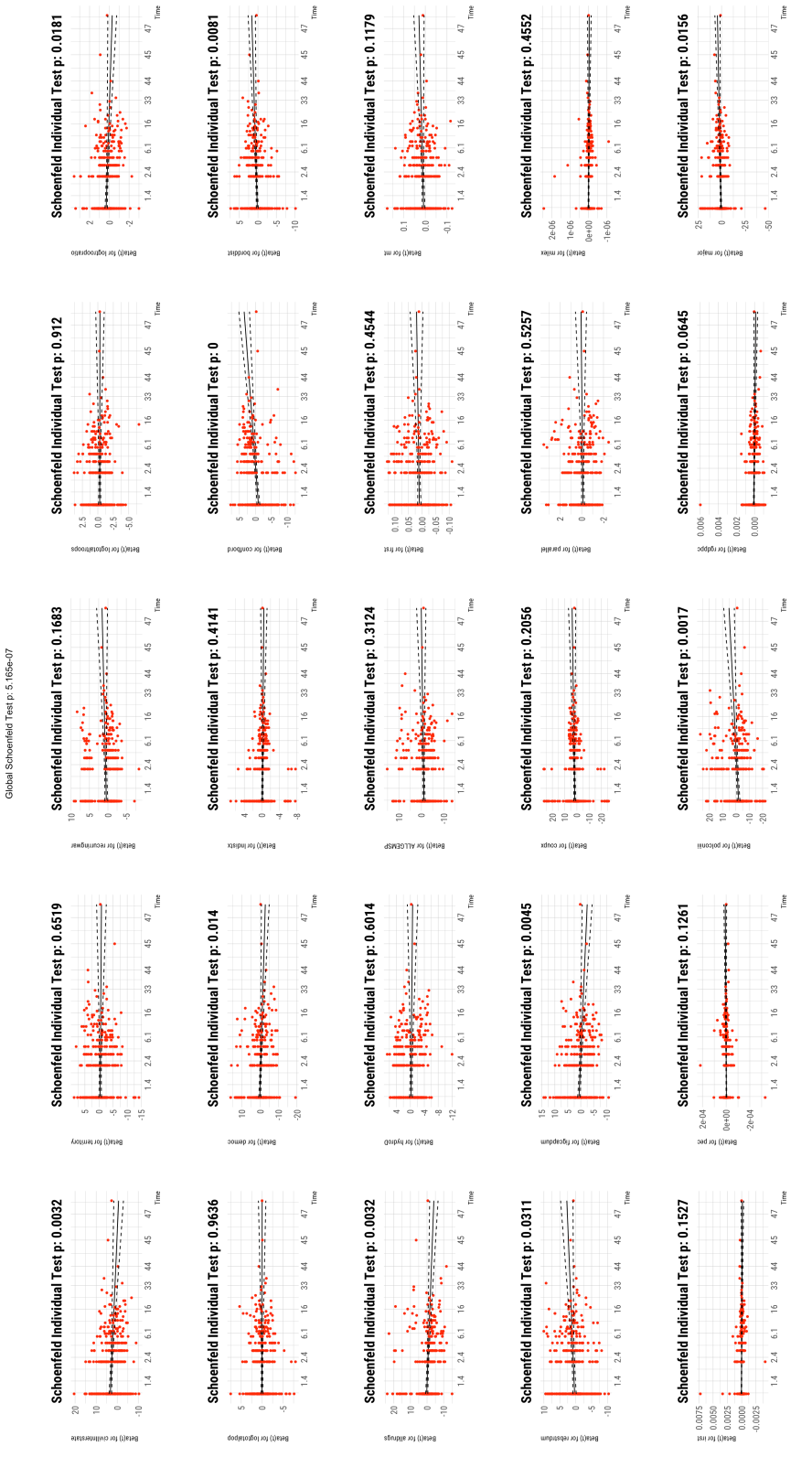

Figure 8.15: Schoenfeld residuals for Cunningham and Lemke 2013 with added covariates

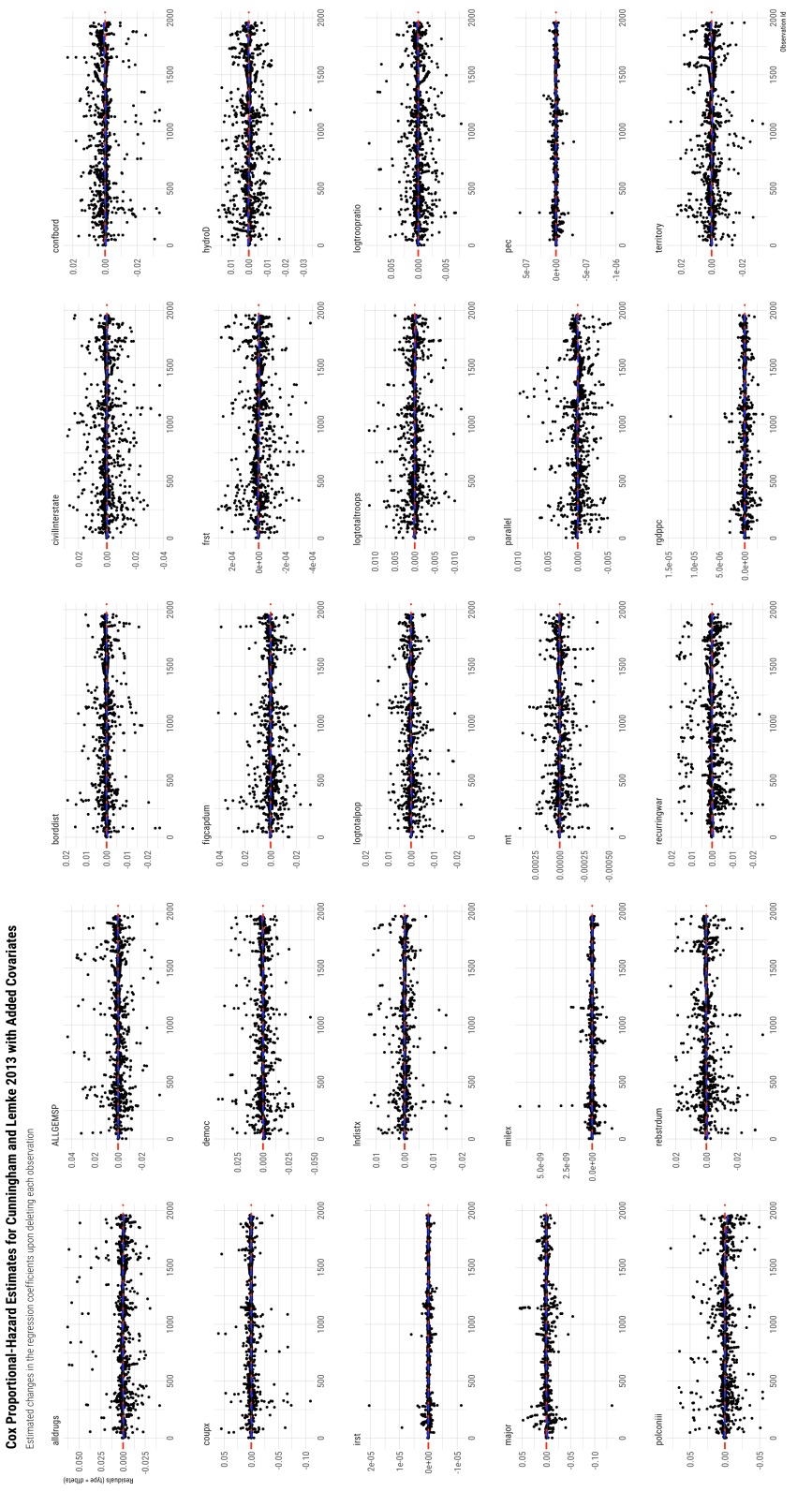

Figure 8.16: Cox-PH diagnostics (estimated change) for Cunningham and Lemke 2013 with added covariates

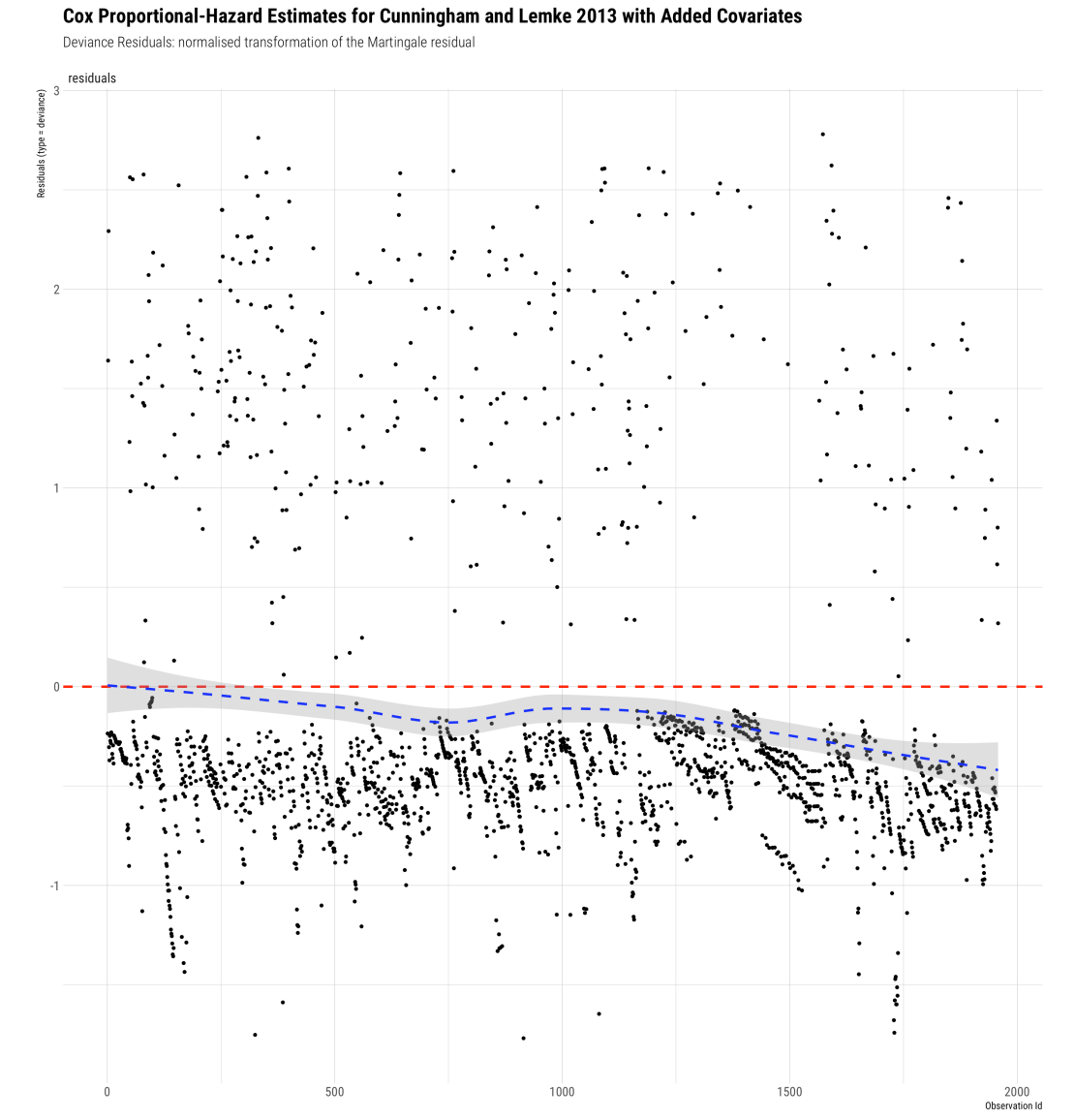

Figure 8.17: Cox-PH diagnostics (deviance) for Cunningham and Lemke 2013 with added covariates

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283.

Abbott, R. D. (1985). Logistic regression in survival analysis. *American Journal of Epidemiology*, 121(3):465–471.

Achen, C. H. (2005). Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science*, 22(4):327–339.

Addison, T., Le Billon, P., and Murshed, S. M. (2002). Conflict in africa: The cost of peaceful behaviour. *Journal of African Economies*, 11(3):365–386.

Adebajo, A. (2002). *Building Peace in West Africa: Liberia, Sierra Leone, and Guinea-Bissau*. Lynne Rienner Publishers.

Akcinaroglu, S. (2012). Rebel interdependencies and civil war outcomes. *Journal of Conflict Resolution*, 56(5):879–903.

Alanyali, M., Preis, T., and Moat, H. S. (2016). Tracking protests using geotagged flickr photographs. *PLOS ONE*, 11(3):1–8.

Aliyev, H. (2017). Pro-regime militias and civil war duration. *Terrorism and Political Violence*, pages 1–21.

Allen, S. H. and Martinez Machain, C. (2017). Understanding the impact of air power. *Conflict management and peace science*, pages 1–14.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1):17–21.

Atran, S., Axelrod, R., and Davis, R. (2007). Sacred barriers to conflict resolution. *Science*, 317:1039–1040.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Aydin, A. and Regan, P. M. (2012). Networks of third-party interveners and civil war duration. *European Journal of International Relations*, 18(3):573–597.

Bagozzi, B. E. (2015). Forecasting civil conflict with zero-inflated count models. *Civil Wars*, 17(1):1–24.

Bagozzi, B. E. (2016). On malaria and the duration of civil war. *Journal of Conflict Resolution*, 60(5):813–839.

Bagozzi, B. E. and Koren, O. (2017). Using machine learning methods to identify atrocity perpetrators. In *Big Data (Big Data), 2017 IEEE International Conference on*, pages 3042–3051. IEEE.

Bailey, K. (1987). *Methods of social research*. Simon and Schuster.

Bakan, D. (1966). The test of significance in psychological research. *Psychological bulletin*, 66(6):423.

Baker, W. D. and Oneal, J. R. (2001). Patriotism or opinion leadership? the nature and origins of the "rally'round the flag" effect. *Journal of Conflict Resolution*, 45(5):661–687.

Bakke, K. M., Cunningham, K. G., and Seymour, L. J. M. (2012). A plague of initials: Fragmentation, cohesion, and infighting in civil wars. *Perspectives on Politics*, 10(02):265–283.

Balcells, L. and Kalyvas, S. N. (2014). Does warfare matter? severity, duration, and outcomes of civil wars. *Journal of Conflict Resolution*, 58(8):1390–1418.

Balch-Lindsay, D. and Enterline, A. J. (2000). Killing time: The world politics of civil war duration, 1820-1992. *International Studies Quarterly*, 44(4):615–642.

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.

Beck, N. (2008). Time-series cross-section methods. In *The Oxford Handbook of Political Methodology*, pages 475–493. Oxford University Press.

Beck, N., Katz, J. N., and Tucker, R. (1998). Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *American Journal of Political Science*, 42(2):1260–1288.

Begg, C. B. and Berlin, J. A. (1988). Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 419–463.

Begley, C. G. and Ioannidis, J. P. (2015). Reproducibility in science: improving the standard for basic and preclinical research. *Circulation research*, 116(1):116–126.

Belkin, A. and Schofer, E. (2005). Coup risk, counterbalancing, and international conflict. *Security Studies*, 14(1):140–177.

Bell, S. R., Cingranelli, D., Murdie, A., and Caglayan, A. (2013). Coercion, capacity, and coordination: Predictors of political violence. *Conflict Management and Peace Science*, 30(3):240–262.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1):6.

Bennett, D. S. and Stam, A. C. (1996). The duration of interstate wars, 1816–1985. *American Political Science Review*, 90(2):239–257.

Bennett, D. S. and Stam, A. C. (2009). Revisiting predictions of war duration. *Conflict Management and Peace Science*, 26(3):256–267.

Bennett, S. D. and Stam III, A. C. (1998). The declining advantages of democracy: A combined model of war outcomes and duration. *Journal of Conflict Resolution*, 42(3):344–366.

Benoit, K. (1996). Democracies really are more pacific (in general) reexamining regime type and war involvement. *Journal of Conflict Resolution*, 40(4):636–657.

Benoit, K. (2018). *quanteda: Quantitative Analysis of Textual Data*. R package version 1.3.4.

Bessler, D. A., Kibriya, S., Chen, J., and Price, E. (2016). On forecasting conflict in the sudan: 2009–2012. *Journal of Forecasting*, 35(2):179–188.

Bethell, L. (1996). The paraguayan war (1864-1870). *ISA Research Papers*, (46).

Biau, D. J., Kernéis, S., and Porcher, R. (2008). Statistics in brief: the importance of sample size in the planning and interpretation of medical research. *Clinical orthopaedics and related research*, 466(9):2282–2288.

Bishop, D. and Cannings, C. (1978). A generalized war of attrition. *Journal of Theoretical Biology*, 70(1):85–124.

Blainey, G. (1973). *The causes of war*. The Free Press, New York.

Boulding, K. E. (1962). Conflict and defense: A general theory.

Box, P. H. (1967). *The Origins of the Paraguayan War*. Russell & Russell, New York.

Branch, M. (2014). Malignant side effects of null-hypothesis significance testing. *Theory & Psychology*, 24(2):256–277.

Brandt, P. T., Colaresi, M., and Freeman, J. R. (2008). The dynamics of reciprocity, accountability, and credibility. *Journal of Conflict Resolution*, 52(3):343–374.

Brandt, P. T. and Freeman, J. R. (2006). Advances in bayesian time series modeling and the study of politics: Theory testing, forecasting, and policy analysis. *Political Analysis*, 14(1):1–36.

Brandt, P. T., Freeman, J. R., and Schrodt, P. A. (2011). Real time, time series forecasting of political conflict. *Conflict Management and Peace Science*, 28(1):41–64.

Brandt, P. T., Freeman, J. R., and Schrodt, P. A. (2014). Evaluating forecasts of political conflict dynamics. *International Journal of Forecasting*, 30(4):944–962.

Breiman, L. (1984). *Classification and regression trees*. Routledge.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brenner, D. (2015). Ashes of co-optation: from armed group fragmentation to the rebuilding of popular insurgency in myanmar. *Conflict, Security & Development*, 15(4):337–358.

Briffa, M. (2014). What determines the duration of war? insights from assessment strategies in animal contests. *PloS one*, 9(9):e108491.

Buhaug, H. (2010). Dude, where's my conflict? lsg, relative strength, and the location of civil war. *Conflict Management and Peace Science*, 27(2):107–128.

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

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

Buhaug, H. and Gleditsch, K. S. (2008). Contagion or confusion? Why conflicts cluster in space. *International Studies Quarterly*, 52(2):215–233.

Bulow, J. and Klemperer, P. (1999). The generalized war of attrition. *American Economic Review*, 89(1):175–189.

Burgoon, B., Ruggeri, A., Schudel, W., and Manikkalingam, R. (2015). From media attention to negotiated peace: human rights reporting and civil war duration. *International Interactions*, 41(2):226–255.

Carroll, R. J. and Kenkel, B. (2016). Prediction, proxies, and power. *American Journal of Political Science*, page Conditionally accepted.

Carruthers, J. (1990). A rationale for the use of semi-structured interviews. *Journal of Educational Administration*, 28(1).

Carter, D. and Signorino, C. S. (2010). Back to the future: Modeling time dependence in binary data. *Political Analysis*, 18(3):271–292.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM.

Caverley, J. D. and Sechser, T. S. (2017). Military technology and the duration of civil conflict. *International Studies Quarterly*, 61(3):704–720.

Cawley, G. C. and Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul):2079–2107.

Cederman, L.-E. (2002). Endogenizing geopolitical boundaries with agent-based modeling. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7296–7303.

Cederman, L.-E. and Weidmann, N. B. (2017). Predicting armed conflict: Time to adjust our expectations? *Science*, 355(6324):474–476.

Centeno, M. A. (2002). *Blood and debt: War and the nation-state in Latin America*. Penn State Press.

Chadefaux, T. (2014). Early warning signals for war in the news. *Journal of Peace Research*, 51(1):5–18.

Chadefaux, T. (2017). Conflict forecasting and its limits. *Data Science*, (Preprint):1–11.

Chartrain, F. (1972). *L'Eglise et les partis dans la vie politique du Paraguay depuis l'indépendance*. PhD thesis.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Chiba, D. and Gleditsch, K. S. (2017). The shape of things to come? expanding the inequality and grievance model for civil war forecasts with event data. *Journal of Peace Research*, 54(2):275–297.

Chiozza, G. and Goemans, H. E. (2004). International conflict and the tenure of leaders: Is war still ex post inefficient? *American Journal of Political Science*, 48(3):604–619.

Choi, K., Joo, D., and Kim, J. (2017). Kapre: On-gpu audio preprocessing layers for a quick implementation of deep neural network models with keras. *arXiv preprint arXiv:1706.05781*.

Chollet, F. and Allaire, J. (2018). *Deep Learning with R*. Manning Publications.

Christensen, R. (2005). Testing fisher, neyman, pearson, and bayes. *The American Statistician*, 59(2):121–126.

Clapham, C. (1998). *African Guerrillas*. James Currey.

Clausewitz, v. C. (1832). *Vom Kriege*. Ullstein, Berlin.

Clayton, G. (2013). Relative rebel strength and the onset and outcome of civil war mediation. *Journal of Peace Research*, 50(5):609–622.

Clodfelter, M. (2002). Warfare and armed conflicts: A statistical reference to casualty and other figures, 1500–2000.

Cohen, J. (1994). The earth is round (p<. 05). In *What if there were no significance tests?*, pages 69–82. Routledge.

Colaresi, M. and Mahmood, Z. (2017). Do the robot: Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research*, 54(2):193–214.

Collier, D. and Mahoney, J. (1996). Insights and pitfalls: Selection bias in qualitative research. *World Politics*, 49(1):56–91.

Collier, D., Mahoney, J., and Seawright, J. (2004a). Claiming too much: Warnings about selection bias. *Rethinking social inquiry: Diverse tools, shared standards*, pages 85–102.

Collier, P. (2000a). Doing well out war: an economic perspective. In Berdal, M. R. and Malone, D., editors, *Greed & grievance: Economic agendas in civil wars*, pages 91–112. Lynne Rienner, Boulder.

Collier, P. (2000b). Rebellion as a quasi-criminal activity. *Journal of Conflict resolution*, 44(6):839–853.

Collier, P. and Hoeffler, A. (2002). On the incidence of civil war in africa. *Journal of conflict resolution*, 46(1):13–28.

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

Collier, P., Hoeffler, A., and Söderbom, M. (2004b). On the duration of civil war. *Journal of peace research*, 41(3):253–273.

Conrad, J. M., Greene, K. T., Walsh, J. I., and Whitaker, B. E. (2018). Rebel natural resource exploitation and conflict duration. *Journal of Conflict Resolution*, pages 1–26.

Conybeare, J. A. (1992). Weak cycles, length and magnitude of war: Duration dependence in international conflict. *Conflict Management and Peace Science*, 12(1):99–116.

Corbetta, P. (2003). *Social research: Theory, methods and techniques*. Sage.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cranmer, S. J. and Desmarais, B. A. (2017). What can we learn from predictive modeling? *Political Analysis*, 25(2):145–166.

Croicu, M. and Hegre, H. (2018). A fast spatial multiple imputation procedure for imprecise armed conflict events. pages Unpublished manuscript, Uppsala University.

Croicu, M. and Weidmann, N. B. (2015). Improving the selection of news reports for event coding using ensemble classification. *Research & Politics*, 2(4):2053168015615596.

Cruz, J. A. and Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2:59–77.

Cunningham, D., Gleditsch, K. S., and Salehyan, I. (2009). It takes two: A dyadic analysis of civil war duration and outcome. *Journal of Conflict Resolution*, 53(4):570–597.

Cunningham, D. E. (2006). Veto players and civil war duration. *American Journal of Political Science*, 50(4):875–892.

Cunningham, D. E. (2010). Blocking resolution: How external states can prolong civil wars. *Journal of Peace Research*, 47(2):115–127.

Cunningham, D. E., Gleditsch, K. S., and Salehyan, I. (2013). Non-state actors in civil wars: A new dataset. *Conflict Management and Peace Science*, 30(5):516–531.

Cunningham, D. E. and Lemke, D. (2013). Combining civil and interstate wars. *International Organization*, 67(3):609–627.

Cunningham, K. G., Bakke, K. M., and Seymour, L. J. M. (2012). Shirts today, skins tomorrow. *Journal of Conflict Resolution*, 56(1):67–93.

Dafoe, A. (2011). Statistical critiques of the democratic peace: Caveat emptor. *American Journal of Political Science*, 55(2):247–262.

Dafoe, A., Oneal, J. R., and Russett, B. (2013). The democratic peace: Weighing the evidence and cautious inference. *International Studies Quarterly*, 57(1):201–214.

Darwin, C. (1859). *On the origin of species*. Routledge.

Day, C. (2015). Bush path to self-destruction: Charles taylor and the revolutionary united front. *Small Wars & Insurgencies*, 26(5):811–835.

De Mesquita, B. B. (1978). Systemic polarization and the occurrence and duration of war. *Journal of Conflict Resolution*, 22(2):241–267.

De Mesquita, B. B. (2010). *The Predictioneer's Game: Using the logic of brazen self-interest to see and shape the future*. Random House LLC.

de Rouen Jr, K. R. and Sobek, D. (2004). The dynamics of civil war duration and outcome. *Journal of Peace Research*, 41(3):303–320.

de Soysa, I., Gartzke, E., and Lin, T. G. (2009). Oil, blood, and strategy: How petroleum influences interstate disputes. *Typescript. The Norwegian University of Science and Technology and the University of California, San Diego*.

Deane-Mayer, Z. A. and Knowles, J. E. (2016). *caretEnsemble: Ensembles of Caret Models*. R package version 2.0.0.

Deng, L., Yu, D., et al. (2014). Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387.

DeRouen Jr., K. and Sobek, D. (2004). The dynamics of civil war: Duration and outcome. *Journal of Peace Research*, 41(3):303–320.

Doyle, M. W. and Sambanis, N. (2000). International peacebuilding: A theoretical and quantitative analysis. *American Political Science Review*, 94(4):779–801.

Drever, E. (1995). *Using Semi-Structured Interviews in Small-Scale Research. A Teacher's Guide.* ERIC.

Driscoll, J. (2012). Commitment problems or bidding wars? rebel fragmentation as peace building. *Journal of Conflict Resolution*, 56(1):118–149.

Duursma, A. (2015). *African solutions to African challenges: explaining the role of legitimacy in mediating civil wars in Africa.* PhD thesis, University of Oxford.

Džeroski, S. and Ženko, B. (2004). Is combining classifiers with stacking better than selecting the best one? *Machine learning*, 54(3):255–273.

Eck, K. (2005). *A beginner's guide to conflict data: finding and using the right dataset.* Department of Peace and Conflict Research, Uppsala University.

Eck, K. (2012). In data we trust? a comparison of ucdp, ged and, acled conflict events datasets. *Cooperation and Conflict*, 47(1):124–141.

Elbadawi, I. A. and Nicholas, S. (2000). External interventions and the duration of civil wars. *World Bank Policy Research Working Paper*, (No. 2433).

Ellis, A. K. (1970). *Teaching and learning elementary social studies.* ERIC.

Ero, C. (2000). Vigilantes, civil defence forces and militia groups. the other side of the privatisation of security in africa. *Conflict trends*, 2000(1):25–29.

Escribà-Folch, A. (2010). Economic sanctions and the duration of civil conflicts. *Journal of Peace Research*, 47(2):129–141.

Esposito, G. and Rava, G. (2015). *Armies of the War of the Triple Alliance 1864–70: Paraguay, Brazil, Uruguay & Argentina*. Men-at-Arms. Bloomsbury Publishing.

Fearon, J. D. (1995). Rationalist explanations for war. *International Organization*, 49(3):379–414.

Fearon, J. D. (2004). Why some civil wars last so much longer than others? *Journal of Peace Research*, 41(3):275–301.

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

Fearon, J. D. and Laitin, D. D. (2008). Integrating qualitative and quantitative methods. In *The Oxford Handbook of Political Science*.

Fearon, J. D. and Laitin, D. D. (2011). Sons of the soil, migrants, and civil war. *World Development*, 39(2):199–211.

Feldman, R. L. and Arrous, M. B. (2013). Confronting africa's sobels. *Parameters*, 43(4):67–75.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems. *J. Mach. Learn. Res*, 15(1):3133–3181.

Fey, M. and Ramsay, K. W. (2007). Mutual optimism and war. *American Journal of Political Science*, 51(4):738–754.

Filson, D. and Werner, S. (2004). Bargaining and fighting: The impact of regime type on war onset, duration, and outcomes. *American Journal of Political Science*, 48(2):296–313.

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 69–78.

Fisher, R. A. (1937). *The design of experiments*. Oliver And Boyd; Edinburgh; London.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Oxford, England: Hafner Publishing Co.

Fitzsimmons, S. (2013). When few stood against many: Explaining executive outcomes' victory in the sierra leonean civil war. *Defence Studies*, 13(2):245–269.

Fjelde, H. and Nilsson, D. (2012). Rebels against rebels: Explaining violence between rebel groups. *Journal of Conflict Resolution*, 56(4):604–628.

Florkowski, C. M. (2008). Sensitivity, specificity, receiver-operating characteristic (roc) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical Biochemist Reviews*, 29(Suppl 1):S83.

Forsberg, E. (2016). Transnational dimensions of civil wars. *What Do We Know About Civil Wars?*, page 75.

Freeman, E. A., Moisen, G. G., Coulston, J. W., and Wilson, B. T. (2015). Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. *Canadian Journal of Forest Research*, 46(3):323–339.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics New York.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

Fuhrmann, M. and Horowitz, M. C. (2014). When leaders matter: Rebel experience and nuclear proliferation. *The Journal of Politics*, 77(1):72–87.

Fukumoto, K. (2015). What happens depends on when it happens: Copula-based ordered event history analysis of civil war duration and outcome. *Journal of the American Statistical Association*, 110(509):83–92.

Gartzke, E. (1999). War is in the error term. *International Organization*, 53(3):567–587.

Gartzke, E. and Braithwaite, A. (2011). Power, parity and proximity. *How Distance and Uncertainty*, page Unpublished manuscript.

George, A. L. and Bennett, A. (2005). *Case studies and theory development in the social sciences*. MIT Press.

Gerber, A., Malhotra, N., et al. (2008). Do statistical reporting standards affect what is published? publication bias in two leading political science journals. *Quarterly Journal of Political Science*, 3(3):313–326.

Gevrey, M., Dimopoulos, I., and Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, 160(3):249–264.

Gibler, D. M. (2008). The costs of reneging: Reputation and alliance formation. *Journal of Conflict Resolution*, 52(3):426–454.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5):587–606.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, 52(3):647–674.

Gill, J. (2018). Comments from the new editor. *Political Analysis*, 26(1):1–2.

Gilmore, E., Gleditsch, N. P., Lujala, P., and Rød, J. K. (2005). Conflict diamonds: A new dataset. *Conflict Management and Peace Studies*, 22(3):257–272.

Gleditsch, K. S. (2002). Expanded trade and gdp data. *Journal of Conflict Resolution*, 46(5):712–724.

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

Gleditsch, K. S. (2017). Ornithology and varieties of conflict: A personal retrospective on conflict forecasting. *Peace Economics, Peace Science and Public Policy*, 23(4).

Gleditsch, K. S. and Ward, M. D. (2013). Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes. *Journal of Peace Research*, 50(1):17–31.

Goemans, H. E. (2000). Fighting for survival: The fate of leaders and the duration of war. *Journal of Conflict Resolution*, 44(5):555–579.

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

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. In *Seminars in hematology*, volume 45, pages 135–140. Elsevier.

Gorden, R. L. (1975). *Interviewing: Strategy, techniques, and tactics*. "The" Dorsey Press.

Greenhill, B., Ward, M. D., and Sacks, A. (2011). The separation plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science*, 55(4):991–1002.

Gurr, T. R. and Lichbach, M. I. (1986). Forecasting internal conflict a competitive evaluation of empirical theories. *Comparative Political Studies*, 19(1):3–38.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., and Drummond, G. B. (2015). The fickle p value generates irreproducible results. *Nature methods*, 12(3):179.

Harding, J. (1997). The mercenary business:'executive outcomes'. *Review of African Political Economy*, 24(71):87–97.

Hartzell, C. A. (2009). Settling civil wars: armed opponents' fates and the duration of the peace. *Conflict Management and Peace Science*, 26(4):347–365.

Hartzell, C. A. and Hoddie, M. (2007). *Crafting peace: Power-sharing institutions and the negotiated settlement of civil wars*. Penn State Press.

Hegre, H. (2004). The duration and termination of civil war. *Journal of Peace Research*, 41(3):243–252.

Hegre, H. (2014). Democracy and armed conflict. *Journal of Peace Research*, 51(2):159–172.

Hegre, H., Buhaug, H., Calvin, K. V., Nordkvelle, J., Waldhoff, S. T., and Gilmore, E. (2016). Forecasting civil conflict along the shared socioeconomic pathways. *Environmental Research Letters*, 11(5):054002.

Hegre, H., Karlsen, J., Nygård, H. M., Strand, H., and Urdal, H. (2013). Predicting armed conflict, 2010–2050. *International Studies Quarterly*, 57(2):250–270.

Hegre, H. and Sambanis, N. (2006). Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution*, 50(4):508–535.

Hendrix, C. S. (2010). Measuring state capacity: Theoretical and empirical implications for the study of civil conflict. *Journal of Peace Research*, 47(3):273–285.

Hendrix, C. S. (2017). Oil prices and interstate conflict. *Conflict Management and Peace Science*, 34(6):575–596.

Henisz, W. J. (2017). The political constraint index (polcon) dataset.

Hill, R. C. and Judge, G. (1987). Improved prediction in the presence of multicollinearity. *Journal of Econometrics*, 35(1):83–100.

Hoffman, D. (2007). The meaning of a militia: Understanding the civil defence forces of sierra leone. *African Affairs*, 106(425):639–662.

Honaker, J. and King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, 54(2):561–581.

Hooker, T. (2008). *Paraguayan War: Armies of the Nineteenth Century–The Americas*. Nottingham: Foundry Books.

Horowitz, D. L. (1985). *Ethnic groups in conflict*. University of California Press, Berkeley.

Howe, H. M. (1998). Private security forces and african stability: the case of executive outcomes. *The Journal of Modern African Studies*, 36(2):307–331.

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational researcher*, 17(8):10–16.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New phytologist*, 11(2):37–50.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., and Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, 50(2):105–115.

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative science quarterly*, 24(4):602–611.

Johnstone, I. (2006). Dilemmas of robust peace operations. *Annual Review of Global Peace Operations*, pages 1–18.

Kahneman, D. and Egan, P. (2011). *Thinking, fast and slow*, volume 1. Farrar, Straus and Giroux New York.

Kajornboon, A. B. (2005). Using interviews as research instruments. *E-journal for Research Teachers*, 2(1):1–9.

Kalyvas, S. N. and Balcells, L. (2010). International system and technologies of rebellion: How the end of the cold war shaped internal conflict. *American Political Science Review*, 104(3):415–429.

Kane, T. M. (2012). *Military logistics and strategic performance*. Routledge.

Kassambara, A. and Kosinski, M. (2018). *survminer: Drawing Survival Curves using 'ggplot2'*. R package version 0.4.3.

Kaufman, S. J. (2006). Symbolic politics or rational choice? testing theories of extreme ethnic violence. *International Security*, 30(4):45–86.

Kaufmann, C. (1996). Possible and impossible solutions to ethnic civil wars. *International Security*, 20(4):136–175.

Kaufmann, C. D. (1998). When all else fails: Ethnic population transfers and partitions in the twentieth century. *International security*, 23(2):120–156.

Keen, D. (2005). *Conflict and collusion in Sierra Leone*. James Currey (imprint of Boydell & Brewer Ltd.).

Kennedy, P. (1987). *The rise and fall of the great powers*. Vintage.

King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, pages 666–687.

King, G., Keohane, R. O., and Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirschner, S. A. (2010). Knowing your enemy: Information and commitment problems in civil wars. *Journal of Conflict Resolution*, 54(5):745–770.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017). Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, pages 972–981.

Kleinpenning, J. M. (2002). Strong reservations about" new insights into the demographics of the paraguayan war". *Latin American Research Review*, pages 137–142.

Koch, M. T. (2009). Governments, partisanship, and foreign policy: The case of dispute duration. *Journal of Peace Research*, 46(6):799–817.

Kocher, M. A., Pepinsky, T. B., and Kalyvas, S. N. (2011). Aerial bombing and counterinsurgency in the vietnam war. *American Journal of Political Science*, 55(2):201–218.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Korjus, K., Hebart, M. N., and Vicente, R. (2016). An efficient data partitioning to improve classification performance while keeping parameters interpretable. *PloS one*, 11(8):e0161788.

Krepinevich, A. (1986). *The army and Vietnam*. Johns Hopkins paperback. Johns Hopkins University Press.

Krustev, V. L. (2006). Interdependence and the duration of militarized conflict. *Journal of Peace Research*, 43(3):243–260.

Kuhn, M. (2018). *caret: Classification and Regression Training*. R package version 6.0-80.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York.

Kukull, W. A. and Ganguli, M. (2012). Generalizability the trees, the forest, and the low-hanging fruit. *Neurology*, 78(23):1886–1891.

Kuncheva, L. I. and Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207.

Lake, D. A. (2002). Rational extremism: Understanding terrorism in the twenty-first century. *Dialogue IO*, 1(1):15–28.

Lake, D. A. (2003). International relations theory and internal conflict: insights from the interstices. *International Studies Review*, 5(4):81–89.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Langlois, C. C. and Langlois, J.-P. P. (2009). Does attrition behavior help explain the duration of interstate wars? a game theoretic and empirical analysis. *International Studies Quarterly*, 53(4):1051–1073.

Le Billon, P. (2008). Diamond wars? conflict diamonds and geographies of resource wars. *Annals of the Association of American Geographers*, 98(2):345–372.

Licklider, R. (1995). The consequences of negotiated settlements in civil wars, 1945-1993. *American Political Science Review*, 89(3):681–690.

Lo, A., Chernoff, H., Zheng, T., and Lo, S.-H. (2015). Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences*, 112(45):13892–13897.

Lujala, P. (2009). Deadly combat over natural resources: Gems, petroleum, drugs, and the severity of armed civil conflict. *Journal of Conflict Resolution*, 53(1):50–71.

Lunardon, N., Menardi, G., and Torelli, N. (2014). ROSE: a Package for Binary Imbalanced Learning. *R Journal*, 6(1):82–92.

Lyall, J. (2009). Does indiscriminate violence incite insurgent attacks? evidence from chechnya. *Journal of Conflict Resolution*, 53(3):331–362.

Lyall, J. (2010). Do democracies make inferior counterinsurgents? reassessing democracy's impact on war outcomes and duration. *International Organization*, 64(1):167–192.

Lykken, D. T. (1991). What's wrong with psychology anyway. *Thinking clearly about psychology*, 1:3–39.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, 13(3):e0194889.

Maoz, Z. (1983). Resolve, capabilities, and the outcomes of interstate disputes, 1816-1976. *Journal of Conflict Resolution*, 27(2):195–229.

Maoz, Z. and Abdolali, N. (1989). Regime types and international conflict, 1816-1976. *Journal of Conflict Resolution*, 33(1):3–35.

Maoz, Z. and Russett, B. (1993). Normative and structural causes of democratic peace, 1946–1986. *American Political Science Review*, 87(3):624–638.

Marascuilo, L. A. and McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, Calif.(USA) Brooks/Cole Pub.

Markowitz, J. N. and Fariss, C. J. (2018). Power, proximity, and democracy: Geopolitical competition in the international system. *Journal of Peace Research*, 55(1):78–93.

Marron, J. S., Todd, M. J., and Ahn, J. (2007). Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271.

Masson, M. E. (2011). A tutorial on a practical bayesian alternative to null-hypothesis significance testing. *Behavior research methods*, 43(3):679–690.

Mayer, T. and Zignago, S. (2011). Notes on cepii's distances measures: The geodist database. *CEPII Working Paper No. 2011-25*.

McCallum, A. and Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In *Learning for Text Categorization: Papers from the 1998 AAAI Workshop*, pages 41–48.

McGillivray, F. and Smith, A. (2000). Trust and cooperation through agent-specific punishments. *International Organization*, 54(4):809–824.

McGillivray, F. and Smith, A. (2004). The impact of leadership turnover on trading relations between states. *International Organization*, 58(3):567–600.

Meernik, J. and Brown, C. (2007). The short path and the long road: Explaining the duration of us military operations. *Journal of Peace Research*, 44(1):65–80.

Mertens, D. M. and Hesse-Biber, S. (2012). *Triangulation and mixed methods research: Provocative positions*. SAGE Publications Sage CA: Los Angeles, CA.

Metternich, N. (2011). Expecting elections: Interventions, ethnic support, and the duration of civil wars. *Journal of Conflict Resolution*, 55(6):909–937.

Mhaskar, H. N. and Poggio, T. (2016). Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14(06):829–848.

Miller, R. A. (1999). Regime type, strategic interaction, and the diversionary use of force. *Journal of Conflict Resolution*, 43(3):388–402.

Miller, T. W. (2014). *Modeling techniques in predictive analytics: business problems and solutions with R*. FT Press Analytics.

Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT press.

Moisen, G. G., Freeman, E. A., Blackard, J. A., Frescino, T. S., Zimmermann, N. E., and Edwards Jr, T. C. (2006). Predicting tree species presence and basal area in utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological modelling*, 199(2):176–187.

Montgomery, J. M., Hollenbach, F. M., and Ward, M. D. (2012). Improving predictions using ensemble bayesian model averaging. *Political Analysis*, 20(3):271–291.

Moonesinghe, R., Khoury, M. J., and Janssens, A. C. J. (2007). Most published research findings are false—but a little replication goes a long way. *PLoS medicine*, 4(2):e28.

Moore, M. (2012). Selling to both sides: The effects of major conventional weapons transfers on civil war severity and duration. *International Interactions*, 38(3):325–347.

Morrison, D. G. and Schmittlein, D. C. (1980). Jobs, strikes, and wars: Probability models for duration. *Organizational Behavior and Human Performance*, 25(2):224–251.

Morrow, J. D. (1985). A continuous-outcome expected utility theory of war. *Journal of Conflict Resolution*, 29(3):473–502.

Muchlinski, D., Siroky, D., He, J., and Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, 24(1):87–103.

Mueller, J. E. et al. (1973). *War, presidents, and public opinion*. Wiley New York.

Mukherjee, S. (2014). Why are the longest insurgencies low violence? politician motivations, sons of the soil, and civil war duration. *Civil Wars*, 16(2):172–207.

Neunhoeffer, M. and Sternberg, S. (2018). How cross-validation can go wrong and what to do about it. *Political Analysis*, Forthcoming.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, 5(2):241.

Nilsson, M. (2012). Offense–defense balance, war duration, and the security dilemma. *Journal of Conflict Resolution*, 56(3):467–489.

Ogutu, J. O., Piepho, H.-P., and Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. In *BMC proceedings*, volume 5, page S11. BioMed Central.

Olonisakin, F. (2008). *Peacekeeping in Sierra Leone: the story of UNAMSIL*. Lynne Rienner Publishers Boulder, Colo.

Olsen, W. (2004). Triangulation in social research: qualitative and quantitative methods can really be mixed. *Developments in sociology*, 20:103–118.

Owsiak, A. P. (2015). Forecasting conflict management in militarized interstate disputes. *Conflict Management and Peace Science*, 32(1):50–75.

Paluszynska, A. and Biecek, P. (2017). *randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance*. R package version 0.9.

Pape, R. A. (1997). Why economic sanctions do not work. *International security*, 22(2):90–136.

Pape, R. A. (1998). Why economic sanctions still do not work. *International Security*, 23(1):66–77.

Pearce, S. (1992). Introduction to fisher (1925) statistical methods for research workers. In *Breakthroughs in statistics*, pages 59–65. Springer.

Pearlman, W. and Cunningham, K. G. (2012). Nonstate actors, fragmentation, and conflict processes. *Journal of Conflict Resolution*, 56(1):3–15.

Penfold, P. (2013). *Atrocities, Diamonds and Diplomacy: The Inside Story of the Conflict in Sierra Leone*. Pen & Sword Military.

Peterson, R. H. (1967). Letter to the editor—on the "logarithmic law" of attrition and its application to tank combat. *Operations Research*, 15(3):557–558.

Powell, R. (1996). Bargaining in the shadow of power. *Games and Economic Behavior*, 15(2):255–289.

Powell, R. (2002). Bargaining theory and international conflict. *Annual Review of Political Science*, 5(1):1–30.

Powell, R. (2004). The inefficient use of power: Costly conflict with complete information. *American Political Science Review*, 98(2):231–241.

Powell, R. (2006). War as a commitment problem. *International Organization*, 60(1):169–203.

Powell, R. (2012). Persistent fighting and shifting power. *American Journal of Political Science*, 56(3):620–637.

Prorok, A. K. (2018). Led astray: Leaders and the duration of civil war. *Journal of Conflict Resolution*, 62(6):1179–1204.

Quek, K. (2017). Rationalist experiments on war. *Political Science Research and Methods*, 5(1):123–142.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raknerud, A. and Hegre, H. (1997). The hazard of war: Reassessing the evidence for the democratic peace. *Journal of Peace Research*, 34(4):385–404.

Ray, J. L. (2003). Explaining interstate conflict and war: What should be controlled for? *Conflict Management and Peace Science*, 20(2):1–31.

Ray, J. L. (2005). Constructing multivariate analyses (of dangerous dyads). *Conflict Management and Peace Science*, 22(4):277–292.

Reber, V. B. (1988). The demographics of paraguay: A reinterpretation of the great war, 1864-70. *The Hispanic American Historical Review*, 68(2):289–319.

Reber, V. B. (2002). Comment on" the paraguayan rosetta stone". *Latin American Research Review*, 37(3):129–136.

Regan, P. M. (2002). Third-party interventions and the duration of intrastate conflicts. *Journal of Conflict Resolution*, 46(1):55–73.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

Richards, P. et al. (1998). *Fighting for the rain forest: war, youth & resources in Sierra Leone.* Number Reprinted Ed. James Currey Ltd.

Richardson, S. A., Dohrenwend, B. S., and Klein, D. (1965). *Interviewing: Its forms and functions*. Basic Books.

Ripley, B. D. and Ripley, R. M. (2001). Neural networks as statistical methods in survival analysis. *Clinical applications of artificial neural networks*, pages 237–255.

Rish, I. (2001). An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46.

Robson, C. (2002). *Real World Research: A Resource for Social Scientists and Practitioner-Researchers*. Regional Surveys of the World Series. Wiley.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39.

Rosa, J. M. (1968). *La guerra del Paraguay y las montoneras argentinas*. A. Peña Lillo.

Rose, W. (2000). The security dilemma and ethnic conflict: Some new hypotheses. *Security Studies*, 9(4):1–51.

Ross, M. L. (2004). How do natural resources influence civil war? evidence from thirteen cases. *International Organization*, 58(Winter):35–67.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological bulletin*, 57(5):416.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Ruggeri, A., Dorussen, H., and Gizelis, T.-I. (2016). On the frontline every day? subnational deployment of united nations peacekeepers. *British Journal of Political Science*, pages 1–21.

Rummel, R. J. (1969). Forecasting international relations: A proposed investigation of three-mode factor analysis. *Technological Forecasting*, 1(2):197–216.

Rummel, R. J. (1979). Understanding conflict and war: Vol. 4: War, power, peace. *B everly H ills: Sage*.

Russett, B. (1994). *Grasping the democratic peace: Principles for a post-Cold War world*. Princeton university press.

Russett, B. M. and Oneal, J. R. (2001). *Triangulating peace: Democracy, interdependence, and international organizations*, volume 9. Norton.

Saideman, S. M. (2017). Elf must die: Institutions, concentration, the international relations of ethnic conflict and the quest for better data. *Ethnopolitics*, 16(1):66–73.

Sakaguchi, D. (2011). Distance and military operations: Theoretical background toward strengthening the defense of offshore islands. *NIDS Journal of Defense and Security*, 12:83–105.

Salehyan, I. (2009). *Rebels without borders: transnational insurgencies in world politics*. Cornell University Press.

Sambanis, N. (2002). A review of recent advances and future directions in the quantitative literature on civil war. *Defence and Peace Economics*, 13(3):215–243.

Schelling, T. C. (1960). *The strategy of conflict*. Havard University Press, Cambridge, MA.

Schneider, G., Gleditsch, N. P., and Carey, S. (2011). Forecasting in international relations: One quest, three approaches. *Conflict Management and Peace Science*, 28(1):5–14.

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

Schrodt, P. A. (2009). Tabari: Textual analysis by augmented replacement instructions. *Dept. of Political Science, University of Kansas, Blake Hall, Version 0.7. 3B3*, pages 1–137.

Schrodt, P. A. and Gerner, D. J. (2000). Cluster-based early warning indicators for political change in the contemporary levant. *American Political Science Review*, pages 803–817.

Schrodt, P. A., Yonamine, J., and Bagozzi, B. E. (2013). Data-based computational approaches to forecasting political violence. In *Handbook of Computational Approaches to Counterterrorism*, pages 129–162. Springer.

Seawright, J. (2016). *Multi-method social science: Combining qualitative and quantitative tools*. Cambridge University Press.

Seawright, J. and Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political Research Quarterly*, 61(2):294–308.

Shannon, M., Morey, D., and Boehmke, F. J. (2010). The influence of international organizations on militarized dispute initiation and duration. *International Studies Quarterly*, 54(4):1123–1141.

Shirkey, Z. C. (2012). When and how many: The effects of third party joining on casualties and duration in interstate wars. *Journal of Peace Research*, 49(2):321–334.

Singer, J. D. (1972). The ?correlates of war? project: Interim report and rationale. *World Politics*, 24(02):243–270.

Singer, J. D., Bremer, S., and Stuckey, J. (1972). Capability distribution, uncertainty, and major power war, 1820-1965. *Peace, war, and numbers*, 19.

Singer, P. (2011). *Corporate Warriors: The Rise of the Privatized Military Industry*. Cornell Studies in Security Affairs. Cornell University Press.

Slantchev, B. L. (2003). The principle of convergence in wartime negotiations. *American Political Science Review*, 97(4):621–632.

Slantchev, B. L. (2004). How initiators end their wars: The duration of warfare and the terms of peace. *American Journal of Political Science*, 48(4):813–829.

Slantchev, B. L. and Tarar, A. (2011). Mutual optimism as a rationalist explanation of war. *American Journal of Political Science*, 55(1):135–148.

Sollich, P. and Krogh, A. (1996). Learning with ensembles: How overfitting can be useful. In *Advances in neural information processing systems*, pages 190–196.

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., Hing, C., Kwok, C. S., Pang, C., and Harvey, I. (2010). Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess*, 14(8):1–193.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Stam, A. C. (1996). *Win, lose, or draw: Domestic politics and the crucible of war*. University of Michigan Press.

Stanley, E. A. and Sawyer, J. P. (2009). The equifinality of war termination: Multiple paths to ending war. *Journal of Conflict Resolution*, 53(5):651–676.

Stedman, S. J. (2001). *Implementing peace agreements in civil wars: lessons and recommendations for policymakers*. International Peace Academy New York.

Stekhoven, D. J. and Bühlmann, P. (2011). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Stone, C. J. (1996). *A course in probability and statistics*. Duxbury Press Belmont:.

Sullivan, P. L. (2008a). At what price victory? the effects of uncertainty on military intervention duration and outcome. *Conflict Management and Peace Science*, 25(1):49–66.

Sullivan, P. L. (2008b). Sustaining the fight: A cross-sectional time-series analysis of public support for ongoing military interventions. *Conflict Management and Peace Science*, 25(2):112–135.

Svensson, I. (2007). Bargaining, bias and peace brokers: How rebels commit to peace. *Journal of Peace Research*, 44(2):177–194.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. Random house.

Taliaferro, J. W. (1998). Quagmires in the periphery: Foreign wars and escalating commitment in international conflict. *Security Studies*, 7(3):94–144.

Tarrow, S. (1995). Bridging the quantitative-qualitative divide in political science. *American Political Science Review*, 89(2):471–474.

Tetlock, P. (2005). *Expert political judgment: How good is it? How can we know?* Princeton University Press.

Tetlock, P. E. and Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Thomas, J. (2014). Rewarding bad behavior: How governments respond to terrorism in civil war. *American Journal of Political Science*, 58(4):804–818.

Thyne, C. (2017). The impact of coups d'etat on civil war duration. *Conflict management and peace science*, 34(3):287–307.

Thyne, C. L. (2012). Information, commitment, and intra-war bargaining: The effect of governmental constraints on civil war duration. *International Studies Quarterly*, 56(2):307–321.

Tiernay, M. (2015). Killing kony: Leadership change and civil war termination. *Journal of Conflict Resolution*, 59(2):175–206.

Tollefsen, A. F. and Buhaug, H. (2015). Insurgency and inaccessibility1. *International Studies Review*, 17(1):6–25.

Toloşi, L. and Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994.

Torgo, L. (2010). *Data Mining with R, learning with case studies*. Chapman and Hall/CRC.

Tsebelis, G. and Yataganas, X. (2002). Veto players and decision-making in the eu after nice. *Journal of Common Market Studies*, 40(2):283–307.

Tukey, J. W. (1960). Conclusions vs decisions. *Technometrics*, 2(4):423–433.

Turkheimer, F. E., Aston, J. A., and Cunningham, V. J. (2004). On the logic of hypothesis testing in functional imaging. *European journal of nuclear medicine and molecular imaging*, 31(5):725–732.

Ucko, D. H. (2016). Can limited intervention work? lessons from britain's success story in sierra leone. *Journal of Strategic Studies*, 39(5-6):847–877.

Uzonyi, G. and Wells, M. (2016). Domestic institutions, leader tenure and the duration of civil war. *Conflict management and peace science*, 33(3):294–310.

Van Evera, S. (2001). Primordialism lives! *APSA-CP: Newsletter of the organized section in comparative politics of the American Political Science Association*, 12(1):20–22.

Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91.

Vogt, M., Bormann, N.-C., Rüegger, S., Cederman, L.-E., Hunziker, P., and Girardin, L. (2015). Integrating data on ethnicity, geography, and conflict: The ethnic power relations data set family. *Journal of Conflict Resolution*, 59(7):1327–1342.

Vuchinich, S. and Teachman, J. (1993). Influences on the duration of wars, strikes, riots, and family arguments. *Journal of Conflict Resolution*, 37(3):544–568.

Wagner, R. H. (2000). Bargaining and war. *American Journal of Political Science*, 44(3):469–484.

Walter, B. F. (1997). The critical barrier to civil war settlement. *International organization*, 51(3):335–364.

Walter, B. F. (1999). Designing transitions from civil war: Demobilization, democratization, and commitments to peace. *International Security*, 24(1):127–155.

Walter, B. F. (2002). *Committing to peace: The successful settlement of civil wars*. Princeton University Press.

Waltz, K. N. (1979). *Theory of International Politics*. Addison-Wesley, Reading.

Wang, B. and Zou, H. (2016). Sparse distance weighted discrimination. *Journal of Computational and Graphical Statistics*, 25(3):826–838.

Ward, M. D. and Bakke, K. (2005). Predicting civil conflicts: on the utility of empirical research. In *Conference on Disaggregating the Study of Civil War and Transnational Violence, University of California Institute of Global Conflict and Cooperation*. Citeseer.

Ward, M. D., Greenhill, B. D., and Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375.

Ward, M. D., Metternich, N. W., Dorff, C. L., Gallop, M., Hollenbach, F. M., Schultz, A., and Weschle, S. (2013). Learning from the past and stepping

into the future: Toward a new generation of conflict prediction. *International Studies Review*, 15(4):473–490.

Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*, volume 111. Rand McNally Chicago.

Weeks, J. L. (2008). Autocratic audience costs: Regime type and signaling resolve. *International Organization*, 62(1):35–64.

Weidmann, N. B. (2015). On the accuracy of media-based conflict event data. *Journal of Conflict Resolution*, 59(6):1129–1149.

Weidmann, N. B. and Ward, M. D. (2010). Predicting conflict in space and time. *Journal of Conflict Resolution*, 54(6):883–901.

Weisiger, A. (2016). Learning from the battlefield: Information, domestic politics, and interstate war duration. *International Organization*, 70(2):347–375.

Weiss, S. M. and Kulikowski, C. A. (1991). *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann Publishers Inc.

Whigham, T. (2002). *The Paraguayan War: causes and early conduct*, volume 1. U of Nebraska Press.

Whigham, T. L. and Potthast, B. (1999). The paraguayan rosetta stone: New insights into the demographics of the paraguayan war, 1864-1870. *Latin American Research Review*, pages 174–186.

Witmer, F. D., Linke, A. M., O?Loughlin, J., Gettelman, A., and Laing, A. (2017). Subnational violent conflict forecasts for sub-saharan africa, 2015–65, using climate-sensitive models. *Journal of Peace Research*, 54(2):175–192.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wittman, D. (1979). How a war ends: A rational model approach. *Journal of Conflict Resolution*, 23(4):743–763.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Wolford, S. (2007). The turnover trap: New leaders, reputation, and international conflict. *American Journal of Political Science*, 51(4):772–788.

Wood, R. M. (2010). Rebel capability and strategic violence against civilians. *Journal of Peace Research*, 47(5):601–614.

Wright, M. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software, Articles*, 77(1):1–17.

Wucherpfennig, J., Metternich, N. W., Cederman, L.-E., and Gleditsch, K. S. (2012). Ethnicity, the state, and the duration of civil war. *World Politics*, 64(1):79–115.

Zack-Williams, A. B. (1997). Kamajors,'sober'& the militariat: civil society & the return of the military in sierra leonean politics.

Zupan, B., Demšar, J., Kattan, M. W., Beck, J. R., and Bratko, I. (2000). Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial intelligence in medicine*, 20(1):59–75.