Essays on political and criminal violence

Freire, Danilo Alves Mendes

Awarding institution:
King's College London

KING'S College LONDON

Essays on Political and Criminal Violence

Danilo Alves M. Freire

Submitted for the degree of Doctor of Philosophy

Department of Political Economy

King's College London

December 2018

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Signature:

Danilo Alves M. Freire

Abstract

This thesis addresses three topics in political and criminal violence. The first essay is an empirical evaluation of a broad set of homicide reduction policies implemented in the state of São Paulo, Brazil. I employ the synthetic control method, a generalisation of differences-in-differences, to compare these measures against an artificial São Paulo. The results indicate a large drop in homicide rates in actual São Paulo when contrasted with the synthetic counterfactual, with about 20,000 lives saved during the period.

The second essay offers a rational choice account for the Brazil's *jogo do bicho*, or the 'animal game', possibly the largest illegal gambling game in the world. I investigate the institutions that have caused the *jogo do bicho*'s notable growth and long-term survival outside the boundaries of the Brazilian law. I show how *bicheiros* or bookmakers promote social order, solve information asymmetries, and reduce negative externalities via costly signalling and the provision of club goods. I also explain the emergence of the informal rules that govern the game as well as their enforcement mechanisms.

In the third essay, I employ extreme bounds analysis and distributed random forests to identify the key determinants of state-sponsored violence. Although scholars have suggested a number of potential correlates of mass killings, it remains unclear whether the estimates are robust to different model specifications, or which variables accurately predict the onset of large-scale violence. I employ extreme bounds analysis and random forests to test the sensitivity of 40 variables on a sample of 177 countries from 1945 to 2013. The results help clear the brush around mass killings, as few variables in

this literature are robust determinants of atrocity. However, support for an opportunity logic persists as greater constraints on a government limit its ability to employ barbarous tactics. It appears that the Conflict Trap applies to government atrocity. Atrocity breeds atrocity, while wealthy stable democracies tend to avoid episodes of mass killing.

Acknowledgements

I would like to extend thanks to the many people, in many countries, who so generously contributed to the work presented in this thesis. First and foremost, I want to thank my supervisor, Professor David Skarbek, for the guidance and encouragement he provided me throughout this research. David has been very supportive since the first time we met, and his dynamism, vision, empathy, and great sense of humour have deeply inspired me. I could not have wished for a better supervisor for my PhD studies. Thank you very much.

Similarly, profound gratitude goes to Professor Gabriel Leon, who gave me an incredible amount of support and guidance in all stages of this project. His help in the last semester of my PhD studies was invaluable and I am very thankful for his generosity.

I am heartily grateful to my examiners, Professor Toke Aidt and Professor Graham Denyer Willis, who accepted to dedicate their precious time to evaluate this work.

The Department of Political Economy at King's College London has provided me with a very stimulating environment in what concerns the extraordinary quality of its students and academic staff, and that experience will leave marks beyond this thesis. I express my thanks to Professor Florian Foos, Professor Rubén Ruiz-Rufino, and all my PhD colleagues for the inspirational discussions and suggestions. David Muchlinski has taught me many things about machine learning and I am very thankful for that. Professors Kostas Matakos, John Meadowcroft, Mark Pennington, and Emily Skarbek have supported me in so many ways that I cannot even count. Thank you very much for the opportunities you provided, for your time and skills, and for having faith in me.

I have been very fortunate to work with bright and motivated co-authors during this and other related projects. I am grateful to Rodolpho Bernabel, Gabriel Cepaluni, Diogo Costa, Guilherme Jardim Duarte, Malte Hendrickx, Robert McDonnell, Umberto Mignozzetti, and Gary Uzonyi for having taught me so much about both scientific research and life in general.

One of the best experiences I have had in the previous years was to become affiliated with the Mercatus Center and the Institute for Humane Studies. There, I have met a community of smart, hard-working and innovative scholars that have in many ways contributed to this research. I would like to express my gratitude to Paul Dragos Aligica, Nigel Ashford, Peter Boettke, Don Boudreaux, Brandon Brice, John Buchmann, Jeff Carroll, Chris Coyne, Courtney Dunn, Luke Foster, Bobbi Herzberg, Marcus Hunt, Nick Jacobs, Isaac Jilbert, Peter Lipsey, Ian Madison, Kelly Nelson, Gabriel Oliva, Raj Patel, Pablo Prieto, Blaž Remic, Virgil Storr, Alienor van den Bosch, Richard Wagner, and Larry White for all the engaging conversations. I have learned so much from you.

Completing this work would have been all the more difficult were it not for the support and constructive criticism provided by my friends. Specially, I would like to thank Veronica Adaeva, André Amaro, Paulo Roberto Araujo, Aline Bagatin, Fábio Barros, Michel Hulmann, Guilherme Kerr, Maurício Pantaleão, Letícia Sene, and the participants of academic seminars at King's College London and the Prometheus Institute in Berlin. I am deeply indebted to them for their insightful comments and hard questions.

I am also sincerely thankful to several present and past members of the Friedrich Naumann Foundation for Freedom. In 2009, I received a scholarship to attend a short course in Germany, and that first experience abroad has transformed my life ever since. The present work is, to a large extent, a result of the push I received from the Naumann Foundation almost 10 years ago. I owe a great debt of gratitude to Cidália Achten, Gulmina Bilal, Rainer Erkers, Beate Forbriger, Verena Gierszewski, Birgit Lamm, Monica Lehmann, Stefan Melnik, and Wulf Pabst for their support. As time goes on, it is easy to see the huge impact they have had on my academic career.

A very special word of thanks goes for my mother Rosali and my grandmother Maria. They have been absolutely wonderful over the years and have gone above and beyond to provide me with a good education. This thesis is dedicated to them. *Nós conseguimos*.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The literature on political and criminal violence has increased exponentially over the last decades. Although interstate wars have traditionally occupied a privileged position in political science, scholars have broadened their scope to include a myriad of hitherto understudied phenomena into their research agendas. Civil wars (Collier and Hoeffler 2004; Fearon and Laitin 2003; Kalyvas 2006), genocides (Mamdani 2014; Power 2013), ethnic conflicts (Kaufmann 1996; Montalvo and Reynal-Querol 2005; Sambanis 2001), wartime sexual abuse (Cohen 2013; Wood 2006, 2009), electoral violence (Höglund 2009; Wilkinson 2006), state-sponsored killings (Harff and Gurr 1988; Krain 1997, 2005; Uzonyi 2014), terrorism (Bueno De Mesquita 2005; Bueno de Mesquita and Dickson 2007; Pape 2003), drug-related violence (Holmes et al. 2006; Lessing 2015; Richani 2013; Shirk 2010), street gangs (Franzese et al. 2016; Jones 2009; Rodgers 2006; Sobel 1987), and prison gangs (Dias 2011; Freire 2014; Skarbek 2011a, 2012, 2014) have recently moved from the margins to the centre stage of the discipline. The present dissertation contributes to this expanding field.

In order to clarify crucial aspects of my research topic, I employ an eclectic combination of research designs. The methods range from qualitative case studies to machine learning algorithms. This diversity not only reflects the multiple aspects of organised violence, but it is also a pragmatic response to problems which are common in this area, such as incomplete data, reporting bias, and

model uncertainty. By using an array of methodological tools, I hope to overcome some of these challenges.

Regarding the geographical scope of this dissertation, two of the following chapters deal with issues of violence in Latin America, specially in Brazil. According to the World Bank, Latin America is home to about 8% of the global population, yet it accounts for more than 30% of the world's homicides.[1] Moreover, the yearly ranking by the Citizen's Council for Public Security and Criminal Justice (Consejo Ciudadano para la Seguridad Pública y la Justicia Penal), a Mexican non-governmental organisation, shows that 43 of the 50 most violent cities in the world are located in Latin America, including all of the top 10.[2] Given the severity of violence in the continent, Latin America was expected to be an important part of this study.

Brazil exemplifies many of the challenges of fighting violence in developing nations. The country has the highest absolute number of homicides in the world, about 56,000 per year, and it hosts 19 of the 50 world's deadliest cities according to the above-mentioned ranking (Waiselfisz 2014; Consejo Ciudadano para la Securidad Pública y Justicia Penal 2014; United Nations Office on Drugs and Crime 2013). Homicide rates have increased markedly after democratisation (1985), and whereas Brazil has tried several policies to reduce violence, the results are yet to be evaluated in a consistent fashion.

The next chapter attempts to address this issue. Although Brazil remains notably affected by civil violence, the state of São Paulo has made significant inroads into fighting criminality. In the last decade, São Paulo has witnessed a 70% decline in homicide rates, a result that policy-makers attribute to a series of crime-reducing measures implemented by the state government (Goertzel and Kahn 2009; Kahn and Zanetic 2005). While recent academic studies seem to confirm this downward trend, no estimation of the total impact of state policies on homicide rates currently exists. I fill this gap by employing the synthetic control method (Abadie and Gardeazabal 2003; Abadie et al. 2010, 2014), a generalisation of differences-in-differences (Angrist and Pischke 2008; Bertrand et al. 2004; Imbens

[1] See: https://goo.gl/d2WC3V. Access: April 2017.

[2] For the complete ranking, see http://www.seguridadjusticiaypaz.org.mx/biblioteca/prensa/summary/6-prensa/239-las-50-ciudades-mas-violentas-del-mundo-2016-metodologia. Access: February 2018.

and Wooldridge 2009), to compare these measures against an artificial São Paulo. The results indicate a large drop in homicide rates in actual São Paulo when contrasted with the synthetic counterfactual, with about 20,000 lives saved during the period. The theoretical usefulness of the synthetic control method for public policy analysis, the role of the *Primeiro Comando da Capital*, a local prison gang, as a moderating variable, and the practical implications of the security measures taken by the São Paulo state government are also discussed.

Chapter three offers a rational choice account for Brazil's *jogo do bicho*, or the 'animal game', possibly the largest illegal gambling game in the world. The lottery has been running for over 120 years and according to estimations of Fundação Getúlio Vargas, a Brazilian think tank, it profits up to 800 million dollars per year.[3] The *jogo do bicho* has exerted a significant impact on the Brazilian society. The lottery has been a major sponsor of the Carnival Parade in Rio de Janeiro, which is among the world's most famous popular festivals, and it has remained an important driver of state corruption in the country (Bezerra 2009; Chazkel 2011; DaMatta and Soárez 1999; Labronici 2012; Magalhães 2005; Soares 1993). I investigate the institutions that have caused the *jogo do bicho*'s notable growth and long-term survival outside the boundaries of the Brazilian law. I show how *bicheiros* or bookmakers promote social order, solve information asymmetries, and reduce negative externalities via costly signalling and the provision of club goods. I also explain the emergence of the informal rules that govern the game as well as their enforcement mechanisms.

The last chapter presents an empirical evaluation of explanations for genocides and politicides. Although the literature on state-sponsored killings has grown significantly over the last decades, it remains unclear whether estimates are robust to different model specifications, or which variables accurately predict the onset of large-scale violence. I employ extreme bounds analysis and distributed random forests to test the sensitivity of 40 variables on a sample of 177 countries from 1945 to 2013. The results show that GDP per capita, the post-Cold War period, and stable political regimes

[3] See http://goo.gl/9kNeX8 and http://goo.gl/8FSAZl (in Portuguese). Access: April 2017

are negatively associated with mass killings. In contrast, ethnic diversity, civil wars, and previous political turmoils increase the risk of state-led violence.

Chapter 2

Evaluating the Effect of Homicide Prevention Strategies in São Paulo, Brazil: A Synthetic Control Approach

2.1 Introduction

Brazil has long been ravaged by an undeclared civil war. According to the Citizen Council on Public Security and Criminal Justice, a Mexican think-tank, 19 of the 50 most violent cities in the world are located in Brazil (Consejo Ciudadano para la Securidad Pública y Justicia Penal 2014).[1] The 2014 Violence Map survey shows that 56,337 people were murdered in Brazil in 2012 alone, the highest incidence rates of intentional homicides on the planet (Waiselfisz 2014; United Nations Office on Drugs and Crime 2013). Paradoxically, the sharp rise in lethal violence has occurred during Brazil's longest period of political openness (Ahnen 2003; Pinheiro 2000, 2001). Murder rates have almost doubled over three decades of democracy, jumping from 15 homicides per 100,000 people in 1985 to roughly 29 per 100,000 in 2012 (Waiselfisz 2014).[2]

[1] The study disregards war zones and cities with unavailable data.

[2] Cerqueira (2013) argues that the actual rates may be different from the official statistics. He states that many homicides from 1996 to 2010 were (intentionally or not) misclassified as 'death by undetermined causes.' After performing

São Paulo has traditionally occupied a key position in Brazil's violence statistics. It is the country's richest and most-densely populated state, and in the 1990s its homicide rate was roughly 50% higher than the national average (Barata and Ribeiro 2000, 120). Some areas of the namesake capital city had even worse numbers. Between 1996 and 1999, the ramshackle districts of Jardim São Luiz and Jardim Ângela had respectively 103 and 116 violent deaths per 100,000 residents (Cardia et al. 2003, 8), figures that placed them amongst the deadliest neighbourhoods on the globe (World Health Organization 2015).

Nevertheless, the state of São Paulo has experienced a drastic reduction in homicides during the last years (Camargo 2007). The decline is so remarkable that some authors have called it 'the great homicide drop' (Goertzel and Kahn 2009). The city of São Paulo, which is currently home to about 11 million inhabitants, provides a telling example. Over a span of only seven years (2000–2007), the number of annual violent deaths in the capital fell from 5,979 to 1,311, a 78% decrease.[3] Significantly, São Paulo city became the safest state capital in Brazil (Waiselfisz 2011).

São Paulo's success should be attributed to local factors. From 1999 onwards, the state government created or expanded a number of policies that have arguably contributed to the decrease in criminality. In a move coherent with the basic tenets of the economics of crime (e.g. Becker 1968; Cornish and Clarke 2014), the administration increased the certainty and the intensity of punishment to discourage potential offenders. Amongst other measures, the government implemented strict gun control policies (Goertzel and Kahn 2009), raised incarceration rates (Salla 2007), and imposed harsher sentences on those convicted of a crime (Carvalho and Freire 2005).

But whereas several authors acknowledge the effectiveness of these policies, few quantitative studies have gone beyond statistical correlations to justify their arguments. In the case of São

data correction procedures, the author estimates that the number of homicides in Brazil during that period should be 18.3% higher than the reported figures. Recent criticism about the quality of São Paulo homicide data can also be found at http://goo.gl/x0pHac (in Portuguese). Access: January, 2016. In this article, I avoid these issues by using obituary data instead of police records.

[3]The homicide statistics cited in this paragraph come from the Centre for the Study of Violence, a research group of the University of São Paulo. Their data set can be found at the following electronic address: http://nevusp.org/downloads/bancodedados/homicidios/distritossp/num-homicidios-distritos-2000-2007.htm. Access: March, 2016.

Paulo, a major difficulty is separating the state's particular time trend to that of Brazil. Ideally, one should compare São Paulo to a control case that shares the same characteristics of the existing state, except that it has not been subjected to the specific set of policies implemented by the São Paulo state government. This thought exercise, which emulates the logic of a controlled experiment (Angrist and Pischke 2008; Imbens and Rubin 2015; Holland 1986; Morgan and Winship 2014), would allow practitioners to untangle the effects of homicide reduction programmes from other potential confounders.

In this paper, I employ the synthetic control method (henceforth SCM) to approximate this experimental ideal and measure the total causal effect of post-1999 public policies on São Paulo homicide rates. The method consists of creating an artificial counterfactual to estimate the impact of a given intervention on a unit of interest. SCM has gained widespread acceptance in many fields, having been successfully applied in political science (Abadie et al. 2014; Montalvo 2011), economics (Billmeier and Nannicini 2013; Coffman and Noy 2012; Jinjarak et al. 2013), education studies (Hinrichs 2012), and public health science (Heim and Lurie 2014). However, SCM has rarely, if ever, been used to evaluate homicide prevention strategies in São Paulo, despite being a useful tool for this particular type of question. SCM was specifically designed for situations where there is only one treated unit of interest, no readily-available counterfactual, and no certainty as to whether the treated and the control units follow parallel trends after the intervention (Abadie and Gardeazabal 2003; Abadie et al. 2010, 2014). Moreover, SCM also has some of the desirable properties of popular causal inference tools such as differences-in-differences (Angrist and Pischke 2008; Bertrand et al. 2004) and matching estimators (Dehejia and Wahba 2002; Ho et al. 2007; Rubin 1973; Stuart 2010).

I find that from 1999 to 2009, about 20,000 lives were saved in São Paulo. When compared to a synthetic counterfactual, São Paulo's actual homicide rates were less than 50% of what would be expected in the absence of policy implementation (15 versus 32 homicides per 100,000 people).

Additional tests confirm the robustness of the results and indicate a 96.3% chance of a causal effect in the intervention period.

The article is structured as follows. Section 2.2 discusses how deterrence provides a useful framework to understand the reduction in homicide rates in the state. I also examine an alternative hypothesis for the drop in crime in São Paulo – the rise of the *Primeiro Comando da Capital* – and argue that the prison gang should be regarded as a moderator, but probably not as an independent cause of homicide reduction. Section 4.2 presents a justification for, and a technical explanation of, the synthetic control method. Section 2.4 describes the data used in this paper. Section 2.5 discusses the results of the models and presents several robustness tests. Section 4.5 offers some concluding remarks.

2.2 Theoretical Background

2.2.1 Deterrence, Information and the Drop in Homicides

A myriad of explanations have been proposed for the fall in homicide rates in São Paulo. Some authors have stressed the importance of long-term factors on local levels of violence. Mello and Schneider (2010) claim that the shrinking of the proportion of males in the 15–25 age bracket has led to fewer violent deaths at both state and city levels. Hughes (2004) argues that São Paulo's spatial segregation patterns have had a lasting impact on murder rates. Barata and Ribeiro (2000), in turn, posits that macroeconomic conditions, mainly inequality indicators, are positively correlated with violent crime in São Paulo.

Structural variables have likely been important in reducing violence, but the role of public policies should not be underestimated. The Brazilian Social Democracy Party (*Partido da Social Democracia Brasileira*, PSDB), which has ruled São Paulo since 1995, has repeatedly asserted its commitment to reducing urban crime throughout the state (Bueno 2014). In 1998, former governor Mário Covas –

then running for re-election – set the ambitious goal of "slashing criminality rates in half" during his second term in office (Santos 2008). This commitment was then followed by his vice-governor and successor, Geraldo Alckmin, who has expanded those measures and taken a notoriously tough stance on crime (Feltran 2012a).

Methods of crime prevention have received considerable attention from the authorities. Firstly, the São Paulo government significantly increased incarceration rates in the past decade (Salla 2007). The state currently holds around 200,000 convicts in prison (35% of Brazil's inmate population) and adds another 15,000 inmates to the official statistics every year (Brasil de Fato 2013). Furthermore, prisoners have also become subject to harsher legal punishments. The São Paulo administration has also been making large use of the *Regime Disciplinar Diferenciado* (Special Disciplinary Regime), which provides for up to 360 days of solitary confinement for disobeying the law (Carvalho and Freire 2005).

Secondly, the state government has successfully enforced a ban on gun possession in São Paulo. Studies show that this policy has been effective in reducing homicides resulting from both drug-related crimes and domestic disputes (Goertzel and Kahn 2009; Kahn and Zanetic 2005). Furthermore, the impact of the Brazil's 2003 National Disarmament Act was particularly pronounced in São Paulo. Cerqueira and Mello (2013) argue that between 2005 and 2007 the enforcement of the anti-firearm legislation was responsible for saving between 2,000 to 2,750 lives in cities with more than half a million inhabitants in the state of São Paulo.

This set of policies is largely in line with the rational choice theory of crime (e.g. Becker 1968; Ehrlich 1973; Levitt 1996, 1997; Paternoster 2010). The rational choice school posits that criminals are motivated by utilitarian cost-benefit analysis. Individuals calculate what the possible trade-offs are between the benefit of the committing a crime and the risk of being punished for it. Criminal offenders, therefore, are in no way different from non-criminals: the only difference between them

is their *choices* (Nagin 2007). To reduce criminality, policy-makers have to ensure that the costs of committing a crime outweigh the eventual utility an individual derives from it.

Deterrence measures have been complemented by investments in police intelligence. In 1999, the state administration created a new system for crime prevention, Infocrim (Risso 2014, 3) The system gathers geo-coded information on homicides and maps the most important 'hot spots' of criminal activity in the state. The government has also developed a new photo database, Fotocrim, to speed up the process of facial recognition of criminals (Mello and Schneider 2010, 3).

More information improves the effectiveness of police strategies via two mechanisms. On the one hand, police forces can be quickly moved to where they are most needed. This reinforces the role of deterrence as it increases the likelihood of punishment for criminals. On the other hand, the system also makes clear what regions are making progress in reducing crime. This allows police chiefs to monitor local personnel and take measures to improve performance if required.[4]

Recent evidence shows that the intelligence system has effectively lowered the crime statistics in São Paulo. Using a spatial differences-in-differences estimator, Cabral (2016a) argues that Infocrim has had a large negative impact on homicide rates in the municipalities where it was implemented. The author also notes that the effect remains important even after accounting for possible displacement effects. As expected, some criminals did take their activities elsewhere after the creation of Infocrim, but this movement has not offset the benefits of the system.

How well have these policies performed over time? The results suggest a favourable outlook. Compared to other Brazilian states, São Paulo is an outlier when it comes to homicide rates. Despite the fact that crimes against property have declined little over the last decades,[5] the number of violent deaths per 100,000 inhabitants shows a steep downward trend. Figure 2.1 presents the evolution of homicide rates in São Paulo in comparison with the Brazilian average.

[4]See: http://goo.gl/kqLhYb (in Portuguese). Access: August 2016.

[5]Recent data on property crimes in São Paulo can be seen at http://www.ssp.sp.gov.br/novaestatistica/Pesquisa.aspx (in Portuguese). Access: July 2016.

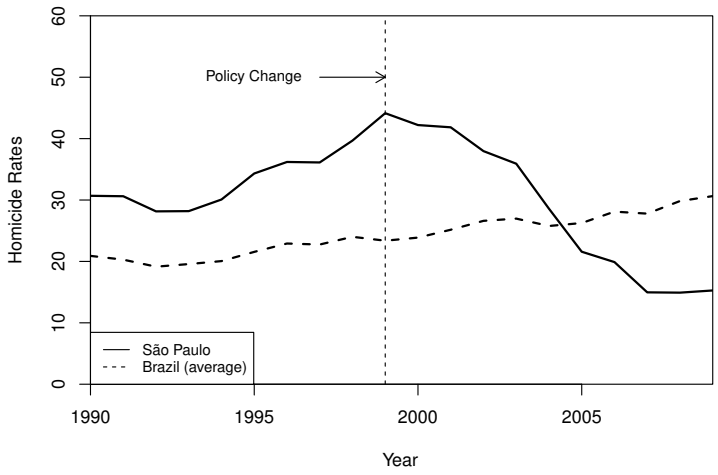

Figure 2.1: Homicide Rates per 100,000 Population – São Paulo and Brazil (Excluding São Paulo)

The trends are even more striking if we consider that deterrence policies are still controversial in the literature. Barbarino and Mastrobuoni (2014), Buonanno and Raphael (2013), Levitt (1996, 2004) and Owens (2009) claim that incapacitation measures effectively reduce crime, but Eck and Maguire (2006) and Beattie and Mole (2007) suggest that increases in police forces and incarceration rates in the United States and in Canada did not lead to expected outcomes.

There is good evidence that incapacitation measures have worked well in São Paulo during the last decade. Gun-related homicides have declined about 74% from 2001 to 2008 (Peres et al. 2011) at the same time when São Paulo has experienced an increase of 770% in the arrests of repeated murderers (Manso 2012, 36). Although there are studies that indicate possible 'hardening effects' of imprisonment, that is, longer sentences may positively affect an individual's tendency to commit further crimes (e.g. Chen and Shapiro 2007; Glaeser et al. 1996; Western et al. 2001), the São Paulo case appears to suggest otherwise. Moreover, violent deaths have decreased in all population strata, but especially amongst males (-74.5%), 15 to 24 year-old men (-78,0%) and those who live in extreme poverty (-79,3%), groups that are generally associated with criminality (Peres et al. 2011).

Nevertheless, it is difficult to know which of the policies have contributed more to this large homicide reduction. Not only we do not have disaggregated data to test preliminary hypotheses, but

there may be large interaction effects amongst different public security measures. Therefore, at the moment it is not possible to disentangle micro-level causes from macro effects. But the aggregated impact of the anti-crime policies can be correctly identified if there is no other variable in the causal path leading from the policies mentioned above to our dependent variable (state homicide rates). I argue below that this type of estimation is feasible for the São Paulo case. To back this claim, I suggest that a competing explanation for the homicide drop in São Paulo – the rise of the PCC – interferes only with the direct effect of the policies on crime, but not with their *total* effect. In this sense, the synthetic control method provides a plausible identification strategy for my question of interest.

2.2.2 Alternative Explanation: The Emergence of the PCC

A recent hypothesis attributes the decrease in violent deaths in São Paulo to the *Primeiro Comando da Capital* (First Command of the Capital, henceforth PCC) (Biondi 2010; Dias 2009a, 2011; Feltran 2010, 2012a; Willis 2015). The PCC is a prison gang that emerged in the early 1990s as a response to the demands of a growing prison population. The PCC provides personal security and financial assistance to their members and affiliates. The gang's internal statute clearly declares that "[…] those who are in liberty [must contribute] to the brothers inside prisons [PCC members] through lawyers, money, help to family members and prison outbreak operations" (Folha de São Paulo 2001).

A group of scholars argue that the PCC significantly contributed to the reduction in violence mainly through the São Paulo prison system. At least since the mid-2000s, these authors argue that the PCC has been able to emerge as an undisputed mediator and solve conflicts between inmates. Dias (2009b, 83) writes that "[…] when unable to constitute a universal source of regulation, the official law leaves gaps which are filled by informal instances – such as the Primeiro Comando da Capital (PCC), in the prisons of São Paulo." The gang has implemented informal courts that resemble state institutions, and those meetings have progressively replaced other forms of popular justice such

as lynchings or the hiring of target killers (Feltran 2012b, 3). Moreover, the *Comando* has developed a series of assertive ways to terrorise inmates. Since the PCC's threats are credible, the group is able to impose discipline within the São Paulo prison system (Biondi 2010; Dias 2009a).

Paradoxically, the PCC might have also helped to reduce crime in São Paulo by collaborating with street-level police. The Brazilian state does not hold a perfect monopoly of force in many areas of the country (Arias 2009; Feltran 2012a; Hughes 2004; Pinheiro 2000), thus access to local knowledge may prove vital for the success of a given operation. In this regard, the PCC and the state may collude if the situation is beneficial to both, and as such there is an informal–but potentially unstable–"killing consensus" in the state (Willis 2014, 2015).

There has been a vigorous debate over whether the PCC has had a significant impact on violence rates. A few authors see the PCC as the sufficient condition behind the homicide rates decline across São Paulo state (Biondi 2010; Dias 2009a, 2011), whilst others take a more nuanced view of the role of the prison gang (e.g. Willis 2015). But both groups of scholars affirm that, based on their first-hand experience, the PCC is the key explanatory variable behind the drop in murders in São Paulo.

Recent econometric works, however, do not seem to confirm that argument. Marcelo Nery has found no convincing results in favour of the 'PCC hypothesis' using geo-referenced data for São Paulo (BBC 2016). Biderman et al. (2016) use anonymous calls to a crime hotline as a proxy for PCC presence in São Paulo city favelas. The authors suggest there is some support for the idea that the criminal syndicate reduces lethal violence in areas under its control, but PCC presence corresponds to only a minor drop in violent crime. Although the PCC impact is not negligible, the gang is not a sufficient condition for the homicide decline.

Another counter-argument to the PCC thesis is that homicides also decreased in areas and groups over which the PCC does not exert control. Firstly, descriptive statistics show that the decline in violent deaths started *before* the PCC's expansion period.[6] Secondly, the drop in crime was evenly

[6] As shown in figure 2.1, São Paulo's homicide rates started to drop in 1999. The PCC consolidated their power in the prison system only in the mid-2000s (Dias 2011).

distributed throughout the state: urban and rural areas, small and large cities alike experienced fewer murders.[7] Finally, as noted above, Peres et al. (2011) point out that violent death rates decreased in *all* age groups and social classes in the city São Paulo. Hence, cohorts that do not correspond to typical PCC members (such as the elderly or middle-age females) are also less affected by violence.

It seems that the influence of the PCC on physical violence has been overstated. It is unlikely that the PCC – which is underfunded for its size[8] – could have achieved such deep penetration into society and lowered the violence levels across all population groups in the whole state.

Yet, the group's importance cannot be fully dismissed either. Data on PCC-controlled areas are likely to contain measurement errors that may bias the coefficients, thus caution is required before making strong causal claims on this discussion. Despite mounting observational evidence that the PCC may not provide a complete explanation to São Paulo's lower crime rates, the argument could only be comprehensively tested in a counterfactual case in which the PCC is present and the state policies are not.[9] Currently-available data do not allow us to evaluate such scenario.

2.2.3 Causal Paths, Moderators, and Total Effects

A methodological issue remains. If we are to estimate the causal effect of the public measures on the crime rates, how should we proceed? I have noted above that the specific impact of micro-level policies cannot be evaluated due to lack of data. Nonetheless, it is theoretically possible to estimate the *total* effect of policies on crime.

The difference between direct and total effects can be understood as follows. The direct effect captures the sensitivity of a dependent variable Y to changes in X when this relationship is not mediated by any other variables in the model. Holding all factors constant, the direct effect is a

[7] See: http://www.fenapef.org.br/27764/ (in Portuguese). Access: July 2016.

[8] A Parliamentary Commission of Inquiry has stated that the PCC earns about 16 million Brazilian Reals per month, which amounts to approximately 60 million US dollars per year. See: http://goo.gl/FwhPa3 (in Portuguese). Access: July 2016. Given the size of the organisation and its undisputed position as the leading crime syndicate in São Paulo, the figures are rather small. As a comparison, Mexico's Sinaloa Cartel profits about 3 billion dollars per year, a sum comparable to the annual earnings of Netflix or Facebook. See: http://nyti.ms/1B09qyV. Access: July 2016.

[9] I would like to thank an anonymous reviewer for highlighting this point.

causal chain of length one (Sobel 1987, 160) and could be described simply as $X \rightarrow Y$. In turn, the total effect can be defined as $P(Y_x = y)$, that is, "the probability that response variable Y would take on the value y when X is set to x by external intervention" (Pearl 2001, 1572). The total effect is the sum of direct and indirect (or mediated) effects.

In our case, gun control, incarceration, and police intelligence have likely had a direct effect on homicides. Combined, these variables comprise a direct aggregate policy effect. The omission of a variable measuring the impact of the PCC could bias such an effect, but not interfere with the *total policy effect*. This point is worthy of further consideration. The total policy effect would be unbiased under the assumption that the PCC is in fact a *moderator* between the public policies and the homicide rates, even if the gang's impact over the violence levels is not particularly large.

Although this argument has rarely been posited in such terms, this position is largely supported by the qualitative literature on the PCC. Fieldwork research generally traces the group's origins and growth to the rising incarceration rates in São Paulo and the need for protection amongst prisoners (Dias 2011; Manso and Godoy 2014). Like other prison groups, the PCC would only mobilise resources to provide welfare and act as an arbitrator under the condition that the certainty of punishment by the state is high (Skarbek 2011b; Freire 2014). Had the state not increased the costs associated with crime, the prison gang would not have expanded their reach, or even been created in the first place. Hence, the impact of the PCC on street-level violent deaths – if it exists – can be safely assumed to be a moderator effect.

Whereas it would be interesting for researchers to separate these types of effects and isolate the PCC from the other causal outcomes, such estimation is not possible at the state level. However, as these measures were implemented throughout São Paulo state at roughly the same time, their combined effect is computable even though their individual direct effects are not. To do so, it is only necessary to contrast the treated unit (São Paulo) with a counterfactual without the time-assigned treatment (1999 onwards) and evaluate the aggregated effect of the public policies.

This analysis can be estimated in a consistent manner with the synthetic control method. In the following sections I describe how the method creates a valid counterfactual case under a certain set of assumptions. The assumptions are: 1) the PCC is an outcome, not a cause, of the crime-targeting policies; 2) the model does not include unnecessary control variables; 3) interpolation bias is not very severe because the cases in the 'donor pool' are relatively similar to the treated unit.

2.3 Methods

The synthetic control approach provides an adequate solution for two enduring problems in the social sciences: the arbitrary selection of comparative cases and the poor estimation of causal effects when few pre-treatment observations are available (Abadie and Gardeazabal 2003; Abadie et al. 2010). With respect to the first issue, scholars often resort to ambiguous criteria in their choice of control units. This practice ends up casting doubts over the validity of their selected counterfactual (Abadie et al. 2011). The synthetic method provides a reliable comparative case by adopting a purely data-driven process in order to select a counterfactual. Also, the researcher can still specify what control cases enter the 'donor pool.' In this sense, qualitative expert knowledge can be incorporated in the estimation via the selection of cases.

Regarding the second issue, the accurate estimation of coefficients from a small number of cases, SCM employs a consistent statistical solution to problems of incorrect data extrapolation and model dependence. SCM can be understood as a combination of matching with differences-in-differences. SCM uses matching as a flexible pre-processing tool to reduce imbalance between treated and control units (Ho et al. 2007; Rubin 1973, 2006). But unlike matching, SCM deals with only one treated unit over time. Therefore, the method can also be interpreted as a semi-parametric extension to differences-in-differences estimators in which both treated and control units are not required to follow parallel trends in the whole period. (Abadie 2005). By combining semi-parametric matching with

differences-in-differences, SCM provides a rigorous yet versatile method to evaluate time-dependent treatment effects.

The method works as follows.[10] SCM starts with the assumption that one case in the sample has received a treatment. The treatment is defined as a time-delimited event that affects the unit of interest, such as the implementation of a new policy or the outbreak of a conflict. SCM also requires a series of control cases to estimate the models, that is, units that did not receive the treatment during the same period. These cases are often related to the treatment case in some meaningful way, and natural choices for the donor pool are provinces within the same country, or states that share important characteristics. These traits can also be more specifically defined and included as quantitative variables in the estimation models.

SCM then selects a few cases from the donor pool to create a new, artificial control for the treated unit of interest. The main goal of SCM is to construct a counterfactual that resembles the treatment unit more closely than any individual control in the donor pool. Cases are combined in way similar to a weighted average, in which controls that are more similar to the treated unit receive more weights. The weights make explicit the contribution of each separate case to the synthetic control, what also increases the transparency and reliability of the method (Abadie et al. 2014). The closer the synthetic control matches the original treated unit before the assignment of the period, the better the quality of the counterfactual.

The method uses an algorithm to minimise the difference between the control cases and the treated unit before the intervention. The authors adopt the mean squared prediction error (MSPE) as a measure of fit (Abadie and Gardeazabal 2003). MSPE is simply the difference between the fitted and the observed trends of the treatment case. A small value means that the two lines are highly correlated and the artificial control is a good approximation of the missing counterfactual in the post-intervention period. In our case, the counterfactual would be São Paulo from 1999 to 2010 without the crime-reducing policies.

[10] Please refer to Appendix 2.7 for a formal presentation of the synthetic control method.

SCM has an intuitive interpretation. Although numeric summaries and other statistics can be obtained from the model, a simple time series graph is usually enough to assess the results. The causal effect is the difference between the treated and the synthetic cohort. The larger the post-treatment gap, the stronger the treatment impact.

As with all types of observational studies, SCM can also suffer from omitted variable bias. One can never be sure whether all required confounders have been included in a given model. However, the graphical output of the SCM helps diagnose the presence of large disparities between treatment and control cases. If the trends follow similar paths during the control period, it provides some indication – albeit only informally – that omitted variable biases are not driving the output. This bias can also be mitigated with expert knowledge. Econometric studies show that the inclusion of a large number of covariates and post-treatment variables to correct for omitted variables bias can actually *worsen* the problem (Achen 1992, 2002; Clarke 2005, 2009; Pearl 2009). This is particularly true for matching methods. Authors have noted that 'over-matching' can lead to severe statistical bias (Baser 2006; Brookhart et al. 2006; Marsh et al. 2002). In this regard, the most plausible solution seems to be attention to the trends and sensible selection of control variables. As I discuss below, the covariates included in this paper are some of the most robust quantitative predictors of homicides.

Furthermore, placebo tests can be run to test the robustness of the findings. For instance, researchers can include 'in-time placebos,' dates under which the treatment *did not* occur. Results should change only in the period when the treatment starts and not at any other point in time. Moreover, scholars can also add 'in-space placebos' to their models. This test consists of adding different members of the donor pools into the models to see if the estimation varies (Abadie et al. 2014). Finally, one can also compare the effects of the treatment of interest by creating a distribution of synthetic cohorts, where every unit (treated or not) is matched with a specific synthetic control case. The parameter of interest should still be relevant. I employ all of these tests in this article and the results can be seen in the following sections.

2.4 Data

I build panel data for the variables *Homicide Rate*, *State GDP per Capita*, *State GDP Growth*, *Years of Schooling*, *Gini Index*, *Natural Logarithm of Population* and *Population Living in Extreme Poverty*. These variables are very common in the specialised literature[11] and represent important social and economic factors I wish to control for.

The unit of analysis is State-Year. I have data from all of the 26 states plus the capital city (Distrito Federal), ranging from 1990 to 2009. The data for years prior to 1990 are scarce and for years after 2009 have not yet been published. All data used in this paper come from the same source, the *Instituto de Pesquisa Econômica e Aplicada* (IPEA), a government-led research group.[12]

My dependent variable measures the number of homicides per 100,000 inhabitants, which is the most commonly used unit of analysis for lethal violence. This variable was coded by the Brazilian Health Ministry from obituary records, therefore it is less likely than police files to suffer from intentional misrepresentation.

There are six control variables in the models. *State GDP per Capita* is adjusted in 2010 Brazilian Reals (at the time 1 Brazilian Real bought roughly 0.5 U.S. dollars). *State GDP Growth* is measured in constant 2010 Brazilian Reals and varies by percentage points. *Years of Schooling* describes the average number of years of formal instruction at educational facilities (males and females, 25 years old or more.) *Gini Index* is a measure of inequality, ranging from 0 to 1 where 0 is the most equal and 1 the most unequal. *Natural Logarithm of Population* represents yearly projections of the state population. Since Brazil only runs a census every 10 years, these projections represent the most accurate data available. I have taken the natural logarithm of this variable to account for size effects. Finally, *Population Living in Extreme Poverty* describes the percentage of the state population which do not meet the minimum intake of 2,000 calories per day. This is the only variable that I created

[11] For overviews of cross-national studies of homicide, see LaFree (1999), Nivette (2011) and Trent and Pridemore (2012).

[12] The data are publicly available at http://www.ipeadata.gov.br/. The original data files have also been added to https://github.com/danilofreire/homicides-sp-synth for reproducibility purposes.

specifically for this study. It was coded by simply taking the number of individuals classified as extremely poor by the IPEA and dividing this number by the state's total population.[13]

2.5 Analysis

2.5.1 Main Model

I construct the synthetic cohort (*Synthetic São Paulo*) by imputing information from all of the Brazilian states plus the Federal District. The synthetic control method outputs a set of weights for states and variables such that the treatment state is approximated optimally by these weighted components. This method not only provides a quantitative way of selecting comparison cases but also gives us a much better baseline to compare with the treatment unit. Synthetic São Paulo is constructed using six states, *i.e.*, the six out of the 27 possible cases that received non-zero weights. Table 1 shows that the states that best synthesize São Paulo are, respectively, Santa Catarina (0.274), Distrito Federal (Brasília) (0.210), Espírito Santo (0.209), Rio de Janeiro (0.169), Roraima (0.137) and Pernambuco, which only accounts for 0.01 of the weights. In this regard the state selection does not appear as a complete surprise. Apart from Roraima, the other members of the federation are richer, more densely populated and better schooled than the country average, thus being indeed similar to São Paulo.

Table 2.1: Synthetic Weights for São Paulo

State	*Synthetic Control Weights*	*Predictor*	*Weights*
Santa Catarina	0.274	*Years of Schooling*	0.469
Distrito Federal	0.210	*State GDP per Capita*	0.275
Espírito Santo	0.209	*Homicide Rate*	0.241
Rio de Janeiro	0.169	*Population Living in Extreme Poverty*	0.009
Roraima	0.137	*Gini Index*	0.005
Pernambuco	0.001	*Ln Population*	0.001

Among the independent variables, only three out of six receive substantial weights. Given the data I could obtain, the predictors that receive more weight are Years of Schooling (0.469), State

[13] *Years of Schooling* and *Gini Index* had a small number of missing observations (about 15 percent) and those cases were imputed with linear interpolation. Both original and imputed variables are available online. See the supplementary appendix for further details on how to replicate this study.

GDP per Capita (0.275) and past Homicide Rate (0.241). The three remaining variables are much less relevant to the model. They are, respectively, the Population Living in Extreme Poverty (0.009), Gini Index (0.005) and Natural Logarithm of the Population (0.001). Table 2 compares characteristics of São Paulo and its synthetic control prior to policy implementation. We see that Synthetic São Paulo has very similar coefficients to those of the treatment unit. Moreover, the synthetic control clearly outperforms the sample means in all of the three relevant predictors. The worst measure is State GDP Growth, whose mean is about 2.6 whereas the figure for São Paulo is roughly 1.3 during that period. However, this outcome does not affect the results since the variables that received zero weight were discarded from the models.

Table 2.2: Homicide Rate Predictor Means Before Policy Implementation

Predictor	*São Paulo*	*Synthetic São Paulo*	*Sample Mean*
Years of Schooling	6.089	6.110	4.963
State GDP Per Capita	23.285	23.079	11.830
Homicide Rate	32.672	32.479	21.843
Population Living in Extreme Poverty	0.054	0.082	0.185
Gini Index	0.536	0.561	0.578
Ln Population	17.335	14.838	14.867
State GDP Growth	1.330	2.585	3.528

The results show that the synthetic control method has successfully created a valid counterfactual to our case of interest. Figure 2.2 depicts the evolution of the dependent variable for the treatment and synthetic control cases. We can see that São Paulo and synthetic São Paulo have very close homicide rates series for the period ranging from 1990 until 1998. From 1999 onwards we observe the trajectories departing sharply from each other. The increase in homicide rates shown in the graph is consistent with previous statistical evidence. It indeed confirms that São Paulo had higher than expected levels of lethal violence, which I noted in the first part of this text.

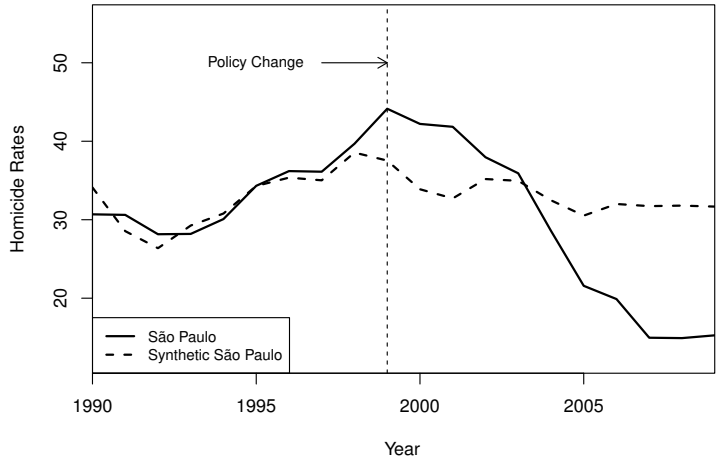

Figure 2.2: Trends in Homicide Rates: São Paulo versus Synthetic São Paulo

Despite the high levels of violence in 1999 – when the new crime-reducing programme was implemented – the number of homicides consistently declined until 2009. The trend is indeed monotonic and there is not a single peak in homicide rates after the policies have been put into practice. I interpret that as strong evidence in favour of the public policies.

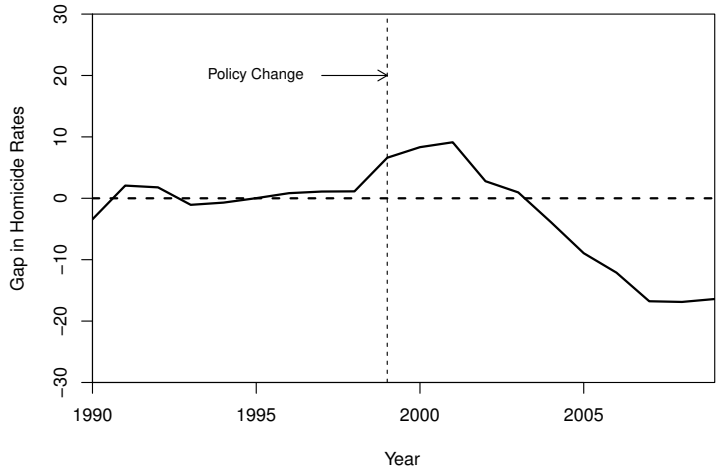

Figure 2.3: Homicide Rates Gap between São Paulo and Synthetic São Paulo

With respect to the size of the effect, in 1998 the homicide rate in São Paulo was around 40 deaths per 100,000 inhabitants. In 2009 – the last year for which data are available – the rate dropped to 15, whereas synthetic São Paulo observed above 30 deaths per 100,000. That means a gap of −20 deaths

for every 100,000 people in São Paulo in 2009, as can be seen in Figure 2.3. I estimate that the new policies implemented in São Paulo saved roughly 20,300 lives in the period from 1999 to 2009.[14] It is important to mention that the homicide rate in São Paulo continues to drop by the year, while the same is not happening in the rest of the country.

2.5.2 Robustness Checks

To further analyse the findings, I run five robustness tests. I first create an 'in-time placebo' synthetic control. This test consists of creating a false starting date for the intervention period to check if one could observe false treatment effects in the pre-treatment years (Abadie et al. 2014). If that were to be the case, the validity of the main results could be put into question. The result of this placebo test can be seen in Figure 2.4. When I run the model with 1994 as the year when there was a supposed policy change, the result shows that there is only a minor gap between both lines. In other words, the method does not indicate a definite departure of trends between treatment and control cases.

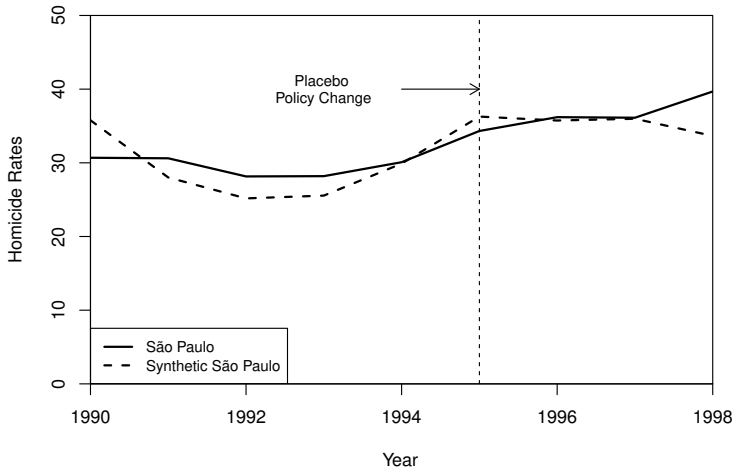

Figure 2.4: Placebo Policy Implementation in 1994: São Paulo versus Synthetic São Paulo

[14] My estimate of lives saved by the policies implemented in São Paulo is done as follows. I consider the years after policy implementation (1999–2009), then I sum the number of homicides in São Paulo in that period. This gives us 124,077 homicides between 1999 and 2009. I do the same procedure for the synthetic São Paulo; I sum the number of homicides in each state that makes the synthetic control in the period, while adjusting the contribution of each of these states by their respective weights in the synthesis. The number of homicides in synthetic São Paulo between 1999 and 2009 is 144,408. Finally, I subtract the number of homicides in the control by the number of homicides in the treatment. The result is 20,331 lives saved.

I also conducted a leave-one-out robustness test. In this test I drop the states composing the synthetic control one at a time. The main goal of this analysis is to evaluate whether a single control state is driving the results. This would suggest that the original synthetic control – which is composed of five states at a time – is probably not a reasonable counterfactual. The results of this analysis can be found in Figure 2.5. We see that the synthetic control (dashed line) is a reasonable amalgam of cases. Also, because the relative positions of treatment and controls are stable across controls, we observe that no control state is biasing the estimates.

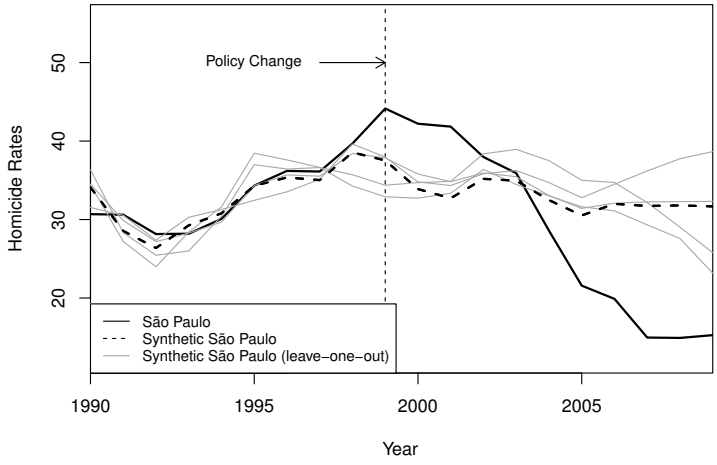

Figure 2.5: Leave-One-Out Distribution of the synthetic Control for São Paulo

Figure 2.6 shows the difference in homicide rates between the treated units and their synthetic controls. Here I estimate a synthetic control case for São Paulo and for each of the other 26 Brazilian states. This test assesses whether there is any previously unobserved national or regional trend that explains the original results. We observe that in São Paulo the homicide rate gap increases consistently during the treatment period, whereas the lines for the other states are moving randomly. Several lines fail to show any substantial difference between the state line and that of its synthetic counterfactual case. This indicates the results for São Paulo are unlikely to be a result of broader trends.

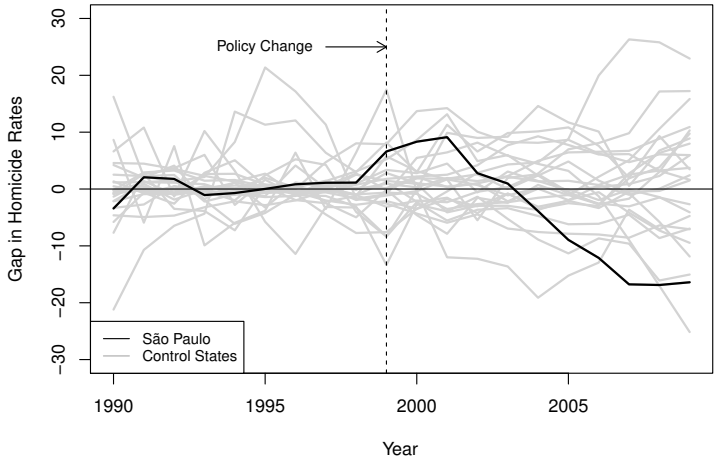

Figure 2.6: Permutation Test: Homicide Rate Gaps in São Paulo and 26 Control States

Figure 2.7 presents the same test displayed in figure 2.6, but it uses a stricter threshold for the simulated synthetic controls. The graph features cases in which the mean squared prediction error, a measure of goodness-of-fit, is no higher than twice that of São Paulo. That is, only placebos that have good synthetic matches were selected for the analysis (Abadie et al. 2010, 503). In this group, the negative gap for the homicide rate São Paulo is by far the most relevant, providing further evidence for the original findings.

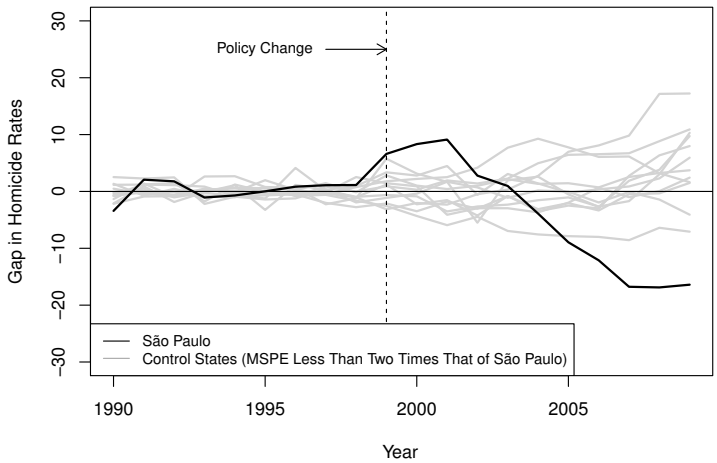

Figure 2.7: Permutation Test: Homicide Rate Gaps in São Paulo and Selected Control States

Lastly, I estimate another synthetic control using a different approach. I employ a Bayesian structural time-series model to verify the stability of the previous results (Brodersen et al. 2015). This inference procedure is similar to that described in section 4.2 and it also consists of matching pre-treatment values of the unit of interest, São Paulo, to other potential control states. However, in this model only the time trends of the dependent variable are matched. In a sense, this is closer to a traditional differences-in-differences approach, but without the restrictive assumption that the treated and the control cases would follow parallel trends over time (Abadie et al. 2010, 494).

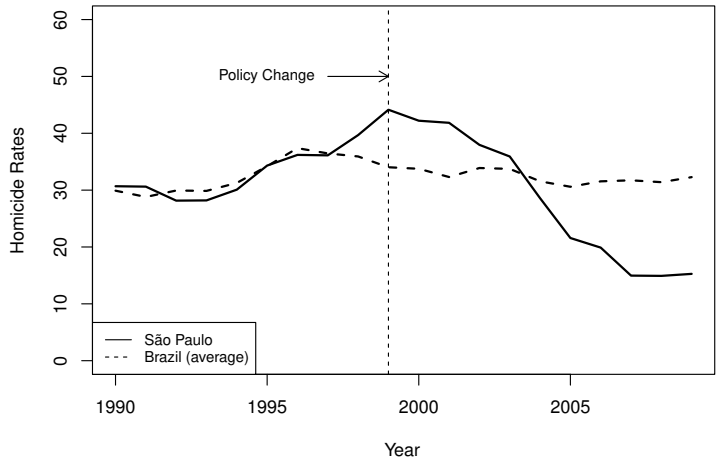

Figure 2.8: Bayesian Structural Time Series Model: São Paulo and Synthetic São Paulo

The model shows that in 2009 we should have expected São Paulo to have a homicide rate equal to 32.3 deaths per 100,000, but we observe only 15.2. Thus, the actual rate in São Paulo corresponds to only 47% of the expected counterfactual. The method also generates an estimate for the probability of causal effect. The calculations indicate a 96.3% chance of a causal impact in the period. In this sense, it is unlikely that the results are a statistical fluke.

2.6 Conclusion

As I have hopefully demonstrated, when compared to a synthetic control case, homicide rates were drastically reduced in São Paulo. Although it is not possible to estimate the treatment effect of each

specific policy implemented during the 1990s and 2000s, I suggest that their aggregate impact is surely not negligible. If the estimation strategy employed in this paper is correct, the state of São Paulo offers an example that it is feasible to fight crime with targeted policies. This as an encouraging result, as it suggests that governments can make progress in reducing crime with the resources they already have at hand and need not rely exclusively upon structural conditions that are largely beyond their control, such as unemployment, per capita income and inequality. Robustness tests provide further evidence for my findings.

I also argue in favour of the synthetic control method as a tool to evaluate government policies. This approach offers an intuitive way to assess causality claims when there is only a single treated unit and it can be easily applied in a great number of situations. Assuming that there is a reasonable number of potential cases in the 'donor pool,' a synthetic control can be meaningfully compared to the actual case. In this way, the technique allows the researcher to use the potential outcomes framework even in unusual conditions.

Future research can extend the present findings in a number of ways. First, it would be interesting to test whether other criminal activities have been affected by the state government policies I mentioned previously. Since property crimes are pervasive in São Paulo, scholars could evaluate the causal link (or lack thereof) between public policies and the incidence of theft or robberies. Unfortunately, several states in Brazil do not publish time-series data for property crime, so I could not use the synthetic control method for that dependent variable. As more data become available, this will create an interesting opportunity for investigation. Secondly, micro-level studies are needed to clarify the mechanisms behind São Paulo's homicide reduction, and isolate direct from indirect effects of each individual policies. Due to the shortage of data on targeted policies, qualitative research may explain what the motivations, successes and shortcomings of São Paulo's recent security measures were. Finally, there are still unresolved questions with regards to the 'PCC hypothesis', and this is a

promising avenue for future academic work. New research could provide insights into how public policies work and, hopefully, help public authorities to design more effective policies against crime.

2.7 Appendix

2.7.1 The Synthetic Control Estimator

This appendix presents a formal presentation of the synthetic control estimator. Let $j = 1, \ldots, J+1$ be a series of units in periods $t = 1, \ldots, T$. In our case, the units are the 27 Brazilian federal states and the time period spans from 1990 to 2009. Assuming that the first unit, São Paulo, has been exposed to the treatment, we have J control units to be included in the case studies donor pool, *i.e.* the 26 remaining states. We define treatment as the series of post-1999 government anti-crime policies implemented in the São Paulo.

Let Y_{it}^N be the homicide rate that would be observed for unit i, São Paulo, at time t with no treatment (1990–1998). Conversely, let Y_{it}^I be the observable outcome for unit i at time t had it been subjected to the treatment in periods $T_0 + 1$ to T (1999–2009). An important assumption is that the treatment has no effect on unit i before the date of intervention, therefore, the values for São Paulo with and without the policy interventions are the same for the pre-treatment period (1990–1998). In formal terms, $Y_{it}^I = Y_{it}^N \; \forall t < T_0$. The observed outcome is defined by $Y_{it}^I = Y_{it}^N + \alpha_{it} D_{it}$, where α_{it} is the effect of crime-reducing policies on homicide rates, and D_{it} is a binary variable that takes the value of 1 if we refer to post-intervention period (after 1999) and 0 otherwise. The goal of this paper is to estimate α_{it}, the effect of the treatment (homicide reduction policies), for the state of São Paulo for all $t \geq T_0$, that is, from 1999 to 2009. However, we cannot observe São Paulo *without* those policies, as there is no way for the state to have and not have the intervention at the same time. This is what Holland (1986) calls the "fundamental problem of causal inference": only one of the outcomes of interest is measurable at any given time.

But although we cannot accurately know how São Paulo would be without the treatment, we can approximate it by using a weighted average of the remaining Brazilian states such that $Y_{it}^N = \delta_t + \theta_t Z_i + \lambda_t \mu_i + \epsilon_{it}$. In this model, δ_t is an unobserved time-dependent factor common to all cases,

Z_i is a $(1 \times r)$ vector of observed control variables not affected by the policy, θ_t is a $(r \times 1)$ vector of unknown time-specific parameters, λ_t is a $(1 \times F)$ vector of unknown common factors to all states, μ_i is a state-specific unobservable variable and ϵ_{it} represents unobserved transitory shocks with mean 0 for all units (error term). Basically, what SCM tries to do is to match Z_i, the control variables, and the pre-treatment Y_{it} of São Paulo (1990–1998) so that μ_i is matched as a result.

Synthetic São Paulo is the weighted average of the other 26 Brazilian states. Thus, it is a $(J \times 1)$ vector of weights $W = (w_2, \ldots, w_{J+1})'$ with $w_j \geq 0$ for $j = 2, \ldots, J+1$ and $w_2 + \cdots + w_{J+1} = 1$. Each of the elements included in W represents a specific weighted average of control states, that is, a potential synthetic control for São Paulo. The idea is to select a case that resembles São Paulo as closely as possible. Let X_1 be a $(k \times 1)$ vector of pre-1999 predictor variables for São Paulo and let X_0 be a $(k \times J)$ matrix containing the predictor variables for the potential control states. Let $\bar{Y}_i^{K_1}, \ldots, \bar{Y}_i^{K_M}$ be M linear functions of pre-treatment outcomes ($M \geq F$). One can choose w^* such that:

$$\sum_{j=2}^{J+1} w_j^* Z_j = Z_1, \sum_{j=2}^{J+1} w_j^* \bar{Y}_j^{K_1} = \bar{Y}_1^{K_1}, \ldots, \sum_{j=2}^{J+1} w_j^* \bar{Y}_j^{K_M} = \bar{Y}_1^{K_M}$$

Consequently, as noted by Abadie and his collaborators (2010), if T_0 is sufficiently large when compared to the scale of ϵ_{it}, an approximately unbiased estimator for α_{1t}, the effect of public security policies in São Paulo, can be described by:

$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$$

for all $t \in \{T_0 + 1, \ldots, T\}$, that is, after the intervention period (1999–2009). In practice, W^* is chosen non-parametrically as to minimise $||X_1 - X_0 W||$, subject to the weight constrains. Consider $||X_1 - X_0 W||v = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$, where V is a $(k \times k)$ symmetric and semi-definite positive matrix with the relative importance of each assigned homicide rate predictor. From various possible ways of choosing V, here I follow the recommendation of Abadie and Gardeazabal (2003)

and choose V^* as the value of V that minimises the root mean squared prediction error (RMSPE) for homicide rates in the entire pre-treatment period (1990-1998).

2.7.2 R Code

The R code below replicates all statistical analyses and graphs included in this chapter. The original data files as well as the final data set are available at https://github.com/danilofreire/homicides-sp-synth.

```
######################
### Data Wrangling ###
######################

# Please set your working directory to the data/ folder

# Clear the workspace
rm(list = ls())

# Load necessary packages
library(reshape2)  # data manipulation

# Dependent variable:
dep <- read.csv("homicide-rates.csv", header = TRUE, skip = 1)

dep.molten <- melt(dep,
                   id.vars = c("Sigla",
                               "Código",
                               "Estado")
                    )

colnames(dep.molten) <- c("abbreviation",
                          "code",
                          "state",
                          "year",
                          "homicide.rates")

dep.molten$year <- as.numeric(substring(dep.molten$year, 2))
```

```
# Independent variables
ind1 <- read.csv("state-gdp-capita.csv", header = TRUE, skip = 1)

ind1.molten <- melt(ind1,
                    id.vars = c("Sigla",
                                "Código",
                                "Estado")
                    )

colnames(ind1.molten) <- c("abbreviation",
                           "code",
                           "state",
                           "year",
                           "state.gdp.capita")

ind1.molten$year <- as.numeric(substring(ind1.molten$year, 2))

ind2 <- read.csv("state-gdp-growth-percentage.csv", header = TRUE, skip = 1)

ind2.molten <- melt(ind2,
                    id.vars = c("Sigla",
                                "Código",
                                "Estado")
                    )

colnames(ind2.molten) <- c("abbreviation",
                           "code",
                           "state",
                           "year",
                           "state.gdp.growth.percent")

ind2.molten$year <- as.numeric(substring(ind2.molten$year, 2))

ind3 <- read.csv("gini.csv", header = TRUE, skip = 1)

ind3.molten <- melt(ind3,
                    id.vars = c("Sigla",
                                "Código",
                                "Estado")
                    )

colnames(ind3.molten) <- c("abbreviation",
                           "code",
                           "state",
                           "year",
                           "gini")

ind3.molten$year <- as.numeric(substring(ind3.molten$year, 2))

ind4 <- read.csv("population-projection.csv",
                 header = TRUE,
                 skip   = 1)
```

```
ind4.molten <- melt(ind4,
                    id.vars = c("Sigla",
                                "Código",
                                "Estado")
                    )

colnames(ind4.molten) <- c("abbreviation",
                           "code",
                           "state",
                           "year",
                           "population.projection")

ind4.molten$year <- as.numeric(substring(ind4.molten$year, 2))

ind5 <- read.csv("population-extreme-poverty.csv", header = TRUE, skip = 1)

ind5.molten <- melt(ind5,
                    id.vars = c("Sigla",
                                "Código",
                                "Estado")
                    )

colnames(ind5.molten) <- c("abbreviation",
                           "code",
                           "state",
                           "year",
                           "population.extreme.poverty")

ind5.molten$year <- as.numeric(substring(ind5.molten$year, 2))

ind6 <- read.csv("years-schooling.csv", header = TRUE, skip = 1)

ind6.molten <- melt(ind6,
                    id.vars = c("Sigla",
                                "Código",
                                "Estado")
                    )

colnames(ind6.molten) <- c("abbreviation",
                           "code",
                           "state",
                           "year",
                           "years.schooling")

ind6.molten$year <- as.numeric(substring(ind6.molten$year, 2))

# Merges files
data.list <- list(dep.molten,
                  ind1.molten,
                  ind2.molten,
                  ind3.molten,
                  ind4.molten,
                  ind5.molten,
```

```
                  ind6.molten)

data1 <- Reduce(function(...) merge(..., all = TRUE), data.list)

# Subset and sort
data2 <- subset(data1, year >= 1990 & year <= 2009)
data2 <- data2[order(data2$state), ]
rownames(data2) <- NULL

# Count missing observations, calculate their percentage
round(sapply(data2, function(x) length(which(is.na(x)))), 2)
round(sapply(data2, function(x) length(which(is.na(x)))/length(x)), 2)

# Linear imputation of missing values.
data2$gini.imp <- approxfun(seq_along(data2$gini), data2$gini)(seq_along(data2$gini))

data2$population.extreme.poverty.imp <- approxfun(seq_along(data2$population.extreme.poverty),
data2$population.extreme.poverty)(seq_along(data2$population.extreme.poverty))

data2$years.schooling.imp <- approxfun(seq_along(data2$years.schooling),
data2$years.schooling)(seq_along(data2$years.schooling))

# Create proportion.extreme.poverty
data2$proportion.extreme.poverty <- data2$population.extreme.poverty.imp / data2$population.projecti

# Transform variables to improve interpretation
data2$population.projection.ln <- log(data2$population.projection)

# Save data as df.csv
write.table(data2,
            "df.csv",
            row.names = FALSE,
            col.names = TRUE,
            sep       = ",")

######################
### Data Analysis ###
######################

# Load necessary packages
library(dplyr) # data manipulation
library(Synth) # models

# Load data
df <- read.csv("df.csv", header = TRUE)

# Prepare data set
df$state <- as.character(df$state) # required by dataprep()

# Plot: Homicide rates for Sao Paulo and Brazil (average)
df1 <- df %>%
     mutate(homicide.sp = ifelse(homicide.rates & state == "São Paulo", homicide.rates, NA)) %>%
        select(year, homicide.sp)
```

```
df2 <- df %>%
     mutate(homicide.rates1 = ifelse(homicide.rates & state != "São Paulo", homicide.rates, NA)) %>%
        group_by(year) %>%
        summarise(homicide.br = mean(homicide.rates1, na.rm = TRUE))

setEPS()
postscript(file    = "br.eps",
           horiz   = FALSE,
           onefile = FALSE,
           width   = 7,     # 17.8 cm
           height  = 5.25)  # 13.3 cm

plot(x = df1$year,
     y = df1$homicide.sp,
     type = "l",
     ylim = c(0, 60),
     xlim = c(1990, 2009),
     xlab = "Year",
     ylab = "Homicide Rates",
     cex = 3,
     lwd = 2,
     xaxs = "i",
     yaxs = "i"
)

lines(df2$year,
      df2$homicide.br,
      lty = 2,
      cex = 3,
      lwd = 2)

arrows(1997, 50, 1999, 50,
       col    = "black",
       length = .1)

text(1995, 50,
     "Policy Change",
     cex = .8)

abline(v   = 1999,
       lty = 2)

legend(x = "bottomleft",
       legend = c("São Paulo",
                  "Brazil (average)"),
       lty    = c("solid", "dashed"),
       cex    = .8,
       bg     = "white",
       lwdc(2, 2)
)

invisible(dev.off())
```

```
# Prepare data for synth
dataprep.out <-
        dataprep(df,
                 predictors = c("state.gdp.capita",
                                "state.gdp.growth.percent",
                                "population.projection.ln",
                                "years.schooling.imp"
                                ),
                 special.predictors = list(
                         list("homicide.rates", 1990:1998, "mean"),
                         list("proportion.extreme.poverty", 1990:1998, "mean"),
                         list("gini.imp", 1990:1998, "mean")
                         ),
                 predictors.op = "mean",
                 dependent     = "homicide.rates",
                 unit.variable = "code",
                 time.variable = "year",
                 unit.names.variable   = "state",
                 treatment.identifier  = 35,
                 controls.identifier   = c(11:17, 21:27, 31:33, 41:43, 50:53),
                 time.predictors.prior = c(1990:1998),
                 time.optimize.ssr     = c(1990:1998),
                 time.plot             = c(1990:2009)
                 )

# Run synth
synth.out <- synth(dataprep.out)

# Get result tables
print(synth.tables   <- synth.tab(
        dataprep.res = dataprep.out,
        synth.res    = synth.out)
      )

# Plot: Main model
setEPS()
postscript(file    = "trends.eps",
           horiz   = FALSE,
           onefile = FALSE,
           width   = 7,     # 17.8 cm
           height  = 5.25)  # 13.3 cm

path.plot(synth.res    = synth.out,
          dataprep.res = dataprep.out,
          Ylab         = c("Homicide Rates"),
          Xlab         = c("Year"),
          Legend       = c("São Paulo","Synthetic São Paulo"),
          Legend.position = c("bottomleft")
)

abline(v   = 1999,
       lty = 2)
```

```
arrows(1997, 50, 1999, 50,
       col    = "black",
       length = .1)

text(1995, 50,
     "Policy Change",
     cex = .8)

invisible(dev.off())

# Main model: gaps plot
setEPS()
postscript(file    = "gaps.eps",
           horiz   = FALSE,
           onefile = FALSE,
           width   = 7,
           height  = 5.25)

gaps.plot(synth.res    = synth.out,
          dataprep.res = dataprep.out,
          Ylab         = c("Gap in Homicide Rates"),
          Xlab         = c("Year"),
          Ylim         = c(-30, 30),
          Main         = ""
)

abline(v   = 1999,
       lty = 2)

arrows(1997, 20, 1999, 20,
       col    = "black",
       length = .1)

text(1995, 20,
     "Policy Change",
     cex = .8)

invisible(dev.off())

## Calculating how many lives were saved during the treatment period

# Weights below retrieved form dataprep.out
# State Code  State Weight  State Name        State Abbreviation
# 42          0.274         Santa Catarina    SC
# 53          0.210         Distrito Federal  DF
# 32          0.209         Espirito Santo    ES
# 33          0.169         Rio de Janeiro    RJ
# 14          0.137         Roraima           RR
# 14          0.001         Pernambuco        PB
# 35          treat         Sao Paulo         SP

# Get years after policy change
df.2 <- df[which(df$year >= 1999),]
```

```
# Calculate total number of deaths in SP
num.deaths.sp <- sum( (df.2$homicide.rates[which(df.2$abbreviation == "SP")])/100000 *
    (df.2$population.projection[which(df.2$abbreviation == "SP")]))

# Calculate estimated number of deaths in Synthetic São Paulo
num.deaths.synthetic.sp <- sum( (0.274 * (df.2$homicide.rates[which(df.2$abbreviation == "SC")])/1000
(df.2$population.projection[which(df.2$abbreviation == "SP")]))
                      + (0.210 * (df.2$homicide.rates[which(df.2$abbreviation == "DF")])/100000 *
                           (df.2$population.projection[which(df.2$abbreviation == "SP")]))
                      + (0.209 * (df.2$homicide.rates[which(df.2$abbreviation == "ES")])/100000 *
                           (df.2$population.projection[which(df.2$abbreviation == "SP")]))
                      + (0.169 * (df.2$homicide.rates[which(df.2$abbreviation == "RJ")])/100000 *
                           (df.2$population.projection[which(df.2$abbreviation == "SP")]))
                      + (0.137 * (df.2$homicide.rates[which(df.2$abbreviation == "RR")])/100000 *
                           (df.2$population.projection[which(df.2$abbreviation == "SP")]))
                      + (0.001 * (df.2$homicide.rates[which(df.2$abbreviation == "PB")])/100000 *
                           (df.2$population.projection[which(df.2$abbreviation == "SP")]))
                                 )

lives.saved <- num.deaths.synthetic.sp - num.deaths.sp
lives.saved # Between 1999 and 2009

########################
### Robustness Tests ###
########################

# Prepare data set
df$state <- as.character(df$state) # required by dataprep()

## Placebo Test -- Control ends in 1994
dataprep.out1 <-
        dataprep(df,
                 predictors = c("state.gdp.capita",
                                "state.gdp.growth.percent",
                                "population.projection.ln",
                                "years.schooling.imp"
                 ),
                 special.predictors = list(
                         list("homicide.rates", 1990:1994, "mean"),
                         list("proportion.extreme.poverty", 1990:1994, "mean"),
                         list("gini.imp", 1990:1994, "mean")
                 ),
                 predictors.op = "mean",
                 dependent     = "homicide.rates",
                 unit.variable = "code",
                 time.variable = "year",
                 unit.names.variable   = "state",
                 treatment.identifier  = 35,
                 controls.identifier   = c(11:17, 21:27, 31:33, 41:43, 50:53),
                 time.predictors.prior = c(1990:1994),
                 time.optimize.ssr     = c(1990:1994),
                 time.plot             = c(1990:1998))
```

```
# Run synth
synth.out1 <- synth(dataprep.out1)

# Get result tables
print(synth.tables   <- synth.tab(
        dataprep.res = dataprep.out1,
        synth.res    = synth.out1)
      )

# Placebo test: graph
setEPS()
postscript(file    = "placebo.eps",
           horiz   = FALSE,
           onefile = FALSE,
           width   = 7,
           height  = 5.25)

path.plot(synth.res       = synth.out1,
          dataprep.res    = dataprep.out1,
          Ylab            = c("Homicide Rates"),
          Xlab            = c("Year"),
          Legend          = c("São Paulo","Synthetic São Paulo"),
          Legend.position = c("bottomleft"),
          Ylim            = c(0, 50)
)

abline(v   = 1995,
       lty = 2)

arrows(1994, 40, 1995, 40,
       col    = "black",
       length = .1)

text(1993, 40,
     "Placebo \nPolicy Change",
     cex = .8)

invisible(dev.off())

## Leave-one-out

# Loop over leave one outs
storegaps <- matrix(NA, length(1990:2009), 4)

colnames(storegaps) <- c(14, 33, 42, 53) # RR, RJ, SC, DF
co <- unique(df$code)
co <- co[-25]

for(k in 1:4){

        # Data prep for training model
        omit <- c(14, 33, 42, 53)[k]
```

```
        # Prepare data for synth
        dataprep.out2 <-
                dataprep(df,
                         predictors = c("state.gdp.capita",
                                        "state.gdp.growth.percent",
                                        "population.projection.ln",
                                        "years.schooling.imp"
                         ),
                         special.predictors = list(
                                 list("homicide.rates", 1990:1998, "mean"),
                                 list("proportion.extreme.poverty", 1990:1998, "mean"),
                                 list("gini.imp", 1990:1998, "mean")
                         ),
                         predictors.op = "mean",
                         dependent     = "homicide.rates",
                         unit.variable = "code",
                         time.variable = "year",
                         unit.names.variable   = "state",
                         treatment.identifier  = 35,
                         controls.identifier   = co[-which(co==omit)],
                         time.predictors.prior = c(1990:1998),
                         time.optimize.ssr     = c(1990:1998),
                         time.plot             = c(1990:2009)
                )

        # Run synth
        synth.out2 <- synth(dataprep.out2)

        storegaps[,k] <- (dataprep.out2$Y0%*%synth.out2$solution.w)
} # Close loop over leave one outs

# Leave-one-out: graph
setEPS()
postscript(file    = "leave-one-out.eps",
           horiz   = FALSE,
           onefile = FALSE,
           width   = 7,
           height  = 5.25)

path.plot(synth.res    = synth.out,
          dataprep.res = dataprep.out,
          Ylab         = c("Homicide Rates"),
          Xlab         = c("Year"),
          Legend       = c("São Paulo","Synthetic São Paulo"),
          Legend.position = c("bottomleft"))

abline(v   = 1999,
       lty = 2)

arrows(1997, 50, 1999, 50,
       col    = "black",
       length = .1)
```

```
text(1995, 50,
     "Policy Change",
     cex = .8)

for(i in 1:4){
        lines(1990:2009,
              storegaps[,i],
              col = "darkgrey",
              lty = "solid")
}

lines(1990:2009,
      dataprep.out$Y0plot %*% synth.out$solution.w,
      col = "black",
      lty = "dashed",
      lwd = 2)

legend(x = "bottomleft",
       legend = c("São Paulo",
                  "Synthetic São Paulo",
                  "Synthetic São Paulo (leave-one-out)"
       ),
       lty    = c("solid", "dashed", "solid"),
       col    = c("black", "black", "darkgrey"),
       cex    = .8,
       bg     = "white",
       lwdc(2, 2, 1)
)

invisible(dev.off())

## Permutation test
states <- c(11:17, 21:27, 31:33, 35, 41:43, 50:53)

# Prepare data for synth
results <- list()
results_synth <- list()
gaps <- list()

for (i in states) {
    dataprep.out <-
            dataprep(df,
                     predictors = c("state.gdp.capita",
                                    "state.gdp.growth.percent",
                                    "population.projection.ln",
                                    "years.schooling.imp"
                                    ),
                     special.predictors = list(
                             list("homicide.rates", 1990:1998, "mean"),
                             list("proportion.extreme.poverty", 1990:1998, "mean"),
                             list("gini.imp", 1990:1998, "mean")
                             ),
```

```
                    predictors.op = "mean",
                    dependent     = "homicide.rates",
                    unit.variable = "code",
                    time.variable = "year",
                    unit.names.variable   = "state",
                    treatment.identifier  = i,
                    controls.identifier   = states[which(states!=i)],
                    time.predictors.prior = c(1990:1998),
                    time.optimize.ssr     = c(1990:1998),
                    time.plot             = c(1990:2009)
                    )
    results[[as.character(i)]] <- dataprep.out
    results_synth[[as.character(i)]] <- synth(results[[as.character(i)]])
   gaps[[as.character(i)]] <- results[[as.character(i)]]$Y1plot - (results[[as.character(i)]]$Y0plot
    results_synth[[as.character(i)]]$solution.w)

}

## Permutation test
setEPS()
postscript(file    = "permutation-gaps2.eps",
           horiz   = FALSE,
           onefile = FALSE,
           width   = 7,
           height  = 5.25)

plot(1990:2009,
     ylim = c(-30, 30),
     xlim = c(1990,2009),
     ylab = "Gap in Homicide Rates",
     xlab = "Year"
)

for (i in states) {
        lines(1990:2009,
              gaps[[as.character(i)]],
              col = "lightgrey",
              lty = "solid",
              lwd = 2
        )
}

lines(1990:2009,
      gaps[["35"]], # São Paulo
      col = "black",
      lty = "solid",
      lwd = 2
)

abline(v   = 1999,
       lty = 2)
```

```
abline(h   = 0,
       lty = 1,
       lwd = 1)

arrows(1997, 25, 1999, 25,
       col    = "black",
       length = .1)

text(1995, 25,
     "Policy Change",
     cex = .8)

legend(x = "bottomleft",
       legend = c("São Paulo",
                  "Control States"),
       lty  = c("solid", "solid"),
       col  = c("black", "darkgrey"),
       cex  = .8,
       bg  = "white",
       lwdc(2, 2, 1)
)

invisible(dev.off())

# Permutation graph: states with MSPE no higher than 2x São Paulo's
low.mspe <- c(13, 15, 17, 21, 23, 24, 25, 31, 41:43, 53)

setEPS()
postscript(file    = "low-mspe.eps",
           horiz   = FALSE,
           onefile = FALSE,
           width   = 7,
           height  = 5.25)

plot(1990:2009,
     ylim = c(-30, 30),
     xlim = c(1990,2009),
     ylab = "Gap in Homicide Rates",
     xlab = "Year"
)

for (i in low.mspe) {
lines(1990:2009,
      gaps[[as.character(i)]],
      col = "lightgrey",
      lty = "solid",
      lwd = 2
      )
}

lines(1990:2009,
      gaps[["35"]], # São Paulo
      col = "black",
```

```
      lty = "solid",
      lwd = 2
      )

abline(v   = 1999,
       lty = 2)

abline(h   = 0,
       lty = 1,
       lwd = 1)

arrows(1997, 25, 1999, 25,
       col    = "black",
       length = .1)

text(1995, 25,
     "Policy Change",
     cex = .8)

legend(x = "bottomleft",
       legend = c("São Paulo",
                  "Control States (MSPE Less Than Two Times That of São Paulo)"),
       lty    = c("solid", "solid"),
       col    = c("black", "darkgrey"),
       cex    = .8,
       bg     = "white",
       lwdc(2, 2, 1)
)

invisible(dev.off())

## CausalImpact
# Uncomment the lines below to install necessary packages
# install.packages(c("devtools", "dtw"))
# library(devtools)
# install_github("google/CausalImpact")
# install_github("klarsen1/MarketMatching", build_vignettes=TRUE)

# Load packages
library(CausalImpact)
library(MarketMatching)

# Prepare data
df$year2 <- as.Date(paste(df$year, sep = "", "-01-01"))

# Estimate model
mm <- best_matches(data=df,
                   id_variable="code",
                   date_variable="year2",
                   matching_variable="homicide.rates",
                   parallel=TRUE,
                   warping_limit=1, # warping limit=1
                   dtw_emphasis=1, # rely only on dtw for pre-screening
```

```
                  matches=5, # request 5 matches
                  start_match_period="1990-01-01",
                  end_match_period="1998-01-01")

# View best matches
subset(mm$BestMatches, code == 35) # SP

# Results
results <- MarketMatching::inference(matched_markets = mm,
                                     test_market = "35",
                                     end_post_period = "2009-01-01")

# Predictions
results$Predictions

# Plot results
results$PlotActualVersusExpected +
        ggtitle("São Paulo versus Synthetic São Paulo") + theme_bw() +
      geom_line(aes(results$PlotActualVersusExpected$data$test_market),colour="#000099")
results$PlotCumulativeEffect

# Graph
setEPS()
postscript(file    = "causal-impact.eps",
           horiz   = FALSE,
           onefile = FALSE,
           width   = 7,     # 17.8 cm
           height  = 5.25)  # 13.3 cm

plot(x = (1990:2009),
     y = as.numeric(results$Predictions$Response),
     type = "l",
     ylim = c(0, 60),
     xlim = c(1990, 2009),
     xlab = "Year",
     ylab = "Homicide Rates",
     cex = 3,
     lwd = 2)

lines(x = (1990:2009),
      y = as.numeric(results$Predictions$Predicted),
      type = "l",
      lty = 2,
      cex = 3,
      lwd = 2)

arrows(1997, 50, 1999, 50,
       col    = "black",
       length = .1)

text(1995, 50,
     "Policy Change",
     cex = .8)
```

```
abline(v   = 1999,
       lty = 2)

legend(x = "bottomleft",
       legend = c("São Paulo",
                  "Brazil (average)"),
       lty    = c("solid", "dashed"),
       cex    = .8,
       bg     = "white",
       lwdc(2, 2)
)

invisible(dev.off())
```

Chapter 3

Beasts of Prey or Rational Animals? Private Governance in Brazil's *Jogo do Bicho*

3.1 Introduction

In 1892, Baron João Batista de Viana Drummond came up with a new idea to fund his cash-strapped zoo. Situated in a quiet neighbourhood in the north of Rio de Janeiro, the *Jardim Zoológico*, or Zoological Garden, hosted a variety of exotic species and offered breath-taking views of the city, but it lacked visitors. An experienced businessman, Drummond realised the zoo would have to provide other kinds of entertainment to keep itself afloat. One of his plans seemed particularly promising: a lottery raffle.

The rules were straightforward. In the morning, the baron would choose one animal from a list of 25 beasts and put its picture inside a wooden box at the zoo's entrance. Visitors who wanted to join the raffle received a ticket bearing the stamp of one of those 25 animals.[1] The lucky winner would take home a prize worth 20 times the ticket price, an amount higher than a carpenter's monthly wage (Chazkel 2007, 542). The baron called the lottery the *jogo do bicho*, or the animal game, and it was well-received by the public. Eager to capitalise on that initial success, Drummond stated that

[1] At first, the zoo staff distributed the tickets at random, so the game consisted in a simple raffle. However, the zoo soon allowed participants to choose the animals they preferred. This small change made the game considerably more appealing as it introduced an element of divination to the *jogo* (DaMatta and Soárez 1999, 71–74).

visitors could buy tickets not only at the zoo, but also in many stores across Rio de Janeiro. What was once a small raffle soon became a large gambling market of its own.

A *jogo do bicho* craze swept the whole city after independent sellers entered the marketplace. A network of street bookmakers, called *bicheiros*, made the lottery available in every part of Rio by scalping tickets or promoting their own versions of the numbers game (Chazkel 2011, 37). The lottery became so widespread that Olavo Bilac, a major literary figure in nineteenth-century Brazil, summarised the situation as follows: "Today [1985] in Rio de Janeiro, the game is everything. [. . .] Nobody works! Everybody plays" (Pacheco 1957, 43).[2]

But this tolerant state of affairs did not last. Civil servants and police officers criminalised the *jogo do bicho* on the grounds of public safety, and in the late 1890s they launched a country-wide campaign against the lottery (Benatte 2002). The campaign extended for several decades and received considerable support from the *Companhia das Loterias Nacionaes do Brazil*, the National Lottery Company, a public-private partnership founded four years after, and perhaps motivated by, the creation of the animal game (DaMatta and Soárez 1999, 82). The Brazilian government officially banned the *jogo do bicho* in 1941 and the lottery remains illegal to this day.

Yet the game has survived. The *jogo do bicho* has outlasted more than 30 Brazilian presidents and thrived under both military regimes and democratic governments (Jupiara and Otavio 2015). But more than an act of civil disobedience, the *jogo do bicho* is a very successful capitalist enterprise (Labronici 2014; Magalhães 2005). A recent study by Fundação Getúlio Vargas, a Brazilian think tank, affirmed that the *jogo do bicho* earns from BRL 1.3 to BRL 2.8 billion per year (USD 400 to USD 850 million), making it the largest clandestine gambling game in the world (HuffPost Brasil 2015). Schneider (1996, 171) estimated that in the 1990s, the game furnished about 50,000 jobs in Rio de Janeiro, almost as many as the oil giant Petrobras in 2011 (Exame 2016).[3]

In this article I offer a rational choice interpretation of the *jogo do bicho*.[4] More specifically, I use an array of bibliographical sources to show how the game operators, commonly called *bicheiros*,[5]

[2]Unless otherwise noted, all translations from the Portuguese are my own.

[3]In 1966, Time Magazine wrote that the *jogo do bicho* was "the largest single industry in Latin America" and employed about 1% of the total Brazilian workforce (Time Magazine 1966).

[4]I adopt a very broad definition of rationality in this paper. In contrast with stricter versions of rational choice theory, I assume here that individuals are not only constrained by formal and informal institutions, but they also have access only to imperfect information at the moment of their choices. Thus, my analysis employs a "thin" notion of rationality and a 'thick' description of social institutions, where individual action can only be understood with reference to the social environment (Boettke 2001, 253).

[5]As I describe in the next section, the *jogo do bicho* structure can be broadly divided into three occupations: the rich game financiers (*banqueiros*), mid-level managers (*gerentes*), and street bookmakers (*bicheiros*). While there are important

have developed unique strategies to solve collective action problems and to maximise their political strategies. This does not mean their tactics are morally defensible; *bicheiros* regularly employ intimidation and corruption to achieve their goals. Nevertheless, I argue here that such strategies are effective, and while they seem counter-intuitive, they do address the long-term needs of the *jogo do bicho* financiers.

Like any business manager, *bicheiros* have to run their firm with low costs to increase profits. However, the fact that the *jogo do bicho* is clandestine imposes additional difficulties to its managers. In particular, *bicheiros* face two main challenges to keep the lottery running. First, they need to gather public support so gamblers are not discouraged to engage in the lottery despite it being illegal. Second, *bicheiros* also have to ensure the state repression is not prohibitively costly to their business; otherwise they would be better off shutting it down. I argue below that the *bicheiros* have succeeded in both by using carefully-designed reputation strategies and employing costly signals to the communities they serve.

I use the case of Rio de Janeiro to illustrate how the *jogo do bicho* has overcome the obstacles to its expansion. Rio is a particularly interesting case because in no other part of Brazil is the game the financiers established such an effective patronage network. *Bicheiros* have sponsored political campaigns, financed cultural activities and football teams, and sometimes even run in local elections themselves. I discuss the ways by which the *bicheiros* have exploited fragilities of the Brazilian political system to their advantage and how those practices have weakened Brazilian democracy.

My analysis discusses three strands of academic literature. First, this work contributes to the scholarship on extra-legal institutions, mainly to the literature on collective action within criminal organisations. For instance, Gambetta (1996) examines the strategies used by the Sicilian Mafia to settle disputes among their members and enforce rules in the areas they exercise control. Leeson (2009, 2010) affirms that pirate groups employed hard-to-fake signals to increase the profitability of their operations. Skarbek (2011a, 2012, 2014), in turn, highlights the role of written and implicit norms in mitigating rent-seeking behaviour and in coordinating productive activities in California prison gangs. I argue that *bicheiros* have employed reputation strategies and provided club goods to enforce private contracts and to foster trust in the community.

hierarchical differences among these groups, all members of the *jogo* are collectively called *bicheiros* in Brazil. I follow the same practice here, and only refer to their specific role within the animal game structure when necessary for clarification.

Second, this work relates to the literature on signalling theory and asymmetric information (e.g., Akerlof 1970; Spence 1973). I provide evidence that *bicheiros* were aware of their social stigma, and as a response they devised signalling strategies to convey reliable information and to reduce the uncertainty associated with clandestine markets. Their main tool to increase credibility was costly signalling (Gambetta 2009; Kimbrough et al. 2015; Schelling 1960). *Bicheiros* believed that by sacrificing their immediate interests they could gain a reputation of honesty that would benefit them in the long run.

Lastly, this work connects to the literature on state capture, which is one of the most important topics in public choice theory (Rose-Ackerman 1978; Shleifer and Vishny 2002; Tollison 1982). More specifically, I use the Brazilian case to illustrate how politicians and civil servants are co-opted by criminal groups, and how this collusion distorts the electoral process and benefits wealthy members of the *jogo do bicho* network. Queiroz (1992) explored why *bicheiros* turned into patrons of the Carnival's samba schools and affirmed this influence gave them leverage over political authorities. Misse (2007) investigated the links between bicheiros and police officers, and suggested the illegal lottery had been the main cause of police corruption in Rio de Janeiro until the 1970s. In a similar vein, Jupiara and Otavio (2015) analyse the relationship between the *jogo do bicho* and the military regime in Brazil (1964–1985). I supplement this literature by highlighting how asymmetrical information and rent-seeking behaviour offer convincing explanations to the issues presented above. Although those concepts have a long tradition in public choice, scholars have not applied those ideas to understand the dynamics of the *jogo do bicho*. By doing so, I integrate seemingly contradictory historical facts into a single narrative that connects micro-level decisions to macro-level outcomes.

3.2 An Overview of the *Jogo do Bicho*

3.2.1 Historical Background: How the *Bicheiros* Avoided Extinction

The late nineteenth-century Brazil had four characteristics that explain the emergence of the *jogo do bicho*: 1) a growing urban population excluded from the formal labour market; 2) an inflow of immigrants whose extended family networks helped them engage in trade; 3) an expansion of the monetary supply in the first years of the republic (1880s–1890s); and 4) a judicial system that, albeit repressive, had only imperfect law enforcement. I discuss each of these elements below.

I start with the impact of urban poverty on the animal game. Brazil abolished slavery in the late 1880s, a period in which the country was rapidly urbanising. Cities like São Paulo and Rio de Janeiro offered a number of occupations for former slaves who wanted to move away from their former rural masters (Andrews 1991; Skidmore 1993). Increasing numbers of Asian and European immigrants also moved to large cities after arriving in Brazil, as urban areas usually provided better standards of living than the countryside (Hall 1969; Lesser 2013). However, the hopes of the African-Brazilians and the new foreign settlers would be frustrated by a series of downturns in the Brazilian economy. Brazil's labour markets suffered a severe contraction in the wake of the *Encilhamento* financial crisis of 1891, and the economic instability aggravated the already difficult conditions of the working classes (Topik 2014; Triner and Wandschneider 2005).

As a result, large swathes of the urban population turned to the informal economy. As Chazkel (2011, 115) observes, there were few occupations available to lower-class women and foreigners in the 1890s, and a large number of poor workers became street vendors. The profession requires little technical skills and has low barriers of entry, but it can be profitable if vendors are able to quickly identify a growing demand for a particular product. The *jogo do bicho* was one such case; the game had a surge in popularity in the 1890s, and it also offered a high rate of return. As the demand for lottery tickets grew, the *jogo* comprised an important share of the Brazilian extra-legal economy.

Immigration also influenced the *jogo do bicho* via social ties. Most foreigners who moved to Brazil came from countries such as Portugal, Spain or Italy, where extended families were the basic form of social organisation (Lobo 2001; Trento 1989). Family and neighbourhood networks created incentives for immigrants to establish trade relations and to enforce cooperation through community responsibility systems (Roth and Skarbek 2014). Because of these particular social characteristics, in the 1890s foreigners were over-represented in the Brazilian trade in general (Mattos 1991; Oliveira 2001) and in the *jogo do bicho* in particular (Magalhães 2005; Villar 2008). Although kinship bonds became less relevant over time, these links offered an important element of social cohesion in the *jogo do bicho*'s formative years.

Next is the impact of expanded monetary supply. The abolition of slavery and the growing industrialisation of Brazil increased the amount of capital available in the country (Franco 1987; Schulz 2008). Moreover, the 1888 Banking Act gave extra liquidity to local financial markets, which made credit more widely available in cities like São Paulo and Rio de Janeiro. Individuals received a

temporary boost in personal income, a part of which they spent on leisure activities such as the *jogo do bicho*. Moreover, the lottery attracted new entrants as it became more profitable, and in only a few years similar versions of the animal game were available throughout Brazil (DaMatta and Soárez 1999, 79).

The last necessary condition for the emergence of the *jogo do bicho* is weak law enforcement. Chazkel (2011, 69–100) notes that until the 1940s, police district chiefs operated within a large margin of discretion, so official repression against bookmakers was notably inconsistent. In the early years of *jogo do bicho*, lottery "bankers" were allowed to operate virtually free from police interference, which surely collaborated to the game's rapid initial expansion (Chazkel 2007, 544). Prosecution against the *bicheiros* hardened in 1917 after the promulgation of the Civil Code, and in 1941 the animal game was banned. Five years later, the federal government declared that all games of chance were illegal in Brazil.[6] Recent estimations show that the prohibition of the *jogo do bicho* prevented the state from earning up to BRL 20 billion (USD 6 billion) per year in expected taxation revenues, aside from the subjective utility losses for players (Folha de São Paulo 2016).

Since the mid-twentieth century, the *jogo do bicho* has been outlawed but it continues to be ubiquitous in Brazil. There is virtually no Brazilian city that does not have its local group of *bicheiros*. Cross and Peña (2006) suggest a distinction between informal and illegal markets that is useful to understand the current "semi-legal" status of the *jogo*. The animal game started as an informal activity, in which the Baron of Drummond and his associates sold lottery tickets to the public. Although the state did not regulate the lottery market, the product itself was not illegal or allegedly immoral. In fact, many organisations, including the Catholic Church, sold raffles and lottery tickets to sponsor their activities (Torcato 2011, 49). Moreover, the state initially found no contradiction between the animal game and its own lottery, and both coexisted for about 15 years (Chazkel 2007, 559). After 1941, however, selling *jogo do bicho* tickets became a legal offence after a long campaign asking for its criminalisation. Since then, the *jogo* has moved underground and consciously evaded state regulations. This shift from informal to illegal has brought important changes to the whole dynamics of the *jogo*. As an example, wealthy animal game bosses have colluded with shady sectors of the armed forces for protection, and this violence has also been used against potential competitors. Moreover, this move to illegality enabled the game to exploit and subsidise large sectors of Brazil's

[6]The 1946 decree stated that gambling was 'harmful to morality and the good customs', hence '[…] the repression against games of chance [was] an imperative of the universal consciousness'.

formal economy, mainly in poor urban areas. I describe the animal game organisation structure and its impact in the Brazilian economic activity in further detail below.

3.2.2 A Hierarchical Organisational Structure

The animal game operates with three levels of hierarchy. At the bottom level are the *bicheiros*, those in charge of selling *jogo do bicho* tickets (Chazkel 2007; DaMatta and Soárez 1999). *Bicheiros* are the most visible part of the *jogo do bicho*. The name loosely describes all those involved in the lottery organisation, yet their meaning in the *jogo do bicho* structure is more particular and refers to street-level ticker sellers. The *bicheiros* usually build their vending stands inside the premises of a local shop, such as a small grocery store, and are recognisable by their chairs facing the street, stamps and blocks of paper (Chazkel 2011, 259). The street bookmakers usually work alone but may employ up to 10 people depending on how busy their betting site is (Labronici 2014, 69).

The *gerentes* (managers) oversee all *jogo do bicho* stands in a given area. Their task is akin to that of a firm accountant. Gerentes control the cash flow between the *bicheiros* and the bankers, manage the payroll of the employees, and provide financial information to the top members of the organisation. They also supervise individuals who carry menial tasks in the business, transfer money to other gambling branches and double-check the balance sheets of the betting sites (Labronici 2012, 71; Misse 2007, 142).

The *banqueiros* (Portuguese for bankers) occupy the top position in the *jogo do bicho* hierarchy. They comprise the small financial elite of the game. A 2012 report by the Brazilian Federal Police affirmed that 10 *banqueiros* controlled the market throughout the country, and five of them were based in the state of Rio de Janeiro (O Globo 2012b). Apart from funding the game, the bankers provide support for the employees to undertake their activities. The *banqueiros*' attributions include paying bribes to police personnel, bailing out sellers arrested by security forces, and offering judicial assistance to employees in case of legal persecution (Labronici 2012, 75).

The lottery bosses run their businesses from fortified houses in unknown locations called *fortalezas* ("forts"). The first *fortalezas* likely appeared in the 1950s, when the animal game was already well-established across the Brazilian territory. The period coincides with a time when the *jogo do bicho* finances had become increasingly concentrated in fewer hands (Chazkel 2011, 259). Due to the

growing scope of the *jogo do bicho* economy, *banqueiros* decided to move their operations away from the public to avoid police persecution and to make coordination easier.

Banqueiros solve problems of internal cooperation by providing club goods (Buchanan 1965) while simultaneously shunning cheaters through selective punishments (Dal Bó 2005; Roth and Murnighan 1978). The first club good offered to *bicheiros* by their bosses is private security. As the game is illegal, street sellers cannot rely on official institutions to protect themselves. Thus, the game bankers have built an extensive network of gunmen and bribed police officers to protect their employees from other criminals (Chinelli and Machado 1993, 48; Labronici 2012, 51). "Zé" (Joe), a bicheiro interviewed by Labronici (2012, 52), described eloquently the deterring effect of the *jogo do bicho* informal security personnel:

> Bums are scared and they don't mess around with us; they think there's an officer nearby or something like that. Look at all this money here! [Shows the interviewer a handful of cash.] It's not ours [referring to street-corner bookmakers]. And if it's not ours, it's someone else's. When I worked in Penha[7], the owner of a pub close by always asked me to stay at the front door of his pub. People know that bums are afraid of *bicheiros*.

The *banqueiros* do not use violence only against other criminals. They often employ violent methods against competitors and their own staff, too. Jupiara and Otavio (2015) argue that Ailton Guimarães Jorge, a former Army officer, tortured and murdered rival lottery bosses in the late 1970s. One of his former allies, Army Colonel Paulo Malhães, told the Rio de Janeiro State Truth Commission that Guimarães "went on a rampage" to consolidate his power (Belém 2015). Castor de Andrade, Rio's most influential animal game banker, also employed similar methods to run his business. Andrade kept an armed bodyguard of 23 men and allegedly murdered a number of competitors. In a famous case, Andrade shot Euclides Ponar, an old *jogo do bicho* boss known as the "Grey-Headed Chinese" (*China Cabeça Branca*) after Ponar denounced a fraud in lottery draws in 1976. Andrade was likely involved in other assassination plots in the 1990s (O Globo 2017).

Despite these serious events, killings are rare in the animal game. Since the lottery bosses can credibly indicate that violence is a low-cost option for them, the mere threat of punishment is enough to discourage defectors. This is a good strategy for the *banqueiros*. The fact that they had committed violent crimes in the past reduces the need to commit them in the present, and as a result the bosses

[7] Penha is a low middle-class neighbourhood in the city of Rio de Janeiro.

can spend less money on security and increase profits. As it happens in many traditional markets, if a group is able to form a cartel, they can increase the price of their services without fearing immediate competition. The same logic is valid for the *jogo do bicho*, although through unconventional means.

The threat of violence is not the only tool the *bicheiros* have at their disposal. They balance the use of violence with financial benefits to low-rank members of the organisation. For instance, street *bicheiros* keep all tips they receive from players, often have small expenses covered by their bosses, and may request interest-free loans to pay for healthcare treatment or other unexpected bills (Labronici 2012).

The most important financial mechanism implemented by bankers to help *bicheiros* is the *descarga*, loosely translated as "the unloading". The *descarga* is the *jogo do bicho*'s main hedging technique and its purpose is to insure small bookmakers against credit risk (Labronici 2012, 59; Magalhães 2005, 178). Booking agents are sometimes unable to honour expensive bets. The top prize in the animal game pays up to 4,000 times the amount invested; thus, *bicheiros* may have to raise thousands of Brazilian Reals in a single day to pay the lucky winners. To prevent the *quebra da banca* ("bust of the bank"), *bicheiros* and small bankers buy an insurance from wealthier financiers, who offer this service for a fee that ranges from 20% to 25% of the total selling amount (Folha de São Paulo 2006). The *descarga* guarantees that small bookmakers will not have liquidity problems, thus permitting bookmakers to continue investing in the *jogo do bicho*.

The *descarga* has significantly changed the distribution of resources in the *jogo do bicho*, and the richest bankers benefited the most from it. Simple probability dictates that a booking agent rarely pays the highest lottery prize, and yet the bankers receive a commission for *every game* they hedge. Over time, there is a transfer of income from the bottom to the top of the animal game structure due to the fees. This accumulation of capital is one of the reasons in the 1990s bankers started offering other types of entertainment such as slot machines and sports lotteries (Estado de São Paulo 2006; O Globo 2015; Terra 2011). They simply had more capital to invest, and this eventually helped to compensate the downturn in the *jogo do bicho* markets in the last years (O Globo 2017). In sum, while the *descarga* has made the game more resilient at the aggregated level, it increased profits for the richest financiers at the expense of small bookmakers.

This has brought broad consequences to the Brazilian illegal economy. As the number of slot machines increased in Brazil, the opportunity for criminals to use them for money laundering

increased as well. The scheme is easy to carry out. The owner of a slot machine issues a ticket with a winning prize, and the criminal declares the prize as his legitimate wealth. Then he can legally use that money for any purpose without raising suspicion from the authorities. The practice have become more widespread in the last decades, and in 2009 a Federal Police task force arrested about a dozen *jogo do bicho* bosses involved in the so-called "slot machine mafia". Relatives of Castor de Andrade were also involved in the scheme (Estado de São Paulo 2011). In 2012, President Dilma Rousseff sanctioned a law that considered slot machines and the *jogo do bicho* as money laundering crimes (Agência Brasil 2012).

Money from the *jogo* also has other effects on the economy. Some of these effects are indirect. For instance, the *jogo do bicho* gives a boost to the Brazilian economy by providing jobs for unskilled workers who cannot easily join the labour market. By doing so, the *jogo* prevents some of the poorest members of the Brazilian society from demanding more inclusive government policies, although they are the ones who would benefit the most from public assistance. Because of the income generated by the *jogo do bicho*, poor workers are able to consume without resorting to government assistance, so state officials can spend a larger amount of public funds on other, probably wealthier, sectors of the population.

Another indirect economic effect of the *jogo* is the rise in inequality. While formal workers can – at least in theory – demand higher compensations during a market upswing, the same is not true for the animal game employees. The threat of violence reduces the space for collective bargaining with the lottery bosses, and higher profits at the top of the *jogo do bicho* structure do not trickle down to those at the bottom. This is one of the reasons why the game bankers turned the game into an oligopoly: the use of violence guarantees new entrants will not be allowed to join the market and profits will be concentrated in the hands of very few individuals.

The *jogo do bicho's* market formation closely resembles what Fligstein (1996) calls "markets as politics", in which firms create institutions to restrain the competition and organise the labour force. At the formation of the lottery market, competition is fierce and participants are akin to social movements. They are constantly trying to convince others of the viability of their ideas. When markets stabilise, however, incumbents collude to impose their conditions of control to other players and to workers. In the case of the *jogo*, this implies in a mix of financial incentives and violence threats carried by corrupt state agents or private bodyguards.

A more difficult question is how the bankers elicit cooperation from *external* members, such as gamblers, community leaders, or public officers. It is puzzling because *bicheiros* do not use violence to induce individuals to play the lottery, nor have they ever clashed with the Brazilian government. Precisely because violence could shun profits, *bicheiros* devised other mechanisms to create a friendly environment for the illegal lottery. I investigate two of them: costly signalling and reputation building.

3.3 Winning Hearts, Minds, and Pockets: Illegal Market Dynamics

Evolutionary game theory (Axelrod 1984; Axelrod and Keohane 1985; Smith 1982) and empirical case studies (Isaac et al. 1984; Ostrom 1990) have both demonstrated that long-term cooperation is possible even in difficult situations. The main requirement for sustained cooperation is that players believe future pay-offs will be higher than present ones. If that condition is true, fear of retaliation will induce individuals not to cheat.

In theory, the same should also apply to illegal organisations. Yet in practice we see that criminal groups are generally short-term oriented, that is, they tend to discount the future more heavily than most people. This makes cooperative behaviour among criminal rather uncommon, and there is substantial evidence suggesting illegal groups face serious collective action problems (e.g., Gambetta 2009; Leeson 2010; Skarbek 2011a, 2012; Varese 2001).

The *jogo do bicho* is an exception to this rule. The game has been running for more than a century without considerable interruption, attesting to the fact that *bicheiros* have managed to solve collective action issues in one way or another. Importantly, the *jogo* involves moderately low levels of violence – at least when compared to other illegal activities such as drug trafficking. Moreover, Brazilians widely consider the *jogo do bicho* an honest lottery, and reports of cheating are surprisingly uncommon. The popular motto associated with the game testifies in its favour: "*Vale o escrito; ganhou, leva*", or "What is written is what counts; if you win it, you take it" (Magalhães 2005). Here I analyse two means by which the *bicheiros* elicit voluntary cooperation from gamblers and members of the community: costly signalling and reputation strategies.

Signalling theory predicts that when someone cannot easily observe a characteristic she is interested in, she searches for signals that credibly conveys pieces of that required information. In the case of an illegal lottery, the main signal a gambler is looking for is *honesty*, that is, that she will receive her prize if she buys the winning ticket.

Legal lotteries employ many techniques to show this is the case. For instance, the Brazilian official lottery, run by the federal government, are regularly audited by the *Controladoria Geral da União* (Comptroller General of Brazil), the *Tribunal de Contas da União* (General Accounting Office), and by Ernst & Young. The balls are measured and weighted every three months by the National Institute of Metrology, Quality and Technology (Inmetro), the Brazilian equivalent of United Kingdom's National Physical Laboratory or the American National Standards Institute (UOL 2016).

A clandestine lottery, in contrast, cannot provide the same signals. Thus, the game providers need to assure gamblers that their business is honest although it is illegal. *Bicheiros* addressed this issue using a traditional commercial practice: they invested on their reputations.

The *jogo do bicho* entrepreneurs have made considerable efforts to present themselves as honest brokers. The first trust-enhancing mechanism they have employed to foster external cooperation was the use of a *fixed-multiplier formula* for pay-outs. It works as follows. If a player wins the lowest prize of the animal game, he or she receives 18 times his/her investment regardless of the size of the bet. Bigger prizes naturally offer higher returns; a lucky winner of the top prize wins up to 4,000 times the value of his/her bet (Labronici 2012, 89; Magalhães 2005, 20).

This stands in sharp contrast to the common practice of sharing a prize among winners. Lottery pay-outs demand high levels of interpersonal trust: players rely on unverifiable information about the total funds collected by the lottery, and they can never be sure whether the payments are evenly distributed. The fixed-multiplier formula alleviates such problems of adverse selection (Akerlof 1970; Cohen and Siegelman 2010; Levin 2001). As players and vendors know the prize value beforehand, the method provides consumers with complete information about their individual prizes while also binding the *bicheiros* to a contract that can be easily enforced. This technique offers buyers a simple yet effective screening strategy that induces *bicheiros* to provide honest information about the game (Spence 1973; Stiglitz and Weiss 1981).

Bicheiros have addressed information asymmetries in another ways. Since the 1950s, when the *jogo do bicho* bankers had moved their operations to the *fortalezas*, the public could not oversee the lottery

draws (Chazkel 2011, 259). This could lead to a decline in trust among buyers and vendors of lottery tickets and, as a result, to reduced profits. *Bicheiros* have mitigated this problem with a two-pronged strategy. First, they started to utilise the winning numbers from the licit government-run lottery, the *Loteria Federal*, instead of their own draws (Chazkel 2007, 546; Labronici 2012, 89; Mello 1989, 39-40). The federal lottery numbers are public information. The media broadcasts the draws on radio and TV, so any interested player can verify the selected numbers. The Loteria Federal is also audited by two independent state institutions, a private accounting firm, and voluntary members of the public. Hence, *bicheiros* can free ride on the lottery's long-standing reputation of credibility.

Second, they included representatives of all major *jogo do bicho* bankers in every draw and independently publicise the game results. Certain *bicheiros* went as far as publishing the numbers in Rio's newspapers. In the early twentieth century, some tabloids were entirely dedicated to the game (Magalhães 2005, 60). Booking agents see this strategy as a credible signal from the game financiers, as providing contrasting information would indicate game manipulation. Moreover, collusion can also be spotted if the draws show repeated numbers or unusual patterns.

These efforts have proved popular with the game enthusiasts. Such mutual confidence reduces the potential for conflict in the game. As the public does not see the *jogo do bicho* as violent or harmful, the stigma of repugnance associated with gambling becomes less pervasive. By reducing the possibilities of cheating and putting long-term interests first, the *jogo do bicho* bankers have avoided the fate of other repugnant markets (DaMatta and Soárez 1999, 20).

This ability to elicit cooperation from external actors caused the *jogo do bicho* to last longer than collection action theories would anticipate. In that regard, the bankers' investments in costly signalling and reputation has largely paid off: the lottery's continuous profits are a proof of their success. With that continuous flow of income, *banqueiros* were able to extend their influence well beyond the poor communities in which they operate. One of these areas is politics.

3.4 Tropical State Capture: *Jogo do Bicho*, Samba and Politics

The impact of the *jogo do bicho* is not restricted to the Brazilian economy. Since the 1960s, *bicheiros* have been the key sponsors of the country's most important cultural festivity: the Rio de Janeiro Carnival parade (Bezerra 2009; Cavalcanti 2006; Queiroz 1992). The *jogo do bicho* accounts for such large share of the funding of the parade that a famous *banqueiro* once remarked that "without the

jogo do bicho the Carnival would have ended" (O Dia 2016). Based on that support, *bicheiros* have established an extensive patronage network over samba schools and local politicians (Arguello 2012, 4641; Congresso em Foco 2007; Jornal do Brasil 2011; Misse 2011, 16). Such client network brings large material benefits to their members, yet it has created perverse incentives for government officials and has caused several distortions to the Brazilian democratic system.

The *jogo do bicho*'s clientelism is most evident in the state of Rio de Janeiro. Historical factors explain why this is the case. First, Rio de Janeiro city was the capital of Brazil for almost 200 years, and despite losing the position to Brasília in 1960, it has remained one of the country's main cultural and financial centres. Secondly, *jogo do bicho* operators had historical ties with popular movements, especially samba groups, and the animal game elites eventually exploited these connections to their advantage. Thirdly, the emergence of state-sponsored Carnival parades created a window of opportunity for *bicheiros* to expand their influence over public authorities, either via bribing or by funding political campaigns. In this regard, Rio provided a suitable environment for self-interested politicians, community leaders and animal game financiers to collaborate. These illegal networks are crucial to understand why samba and Carnival became constituent features of Brazil's national identity, and how the festival has contributed to Rio's high levels of state corruption.

3.4.1 The Medici of Samba: *Bicheiros* as Patrons of Carnival

In 1930, opposition leader Getúlio Vargas led a bloodless coup d'état that brought Brazil's First Republic to an end. During his first presidency (1930–1945), Vargas promoted a radical shift in Brazilian politics by effectively dismantling federalism in favour of a powerful executive branch and an expanded federal bureaucracy (e.g. Bethell 2008; Fausto 1972; Skidmore 1967). In terms of ideology, Vargas's authoritarian-corporatist *Estado Novo* ("New State") promoted a politicised nationalism designed to transcend the regional aspects of Brazilian culture (Lauerhass 1972; Williams 2001). Popular music, in turn, occupied an important place in Vargas's project of "Brazilianing Brazil". Created in the late 1920s in the shanty towns of Rio de Janeiro, modern samba embodied the idea of the multicultural, racially-tolerant country the government aspired to forge (McCann 2004; Stockler 2011).

By the late 1930s, samba reached a unique position in Brazil's cultural identity. In a period when civil and political rights were limited, Vargas used samba as a means to incorporate ethnic minorities

and the new urban classes into the Brazilian mainstream (Chinelli and Machado 1993, 213). Patriotic sambas exalted the country's natural beauties and the figure of the "friendly, happy, cordial and industrious" mulatto[8] (Vianna 1995, 51). The institutionalisation of the Carnival parade in 1935, and the subsequent increases in public funding to the festival, cemented the relationship between politicians and samba groups (Cabral 2016b; Soihet 1998).

The samba groups were not passive members in this process. Since the 1960s, the Rio Carnival has expanded in scope. Stimulated by growing numbers of spectators, the parades have become more elaborate (Cabral 2016b; Chinelli and Machado 1993, 214; Hertzman 2013, 240). Unable to cope with the rising costs of the show, the samba schools, which are large samba groups that compete in the Carnival, resorted to the *jogo do bicho* financiers to fund their activities (Misse 2007). This informal agreement between samba school organisers and wealthy *bicheiros* remains effective to this day, and many of Rio's most famous samba schools are officially presided by high-profile members of the *jogo do bicho* elite (Bezerra 2009; Cavalcanti 2006; Misse 2011; Queiroz 1992).

The animal game at times faced opposition by the local population. The public often perceived the game as immoral and repugnant. *Bicheiros* were aware of that problem. They decided to finance samba schools hoping to win the support of the population and attach a more positive image of the game among urban classes. Members of the *jogo do bicho* had been involved in the Carnival since the early 1920s, but in 1984, a group of rich bankers founded collectively the LIESA (*Liga Independente das Escolas de Samba*, Independent League of the Samba Schools), a civil association intended to direct and sponsor the Carnival parade in Rio de Janeiro. The LIESA marked a shift in the history of the Carnival Parade. For the first time, the *bicheiros* decided to act as a group rather than individuals who shared an interest in popular festivals. The institution was very effective in expanding the parade, but it also provided other benefits to the *jogo do bicho* bankers. The organisation consolidated their power over the Carnival and provided a formal mechanism for the *banqueiros* to solve disputes (Cavalcanti 2006, 43; Farias 2013, 171; Labronici 2012, 55).

The funding of the samba schools had an indirect effect on the animal game. The patronage also reduced agent-principal problems within the *jogo do bicho*. *Bicheiros* donate to samba school to gather support of the communities, and by doing so they gain access to local information on their business.

[8] A mulatto is a person of mixed white and black ancestry. The etymology of the word is originally derogatory as it alludes to 'mule' (Latin: *mulus*), the infertile offspring of a male donkey and a female horse. However, in the 1930s the word loses its pejorative connotation in Brazil. Mainly due to the work of sociologist Freyre (1933), the idea of a racial democracy becomes pervasive in the government discourse, and as a result the word gains a positive tone (Reiter and Mitchell 2009, 4).

Clients who have a positive image of the *bicheiro* may denounce fraudsters to their superiors, thus monitoring the cost-effectiveness for animal game managers. Thus, street bookmakers have fewer incentives to cheat. In addition, street sellers are often recruited from the poor communities, so they tend to be immediate beneficiaries of *bicheiros*' donations (BBC 2012). Hence, funds donated to samba schools and other charitable organisations help align the interests of different members of the *jogo do bicho* organisation. The patronage can be interpreted as an illegal version of "profit-sharing," a mechanism that has induced effectively cooperative behaviour in both small and large corporations (Cahuc and Dormont 1997; FitzRoy and Kraft 1987; Kruse 1992).

Samba schools have profited from this association too. First, they have gained autonomy from the government. Samba schools do not need to rely exclusively on public funds to organise the parade, and money from the *jogo do bicho* has permitted the schools to act independently (Chinelli and Machado 1993, 209). Second, the support of the *jogo do bicho* has increased the political and social clout of the samba schools. In a country where the state is not present throughout the territory and human right abuses are frequent (O'Donnell 1993; Pinheiro 2000), *jogo do bicho* bankers, and more recently drug traffickers, have provided private governance to poor areas of Rio de Janeiro by enforcing property rights, mediating disputes, and preventing police abuse in the favelas (Arias 2006; Goldstein 2013). In return for funds and protection from the *bicheiros*, samba schools have served as intermediaries between the underworld and the political system. Although the *banqueiros* are interested in weak law enforcement against the animal game, politicians have resorted to samba schools to contact *bicheiros* and use their financial and electoral influence in the shanty towns (Misse 2011, 17). The samba schools, therefore, have increased their bargaining power in the political sphere and have extended their reach within Rio's poor communities (Chinelli and Machado 1993, 215).

3.4.2 Political Support

Politicians were opposed to the *jogo do bicho* in the early twentieth century, but their relationship with the animal game bankers later became more ambivalent. The collaboration between public authorities and *bicheiros* gained prominence during the military dictatorship (1964–1985) (Gaspari 2002; Jupiara and Otavio 2015). The regime effectively dismantled the few checks and balances implemented in the Second Republic (1945–1964), and paramilitaries and police forces had considerable discretion to repress political dissidents. Extortion of civilians was also widespread (Magalhães 1997; Misse 2009).

But *bicheiros* saw the corruption of some members of the military as an opportunity to increase profits. Wealthy *jogo do bicho* bankers hired rogue police officers to work as security guards and to threaten eventual competitors in their regions of influence. The agreement between *bicheiros* and corrupt members of the military was ultimately responsible for the transformation of the *jogo do bicho* into a coercive oligopoly (Jupiara and Otavio 2015). When the *jogo* transitioned from an informal to an illegal market, the use of violence in the game became more widespread. The support of the armed forces meant new groups would be prohibited from entering the market and the illegal lottery could operate undisturbed by the government.

The links between *bicheiros* and the public authorities changed after Brazil became a democracy in 1985. In the military regime, government officials were mainly interested in bribes from the animal game. But in the democratic period, votes were the most sought-after political resource. *Bicheiros* are important in this sense as they have direct influence over a number of poor communities, either because of their role as patrons, or as reliable sources of governance. Their patronage networks ensure that candidates supported by *bicheiros* receive a substantial amount of votes from areas where campaigning is too difficult or too costly (Misse 2011, 17).

Politicians from all spheres of government are involved with *jogo do bicho* bankers. Recent investigations have shown that from local representatives to senators, politicians of every level receive illegal money to fund their campaigns. Carlinhos Cachoeira, a famous animal game banker from the state of Goiás established a large patronage network that included mayors, deputies, senators, judges, and businessmen, many of them linked to the federal government. His brother Marcos ironically noted that Carlinhos was "too dedicated to politics" and he illegally donated about USD 300 million to political candidates in his home state (O Estado de São Paulo 2012).

The Brazilian political system is particularly conductive to client practices. Brazil has one of the most fragmented party systems in the world, which induces political entrepreneurs to run highly individualised campaigns (Figueiredo and Limongi 2000; Geddes and Neto 1992). In addition, Brazil uses an open-list proportional representation electoral system; each of the 27 states of the federation are considered at-large electoral districts (Ames 1995; Samuels 2000, 483). These two elements indicate that Brazilian politicians are often free from the strong requirements of political parties and can run their campaigns with a high degree of independence. Nevertheless, that independence means

candidates rely mostly on themselves to raise funds and mobilise potential voters. Hence, political campaigns in Brazil tend to be expensive and personality-centred.

The support from the *jogo do bicho* mitigates both problems. With respect to the financial costs of campaigns, illegal donations from *bicheiros* help to cover advertising expenses while having the additional benefit of not appearing in the official records of the candidates (Congresso em Foco 2007; O Globo 2012a). This suggests that *jogo do bicho*-funded politicians can circumvent spending limits and have an electoral advantage over their competitors. As candidates do not know whether their competitors receive funding from the *jogo do bicho* nor the amount each one was paid, their dominant position is to contact the *bicheiros* and join their networks. The situation is a prisoner's dilemma in which candidates would be better off running cheaper campaigns and not being dependent on *jogo do bicho* bankers, but asymmetric information prevents them from reaching a cheaper solution.

Votes from poor communities are instrumental for aspiring politicians. Brazil has an enforced compulsory voting system; therefore, turnout rates tend to be higher than in other democracies. Consequently, votes have high marginal utility for politicians. As elections may be decided by a small difference, the *bicheiros*' client ties guarantee a minimum number of votes that politicians can rely upon on election day. Nonetheless, the patronage subverts the preferences of the public and, as such, the democratic process *per se*. Individuals may be punished if the candidate does not receive the expected number of votes, and they are often compelled to vote for politicians who have only loose connections with their communities. Therefore, although voters have the right to choose their representatives, in practice the suffrage is limited for a share of Brazil's lower classes.

In a nutshell, the difficulty in permanently outlawing the *jogo* showcases how the Brazilian state itself is deeply embedded within criminal sectors of the society. While some sectors of the Brazilian bureaucracy may be described as "islands of excellence" and closely resemble the Weberian ideal of public administration (Bersch et al. 2017), local politicians remain dependent of unstable, and often unlawful, connections with social elites. This system of "partial embedded autonomy" (Evans 1995) provides the required stability for the political system to operate, yet it offers significant opportunities for rent-seeking behaviour. This structure tends to perpetuate itself as it brings benefits for both the animal game bankers and local politicians, such as limited competition and the ability to extract resources from poor voters.

3.5 Concluding Remarks

Past research has shown that criminal organisations face considerable challenges to elicit cooperation from their members and to establish close ties with the population. Yet, the *jogo do bicho* offers a convincing example that it is possible for an illegal syndicate to operate with low levels of violence for more than a hundred years. *Bicheiros* employ a number of strategies to obtain reliable information from their subordinates while offering club goods and other selected benefits to workers. Furthermore, by investing in the Carnival parade, *bicheiros* have been able to gather popular and government support. Poor communities have associated with the *bicheiros* to receive welfare provisions, whereas politicians have collaborated with them to reap the financial and electoral benefits the *jogo do bicho*'s networks can provide.

Nevertheless, the *jogo do bicho* has also created negative externalities. Violence is used to punish defectors and to constrain competitors. The client relationship *bicheiros* have with local politicians has lead to undemocratic outcomes, such as predatory political campaigning, distortions in electoral representation, and impunity for human rights violations. These negative externalities have long-term effects and still impact the Brazilian public sphere.

Although the *jogo do bicho* has received increasing attention from scholars, much of its inner workings remain poorly understood. First, the relationship between *bicheiros* and drug dealers is a topic that deserves attention. Brazil has become one of the world's largest consumers of illicit drugs and South America's principal drug trafficking transit route (Miraglia 2015; Misse 2011). The question whether *bicheiros* collaborated or opposed the emergent drug dealing business is still unclear. Second, the extent to which *bicheiros* use other businesses, such as hotels or factories, to launder money has been mentioned by members of the Brazilian judiciary (O Globo 2012a, 2015); however, there is no reliable estimate on its size. Lastly, more research is required to clarify how *bicheiros* from different parts of Brazil coordinate their activities and prevent large-scale conflicts. Cases studies are usually focused on Rio de Janeiro's *bicheiros*, but scholars would benefit from comparative analyses with a larger number of states. This is an important step to elucidate how *bicheiros* continue to influence politics and the public throughout Brazil.

Chapter 4

What Drives State-Sponsored Violence?: Evidence from Extreme Bounds Analysis and Ensemble Learning Models

4.1 Introduction

Since the end of World War II, mass killings, genocides, and politicides have claimed over 34.5 million lives (Marshall et al. 2017).[1] The international community has responded with an effort to prevent further state-sponsored mass murder by strengthening laws against war crimes, genocide, and crimes against humanity. Furthermore, the United Nations established a Special Adviser on the Prevention of Genocide and recognised its members' responsibility to protect civilian populations within and outside their own borders. Yet, such atrocities still occur. Recently, President al-Assad of Syria has massacred tens of thousands of civilians during the Syrian Civil War (Goldman 2017). Similarly, South Sudan's President Kiir is actively starving and killing civilians from dissident and rival tribes (Nichols 2017). While there is some evidence that such atrocities may be declining since the Cold War (Valentino 2014), the international community has been far from successful in realising slogans like "Never Again" and "Not on My Watch" (Cheadle and Prendergast 2007).

[1] Genocide and politicide are the attempted intentional destruction of communal or political groups, respectively (see Harff and Gurr 1988). Mass killing includes these atrocities, as well as attacks against civilians that result in at least 1,000 deaths but are not intended to destroy a particular group (see Ulfelder and Valentino 2008). While some conflate the logic of these types of atrocities (e.g., Rummel 1995; Valentino et al. 2004), others claim genocide and politicide follow a different logic from other forms of government violence (Kalyvas 2006; Stanton 2015). For discussion on these important differences in conceptualisation see Straus (2007) and Finkel and Straus (2012).

Ultimately, effective prevention requires us to understand why these atrocities occur. In this vein, the academic community has laboured tremendously to establish empirically-based theories as to why governments engage in brutality against their civilian populations. Indeed, since 1995, there have been over 45 quantitative political science articles focused on explaining government-sponsored killing of civilians. Overall, the mass violence literature agrees that government atrocity follows an opportunity logic: as threat increases, so does the likelihood of atrocity, if the costs to such violence are not prohibitive. However, there is little consensus on what factors influence the level of threat or costs a regime faces. Part of the reason for this uncertainty is that scholars use very different model specifications when testing their arguments, thus small changes in model parameters could influence the robustness of empirical results and the inferences I draw from these findings.

To overcome these limitations and provide a better understanding of government atrocity, I employ extreme bounds analysis and random forests to identify the most robust determinants of state-sponsored atrocities. My approach is similar to Hegre and Sambanis' (2006) seminal analysis on the causes of civil war onset, but like Bell (2015), Hill and Jones (2014) and Jones and Lupu (2018), I provide additional tests to verify whether complex interactions and nonlinearities are driving the statistical results. More broadly, my goals are: 1) to examine whether the current quantitative scholarship is able to identify robust explanators of mass atrocities; 2) to evaluate how those variables compare to each other in explaining power.

While recent studies of mass killings have progressively adopted stronger identification strategies, the majority of the literature consists of cross-country regressions. While the importance of micro-level causal designs have been largely discussed (e.g. Angrist and Pischke 2008; Imbens and Rubin 2015), large-*n* analyses also have strong benefits that are often overlooked. For instance, quantitative studies allow scholars to assess the external validity of specific explanatory mechanisms. Additionally, these studies provide a safeguard against the perils of selecting on the dependent variable, a bias that can severely distort regression results (Bell 2015; King et al. 1994). Cross-national comparisons also show how structural factors condition immediate causes of mass violence. Atrocities are multicausal phenomena, and large-*n* studies can point to interactions that scholars would otherwise miss. In this specific study, there is the additional advantage of running all the models with the same dependent variable, what constitutes an effective method of replication of the original findings.

In conducting this analysis, I address three debates in the mass violence literature:

1. Why do some governments engage in mass killings, genocides, or politicides? This is the primary question asked by activists, policymakers, and scholars in this field of research.

2. Does the logic underpinning government decision-making follow different patterns during peacetime and wartime? Recent research suggests that government atrocity occurs predominantly during periods of civil unrest (Harff 2003) which has led some scholars to restrict their analyses to only periods of civil war (e.g., Colaresi and Carey 2008; Valentino et al. 2004) or concentrate on predicting both the onset of civil war and atrocity (Goldsmith et al. 2013). Yet, others estimate models of all country-years (e.g., Krain 1997; Montalvo and Reynal-Querol 2008), raising questions of how well these studies speak to each other.

3. Is there a difference in logic between those atrocities labelled as genocide or politicide, compared to other mass killings? While the Political Instability Task Force (Marshall et al. 2017) provides the most widely used data on government atrocity, others provide data with much more lenient inclusion criteria (e.g., Stanton 2015; Ulfelder 2012). These differences in definition of atrocity have led to divergent results, raising questions about important determinants of government behaviour (for discussion, see Straus 2007; Uzonyi 2016; Wayman and Tago 2010).

My analysis tests the sensitivity of 40 variables on a sample of 177 countries from 1945 to 2013. My findings partially confirm previous research – poor, unstable countries are more likely to witness the regime employ atrocity (e.g., Goldsmith et al. 2013; Harff 2003; Krain 1997). However, many of the factors scholars often cite as observable indicators of such instability – regime transitions, coups d'état, the presence of militias, etc. – are not good proxies for instability. Thus, policymakers may be looking for the incorrect signs of impending atrocity when seeking to prevent its onset. Furthermore, I find that the conclusions scholars draw regarding the likelihood of government atrocity largely depend on whether they combine peace and war years or just analyse periods of civil wars, as patterns in mass killings differ dramatically across these contexts. Lastly, I find that genocide and politicide follow vastly different patterns of onset than other forms of state-sponsored mass murder. This is further evidence that different logics govern different forms of political violence (Stanton 2013).

Overall, these findings raise concerns about policy options for preventing violence against civilians. If my conclusion is that poor and unstable countries are violent, then preventing atrocity likely requires significant investments of time and resources in state-building, which is often politically

and practically infeasible (Doyle and Sambanis 2006). This analysis contributes significantly to the political violence literature by highlighting the parsimonious nature of the logic behind government atrocity and clearing away much of the empirical clutter surrounding this conclusion.

4.2 Empirical Methods

To conduct my analysis, I began by surveying the quantitative political science literature on the causes of government mass killing since Rummel's (1995) seminal work on the subject. Counting only published works, I identified 45 articles which employed logit or probit models of mass killing onset in a global sample. I then included all variables that appeared in at least two of these papers in the data set at the country-year unit of analysis for all years from 1945 to 2013. The appendix provides a complete list of the articles I considered and a complete list of the variables I included in my models. Next, I estimated an extreme bounds analysis to determine which variables were the most robust in explaining the onset of government atrocity. Then, I estimated a distributed random forest analysis to see which of the variables best predicted the onset of these atrocities. In this section, I provide more detail on each of these estimation procedures before turning to the results of both analyses in the next section.

4.2.1 Extreme Bounds Analysis

The first method I employ to test the robustness of the potential determinants of state-led violence is the extreme bounds analysis (EBA). Researchers have employed EBA to assess the sensitivity of the determinants of civil war (Hegre and Sambanis 2006), coups d'état (Gassebner et al. 2016), democratisation (Gassebner et al. 2013), economic growth (Levine and Renelt 1992; Sala-i-Martin 1997), nuclear deterrence (Bell 2015), and political repression (Hafner-Burton 2005). The method is particularly useful when there is no consensus about which covariates belong in the "true" regression model (Sala-i-Martin 1997, 178) and scholars worry that omitted or unnecessary predictors can bias the model estimates (Angrist and Pischke 2008; Clarke 2005; Elwert and Winship 2014; Spector and Brannick 2011, 60).

More specifically, the main purpose of EBA is to estimate the distribution of coefficients of each predictor x in an exhaustive combination of regression models with y as a dependent variable. (Leamer 1985, 308) proposed that "sturdy" variables are those whose minimum and maximum of

their coefficient distribution have the same sign and are situated at a distance from zero. If we are to use the conventional value of $p < 0.05$, the mean of the variable coefficients' distribution should be located at least 1.96 standard deviations away from zero.

Leamer's criterion is intuitive, but other authors contend it is too strict for most social science applications. Sala-i-Martin (1997) argued that Leamer's EBA would increase the number of false negatives; in other words, it would classify as fragile covariates that are truly associated with the response. In this paper, I use Sala-i-Martin's more flexible version of EBA and consider the whole range of values of $CDF(0)$. I choose to use the whole distribution because the aggregate $CDF(0)$ allows researchers to move away from a binary indicator of robustness and present the estimations with their appropriate degrees of confidence. My focus is the percentage of the variable's cumulative distribution function that is smaller or greater than zero. I do not assume that the CDFs are normally distributed and use Sala-i-Martin's generic model.[2] I specify the models as follows:

$$Mass\ Killing\ Onset_{it} = \beta_M M_{it} + \beta_F F_{it} + \beta_Z Z_{it} + v_{it} \tag{4.1}$$

My main dependent variable is *Mass Killing Onset*, which denotes the onset of government-sponsored killings. Ulfelder and Valentino (2008, 2) define a mass killing as "any event in which the actions of state agents result in the intentional death of at least 1,000 noncombatants from a discrete group in a period of sustained violence". Respectively, i and t indicate country and year. M is a set of three covariates that are included in every model due to their prominence in the literature (Levine 1992). In my analysis, M includes the natural logarithm of the GDP per capita to control for income, the Polity IV index to control for level of democracy, and a linear time trend since the last episode of government-led atrocity to account for temporal dependence. F denotes a vector of variables of interest, and Z is a vector of other control variables in addition to those included in M. v is the error term. In practice, however, since this research is interested in the effect of all variables in the data set and do not have true control variables, except from M, F and Z are interchangeable. I thus only use this notation to help clarify the connection of my analysis to previous conflict scholars who employed similar extreme bounds analysis (e.g., Hegre and Sambanis 2006; Gassebner et al. 2016). Following Hegre and Sambanis (2006, 514), I lagged the independent variables one year to reduce the risk of endogeneity.

[2]The generic model provides a better fit to the data. Histograms for all coefficients are available in online appendix.

Although the dependent variable is dichotomous, I use linear probability models in my main analysis. Gassebner et al. (2016, 298) argue that linear probability models are less prone to convergence problems and their results can be readily interpreted. Since the data are grouped into countries, I also use cluster-robust standard errors.

As a precaution against collinearity, I place a limit on the Variance Inflation Factor (*VIF*) of all regression coefficients. The VIF estimates how much of the variance of each predictor is dependent on the other covariates in a model. A VIF of 1 indicates that the predictor is uncorrelated with the remaining covariates. VIF limits are often arbitrary (Bell 2015; O'Brien 2007), thus here I use a moderately conservative VIF of 7. As robustness tests, I run the same models without restriction and with different cut-offs.

Two variables were omitted from EBA models but included in the machine learning estimations. The first is *democracy*, a dummy variable that indicates whether the country has a Polity IV score equal or higher than 5. The second is *interstate war*, a binary covariate measuring if the country is at war in a given year (Sarkees and Wayman 2010). I have decided to omit democracy because of its evident correlation with the Polity measure and interstate war due to its correlation with the dependent variables. EBA models do not converge otherwise.[3] Since this problem does not affect machine learning algorithms, the two variables were included in the second set of estimations.

Lastly, I depart slightly from Sala-i-Martin's suggested method and do not assign weights to EBA. Although he recommends using goodness-of-fit measures to construct regression weights, I follow Sturm and de Haan (2002) and Gassebner et al. (2016, 299) and use the unweighted version of the CDF instead. Goodness-of-fit indicators are not equivalent to the probability of a given model being true (Anscombe 1973; King 1986), and the weights constructed this way are not invariant to transformations in the dependent variable. Moreover, the data set has a number of missing observations, so model comparison measures could be misleading (Lall 2016).

4.2.2 Random Forests

I also make use of random forest (Breiman 2001a) to evaluate whether the empirical results in the mass killings literature are driven by parametric assumptions and model specifications. Random

[3] This is a typical case of multicollinearity and conceptual overlap. Hlavac (2016) suggests specifying a set of mutually exclusive variables to avoid the issue. However, as the Polity index is one of the core variables, I decided to drop the binary democracy indicator and use the continuous measure as it provides more details about the effect of political regimes on mass killings. More information available in the appendix.

forest is a machine learning algorithm that consists of a combination of individual decision trees. In a classification problem, each decision tree uses a vector of covariates to split the dependent variable into two increasingly homogeneous parts (Breiman 2001b). However, decision trees are prone to overfitting, i.e., they match the original data set so closely that they tend to perform poorly with new data (Dietterich et al. 1995; Ho 1998). Random forest, in contrast, avoids this issue by growing a decision tree only to a bootstrap sample of the original data, selecting random features at each split, then aggregating the different trees into a single prediction. If the independent variable is continuous, the algorithm will simply choose the average value of the predictions as the best candidate; if the covariate is discrete, the majority class will be employed. The simple procedure of leaving out some data points and growing separate trees with a random subset of covariates is sufficient to eliminate overfitting (Jones and Linder 2015, 9-10).

Random forest has many desirable properties, such as "highly accurate predictions, robustness to noise and outliers, internally unbiased estimate of the generalisation error, efficient computation, and the ability to handle large dimensions and many predictors" (Muchlinski et al. 2015, 7). Thus, random forest allows the researcher to estimate very flexible models with minimal assumptions. Unlike parametric methods such as ordinary least squares or logistic regressions, the analyst does not have to impose any distributional form to the data-generating process. As a result, random forest is able to effectively uncover complex, nonlinear interaction effects in the data without prespecification (Jones and Linder 2015; Jones and Lupu 2018). Random forest models complement the extreme bounds analysis in two ways; first, by providing robust additional tests to the parametric estimations, and also by pointing out eventual limitations of the widely-employed modelling techniques in the mass violence literature.

In this paper I use distributed random forest (DRF) to model the data, a slightly modified version of the original random forest algorithm (The H2O.ai Team 2017). The DRF has two additional features that are useful for the purposes of this chapter. Firstly, DRF is optimised for big data, as it grows decision trees on separate cores to speed up computation time. Secondly, in DRF, non-observed cases are not assumed to be missing at random, but rather as values that contain information in themselves. The algorithm assumes that observations are missing for a reason, what is most likely the case with

social science data (Lall 2016). This is a more conservative approach than assuming that missing cases fit into an underlying parametric distribution.[4]

The DRF has a series of hyperparameters that can be tuned to improve its predictive performance. For instance, users can control the number of decision trees in each iteration, how deep trees should grow, and many other options. The interaction between parameters is generally complex and may involve thousands of potential combinations. As an example, a researcher interested in four parameters with 10 values each would have to estimate 10,000 models before deciding which is the most efficient one. Also, machine learning parameters are sensitive to the data at hand, that is, an optimal solution for one problem cannot be readily implemented in another data set (Genuer et al. 2008; Goldstein et al. 2010; Jones and Linder 2015).

To address these issues, I perform a grid search to select the most accurate random forest model (Cook 2017, 123). I estimate a model for every possible combination of the hyperparameter space to make sure the model results are robust to different specifications. For model selection, I follow the literature on predictive political science and use the area under the ROC curve (AUC) as the model evaluation metric (e.g., Hill and Jones 2014; Ward et al. 2010, 2013). Models with higher AUC values are considered more accurate.

I add several parameters to the grid search. The first is the number of independent trees to grow in each forest. The starting values are 256, 512, and 1024 trees. The machine learning literature does not provide a heuristic on how large a random forest should be, but Oshiro et al. (2012, 166) affirm that "from 128 trees there is no more significant difference between the forests using 256, 512, 1024, 2048 and 4096 trees." I employ a more conservative approach and start from a higher value that the authors suggest as adding more trees do not reduce prediction accuracy (Breiman 2001b, 7).

The depth of each decision tree also influences the algorithm performance. Deeper trees indicate more complex models, and in general they provide a better fit to the data. Nevertheless, this complexity comes at the risk of overfitting, so deeper trees are not necessarily the most adequate solution for every model (Friedman 2001; Segal 2004, 596). In this article, I let the algorithm decide among using 10, 20, or 40 levels for each tree.

I test whether having balanced classes of the dependent variable (mass killing onset) affects the predictive ability of the model. Since the response measure is heavily imbalanced, oversampling

[4]For more information about how the distributed random forest algorithm deals with missing observations, please refer to: `http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html` (access: December 2017).

the positive responses could potentially improve the results (Chawla et al. 2004; Del Río et al. 2014; Japkowicz and Stephen 2002). I also vary how many variables should be considered for each split in the data. The default option is to use $\sqrt{p}$, where p is the number of columns in the data set. As I have 40 covariates of interest, I have selected 5, 6 and 7 variables per split. The DRF uses a majority voting procedure to select which variable is most important. Additionally, the algorithm chooses the percentage of the training set to be modelled by each tree. The default option is 63.2%, but I include the options of using 50% and 100% of the data. Similarly, I give a range of options for choosing how many columns will be included in each tree. The algorithm can randomly choose among 50%, 90% or 100% of the independent variables when estimating a decision tree.

Finally, I use three types of histogram to find optimal split points for each independent variable. Decision trees consider every value of a given independent variable as a potential candidate for a split in the training data. This process is notably time-consuming, and computation time can be significantly reduced at little loss of precision by taking discrete values of the predictor distribution. The DRF algorithm also offers the choice of randomly cycling through all histogram types, including one of the types in each tree estimation. I adopt this "round robin" arrangement as it is both computationally efficient and methodologically parsimonious.

4.3 Results

4.3.1 Main Model

I endeavour to answer three questions in this analysis: 1) what are the robust predictors of government mass killing, 2) do these predictors differ when considering only cases of civil war, and 3) are genocide and politicide different than other forms of atrocity? Table 4.1 summarises the main EBA results in answering Question 1. The table shows the average coefficient estimate of all regressions for each robust variable along with their mean standard deviations.[5] The table also displays the percentage of regressions that are statistically significant at the 90% level. $CDF(0)$ represents the cumulative distribution function, which is the area of the distribution that falls above or below zero.[6] This is my main statistic of interest, and I consider a covariate to be robust if it has a $CDF(0)$ of 0.9 or higher

[5] A list of all independent variables and coding rules are available in the appendix.

[6] I show whichever area is the largest. The sign of the average β coefficient indicates if most of the cumulative distribution is located above or below zero.

(Sala-i-Martin 1997, 181). Lastly, I report the number of estimated regressions models which included each variable.

Variable	Avg. β	Avg. SE	% Sig.	CDF(0)	Models
Base variables					
Log GDP per capita	-0.0091	0.0052	76.055	0.9335	226707
Additional variables					
Post-Cold War years	-0.0133	0.0085	72.845	0.9472	35614
UCDP civil war onset	0.0529	0.0321	52.378	0.9441	20854
Previous riots	0.0140	0.0100	56.242	0.9216	35614
UCDP ongoing civil war	0.0172	0.0115	65.652	0.9092	20854
Ethnic diversity (ELF)	0.0184	0.0137	56.674	0.9050	35614
Polity IV squared	-0.0002	0.0001	61.206	0.9031	35614

Table 4.1: Extreme Bounds Analysis – Mass Killings (Robust Variables Only)

Seven variables pass the EBA criterion and three of them decrease the likelihood of mass killings. First, as widely suggested in the literature, the natural logarithm of GDP per capita is negatively associated with the onset of mass killings (e.g., Besançon 2005; Easterly et al. 2006; Esteban et al. 2015). Second, the post-Cold War years are correlated with lower levels of government violence. Indeed, this finding is in line with several studies that point to a general decline in violence over the last decades, including riots, civil wars, and urban crime (Pinker 2011; Straus 2012b; Valentino 2014). The third robust variable is the squared term of the Polity IV political regime index. This finding points to a nonlinear relationship between political regime and mass killings, thus providing further evidence that democracy reduces state-sponsored violence (Rost 2013; Rummel 1995) and that regimes that mix democratic with autocratic features have the highest risk of conflict (Hegre et al. 2001; Muchlinski 2014).

Four variables are robustly and positively associated with Ulfelder and Valentino's (2008) indicator of government-sponsored violence. Onset and continuation of civil wars are correlated with mass killings, but only when I employ the UCDP measures of violent conflict. I find no effect for the variables compiled by the Correlates of War project or Cederman et al. (2010). Former instances of political turmoil also have a positive coefficient in the models. Moreover, countries with a previous

history of riots are more prone to state violence, which suggests that government repression is path dependent (e.g., Gurr 2000; Harff 2003; Krain 1997; Nyseth Brehm 2017). The results also show that higher levels of ethnic diversity increase the likelihood of atrocities against civilians. Nevertheless, ethnic diversity does not pass all additional tests I implement below and the sturdiness of this finding remains open to question.

Overall, the EBA indicates two patterns in answer to my first question on the causes of mass killing. Atrocity is (1) more likely when violence is already present, reducing the costs of escalating brutality and (2) is less likely as domestic and international constraints increase, increasing the costs of escalating violence further. These patterns support the dominant opportunity narrative in the literature. However, several of the variables commonly used to proxy opportunity, such as military size or regime change, are not robust predictors of atrocity. Thus, this analysis helps clear away much of the brush around the opportunity argument.

Figure 4.1 presents the ten most important predictors of state-sponsored violence in the random forest models. In general, the machine learning estimations have a good fit, with an AUC of about 0.83 in the validation sample. The algorithm confirms some of the main findings of EBA, yet they also show some interesting patterns. Only democracy and state capacity appear to be robust explanators of mass killing in both EBA and random forest models. These patterns further support the refrain that stable states tend to stay stable states.

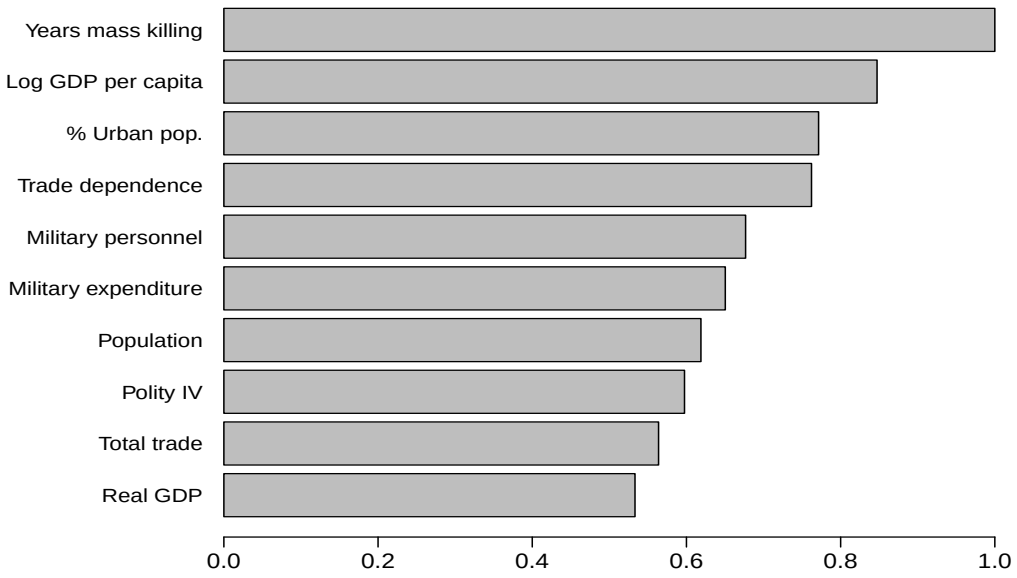

Figure 4.1: Distributed Random Forest – Variable Importance (Scaled)

Interestingly, several variables that are not robust explanators of mass killing in the EBA, have large importance in the machine learning estimates. Variables related to characteristics of the military forces are a good example. As seen below, parametrisation and interactions likely account for this difference. The linear model imposes a parametric structure to the covariate, and the relationship between an independent variable and the response may be a nonlinear one. Also, variables can be relevant predictors only when in interaction with each other. In both cases, those relationships will be captured in the machine learning estimations but not in the extreme bounds analysis. This provides evidence that model specification is driving some of the results in the EBA.

Figure 4.2 displays the partial dependence plots for the ten variables that the distributed random forests highlight as the most important explanators of mass killing onset. These graphs are akin to marginal effect plots in correlation models and help clarify the directional effects of these variables over their entire range. For example, one can see that the effects of Years since Last Mass Killings is highly nonlinear, or that Log GDP per capita does not decrease the likelihood of mass killings after it reaches values close to 9.

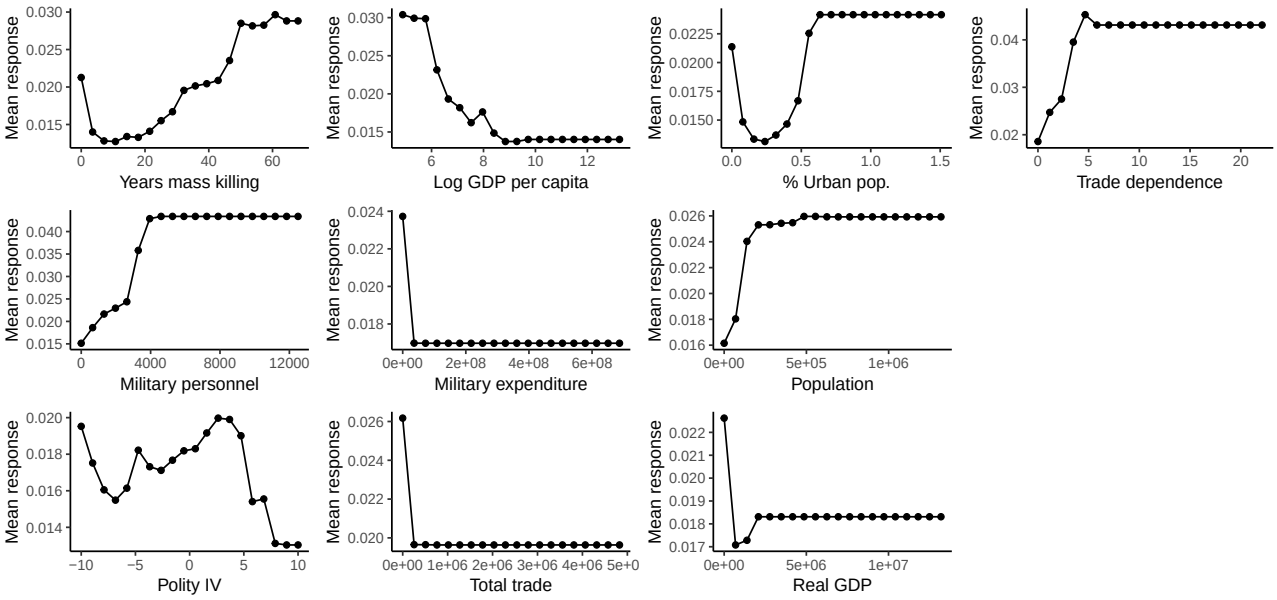

Figure 4.2: Distributed Random Forest – Partial Dependence Plots

One can also infer that authoritarian and mixed political regimes are more likely to engage in mass violence than democratic countries, a result that is also supported by both the EBA and the specialised literature. The number of military personnel positively affects the likelihood of mass killings, yet this increase is counterbalanced by military expenditures. Taken together, these results indicate that countries with large and poorly-funded armed forces have higher risks of mass violence.

4.3.2 Mass Killings during Civil Wars

Table 4.2 presents the EBA results when I restrict the analysis to only civil war years to answer Question 2. I consider three different codings of civil war: 1) the Uppsala Conflict Database Program 2017; 2002), 2) the Correlates of War project (Sarkees and Wayman 2010), and 3) ethnic civil war from Cederman et al. (2010). I find two important patterns. First, considering only civil war years provides a very different understanding of atrocity. Across these models, the only similarity with the full analysis is that mass killing is less likely post-Cold War. Instead, military factors, such military size and militias, and territorial war aims are the most robust predictors of atrocity once war begins. However, contrary to past expectation (Koren 2017), militias have a negative impact on the likelihood of mass killings. Second, there is wide variation in which variables are robust depending on how scholars code civil war. Across the three codings I use here, no variable is robust to all codings and only territorial aims and militias are robust to more than one coding. These results are concerning for scholars using correlation models, as they indicate that our understanding of atrocity, from null hypothesis testing, is largely dependent on which coding of civil war researchers use. For example, only the UCDP data suggests that the post-Cold War years see less barbarism than during the Cold War.

Variable	Avg. β	Avg. SE	% Sig.	CDF(0)	Models
UCDP Data					
Territory aims	-0.044	0.019	74.997	0.9804	17902
Post-Cold War years	-0.038	0.019	66.574	0.9222	17902
COW Data					
Physical integrity	0.024	0.013	66.674	0.9564	17902
Militias	-0.099	0.048	73.104	0.9490	17902
Years since last mass killing	0.006	0.002	88.208	0.9472	101583
Previous riots	0.078	0.041	65.412	0.9348	17902
Ethnic diversity (ELF)	0.095	0.062	48.615	0.9000	17902
Cederman et al. Data					
Territory aims	-0.051	0.026	74.288	0.9167	17902
Militias	-0.050	0.035	52.240	0.9101	17902

Table 4.2: EBA – Mass Killings during Civil Wars (Robust Variables Only)

When I analyse the three codings of civil war using random forest analysis, I find further intricacies in the patterns of mass killing. First, the machine learning estimates highlight a different set of variables than the EBA when analysing the UCDP and Ethnic War data. However, the COW EBA and machine learning analyses both highlight the importance of human rights, previous riots, and the time since the state last engaged in mass killing. Thus, the COW analysis provides the most stable picture of atrocity during civil war. It again highlights the important pattern of the Conflict Trap: violence breeds violence. Second, though, the three codings of civil war each highlight a very similar set of strong predictors of atrocity during conflict. Therefore, the machine learning estimates are not as dependent on the data set employed as are the EBA results. This is good news for scholars of mass killing because it indicates that while the parametric models do not produce robust findings across different civil war data sets, the nonlinear models are able to given us a consistent and clear picture of which factors place a country at the greatest risk for atrocity during civil war.

Figures 4.6–4.5 display the partial dependence plots for the variables with the highest impact in each of the three civil war data sets.

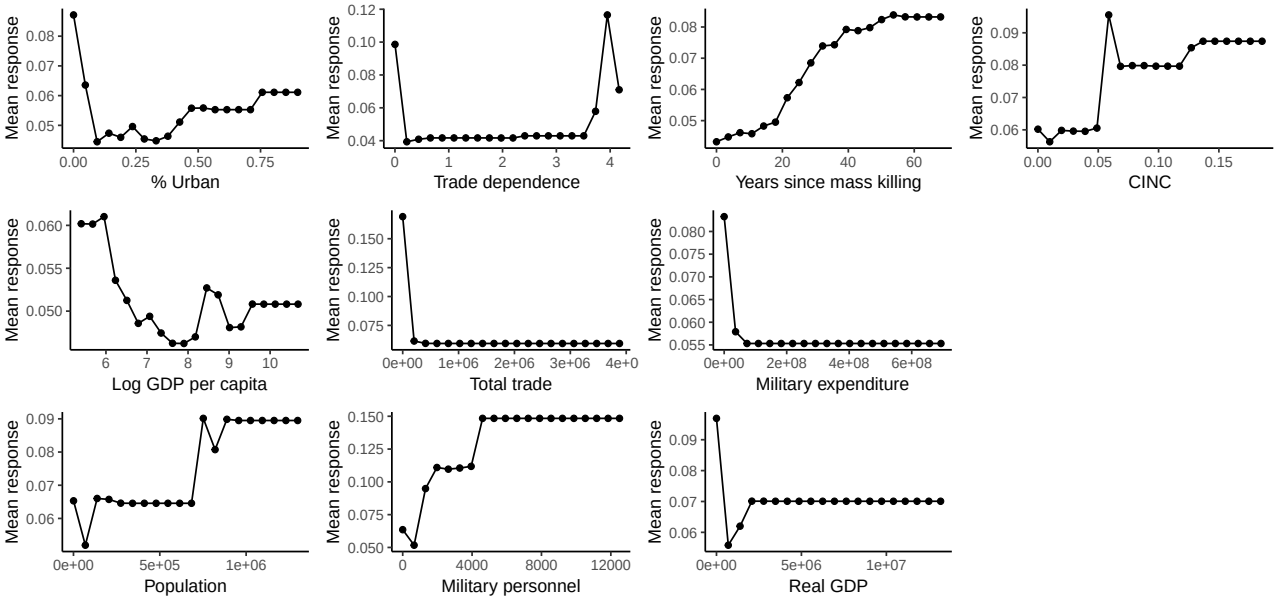

Figure 4.3: Partial Dependence Plots – Mass Killings during Civil Wars (UCDP Data)

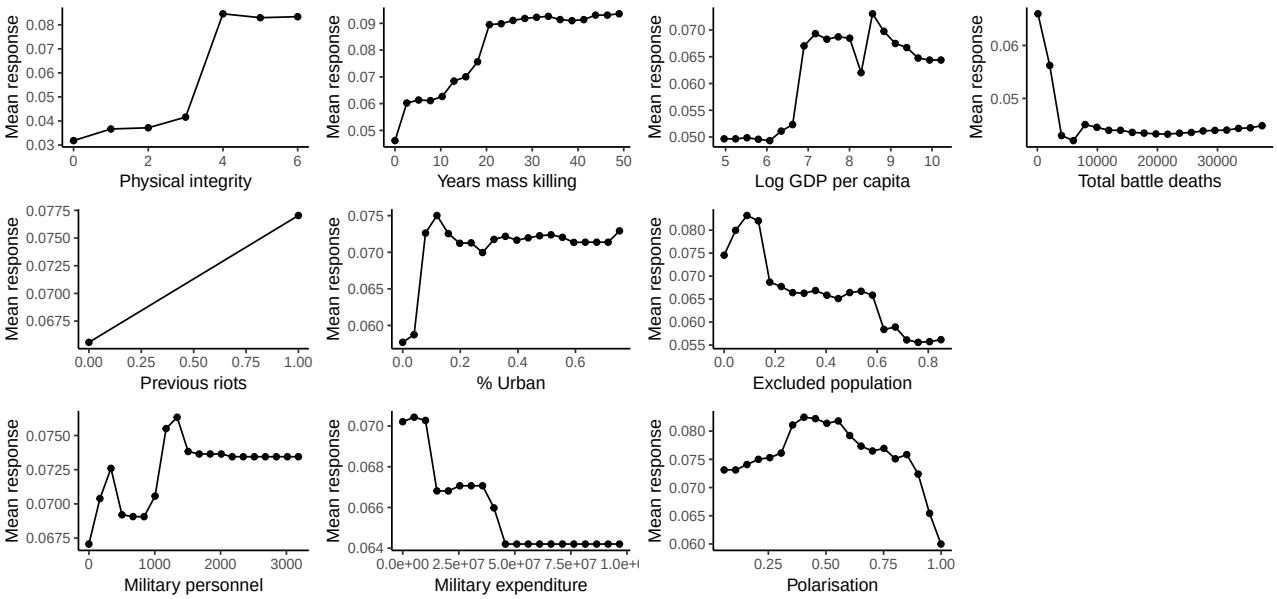

Figure 4.4: Partial Dependence Plots – Mass Killings during Civil Wars (COW Data)

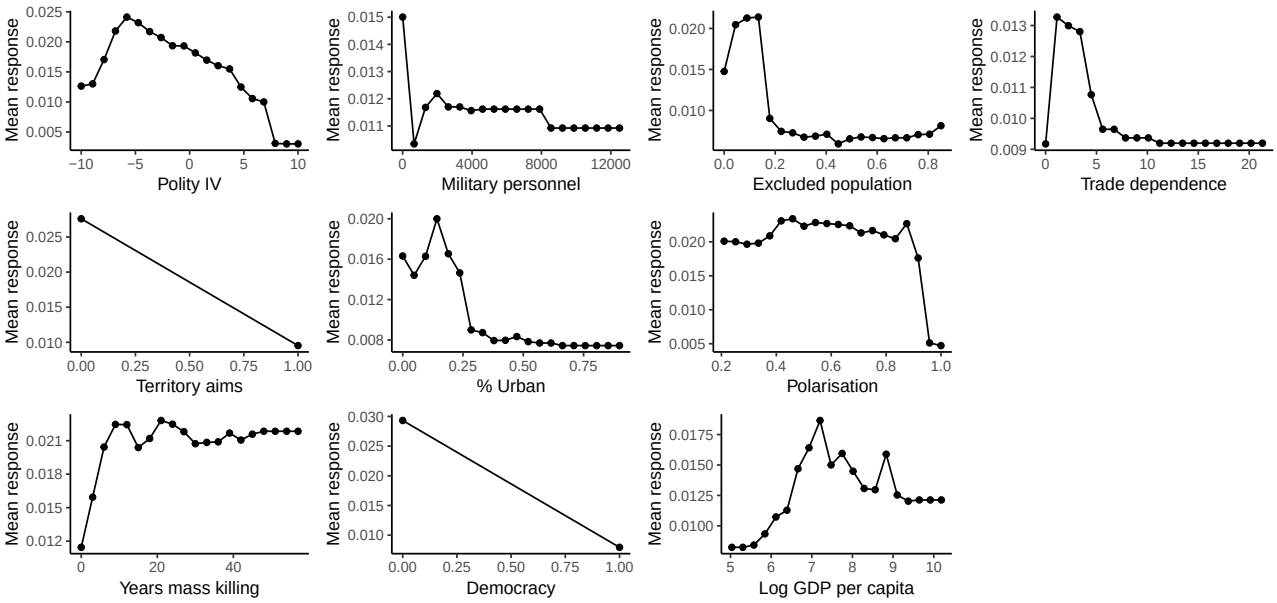

Figure 4.5: Partial Dependence Plots – Mass Killings during Civil Wars (Cederman et al. Data)

4.3.3 Mass Killings in and after the Cold War

Lastly, I test the heterogeneity of the main findings with three sets of models. First, I analyse which factors increase the risk of mass killings during and after the Cold War period. Global dynamics may influence the cost-benefit calculus of state leaders, and consequently affect the likelihood of large-scale responses to internal threats.

Variable	**Avg. β**	**Avg. SE**	**% Sig.**	**CDF(0)**	**Models**
Cold War Period					
Log GDP per capita	-0.018	0.009	83.204	0.9678	50000
Previous riots	0.022	0.014	62.457	0.9031	8278
Post-Cold War Period					
Ethnic war onset	-0.024	0.011	89.608	0.9823	4850
Coup d'état	-0.022	0.011	89.200	0.9822	8602
Territory aims	-0.027	0.014	81.083	0.9653	8775
Displaced Population	-0.048	0.027	58.689	0.9392	8695

Table 4.3: EBA – Mass Killings in and after the Cold War Period (Robust Variables Only)

The EBA models show different patterns for both periods. In the Cold War years, Log GDP per capita has a negative impact on mass killings, while instances of previous riots increase the likelihood of state-led atrocities. The results are in line with those of the pooled model. However, mass killings seem to follow a separate logic in the post-Cold War years. Four independent variables lower the risk of mass killings: ethnic war onset, wars fought for territorial aims, coups d'état, and the share of discriminated population. I interpret the results as showing that ethnic wars are fought by groups with similar capabilities, thus large-scale, one-sided violence is relatively rare. This also explains why atrocities are more likely to occur in countries where the share of discriminated population is not very large. The models show that territorial wars lead to more mass killings than governmental conflicts, a findings which has been previously described in the literature (Eck and Hultman 2007, 240). Coups d'état are correlated with fewer atrocities as well, what stands in contrast with previous research (Wayman and Tago 2010, 10).

The random forests models also display some difference between the two periods, yet several variables appear in both estimations and in the main models presented above. The results for the Cold War model also classify Log GDP per capita and previous riots as important predictors of mass killings. As expected, the effect is negative for income and positive for past social upheavals. The variables that appear in the main model have similar distribution, such as the inverted-U relationship between the Polity IV index and mass killings, and the shape decline in atrocity risk when Log GDP per capita has a value of of 10.

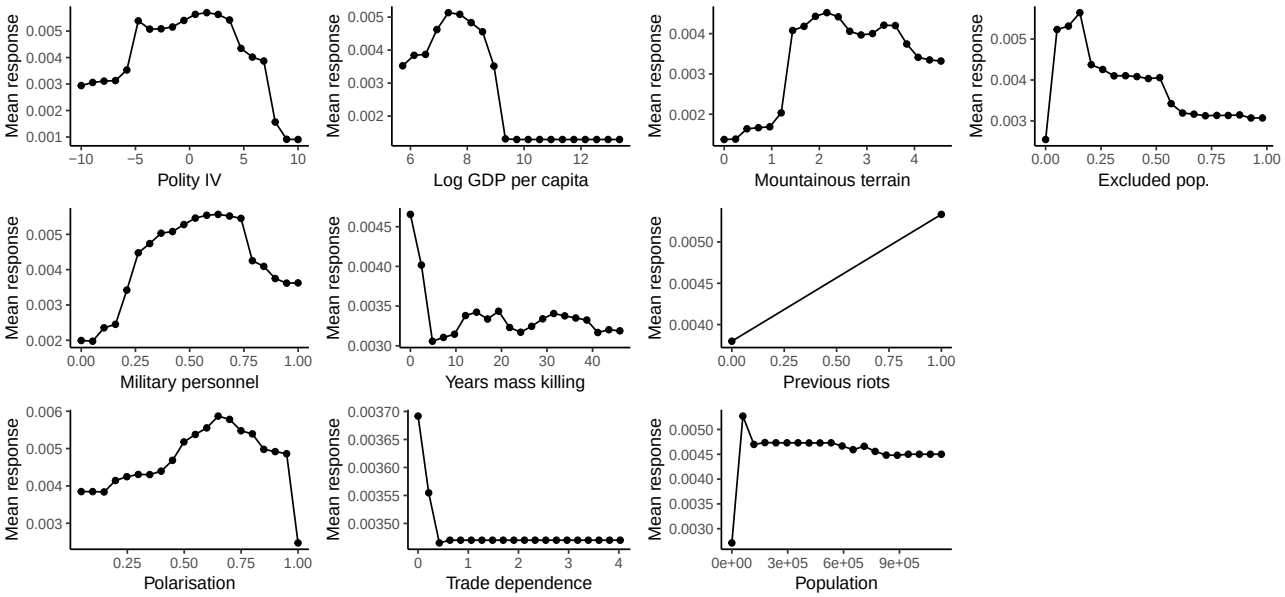

Figure 4.6: Partial Dependence Plots – Mass Killings during the Cold War Period

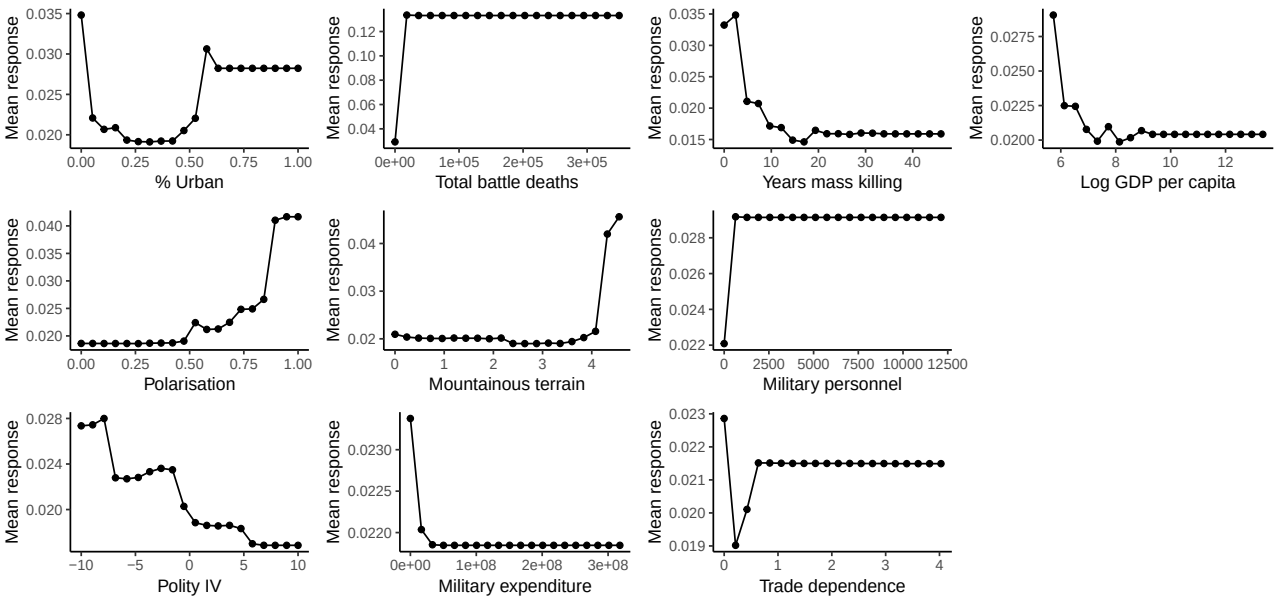

Figure 4.7: Partial Dependence Plots – Mass Killings after the Cold War Period

4.3.4 Genocides and Politicides

To answer Question 3, I estimate the same regressions using Harff's (2003) indicator of genocide and politicide. No variable appears significant in the EBA models for genocide or politicide onset in the full data set. When I limit the sample to civil war years, the Post-Cold War period is negatively correlated with the outcome when using the Correlates of War data set. Excluded population has a negative sign in more than 90% of the models using both Correlates of War's and Cederman et al's (2010) indicators of conflict. Displaced population also has a negative effect in the Correlates of War data set. During ethnic conflicts, the dummy variable for political assassinations has a negative impact on the onset of genocides. Overall, from these EBA analyses, one can conclude then that the significant covariates of genocide and politicide onset differ significantly from those of more general forms of government mass violence. Though, the opportunity story still receives some limited support in these models. However, the machine learning models using Harff's genocide and politicide data are comparable to the ones I present above, with a similar set of variables appearing in the random forest estimations. These results once more highlight that while the mass killing literature struggles to identify correlates of atrocity that are robust across model specification, scholars have done a much better job at identifying variables that help predict both the onset of genocide/politicide and mass killings, more broadly.

4.4 Additional Tests

I estimate a set of additional regressions to assess the robustness of the main findings.[7] In regard to EBA, I include 10 variants of the original model. They largely confirm the prior results. First, I varied the number of covariates included in each regression to 3 and 5 while keeping the M set of 3 control variables. The results are the same as those of the main model, except that ethnic fractionalisation and Polity IV squared become marginally significant with a CDF(0) of about 0.88. Second, I place different restrictions on the variance inflation factor (VIF) to test whether multicollinearity is driving the results. The two models with different values of VIF are identical to the model reported here, while in the model with no VIF restriction ethnic fractionalisation again fails to meet the threshold by a very small margin.

I also reestimate the models using logit and probit regressions. In order to deal with the issue of complete separation (Bell and Miller 2015; Zorn 2005) I follow Gelman et al. (2008) and add a weakly informative prior distribution to the coefficients. In both cases, the logarithm of GDP per capita, post-Cold War period, previous riots, and Polity IV squared remain significant.

As a last robustness test for the EBA, I ran the main model with peace years only; that is, only country-years in which the UCDP, COW and Cederman et al's dichotomous measures of civil conflicts are equal to zero. Despite some issues of multicollinearity,[8] the results are similar to the original model, what indicates that the difference in the estimations is conditional on civil war years.

In regard to random forests, grid searches are themselves a data-driven selection of many possible machine learning models, thus it is not strictly necessary to run a batch of additional tests. Nevertheless, I performed a series of grid searches using three different seeds obtained from `http://random.org` to estimate how different starting numbers influence the model outcomes. The output of those models are largely comparable. The results of each of these analyses are available in the appendix.

[7] For computational purposes, I conducted all additional tests on 50,000 random draws from EBA's posterior distribution. Sala-i-Martin et al. (2004, 819) argue that random draws from the full EBA models are unbiased.

[8] Some independent variables were dropped from the models due to problems of collinearity. The bounds for my indicator of "wars fought over territory" could not be estimated, and the coefficients for "number of battle deaths" and "presence of guerrillas" are unreliable due to their sample sample size. The appendix contains the distribution of the coefficients.

4.5 Conclusion

In this chapter, I apply extreme bounds analysis and distributed random forests to estimate the robustness and predictive ability of 40 variables that have been pointed out as potential determinants of mass killings. I find strong evidence that mass killings are unlikely to happen in rich, stable countries. Nevertheless, there is considerable heterogeneity in some of the results. The findings point out that mass killings have different causes according to the context in which they erupt, so a general theory of state atrocities may obscure important details in our understanding of state killings. Moreover, mass killings are rare events, so local factors likely play an important role in their onset (Straus 2007, 2012a).

Yet one can see this diversity of outcomes under a positive light. The results above suggest new avenues for research, and they also highlight the importance of scholars moving from simple cross-country regressions to methods that can yield more robust predictions. For instance, why are mass killings in ethnic conflicts correlated with a different set of variables than in armed conflicts in general? Would the results remain robust had scholars decided to code ethnic conflicts in another way? More theoretical advancement would also be welcome. Given that GDP per capita is negatively correlated to state atrocities in virtually every model, it would be interesting to unpack the causal mechanisms by which it operates by testing more specific mechanisms.

In terms of practical implications, the results indicate that democratisation and pro-growth economic policies are the most efficient ways to prevent mass killings. The international community can therefore play a role in deterring leaders from using force against their own population, either by offering support for domestic opposition groups, intervening, or by fostering economic development. Although costly in the short run, and sometimes violent during the transition, these measures would substantially decrease the likelihood of state violence by breaking the "conflict trap" in which past conflicts create the condition for new ones (Collier 2003).

4.6 Appendix

This appendix contains all required information to replicate the numerical analyses presented in sections 4.3 and 4.4. R code can be found in subsection **??** and the data are available on the following GitHub repository: https://github.com/danilofreire/mass-killings. I used R version 3.4.4 (15-03-2018) and Ubuntu 16.04.4 LTS to perform all statistical calculations.

4.6.1 Variable Selection

I employ some criteria to select our explanatory variables. First, I included only published articles in the sample. Although working papers and policy may also provide important insights about the onset of mass killings, peer-reviewed research is probably better suited for our purposes. Also, I included only papers that use regression methods on a global sample and were published from 1995 to 2015. The final sample comprises 45 articles: Anderton and Carter (2015), Balcells (2010, 2011), Besançon (2005), Bulutgil (2015), Bundervoet (2009), Clayton and Thomson (2016), Colaresi and Carey (2008), Downes (2006, 2007), Easterly et al. (2006), Eck and Hultman (2007), Esteban et al. (2015), Fazal and Greene (2015), Fjelde and Hultman (2014), Goldsmith et al. (2013), Harff (2003), Joshi and Quinn (2017), Kim (2010), Kim (2016), Kisangani and Wayne Nafziger (2007), Koren (2017), Krain (1997), Manekin (2013), McDoom (2013, 2014), Melander et al. (2009), Montalvo and Reynal-Querol (2008), Pilster et al. (2016), Querido (2009), Raleigh (2012), Rost (2013), Rummel (1995), Schneider and Bussmann (2013), Siroky and Dzutsati (2015), Stanton (2015), Sullivan (2012), Tir and Jasinski (2008), Ulfelder and Valentino (2008), Ulfelder (2012), Uzonyi (2015, 2016) Valentino et al. (2004), Valentino et al. (2006), Verpoorten (2012), Wayman and Tago (2010), Wig and Tollefsen (2016), and Yanagizawa-Drott (2014).

In those 45 studies, scholars made use of nearly 180 measurements to capture roughly 30 key concepts related to threat and costs of mass killings. To be added to our models, a variable should appear in at least two articles. The covariates are summarised in table 4.4. A complete list of variables is available at https://github.com/danilofreire/mass-killings.

Table 4.4: Independent Variables

Variable	Coded	Source
Assassination	Dichotomous	Banks (1999)
CINC	Continuous	Singer et al. (1972)
Coup d'état	Dichotomous	Marshall et al. (2017)
COW civil war onset	Dichotomous	Singer et al. (1972); Singer (1988)
COW civil war ongoing	Dichotomous	Singer et al. (1972); Singer (1988)
Democracy (Polity IV $\geq$ 6)	Dichotomous	Authors' own calculations
Discriminated dummy	Dichotomous	Cederman et al. (2010)
Discriminated population	Continuous	Cederman et al. (2010)
Ethnic diversity (ELF)	Continuous	Fearon and Laitin (2003)
Ethnic war start	Dichotomous	Cederman et al. (2010)
Ethnic war ongoing	Dichotomous	Cederman et al. (2010)
Excluded population	Continuous	Cederman et al. (2010)
Interstate war	Dichotomous	Singer (1988); Singer et al. (1972)
Guerrilla	Dichotomous	Balcells and Kalyvas (2014)
Military expenditure	Continuous	Singer et al. (1972)
Military personnel	Continuous	Singer et al. (1972)
Militias	Dichotomous	Carey et al. (2013)
Mountainous Terrain	Continuous	Fearon and Laitin (2003)
Physical integrity	Continuous	Cingranelli and Richards (2010)
Polarisation (all groups/main group)	Continuous	Authors' own calculations
Polarisation (all groups/population)	Continuous	Authors' own calculations
Polarisation (included groups/population)	Continuous	Authors' own calculations
Polarisation (included groups/main group)	Continuous	Authors' own calculations
Polity IV	Continuous	Marshall et al. (2017)
Polity IV squared	Continuous	Authors' own calculations
Population	Continuous	Gleditsch (2002)
Post-Cold War	Dichotomous	Authors' own calculations
Real GDP	Continuous	Gleditsch (2002)
Real GDP per capita	Continuous	Gleditsch (2002)
Real GDP per capita (log)	Continuous	Authors' own calculations
Regime transition	Continuous	Authors' own calculations
Riot	Dichotomous	Banks (1999)
Total battle deaths	Continuous	Lacina and Gleditsch (2005)
Total trade	Continuous	Singer et al. (1972)
Trade dependence (total trade/real GDP)	Continuous	Authors' own calculations
UCDP civil war onset	Dichotomous	Allansson et al. (2017); Gleditsch et al. (2002)
UCDP civil war ongoing	Dichotomous	Allansson et al. (2017); Gleditsch et al. (2002)
Urban population (percentage)	Continuous	Singer et al. (1972)
Years since last mass killing	Continuous	Authors' own calculations
War with territory aims	Dichotomous	Allansson et al. (2017); Gleditsch et al. (2002)

4.6.2 Descriptive Statistics

Table 4.5: Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Country code	9,162	452.84	247.74	2	950
Year	9,162	1,983.56	18.77	1,945	2,013
Genocide/politicide onset	8,933	0.005	0.07	0	1
Mass killing onset	9,162	0.01	0.11	0	1
Independent Variables					
Assassination dummy	8,991	0.08	0.27	0	1
CINC	8,767	0.01	0.02	0.00	0.38
Coup dummy	8,587	0.05	0.21	0	1
COW civil war onset	8,160	0.01	0.12	0	1
COW civil war ongoing	8,160	0.07	0.25	0	1
Democracy dummy	8,991	0.37	0.48	0	1
Discriminated dummy	6,981	0.35	0.48	0	1
Discriminated population	6,981	0.06	0.15	0.00	0.98
Ethnic diversity (ELF)	6,981	0.41	0.31	0	1
Ethnic war start	7,760	0.01	0.12	0	1
Ethnic war ongoing	7,760	0.11	0.31	0	1
Excluded population	6,981	0.16	0.22	0.00	0.98
Interstate war	8,159	0.04	0.19	0	1
Guerrilla dummy	714	0.81	0.40	0	1
Military expenditure	8,290	4,607,120	27,785,906	0	693,600,000
Military personnel	8,620	176.70	520.90	0	12,500
Militias	4,097	0.22	0.42	0	1
Mountainous Terrain	7,358	2.14	1.43	0.00	4.56
Physical integrity	4,499	4.73	2.31	0	8
Polarisation (all groups/main group)	6,981	0.70	0.26	0.05	1
Polarisation (all groups/population)	6,981	0.63	0.32	0	1
Polarisation (included groups/population)	5,610	0.64	0.32	0	1
Polarisation (included groups/main group)	6,981	0.23	0.35	0	1
Polity IV	8,558	0.42	7.50	−10	10
Polity IV squared	8,558	56.35	32.59	0	100
Population	8,293	32,993.61	112,886.40	118.21	1,324,353.00
Post-Cold War	8,991	0.40	0.49	0	1
Real GDP	8,293	215,317.70	804,827.20	129.68	13,193,478.00
Real GDP per capita	8,293	8,104.20	18,376.73	132.82	632,239.50
Real GDP per capita (log)	8,293	8.25	1.20	4.89	13.36
Regime transition	1,221	−4.24	41.50	−77	99
Riot dummy	8,991	0.16	0.36	0	1
Total battle deaths	714	6,050.86	24,404.78	100	350,000
Total trade	8,174	53,804.01	222,209.90	0.80	4,825,363.00
Trade dependence	7,670	0.26	0.69	0.0001	22.11
UCDP civil war onset	8,733	0.02	0.14	0	1
UCDP civil war ongoing	8,733	0.15	0.36	0	1
Urban population (percentage)	8,767	0.22	0.17	0.00	1.51
Years since last mass killing	9,162	23.81	17.71	0	68
War with territory aims	8,924	0.07	0.26	0	1

Note: All independent variables were lagged one year.

4.6.3 Extreme Bounds Analysis Extensions

Main Model

I present a series of histograms with the coefficients' distribution of all variables in the main EBA model. There are 36 variables in total, seven of which are robust: Log GDP per capita, post-Cold War period, onset and ongoing civil wars (measured by the UCDP), previous riots, ethnic diversity and the squared term of the Polity IV index.

Variable	**Avg. β**	**Avg. SE**	**% Sig.**	**CDF(0)**	**Models**
Base variables					
Log GDP per capita	-0.0091	0.0052	76.055	0.9335	226707
Additional variables					
Post-Cold War years	-0.0133	0.0085	72.845	0.9472	35614
UCDP civil war onset	0.0529	0.0321	52.378	0.9441	20854
Previous riots	0.0140	0.0100	56.242	0.9216	35614
UCDP ongoing civil war	0.0172	0.0115	65.652	0.9092	20854
Ethnic diversity (ELF)	0.0184	0.0137	56.674	0.9050	35614
Polity IV squared	-0.0002	0.0001	61.206	0.9031	35614

Table 4.6: Extreme Bounds Analysis – Mass killings

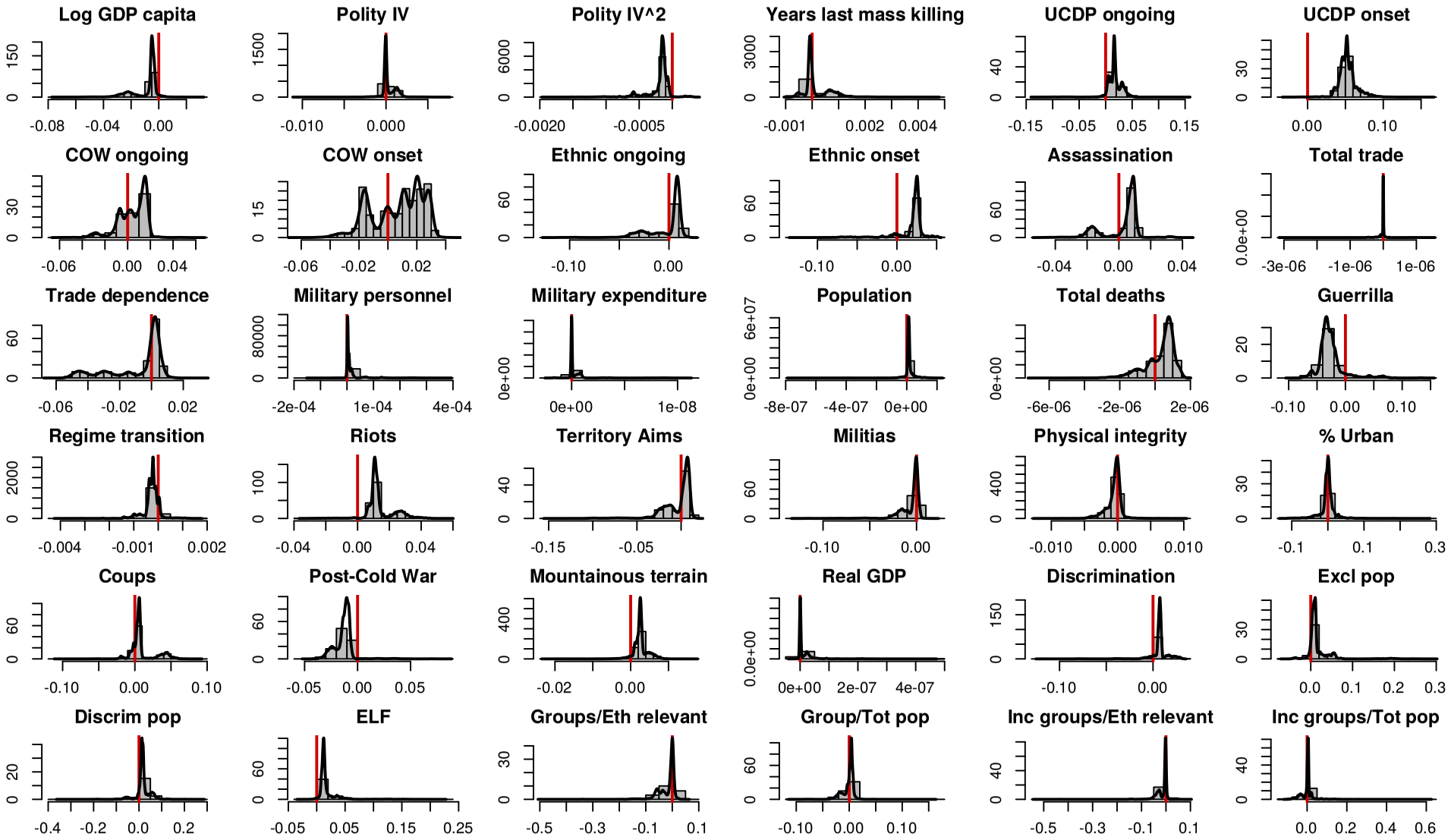

Figure 4.8: Extreme Bounds Analysis – Mass Killings

Genocides during Civil Wars

Next, I discuss genocides that occur during wartime. I use three covariates that denote ongoing civil conflicts: one by the Uppsala Conflict Data Program (Allansson et al. 2017; Gleditsch et al. 2002), another by the Correlates of War (Sarkees and Wayman 2010), and a third indicating the onset of ethnic conflict as coded by Cederman et al. (2010). The variables that reach significance in this set of models below are notably different from those obtained in the main estimation. This result provides evidence that mass violence during wartime time follows a separate logic from state killings in peacetime.

Variable	**Avg. β**	**Avg. SE**	**% Sig.**	**CDF(0)**	**Models**
UCDP data					
Territory aims	-0.044	0.019	74.997	0.9804	17902
Post-Cold War years	-0.038	0.019	66.574	0.9222	17902
COW data					
Physical integrity	0.024	0.013	66.674	0.9564	17902
Militias	-0.099	0.048	73.104	0.9490	17902
Years since last mass killing	0.006	0.002	88.208	0.9472	101583
Previous riots	0.078	0.041	65.412	0.9348	17902
Ethnic diversity (ELF)	0.095	0.062	48.615	0.9000	17902
Cederman et al. data					
Territory aims	-0.051	0.026	74.288	0.9167	17902
Militias	-0.050	0.035	52.240	0.9101	17902

Table 4.7: EBA – Mass Killings during Civil Wars

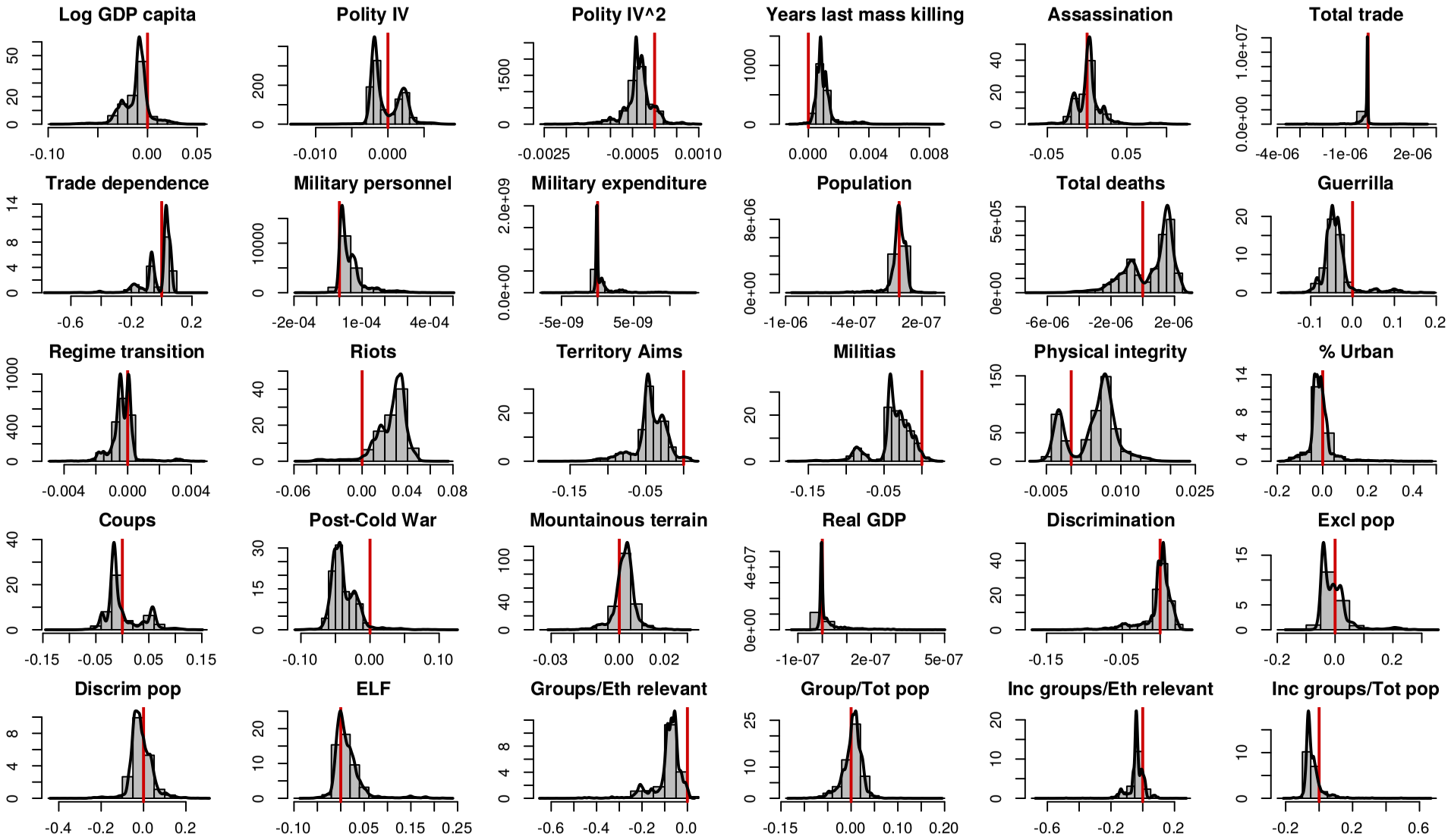

Figure 4.9: EBA – Mass Killings during Civil Wars (UCDP Data)

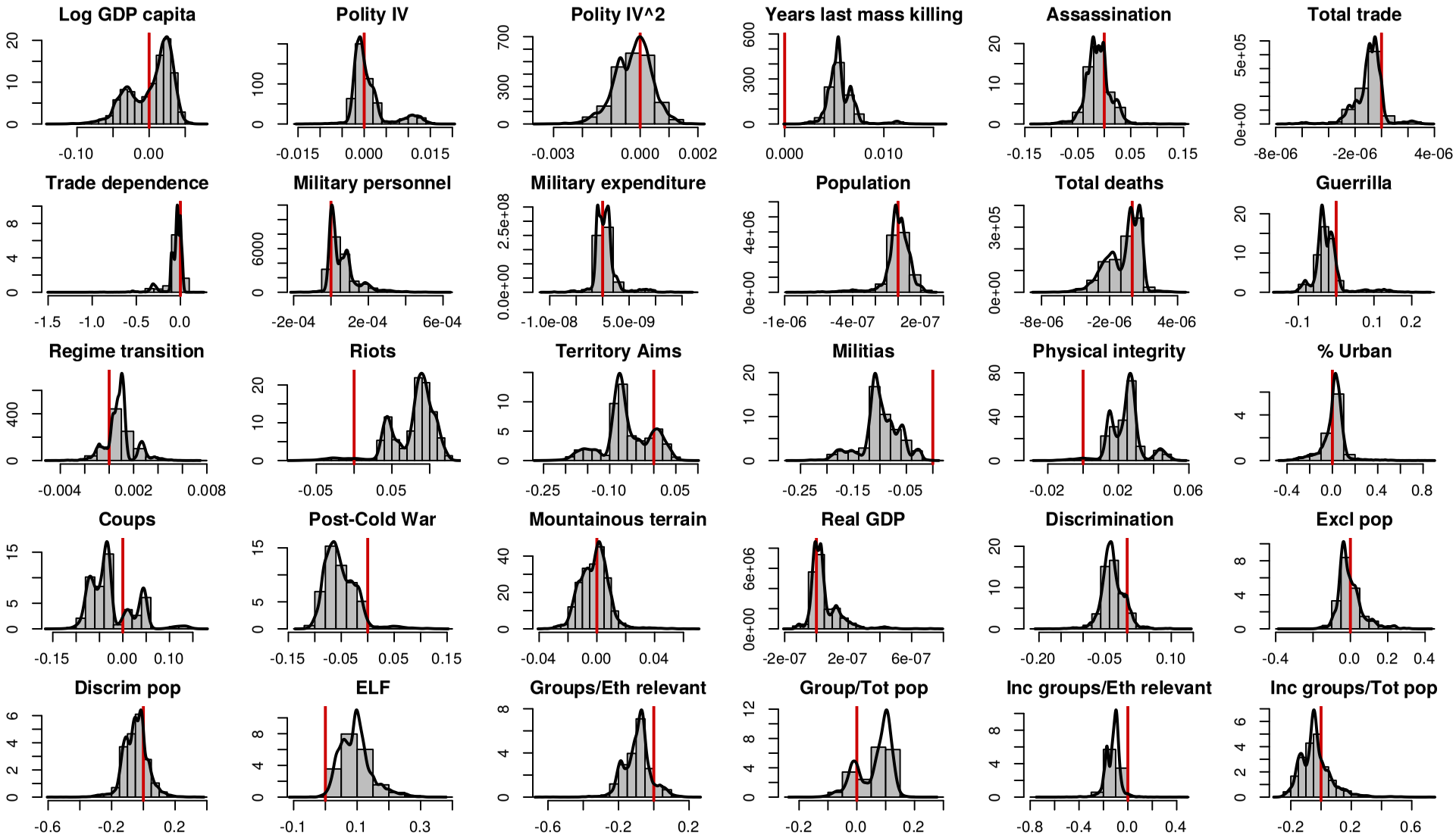

Figure 4.10: EBA – Mass Killings during Civil Wars (COW Data)

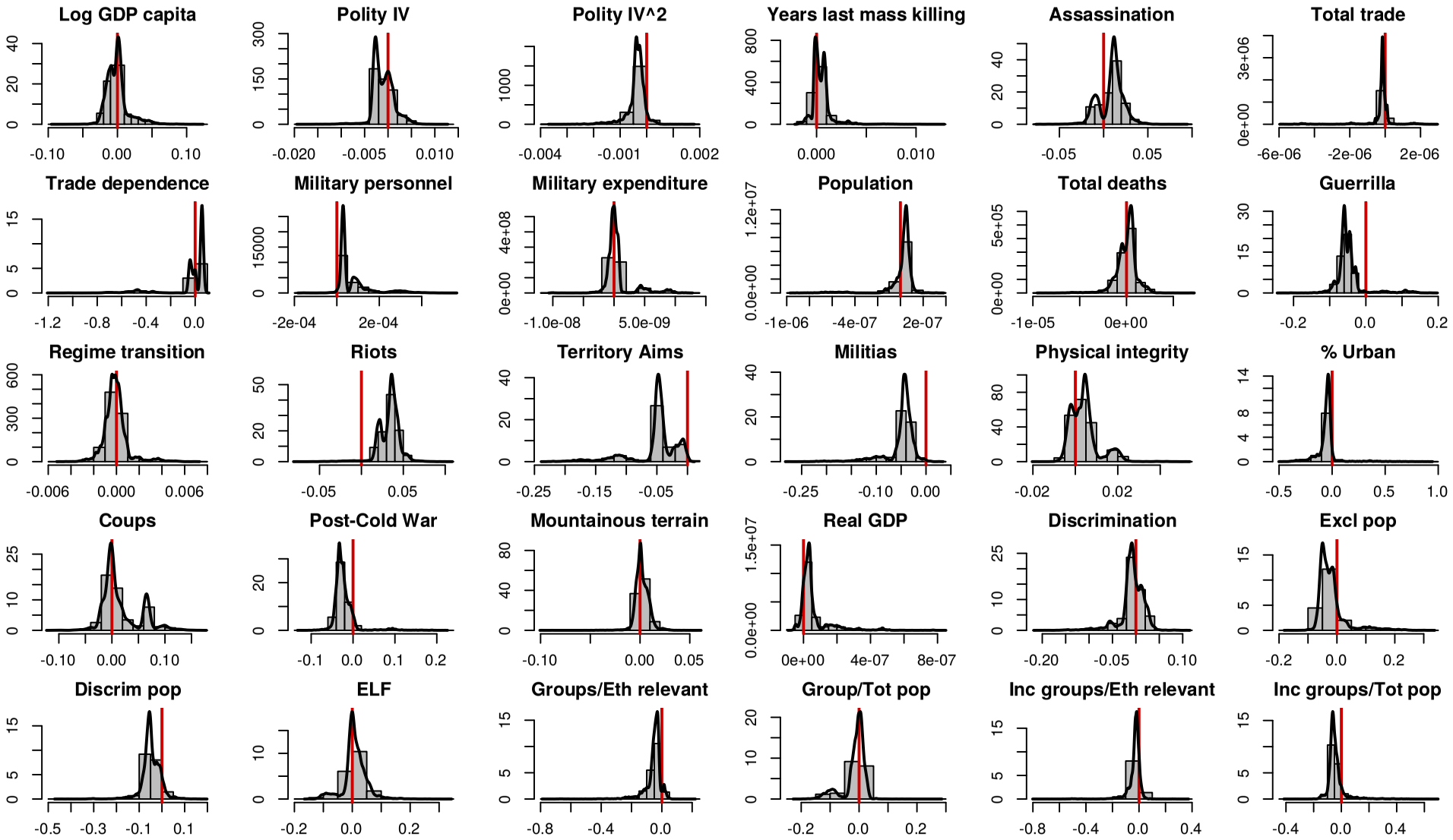

Figure 4.11: EBA – Mass Killings during Ethnic Civil Wars (Cederman et al. Data)

Alternative Number of Variables

The models below are based on 50,000 random draws from the full set of all possible regression models. Sala-i-Martin et al. (2004, 819) argue that random sampling produces unbiased estimates of the regression coefficients with low computational time. The models presented in section 4.3, however, include the full set of possible regressions.

The following table shows the results of an EBA with 3 variable combinations per model. The results are very similar to those reported above.

Variable	**Avg. β**	**Avg. SE**	**% Sig.**	**CDF(0)**	**Models**
Base variables					
Log GDP per capita	0.0082	0.0043	81.439	0.9504	40677
Additional variables					
Post-Cold War years	-0.0121	0.0069	77.804	0.9609	5064
UCDP civil war onset	0.0523	0.0292	62.561	0.9574	3304
Previous riots	0.0134	0.0084	65.936	0.9401	5064
UCDP ongoing civil war	0.0177	0.0094	72.367	0.9372	3304
Polity IV squared	-0.0002	0.0001	66.035	0.9268	5064
Ethnic diversity (ELF)	0.0162	0.0110	70.794	0.9266	5064

Table 4.8: EBA – 3 Variables

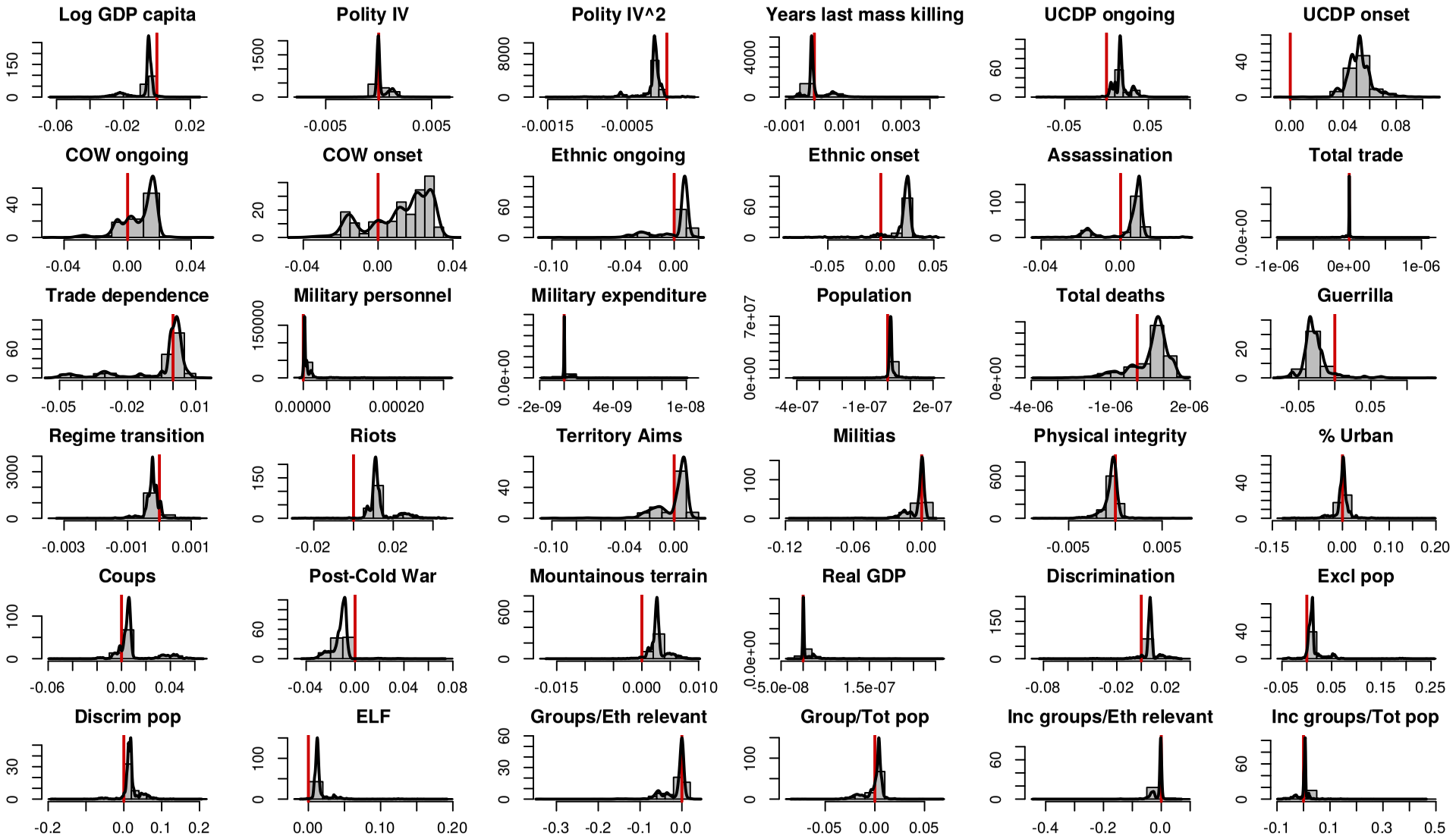

Figure 4.12: EBA – 3 Variables

Table 4.9 presents the results for models with up 5 variables in each regressions. In contrast with the main EBA model, the indicators of UCDP ongoing civil wars, ethnic diversity, and Polity IV square drop out of significance. Their individual CDFs(0) are about 0.88, just marginally below our specified threshold of 0.9.

Variable	**Avg. β**	**Avg. SE**	**% Sig.**	**CDF(0)**	**Models**
Base variables					
Log GDP per capita	-0.010	0.006	70.806	0.9161	50000
Additional variables					
Post-Cold War years	-0.014	0.010	68.496	0.9336	9532
UCDP civil war onset	0.053	0.035	44.784	0.9308	5100
Previous riots	0.015	0.012	47.988	0.9047	9569

Table 4.9: EBA – 5 Variables

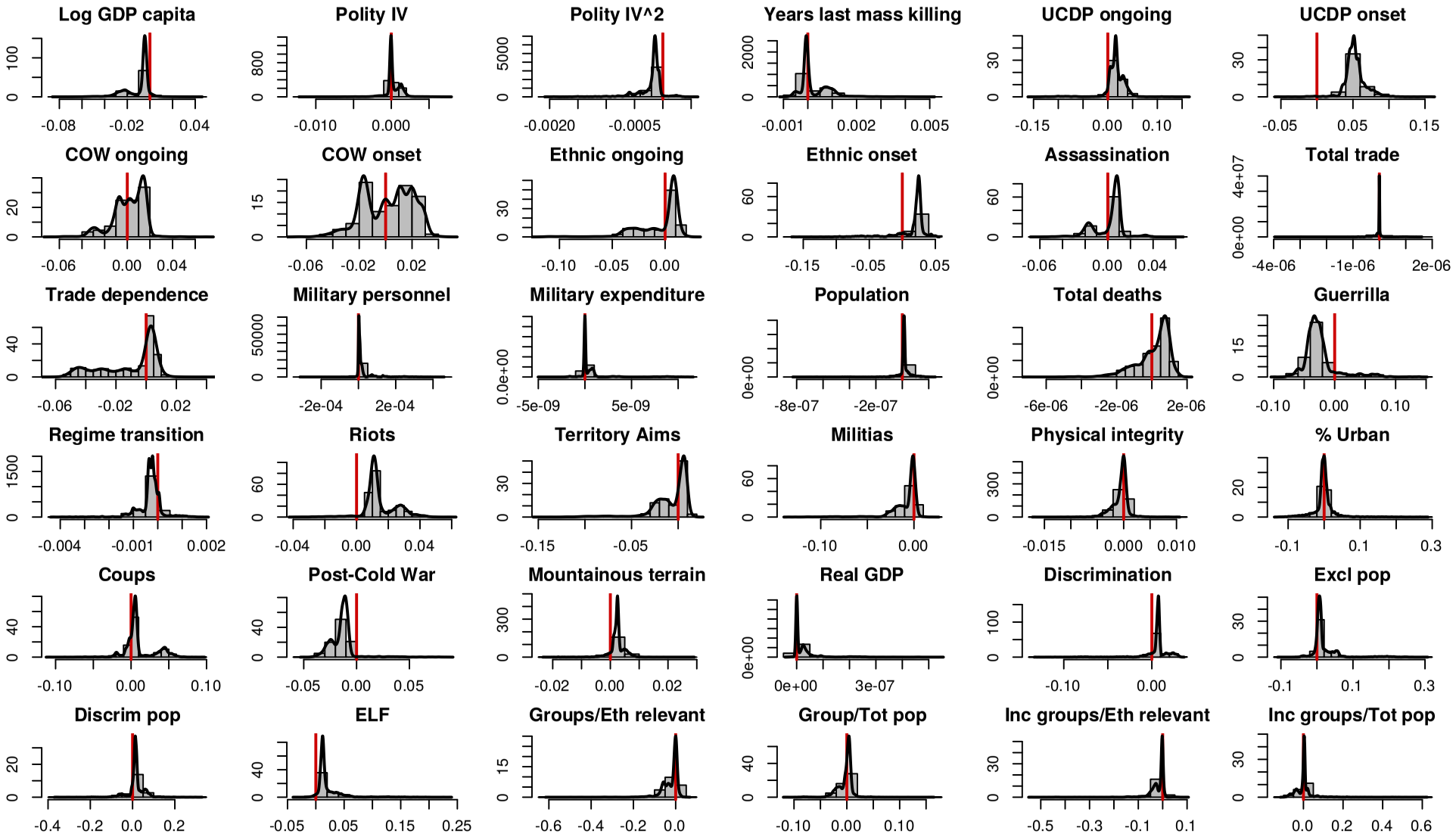

Figure 4.13: EBA – 5 Variables

Alternative Variance Inflation Factors

In this subsection, I estimate EBA models with different values of Variance Inflation Factor (VIF), which is a measure of multicollinearity. There is no standard definition about what constitutes an acceptable VIF value, although researchers often use 10 as rule of thumb to indicate strong multicollinearity (O'Brien 2007, 674). My original model used a slightly more conservative value of 7 as a cutoff. Here, I test the same model with VIF = 10 (less strict), 2.5 (more conservative), and a model without VIF restrictions. The results are essentially identical to those of the main model. In the model with no VIF restriction, however, ethnic fractionalisation fails to meet the threshold by a very small margin. The CDF(0) of that covariate is 0.897, very close to the required value of 0.9.

Variable	**Avg. β**	**Avg. SE**	**% Sig.**	**CDF(0)**	**Models**
Base variables					
Log GDP per capita	-0.0091	0.0052	76.354	0.9343	50000
Additional variables					
Post-Cold War years	-0.0134	0.0084	73.540	0.9495	7929
UCDP civil war onset	0.0529	0.0322	52.141	0.9438	4553
Previous riots	0.0140	0.0100	56.433	0.9216	7772
UCDP ongoing civil war	0.0172	0.0113	66.013	0.9113	4587
Ethnic diversity (ELF)	0.0182	0.0136	56.872	0.9056	8076
Polity IV squared	-0.0002	0.0001	60.791	0.9021	7835

Table 4.10: EBA – VIF 10

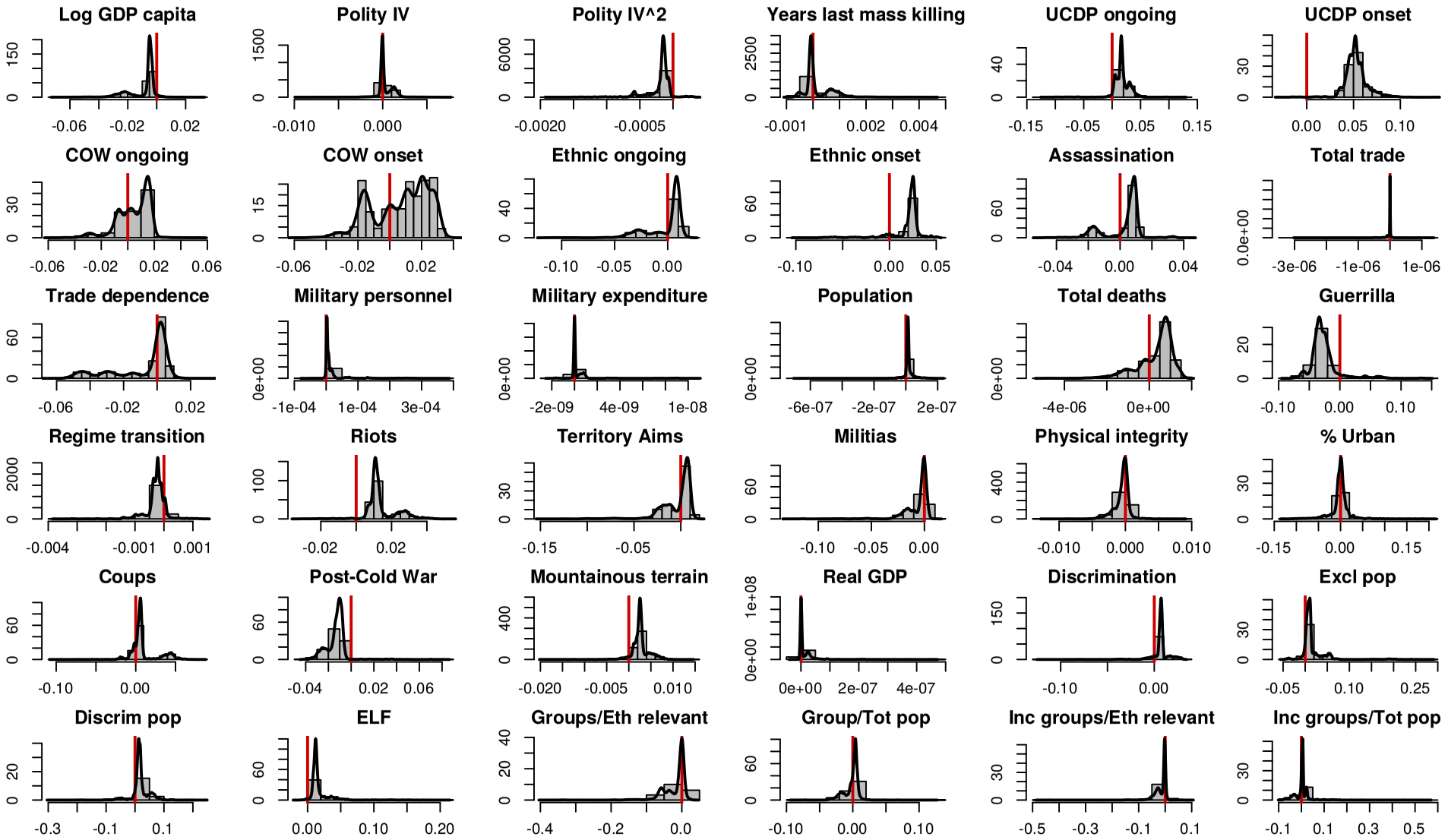

Figure 4.14: EBA – VIF 10

Variable	Avg. β	Avg. SE	% Sig.	CDF(0)	Models
Base variables					
Log GDP per capita	-0.0090	0.0051	76.055	0.9343	49620
Additional variables					
Post-Cold War years	-0.0132	0.0084	72.845	0.9490	7929
UCDP civil war onset	0.0529	0.0322	52.378	0.9438	4553
Previous riots	0.0141	0.0101	56.242	0.9199	7772
UCDP ongoing civil war	0.0174	0.0114	65.652	0.9103	4587
Ethnic diversity (ELF)	0.0184	0.0137	56.674	0.9054	8076
Polity IV squared	-0.0002	0.0001	61.206	0.90267	7835

Table 4.11: EBA – VIF 2.5

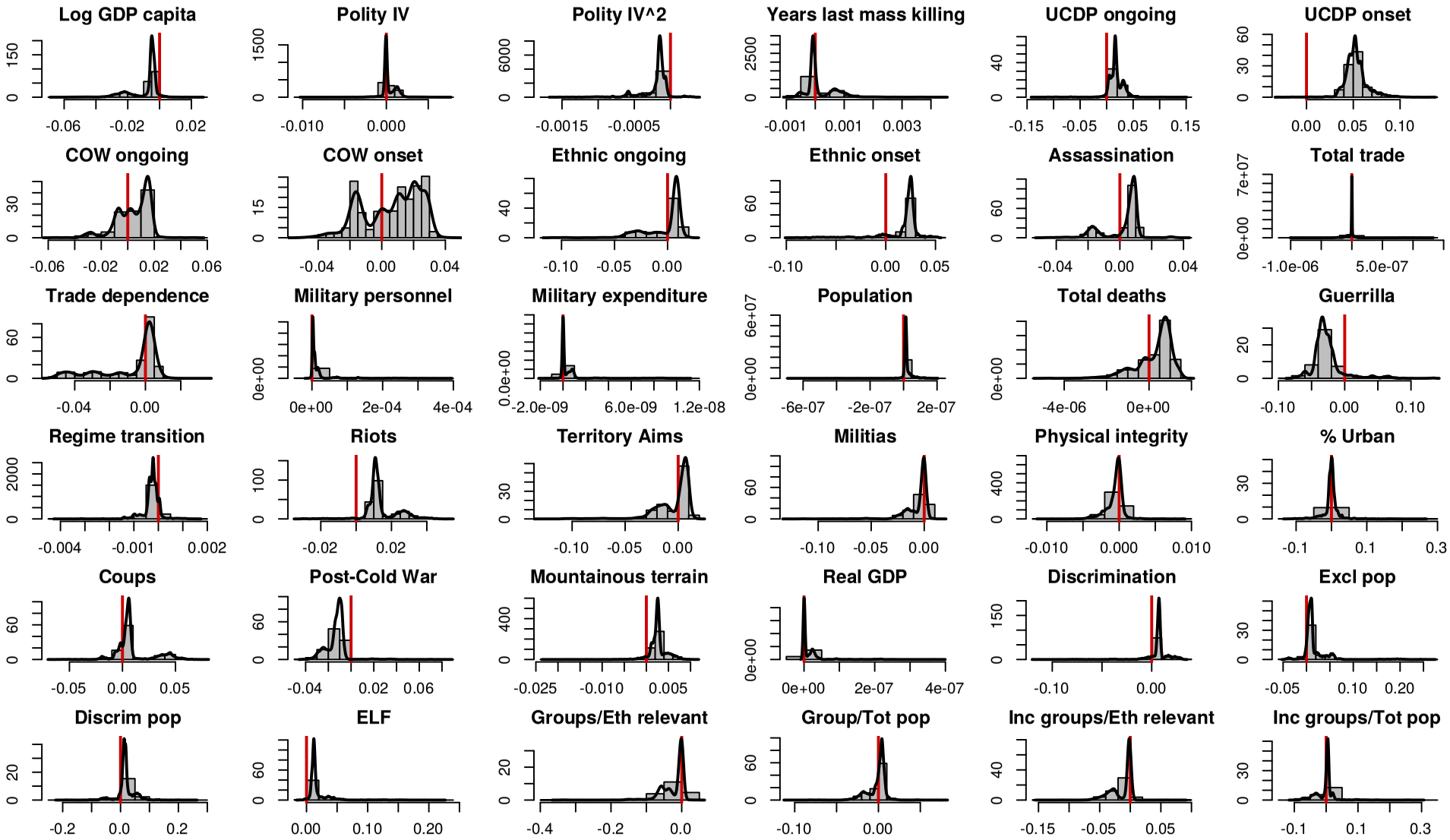

Figure 4.15: EBA – VIF 2.5

Variable	Avg. β	Avg. SE	% Sig.	CDF(0)	Models
Base variables					
Log GDP per capita	-0.0091	0.0052	75.940	0.9343	50000
Additional variables					
Post-Cold War years	-0.0133	0.0085	72.756	0.9469	7800
UCDP civil war onset	0.0531	0.0321	53.068	0.9452	4596
Previous riots	0.0140	0.0101	56.139	0.9200	7811
UCDP ongoing civil war	0.0170	0.0116	64.487	0.9057	4497
Ethnic diversity (ELF)	0.0184	0.0137	56.814	0.9056	7808
Polity IV squared	-0.0002	0.0001	60.825	0.9009	7903

Table 4.12: EBA – No VIF Restriction

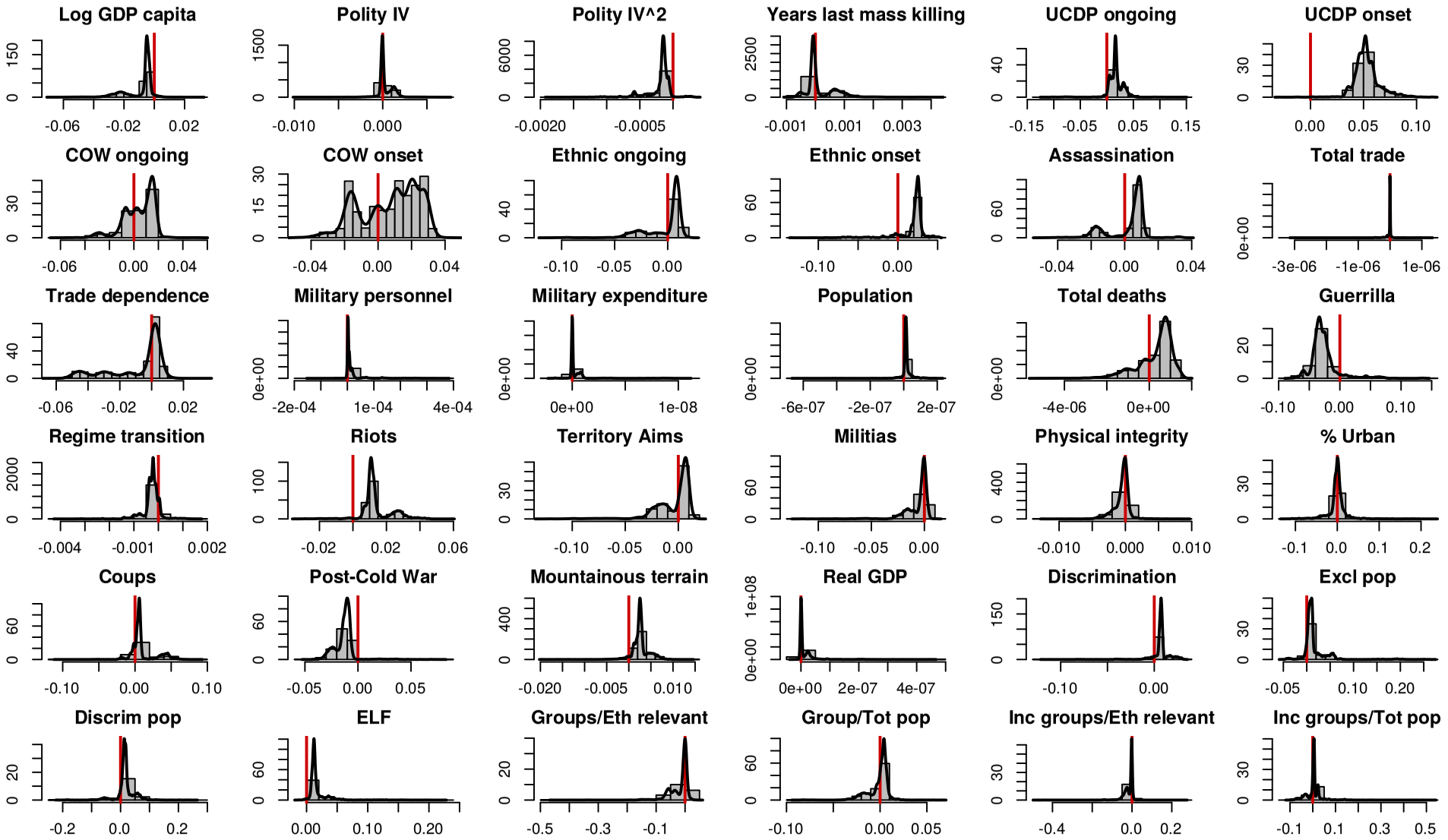

Figure 4.16: EBA – No VIF restriction

Generalised Linear Models

I reestimate the main EBA model with logit and probit models. Nevertheless, logistic and probit regressions may have issues of complete separation, that is, some covariates may perfectly separate zeros and ones in the outcome variable. In that case, the estimations fail to converge. We address this problem by adding a weak prior to the regression coefficients as suggested by Gelman et al. (2008).[9] First, we scaled the non-binary variables to have a mean of 0 and a standard deviation of 0.5, then added a Cauchy distribution with centre 0 and scale 2.5. The probit regressions use a scale of 2.5×1.6, which is also recommended by the authors (Gelman and Su 2016). Ethnic diversity and ongoing civil wars come close to meeting our threshold values (0.88 and 0.84, respectively), and civil war onset (UCDP) has a higher percentage of significant coefficients and a high CDF(0) area than in the linear probability models.

Variable	**Avg. β**	**Avg. SE**	**% Sig.**	**CDF(0)**	**Models**
Base variables					
Log GDP per capita	0.434	0.223	75.570	0.9267	50000
Additional variables					
UCDP civil war onset	1.308	0.530	87.261	0.9742	4506
Post-Cold War years	-0.911	0.428	70.456	0.9448	7890
Previous riots	0.744	0.38	66.778	0.9383	7805
Polity IV squared	-0.015	0.008	68.038	0.9285	7975

Table 4.13: EBA – Logistic Regression

[9] I thank Mark Bell for sharing R code to estimate penalised-likelihood models.

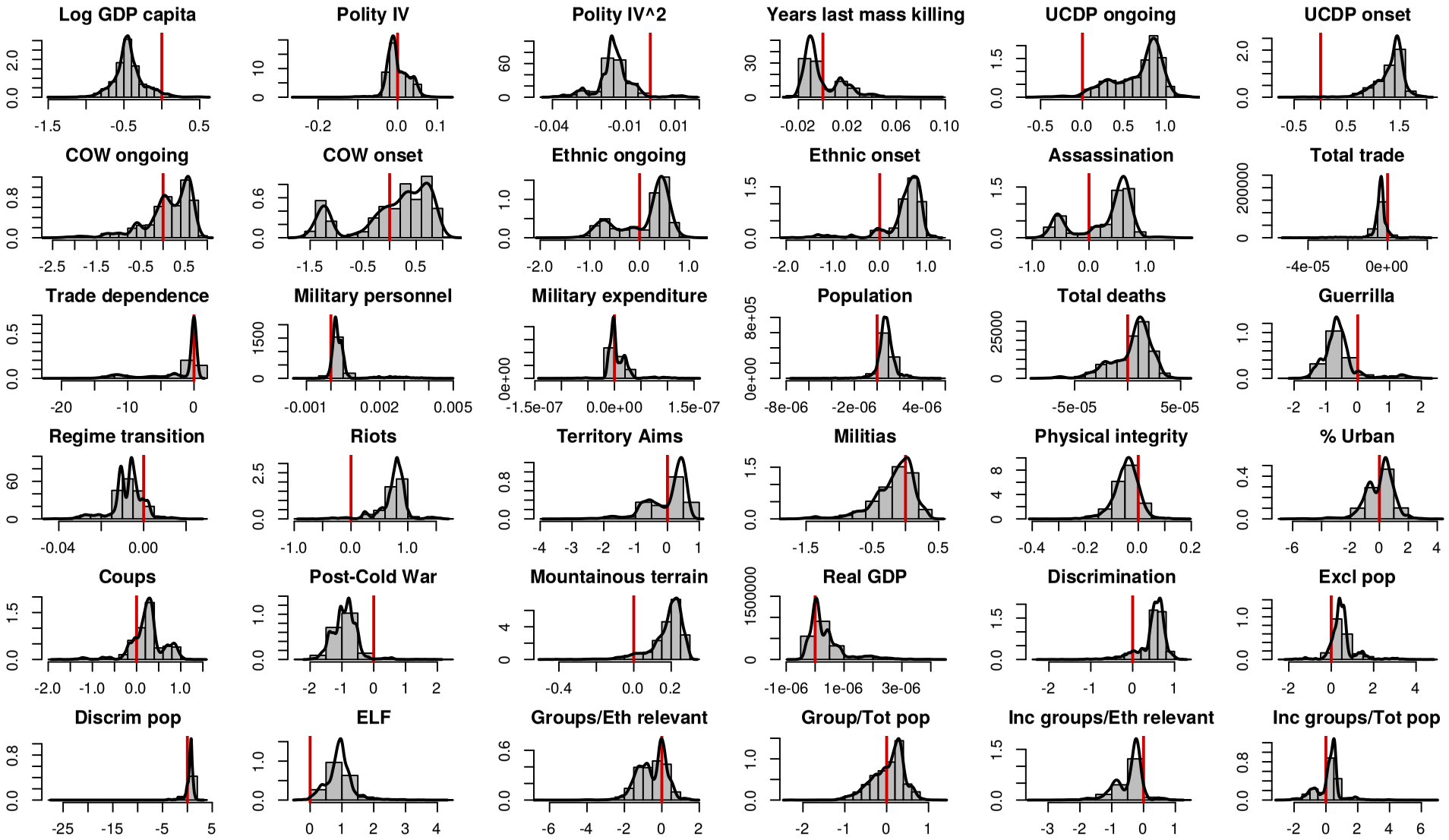

Figure 4.17: EBA – Logistic Regression

Variable	Avg. β	Avg. SE	% Sig.	CDF(0)	Models
Base variables					
Log GDP per capita	-0.1924	0.1031	76.118	0.9258	50000
Additional variables					
UCDP civil war onset	0.6422	0.2582	89.225	0.9772	4501
Previous riots	0.3367	0.1743	71.813	0.9436	7851
Post-Cold War years	-0.3709	0.1830	71.465	0.9404	7836
Polity IV squared	-0.0061	0.0032	70.155	0.9315	7931

Table 4.14: EBA – Probit Regression

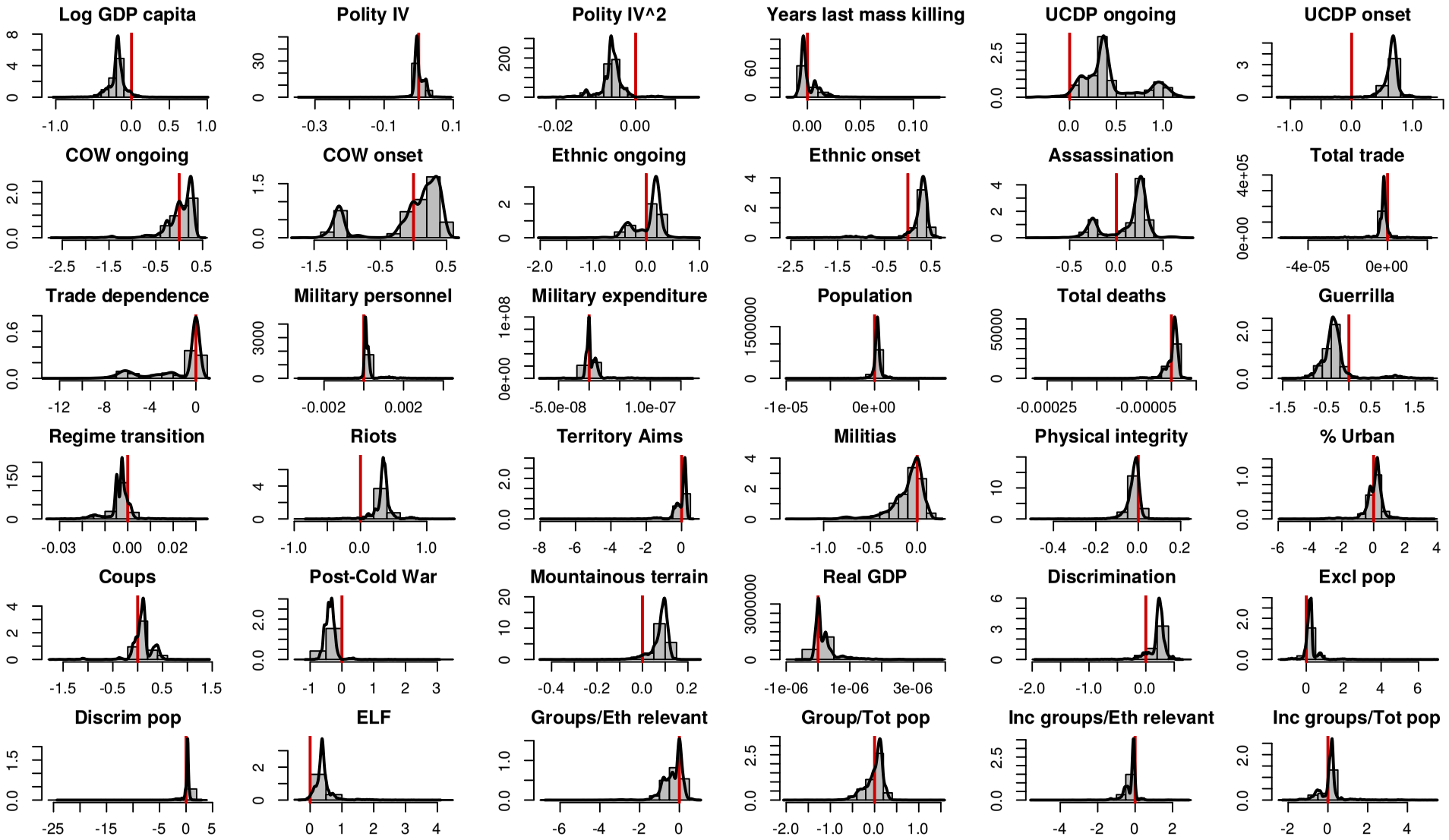

Figure 4.18: EBA – Probit Regression

Mass Killings in and after the Cold War

I also test the heterogeneity of the findings by running one model including only the Cold War years (1945–1991) and another with the post-Cold War period (1991–2013). The results vary in both periods, and there is no overlap between significant variables.

Variable	**Avg. β**	**Avg. SE**	**% Sig.**	**CDF(0)**	**Models**
Cold War Period					
Log GDP per capita	-0.018	0.009	83.204	0.9678	50000
Previous riots	0.022	0.014	62.457	0.9031	8278
Post-Cold War Period					
Ethnic war onset	-0.024	0.011	89.608	0.9823	4850
Coup d'état	-0.022	0.011	89.200	0.9822	8602
Territory aims	-0.027	0.014	81.083	0.9653	8775
Displaced Population	-0.048	0.027	58.689	0.9392	8695

Table 4.15: EBA – Mass Killings in and after the Cold War Period (Robust Variables Only)

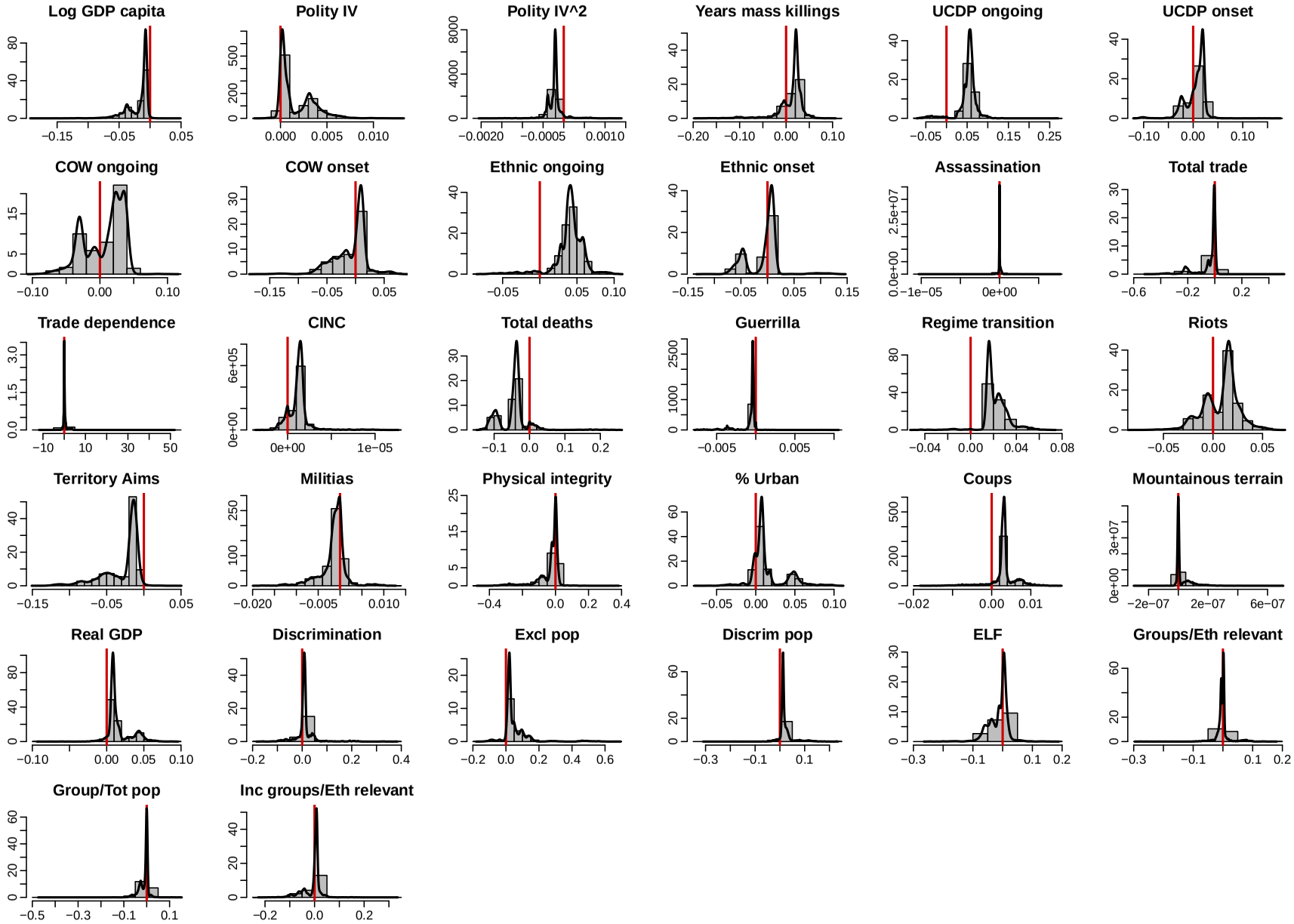

Figure 4.19: EBA – Cold War Period

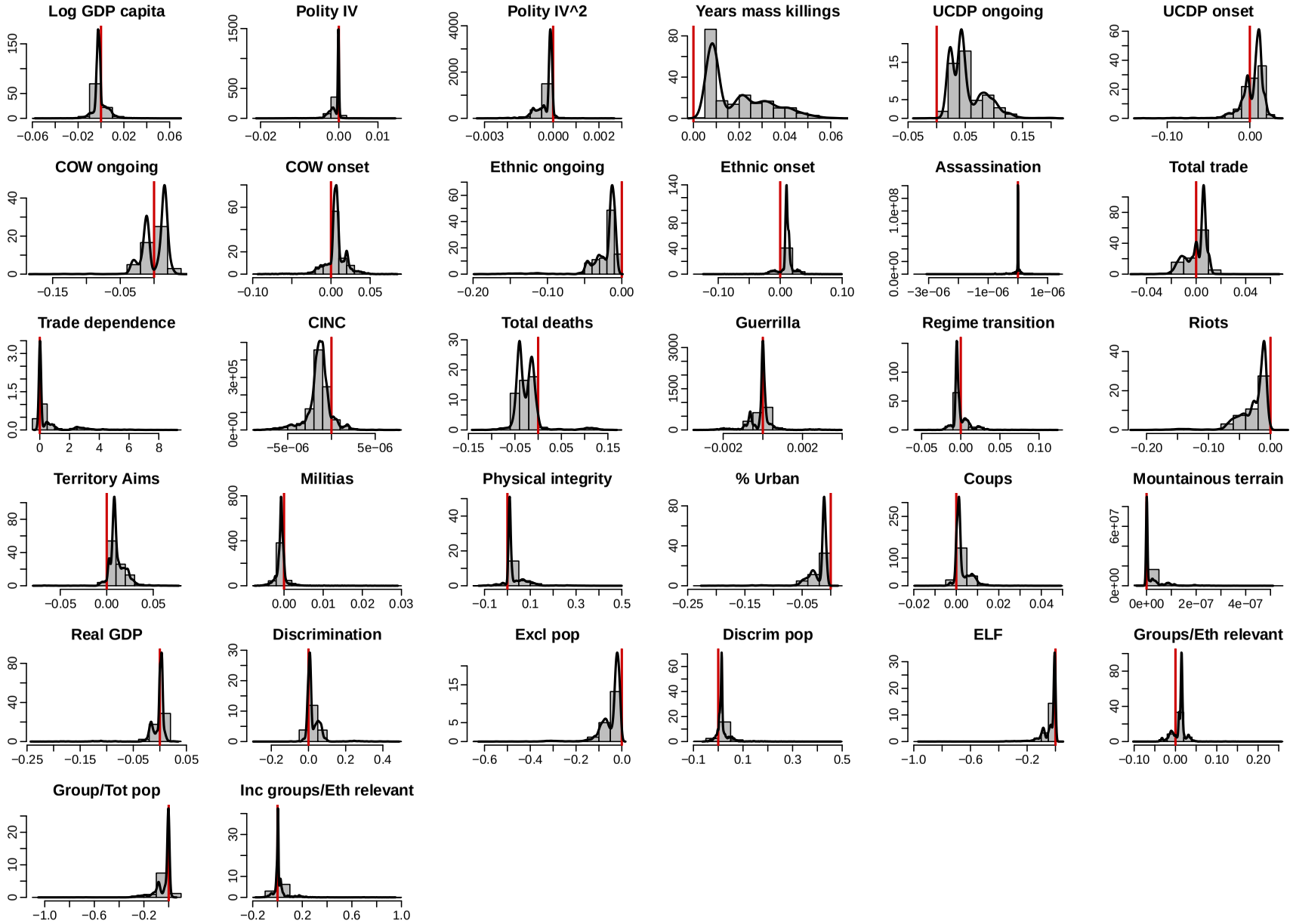

Figure 4.20: EBA – Post-Cold War Period

Mass Killings during Peacetime

I have also tested whether the main EBA findings differ if the sample is restricted to peace years. That is, all observations denoted as 1 in the three conflict indicators mentioned above (UCDP, COW, Cederman et al.) were removed from the dataset. The results are similar to the main model, yet "wars fought over territory" was removed from the EBA due to multicollinearity issues. Moreover, two other variables (total battle deaths and presence of guerrillas) have very small sample sizes and their estimates were should not be interpreted as reliable. The significant variables are presented below.

Variable	**Avg. β**	**Avg. SE**	**% Sig.**	**CDF(0)**	**Models**
Base variables					
Log GDP per capita	-0.004	0.002	86.608	0.9748	26620
Additional variables					
Post Cold War	-0.011	0.003	100	0.9984	5459
Polity IV squared	-1.63e-04	6.27e-05	98.309	0.9914	5441
Discriminated pop	0.009	0.005	92.516	0.9750	5438
Mountainous terrain	0.002	0.001	74.171	0.9718	5428
Population	-9.89e-09	6.20e-09	61.776	0.9280	5418
Previous riots	0.008	0.006	28.315	0.9119	5400

Table 4.16: EBA – Peace Years

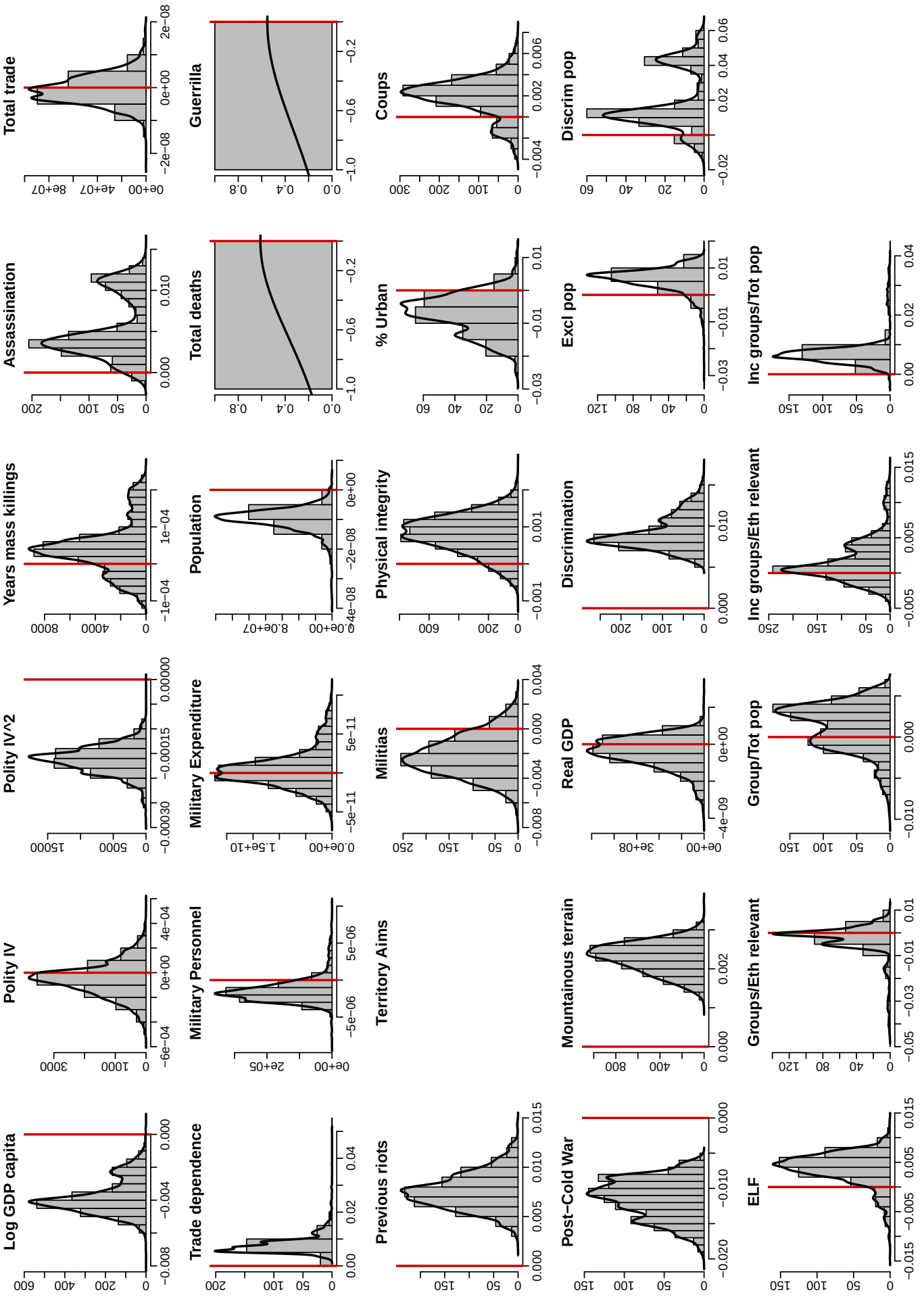

Figure 4.21: EBA – Peace Years

4.6.4 Harff's Genocides and Politicides Data

Main Model

In this section, we evaluate the models presented above with a measure of genocide and politicide by Harff (2003). The results show important contrasts with the previous analyses. First, no variable appear as significant in the main extreme bounds analysis. That is, none of the 36 predictors reached the threshold of CDF(0) > 0.9. Thus, we do not present a table with the results. The variable that came closest to significance was a dummy indicator of coups d'état, which has a CDF(0) of 0.897 and, as expected, is positively correlated with the onset of genocides. The distribution of the covariates' coefficients are available in figure 4.22.

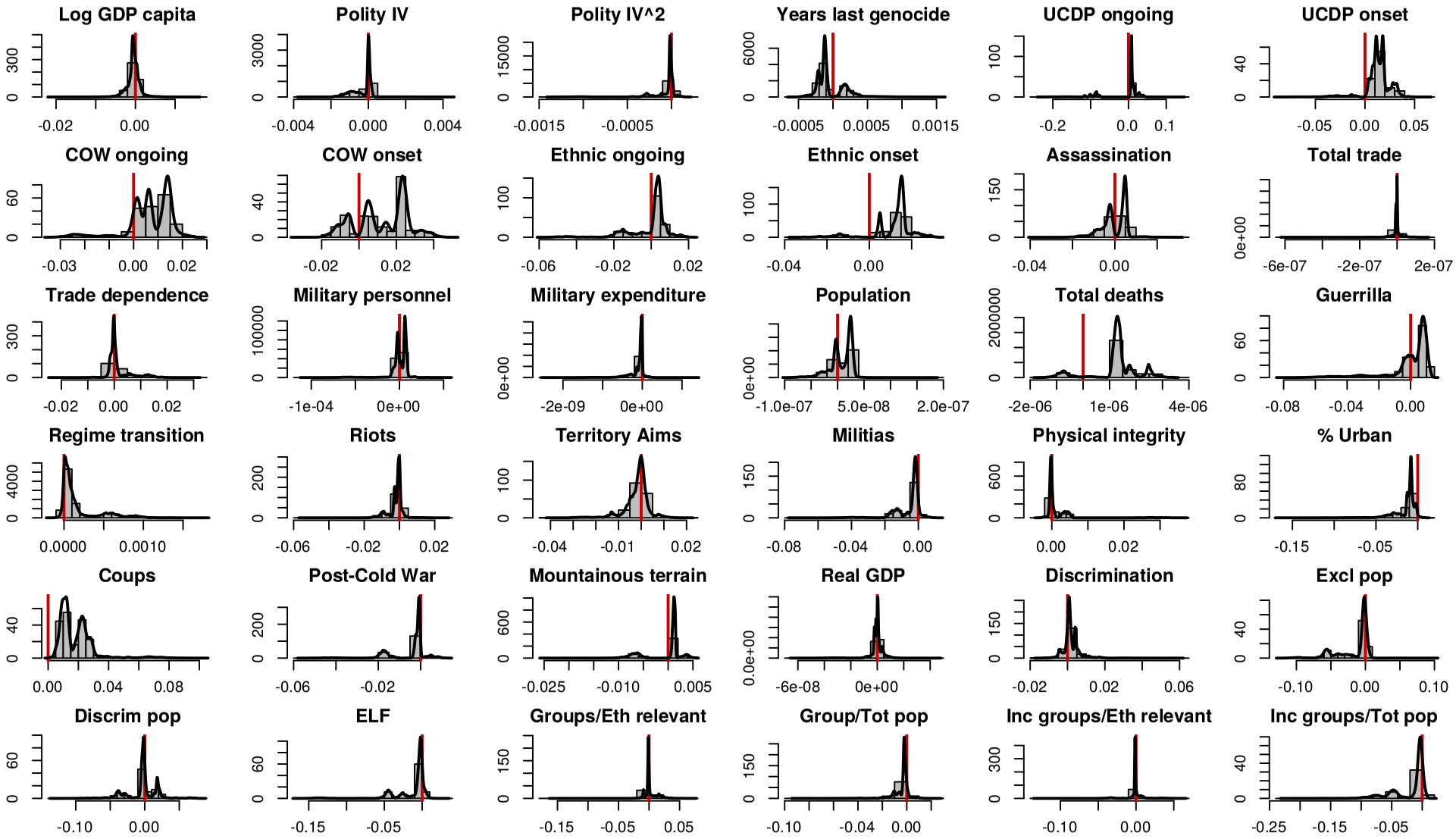

Figure 4.22: EBA – Genocides and Politicides

Genocides and Politicides during Civil Wars

Next, we evaluate what covariates are robust when considering only genocides and politicide that occur during civil conflicts. Post-Cold War years again appear as a significant variable and with a negative sign; excluded population also has a negative impact on the outcome variable in two analyses.

Variable	**Avg. β**	**Avg. SE**	**% Sig.**	**CDF(0)**	**Models**
UCDP data					
Excluded population	-0.037	0.022	64.524	0.9176	8758
COW data					
Excluded population	-0.057	0.031	65.703	0.9570	8820
Discriminated population	-0.050	0.029	53.850	0.9367	8767
Post-Cold War years	-0.019	0.013	42.531	0.9203	8904
Cederman et al. data					
Assassination dummy	-0.009	0.006	47.723	0.9232	8828

Table 4.17: EBA – Genocides/Politicides

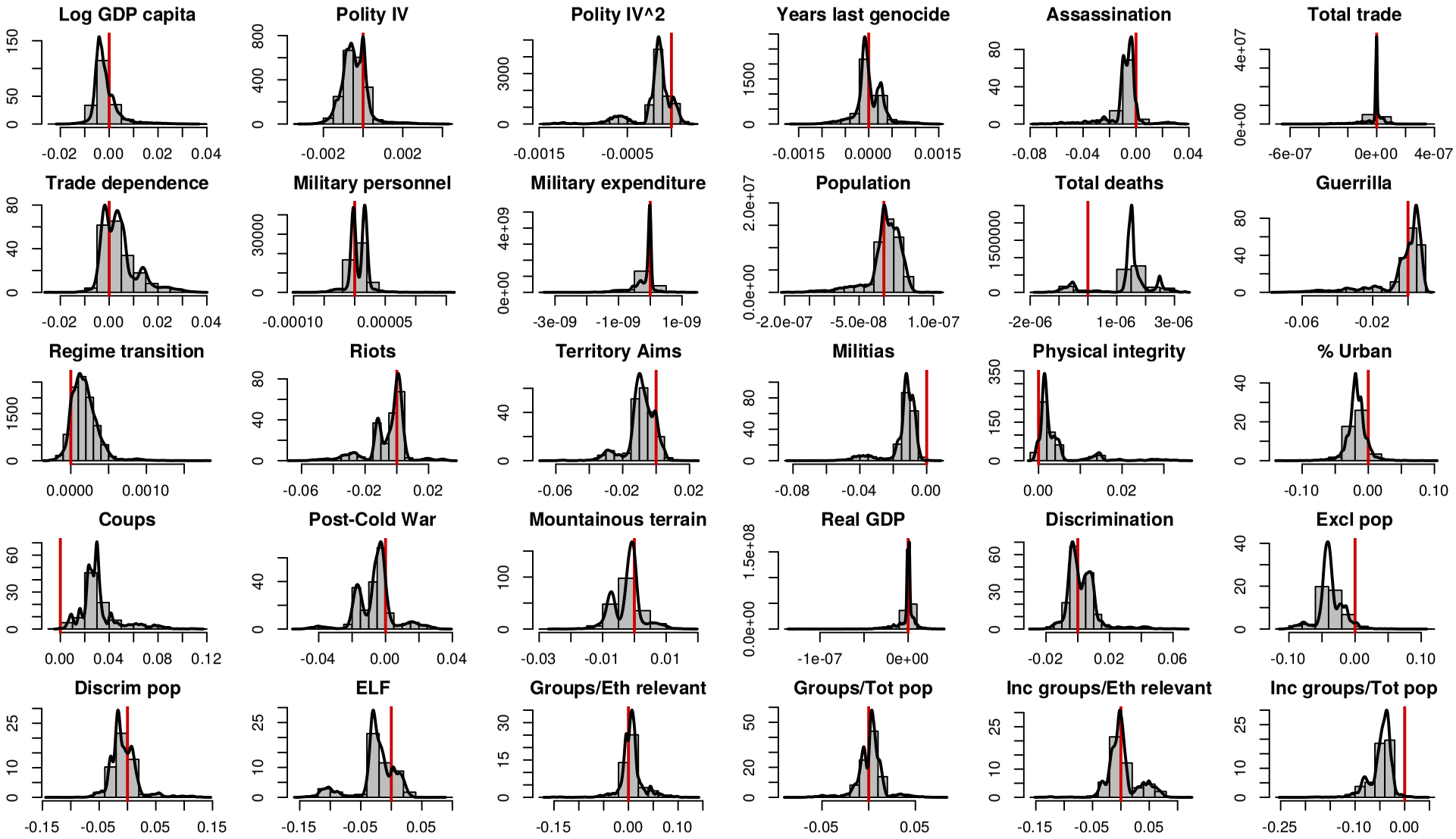

Figure 4.23: EBA – Genocides and Politicides during Civil Wars (UCDP Data)

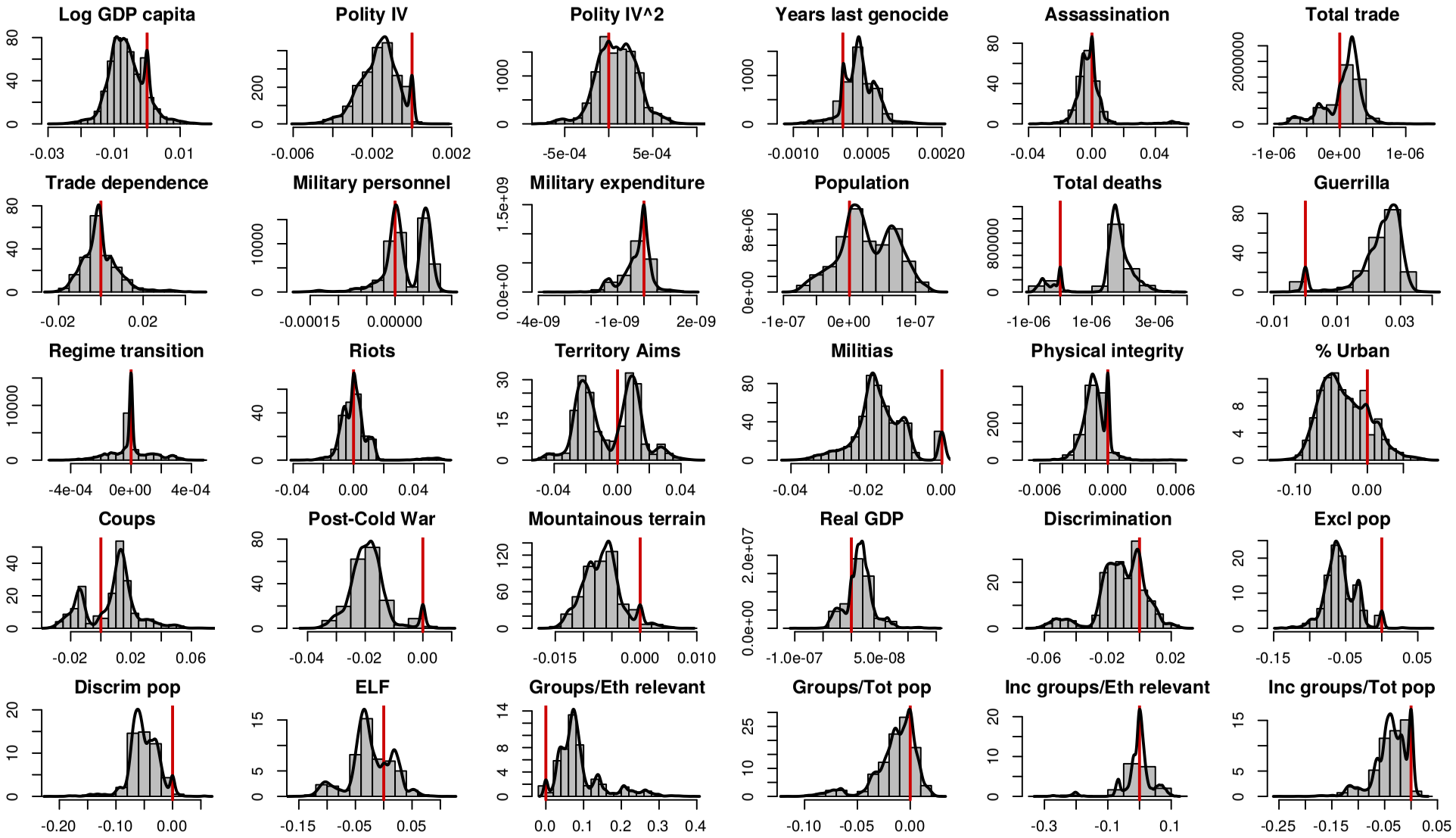

Figure 4.24: EBA – Genocides and Politicides during Civil Wars (COW Data)

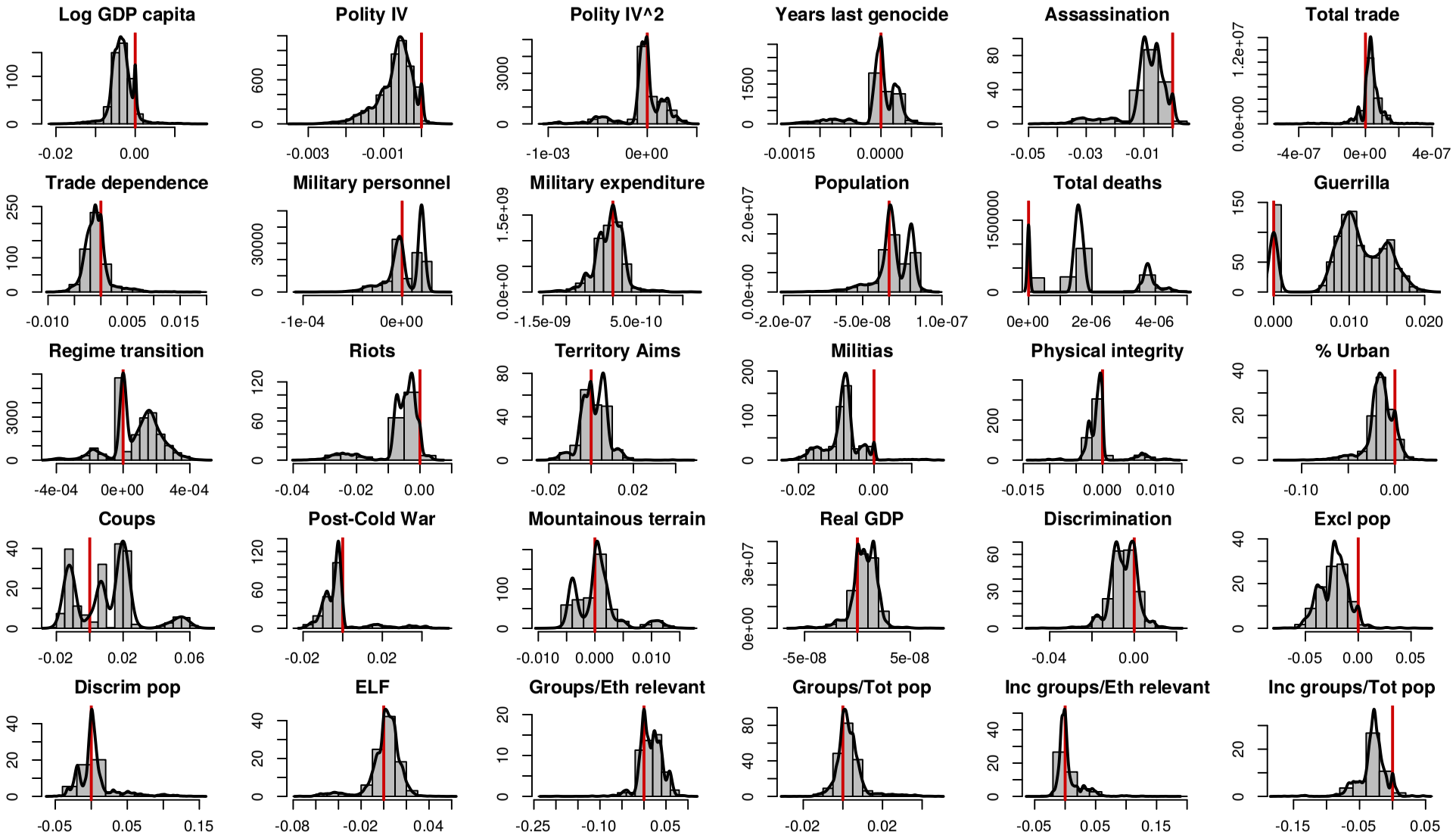

Figure 4.25: EBA – Genocides and Politicides during Ethnic Civil Wars (Cederman et al. Data)

4.6.5 Random Forest

Main Model

We employed the `H2O` machine learning platform (The H2O.ai Team 2017) to estimate the models. `H2O` is open-source, optimised for big data and estimates a large number of models with only a few lines of code. We run the algorithms on 75% of our dataset, and use the remaining 25% as a validation set. That is, we use a percentage of the data to assess the main model's accuracy.[10] Our measure of accuracy is the area under the curve (AUC). All models score well in that regard, and measures of about 0.8 accuracy in our validation sample are common.

The next two plots show the results of the main random forest models.

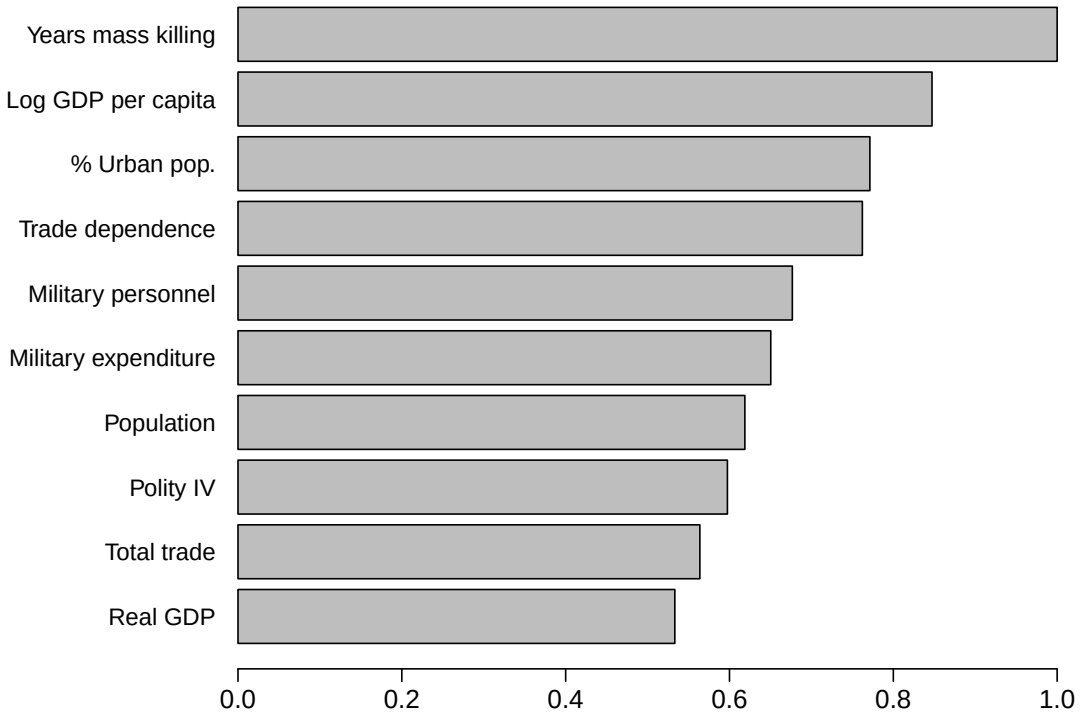

Figure 4.26: Variable Importance – Main Model

[10] For more information about training and validation samples, please refer to http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/algo-params/validation_frame.html.

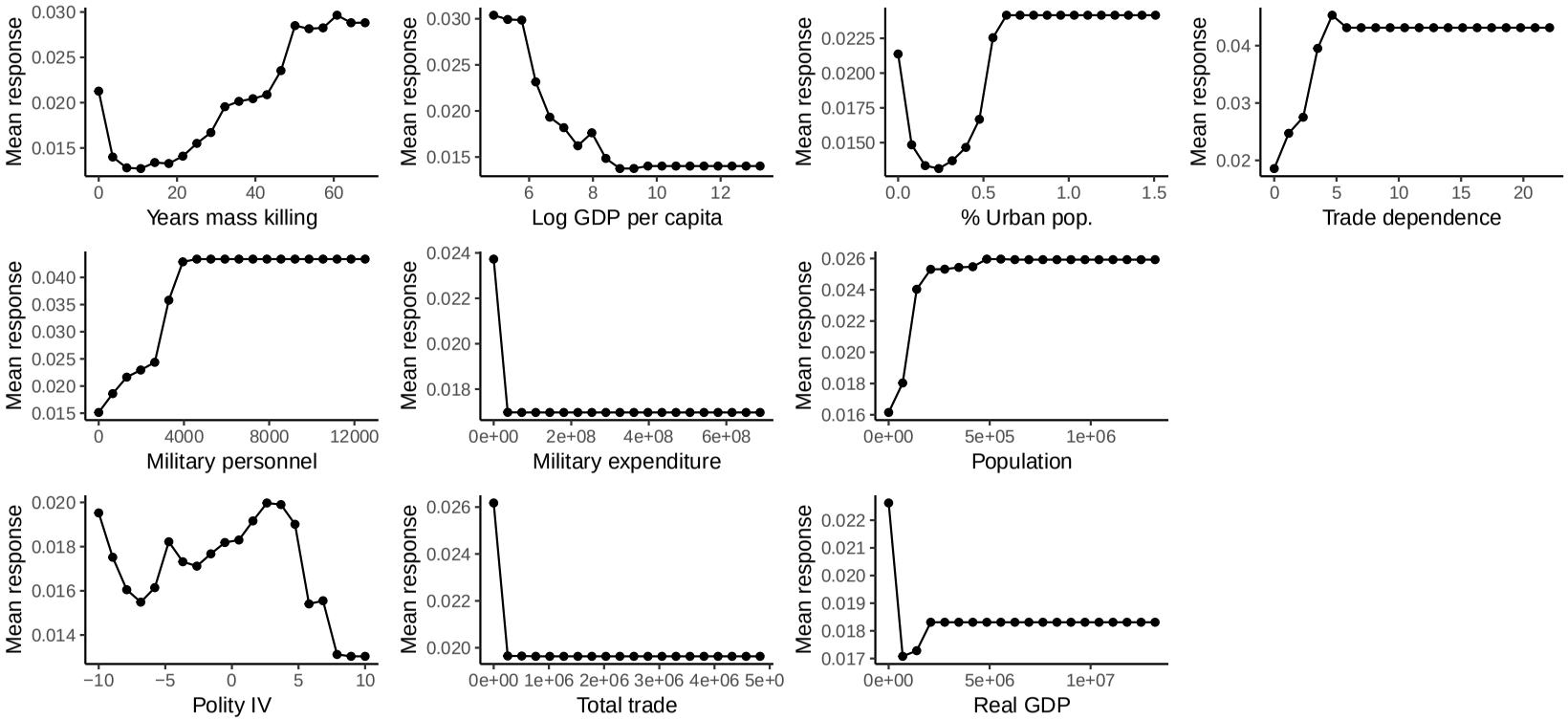

Figure 4.27: Partial Dependence Plot – Main Model

Mass Killings During Civil Wars

The following graphs display the most important predictors of mass killings when we restrict our sample to cases that occur during civil wars. As we note in section 4.6.3, we employ three different measures of civil conflicts. The first one is provided by the Uppsala Conflict Data Program (Allansson et al. 2017; Gleditsch et al. 2002), the second is offered by the Correlates of War (Sarkees and Wayman 2010), and a third indicating the onset of ethnic conflict as coded by Cederman et al. (2010).

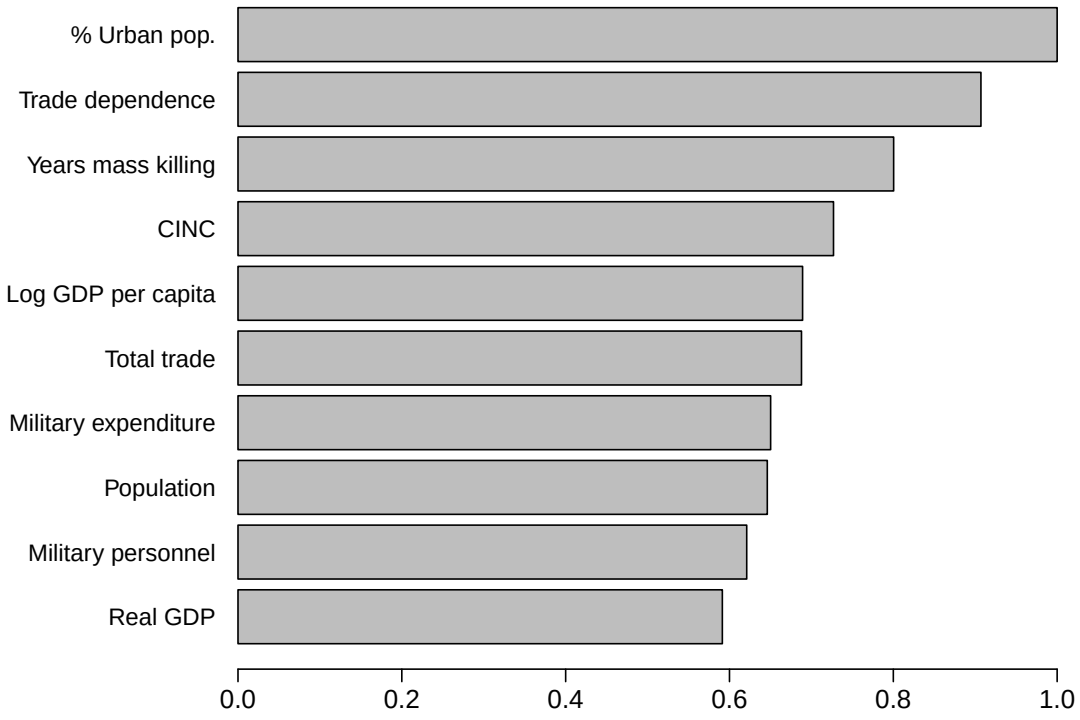

Figure 4.28: Variable Importance – Mass Killings during Civil Wars (UCDP Data)

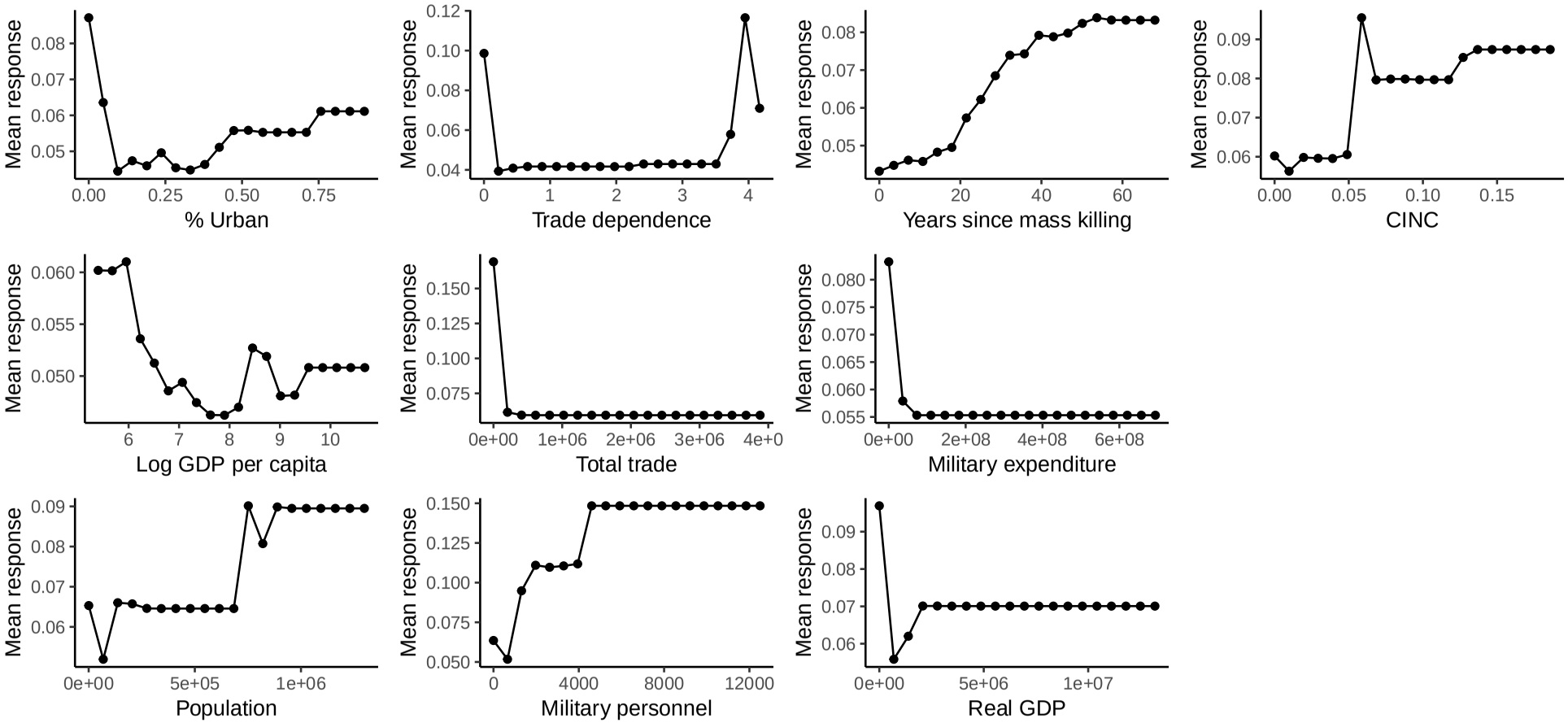

Figure 4.29: Partial Dependence Plot – Mass Killings during Civil Wars (UCDP Data)

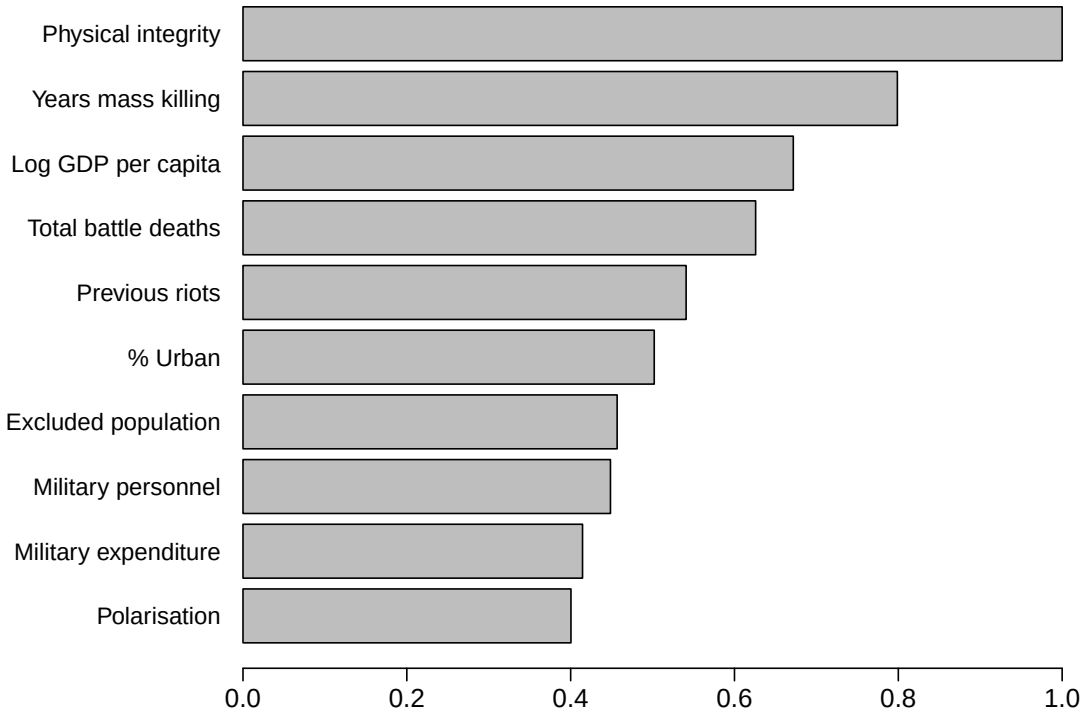

Figure 4.30: Variable Importance – Mass Killings during Civil Wars (COW Data)

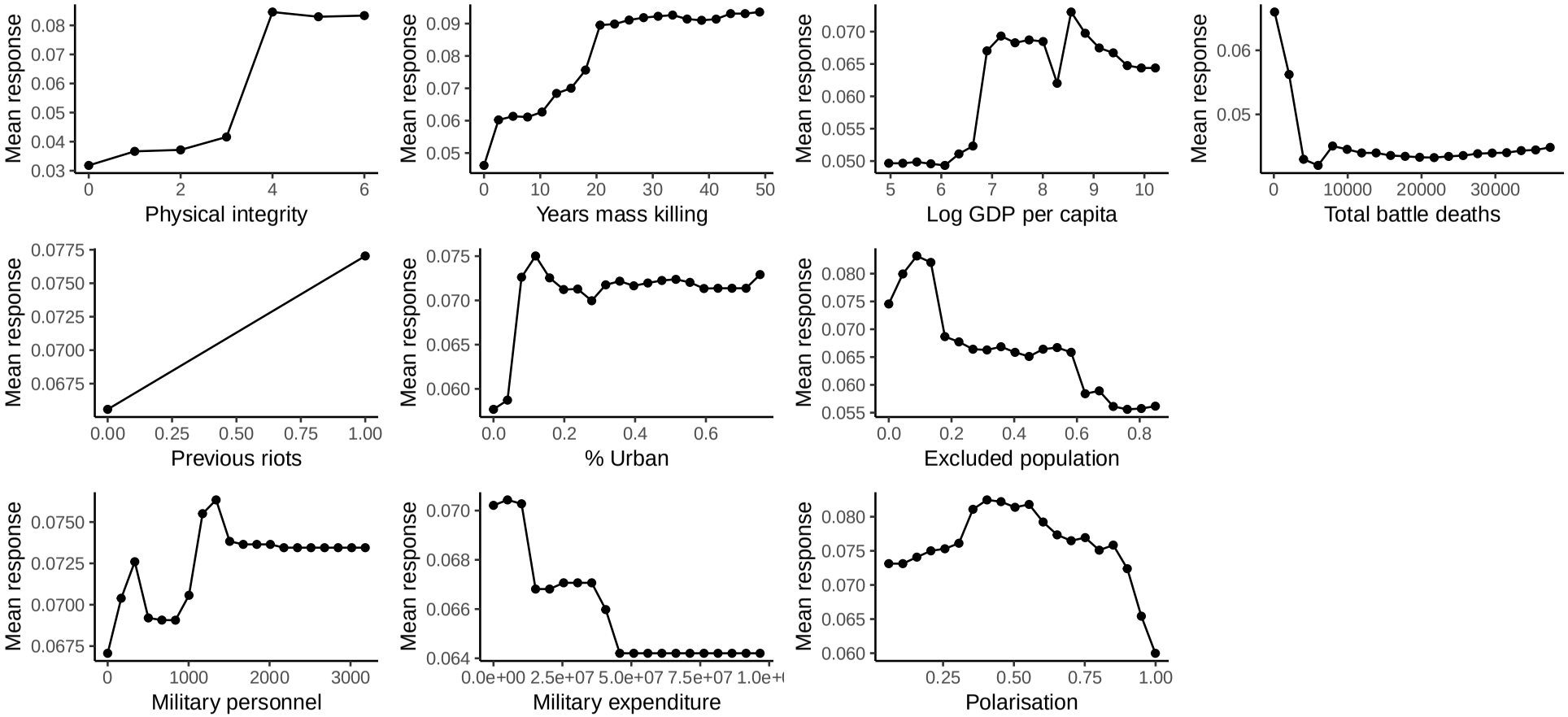

Figure 4.31: Partial Dependence Plot – Mass Killings during Civil Wars (COW Data)

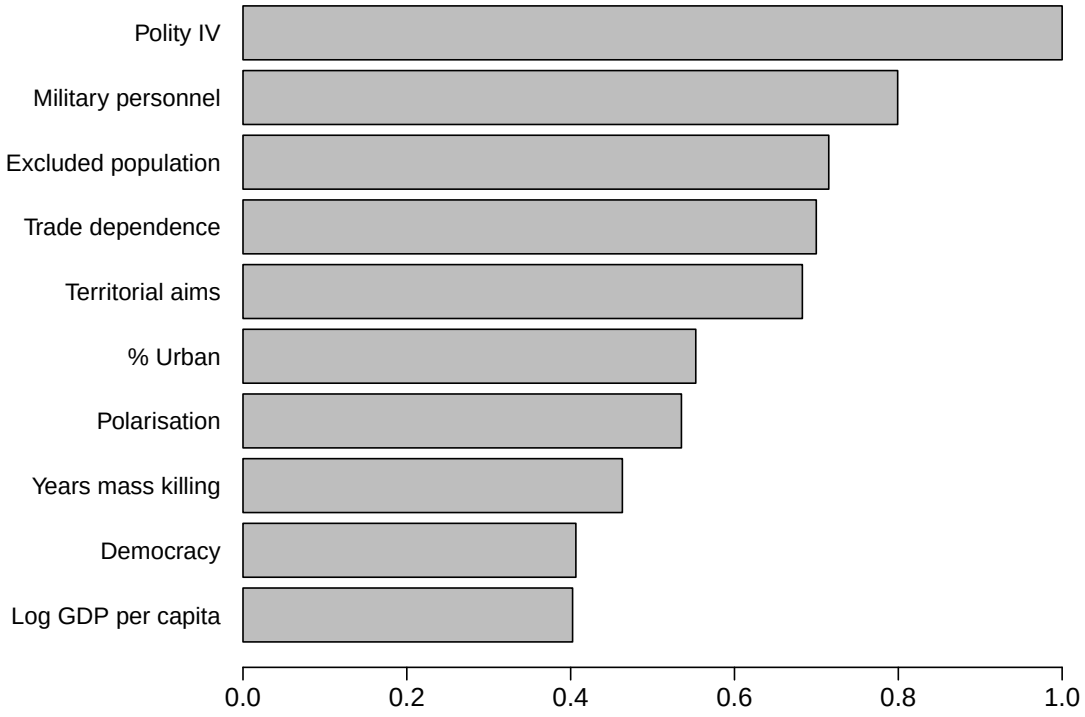

Figure 4.32: Variable Importance – Mass Killings during Ethnic Civil Wars (Cederman et al. Data)

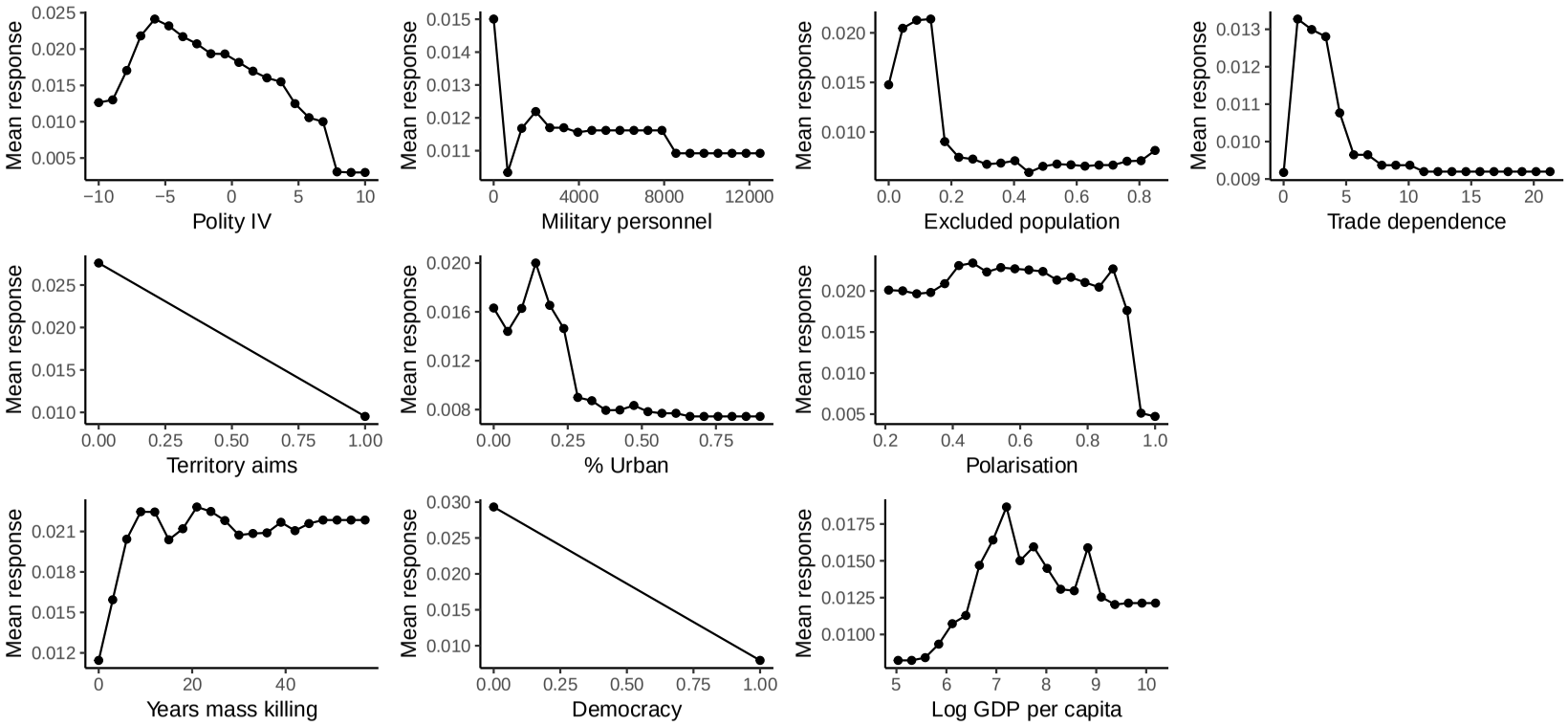

Figure 4.33: Partial Dependence Plot – Mass Killings during Ethnic Civil Wars (Cederman et al. Data)

Alternative Random Seeds

As random forests themselves are an approximation to a number of possible parameter combinations, changes in seed numbers may influence the model output. Thus, we start the main model with two different random seed numbers to check if the results are robust.[11] The main findings hold well; although variable importance changes from one model to another, the most significant variables appear repeatedly in the estimations. The marginal plots also show that the effect of the independent variables remain roughly similar despite the nonlinearities. The graphs below display the ten most significant predictors of mass killings and their respective partial dependence plots.

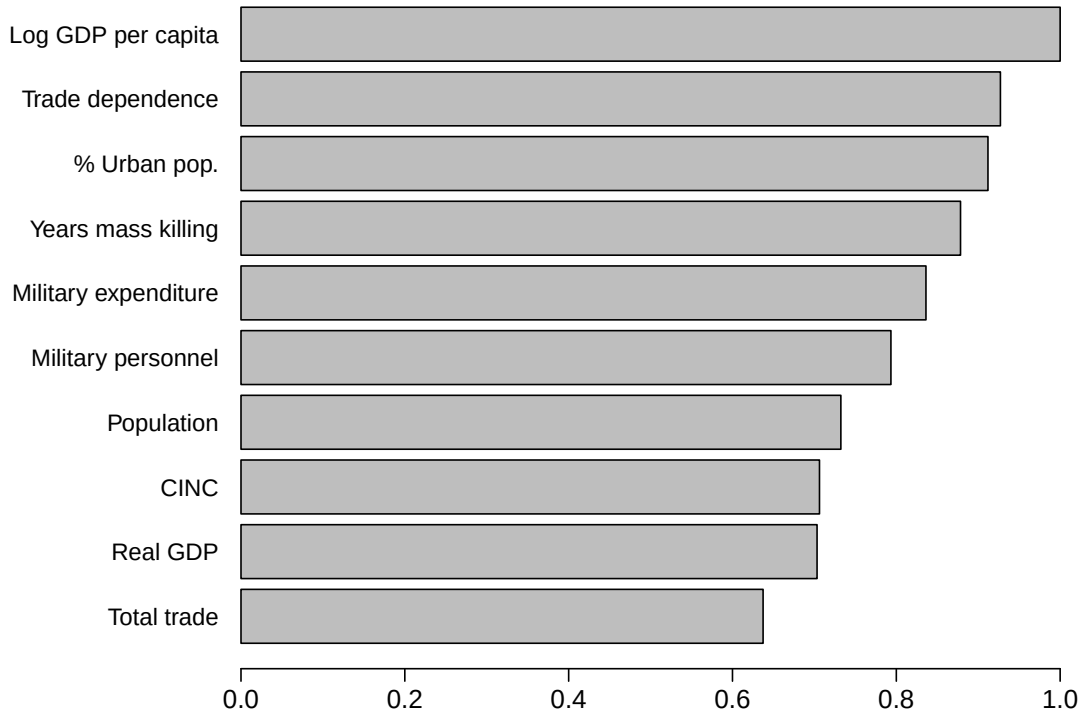

Figure 4.34: Variable Importance – Seed 4363

[11] The numbers were generated at https://www.random.org/.

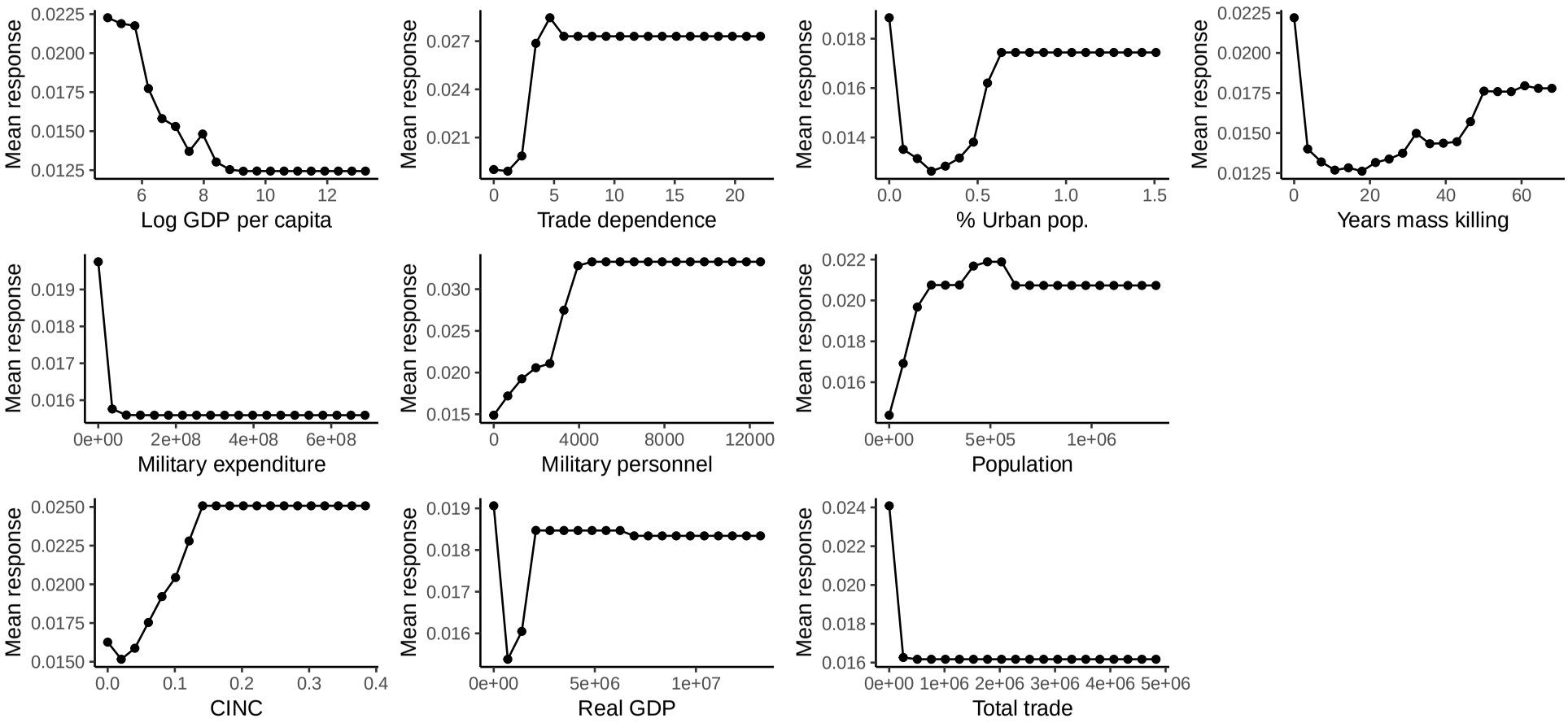

Figure 4.35: Partial Dependence Plot – Seed 4363

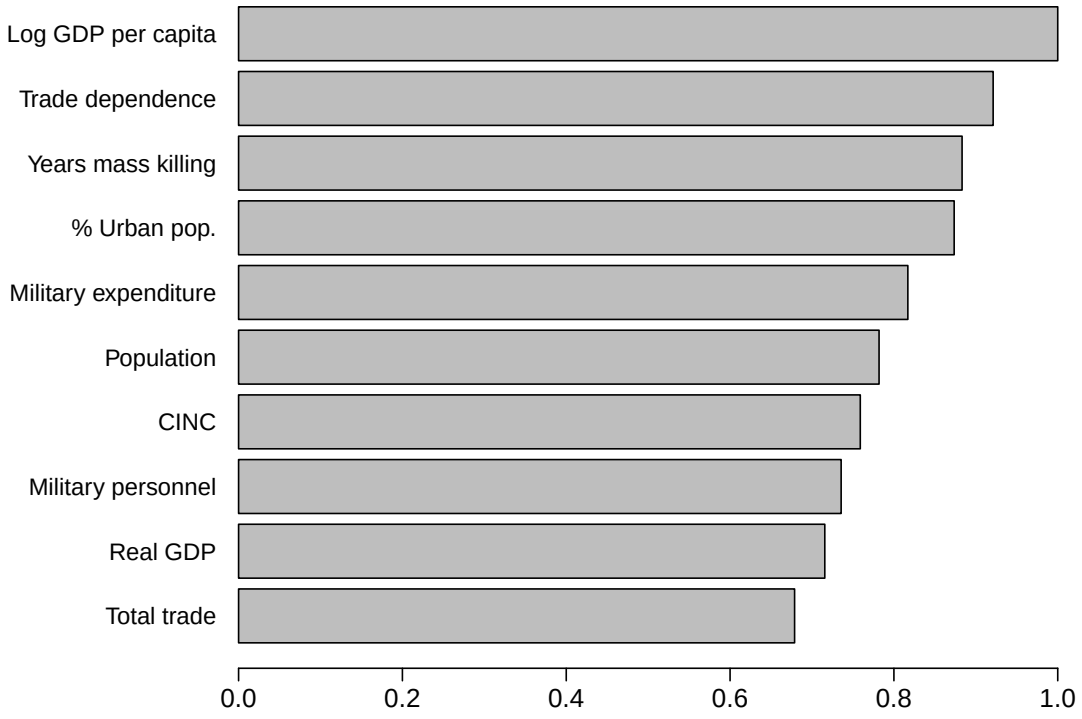

Figure 4.36: Variable Importance – Seed 7015

Mass Killings in and after the Cold War

This last set of models splits the sample into two periods, the Cold War years and the post-Cold War years. A similar set of variables are significant in both periods, and most of them also appear in the main model shown above.

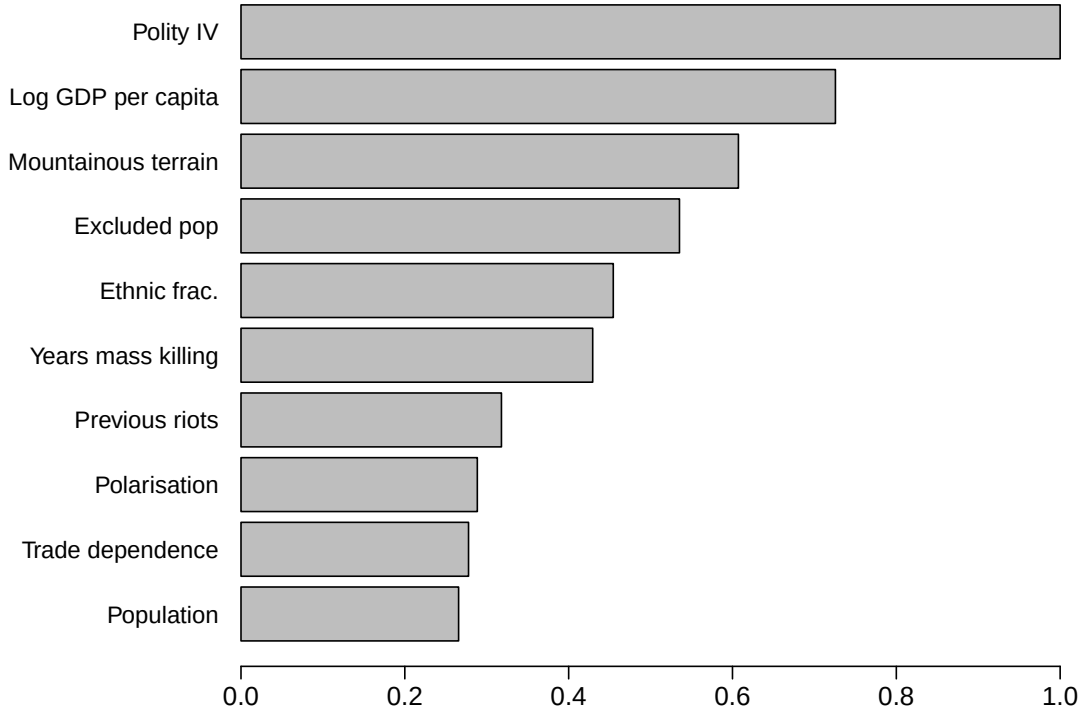

Figure 4.38: Variable Importance – Cold War Period

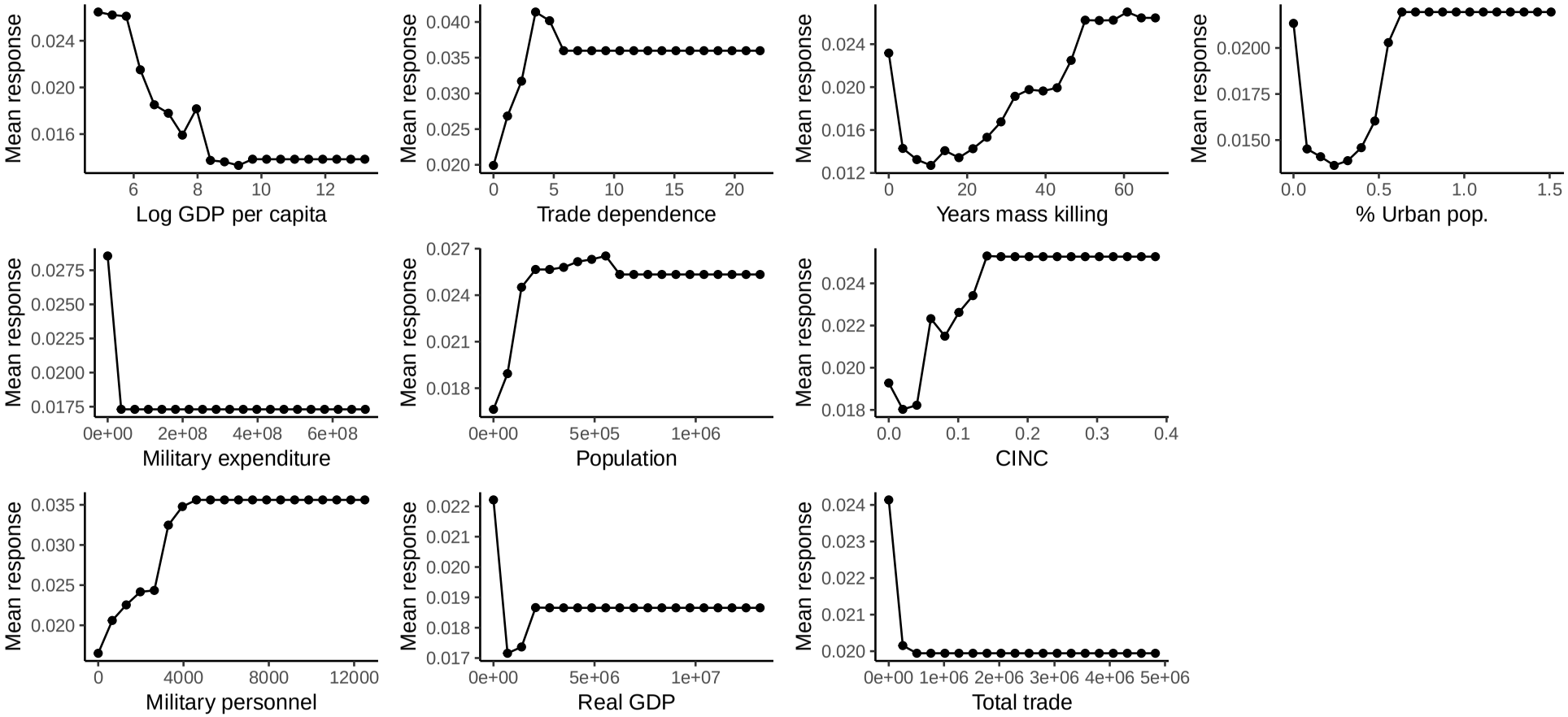

Figure 4.37: Partial Dependence Plot – Seed 7015

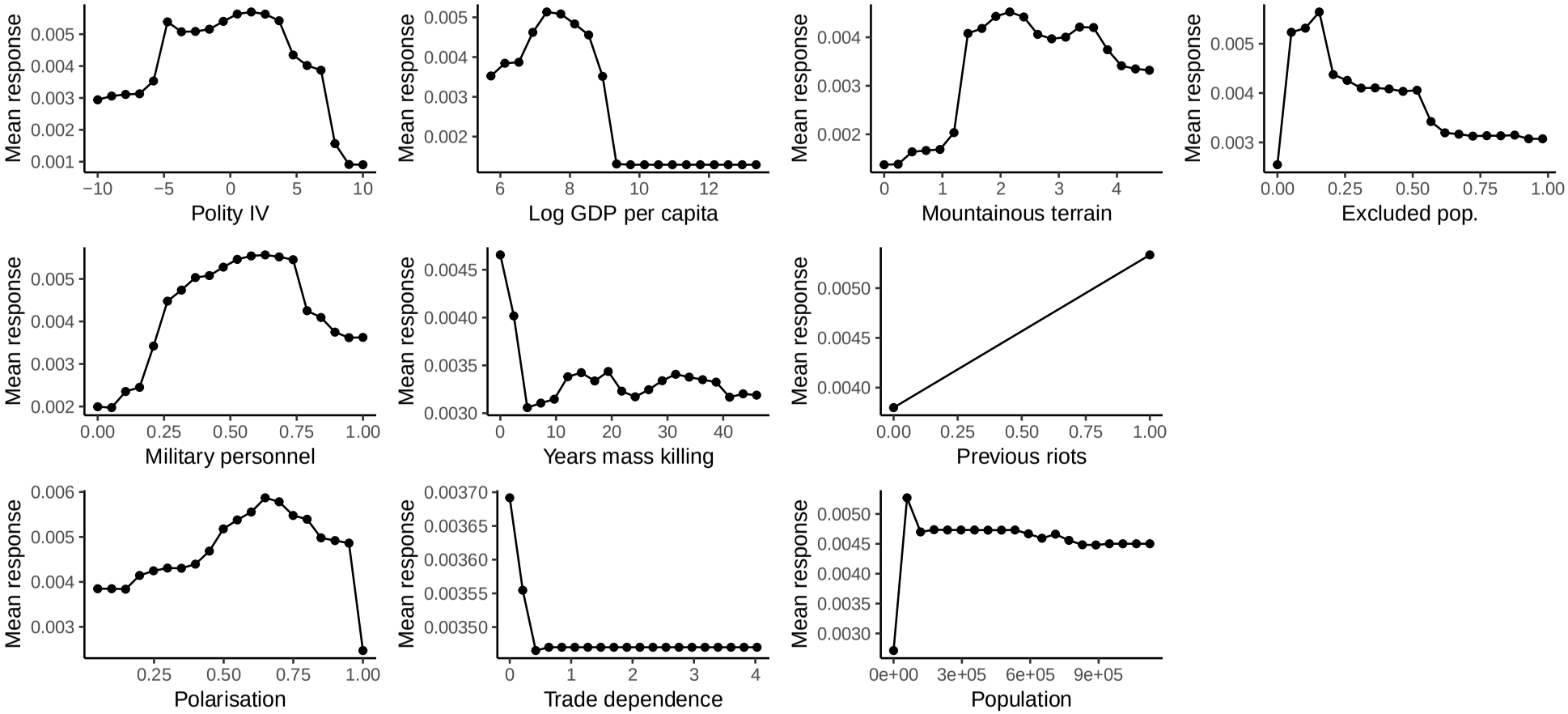

Figure 4.39: Partial Dependence Plot – Cold War Period

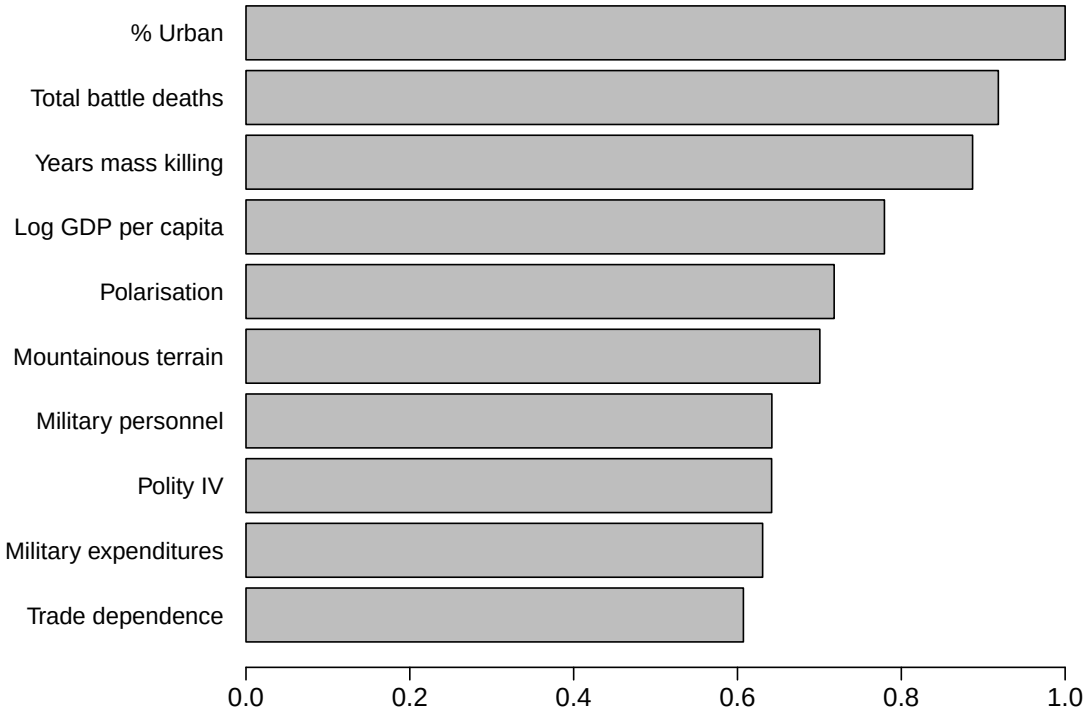

Figure 4.40: Variable Importance – Post-Cold War Period

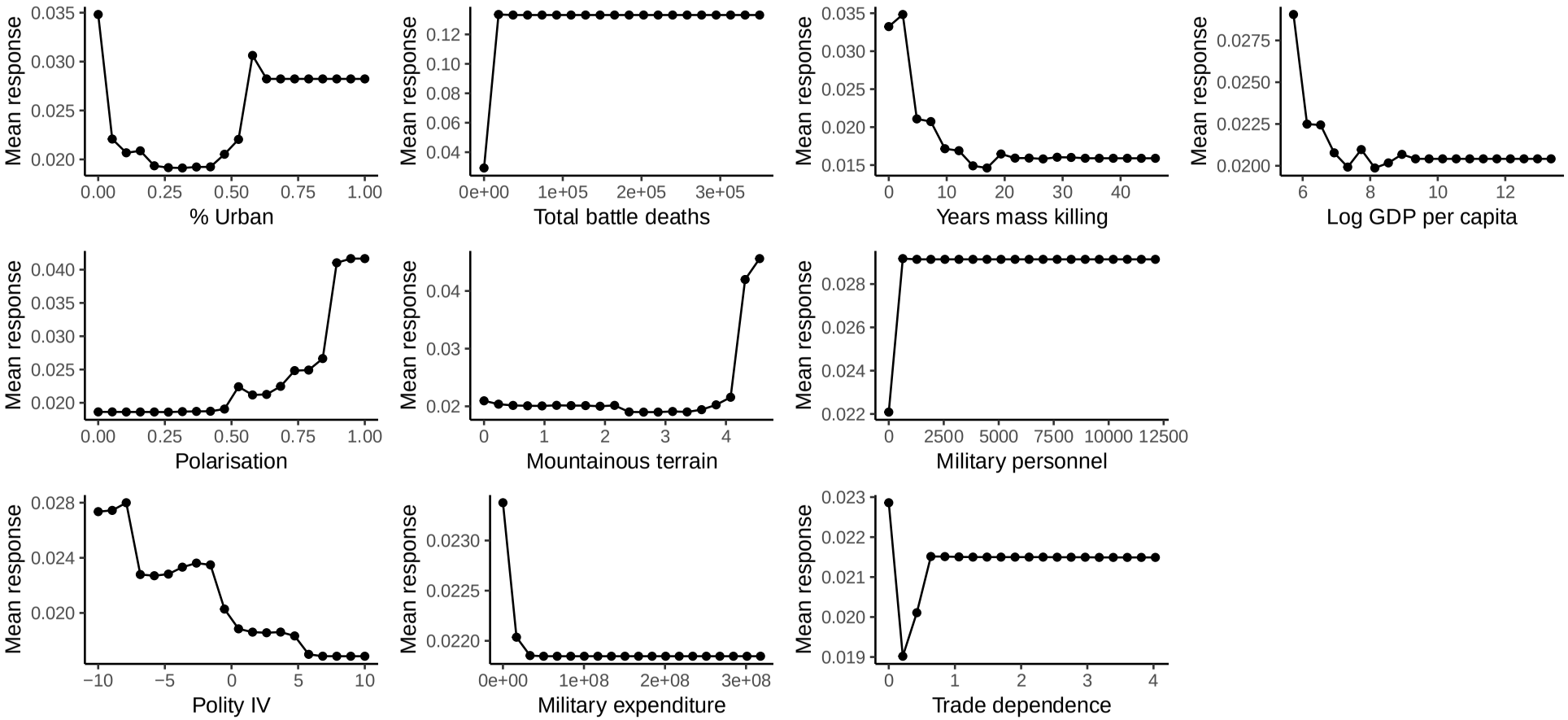

Figure 4.41: Partial Dependence Plot – Post-Cold War Period

Mass Killings in Peacetime

The figures below describe the results of the random forest estimations when I restrict the sample to only peace years. All cases coded as conflicts by the UCDP, COW or Cederman and his colleagues were removed from the data, and I estimate the model only with observations where the three sourced coded as peace years. The results are almost identical to the main model, with only small variations in the importance of the explanator variables.

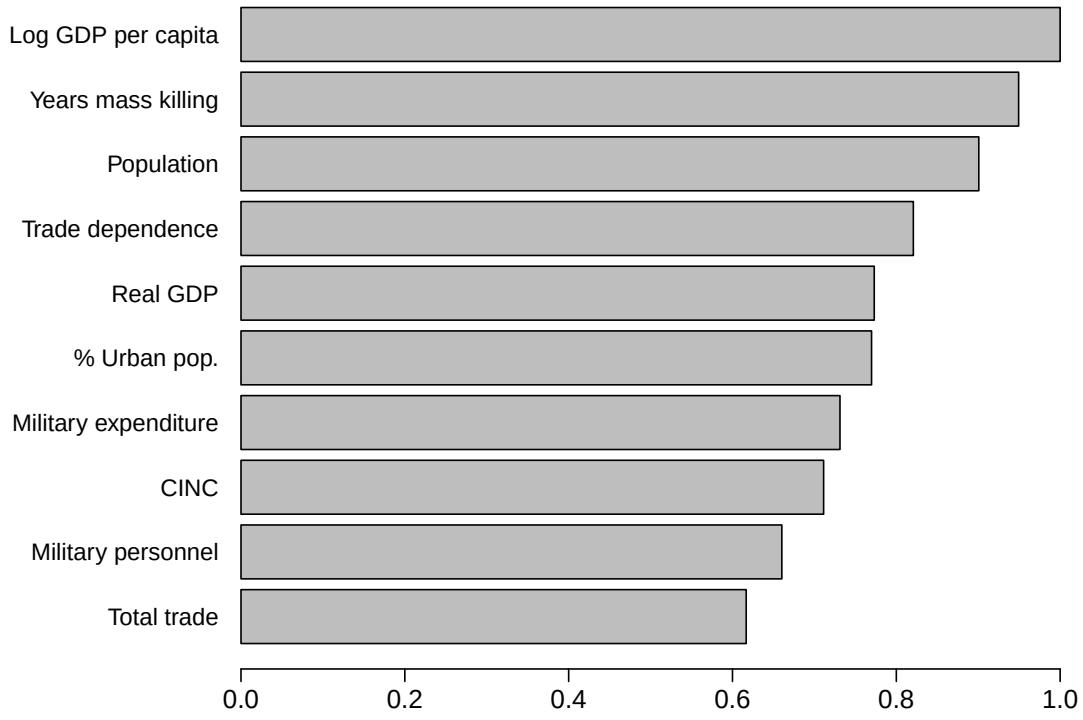

Figure 4.42: Variable Importance – Mass Killings during Peacetime

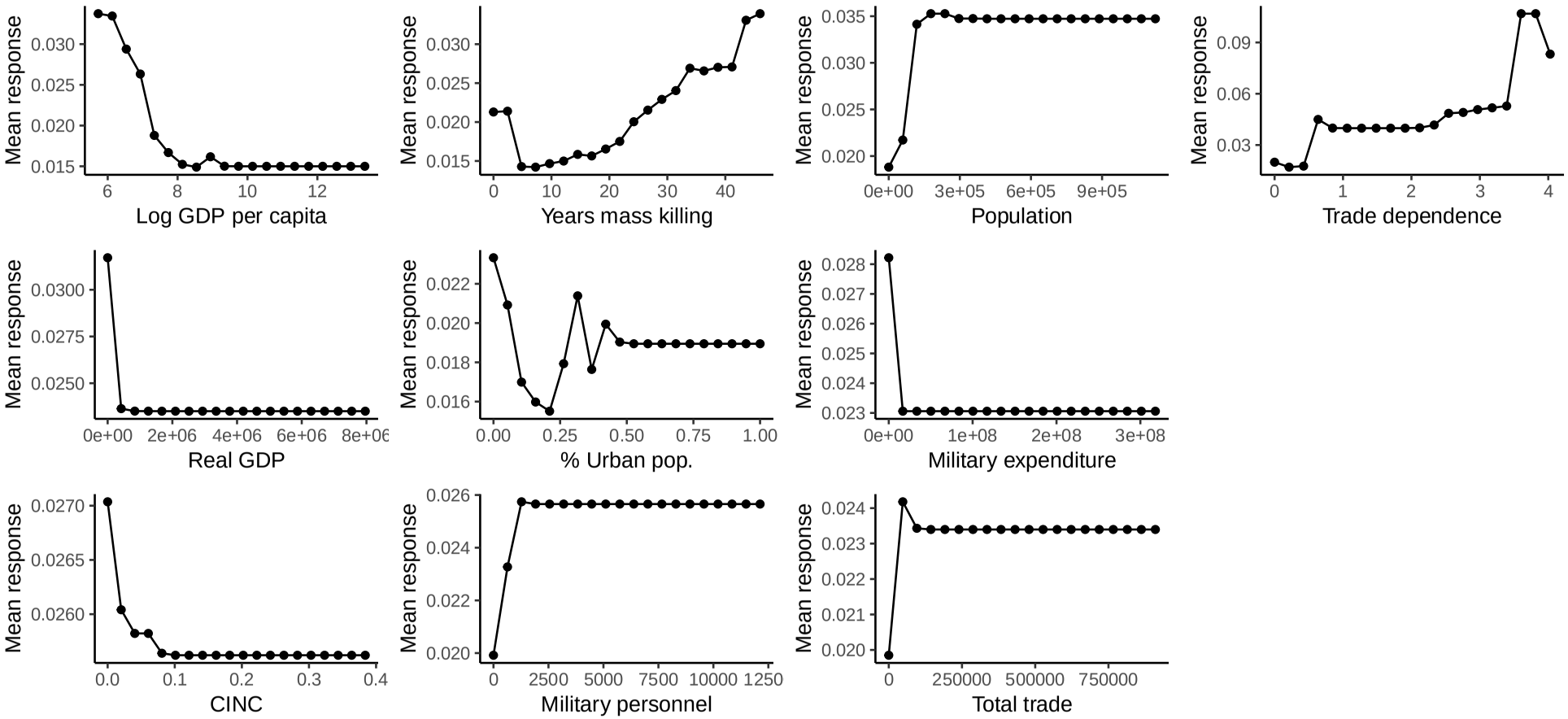

Figure 4.43: Partial Dependence Plot – Mass Killings during Peacetime

4.6.6 Harff's Genocides and Politicides Data

Main Model

We replicate the same analysis using Harff's 2003 data. The results are comparable to the ones presented above. A similar set of variables appear in this model.

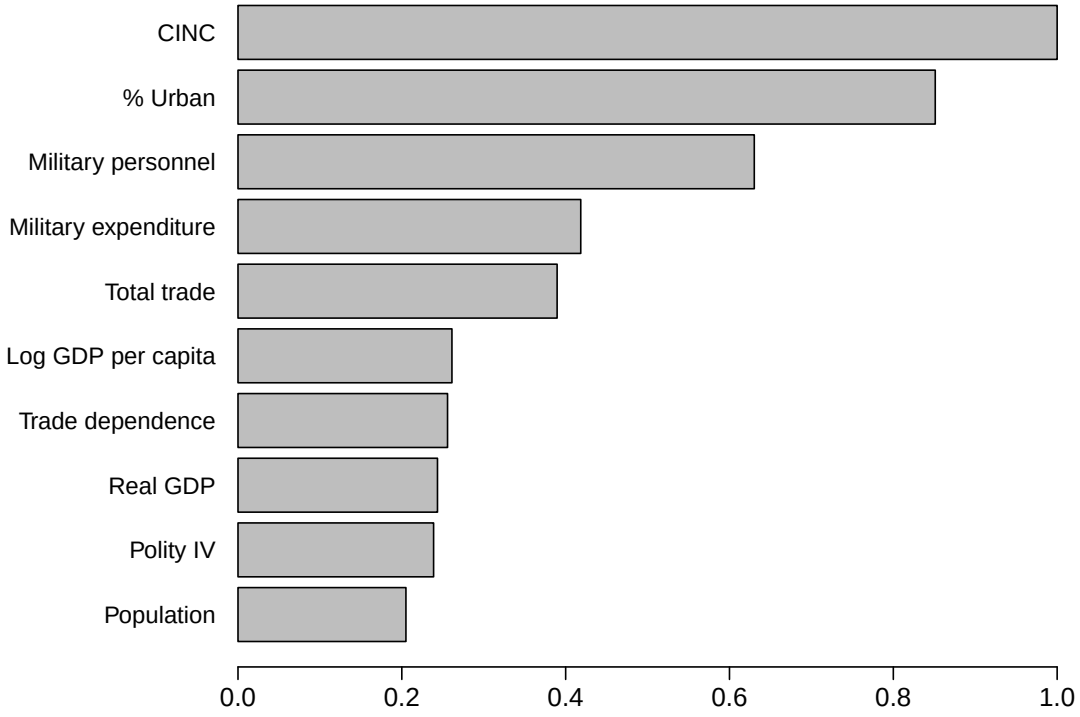

Figure 4.44: Variable Importance – Genocides and Politicides

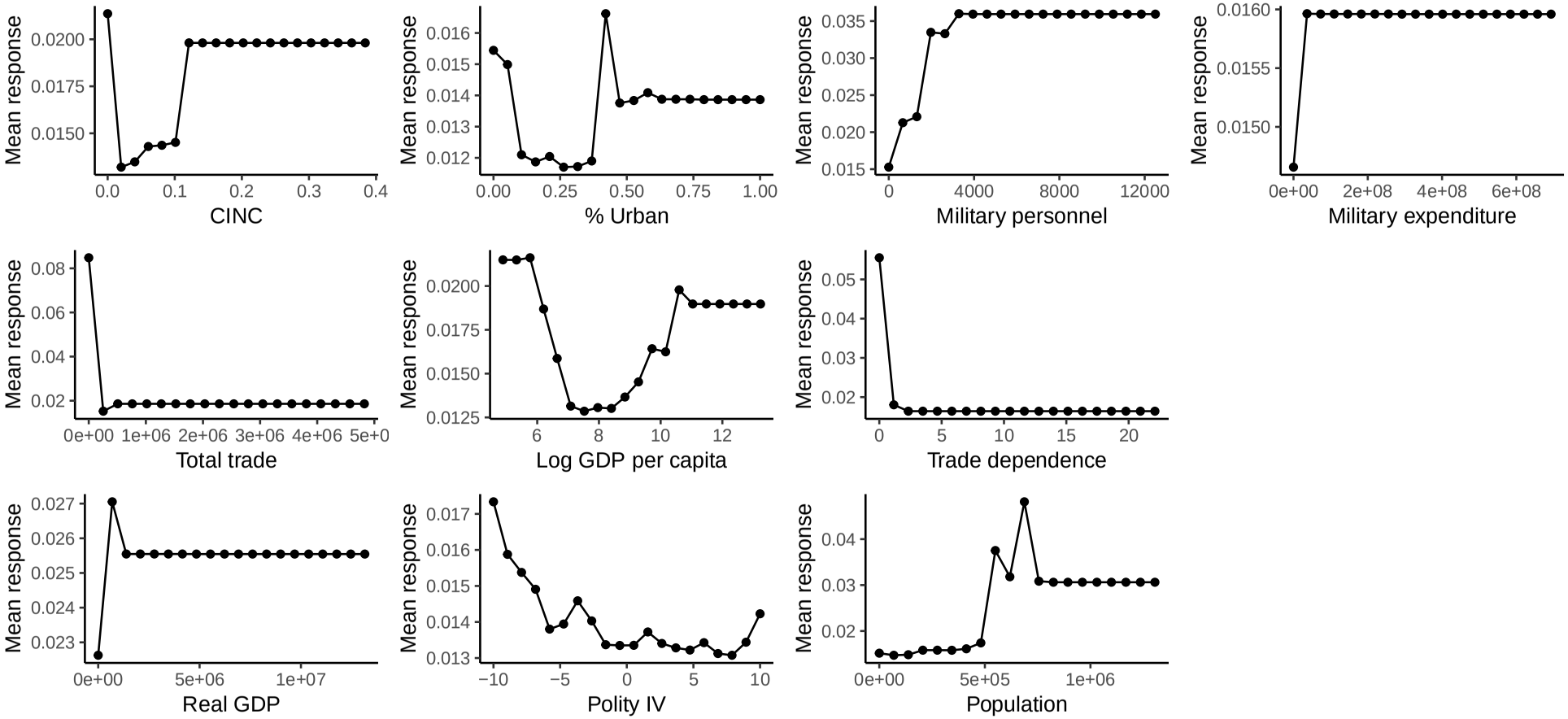

Figure 4.45: Partial Dependence Plot – Genocides and Politicides

Genocides and Politicides during Civil Wars

Lastly, the graphs below show the results of the grid search when we only include civil war years.

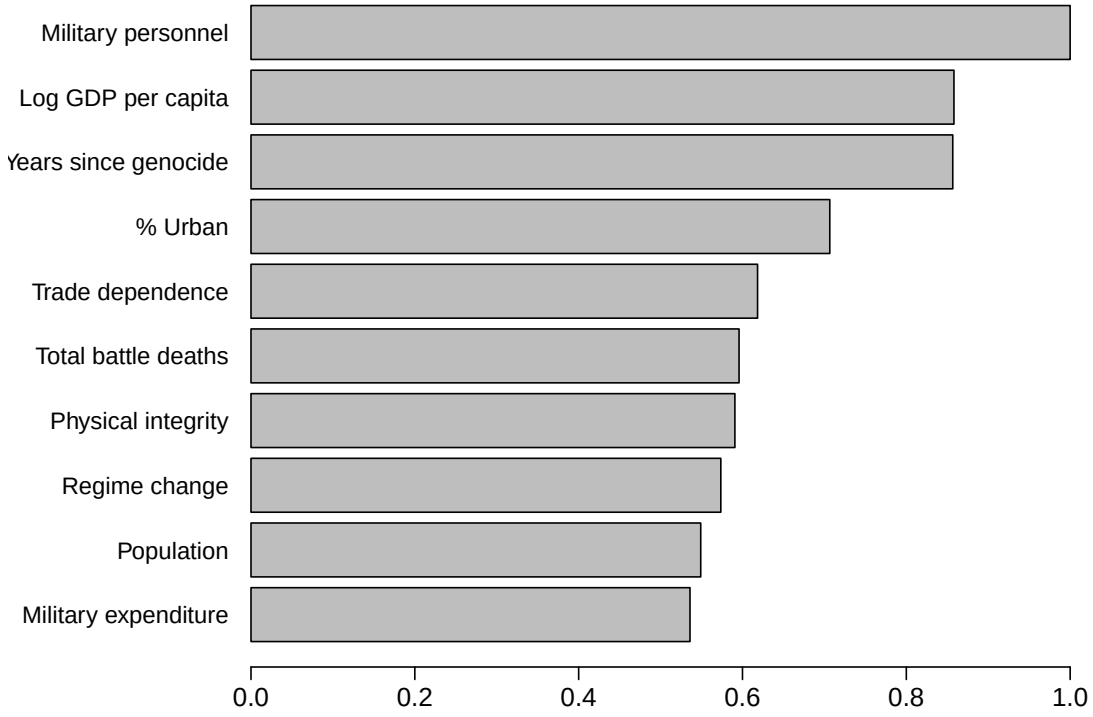

Figure 4.46: Variable Importance – Genocides and Politicides during Civil Wars (UCDP Data)

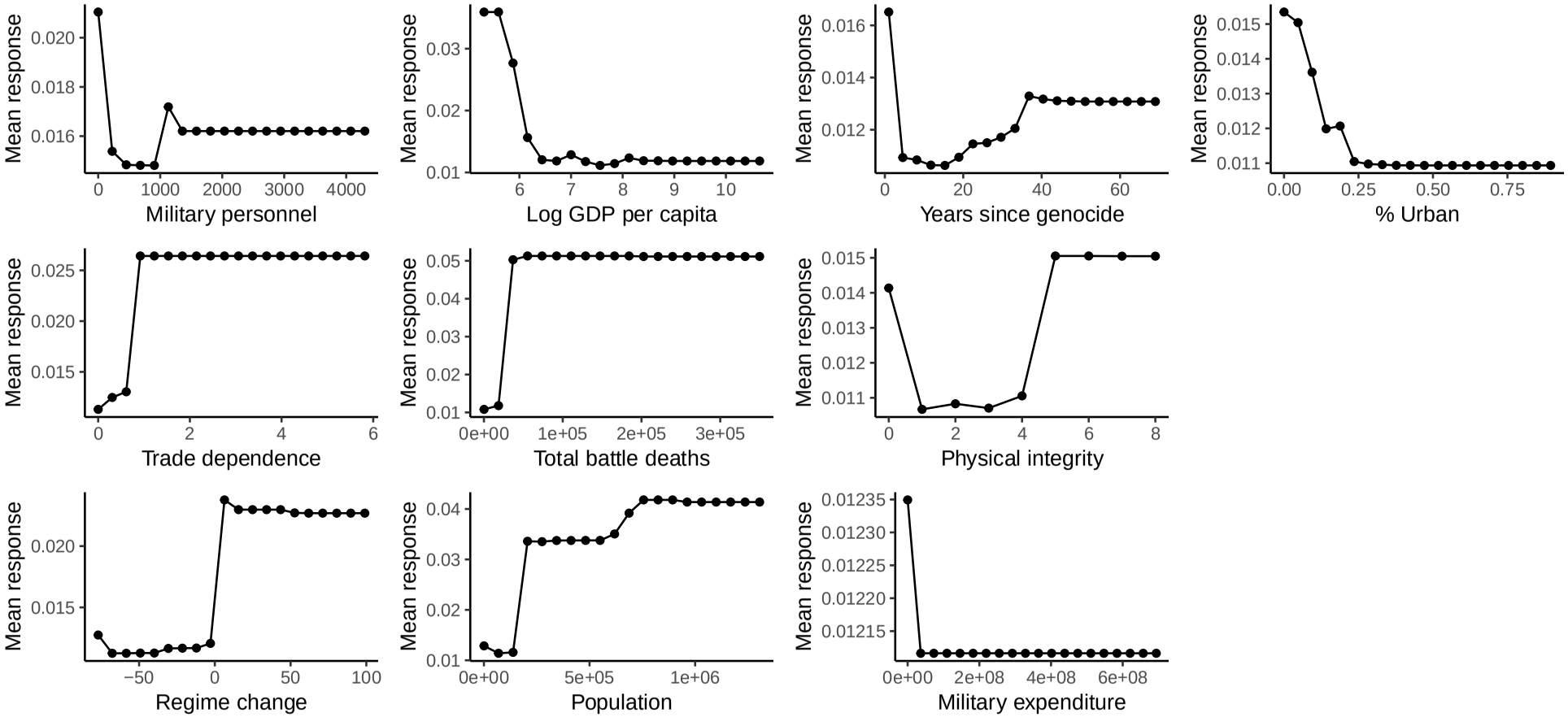

Figure 4.47: Partial Dependence Plot – Genocides and Politicides during Civil Wars (UCDP Data)

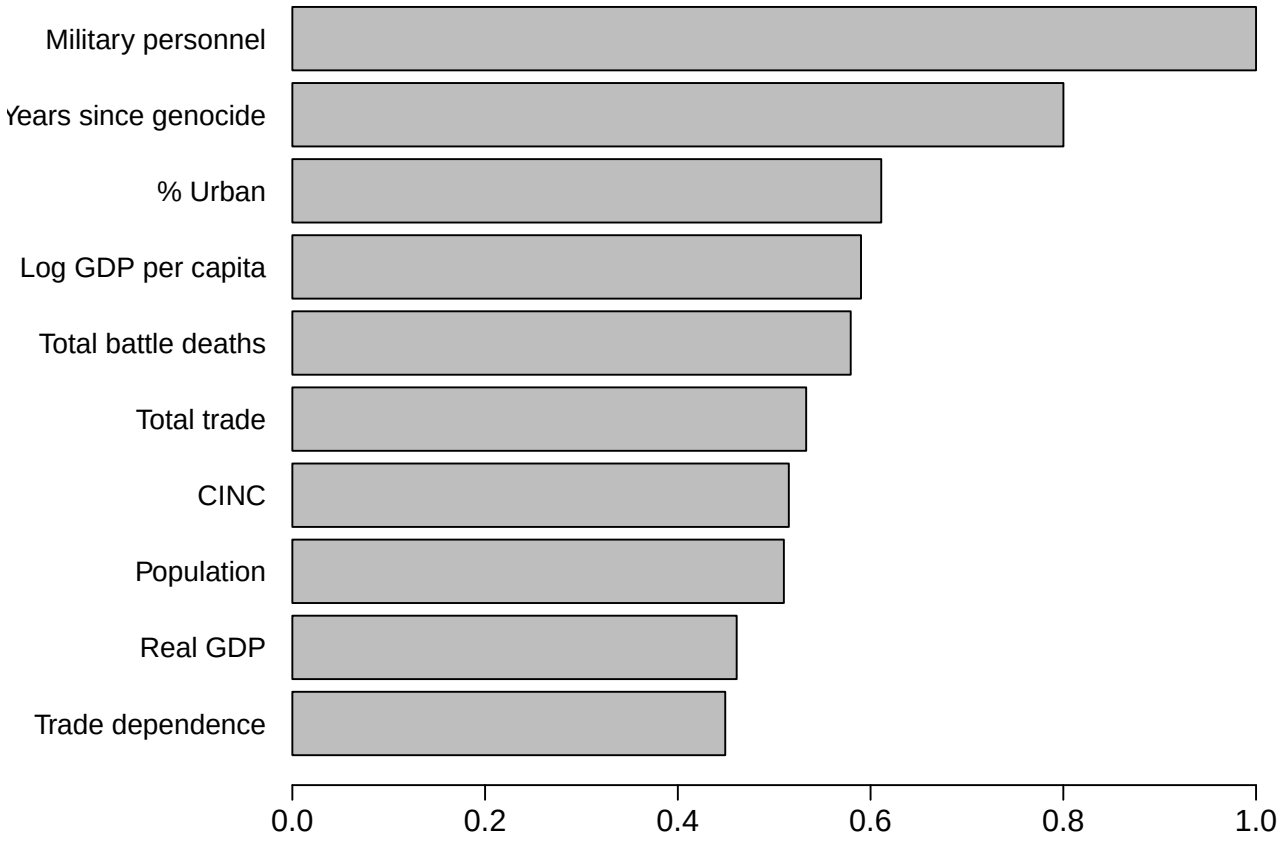

Figure 4.48: Variable Importance – Genocides and Politicides during Civil Wars (COW Data)

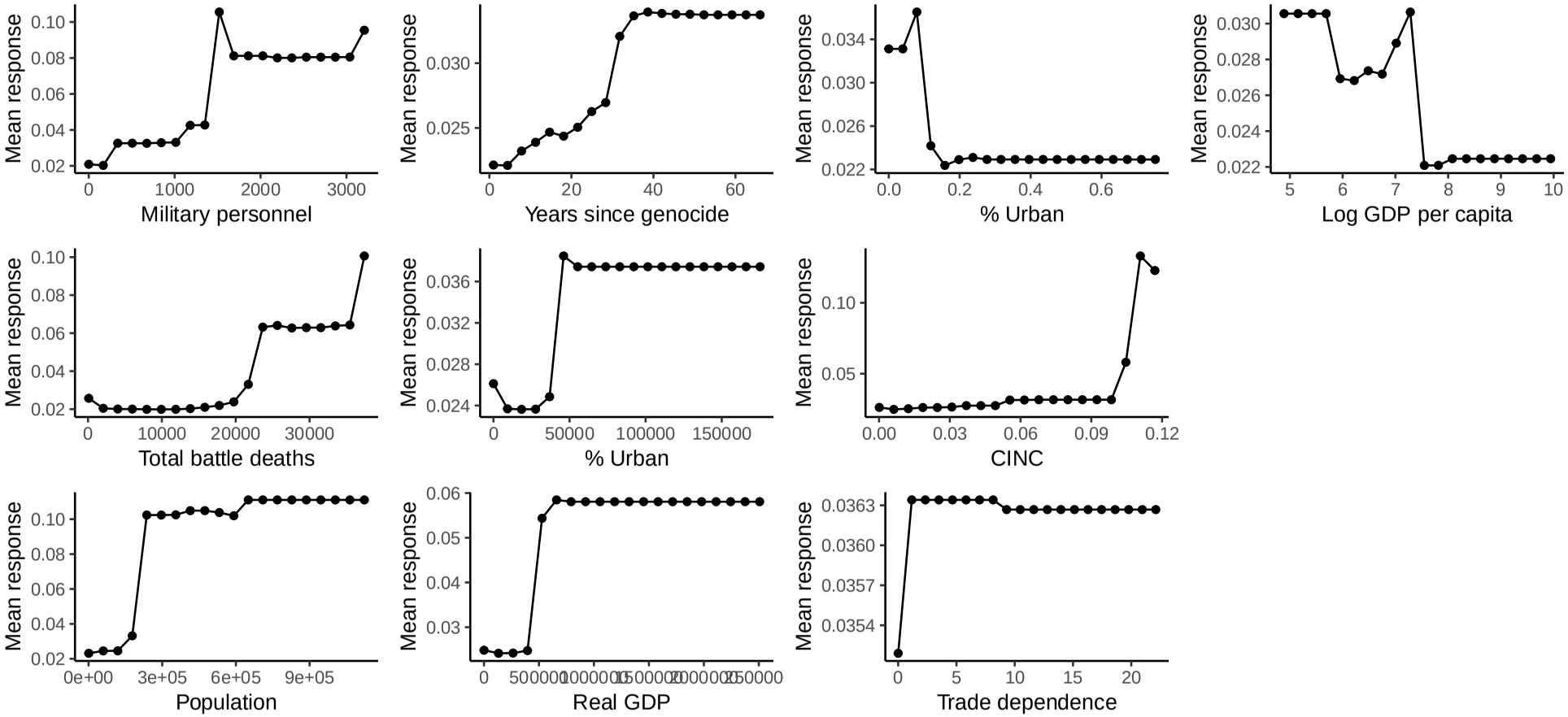

Figure 4.49: Partial Dependence Plot – Genocides and Politicides during Civil Wars (COW Data)

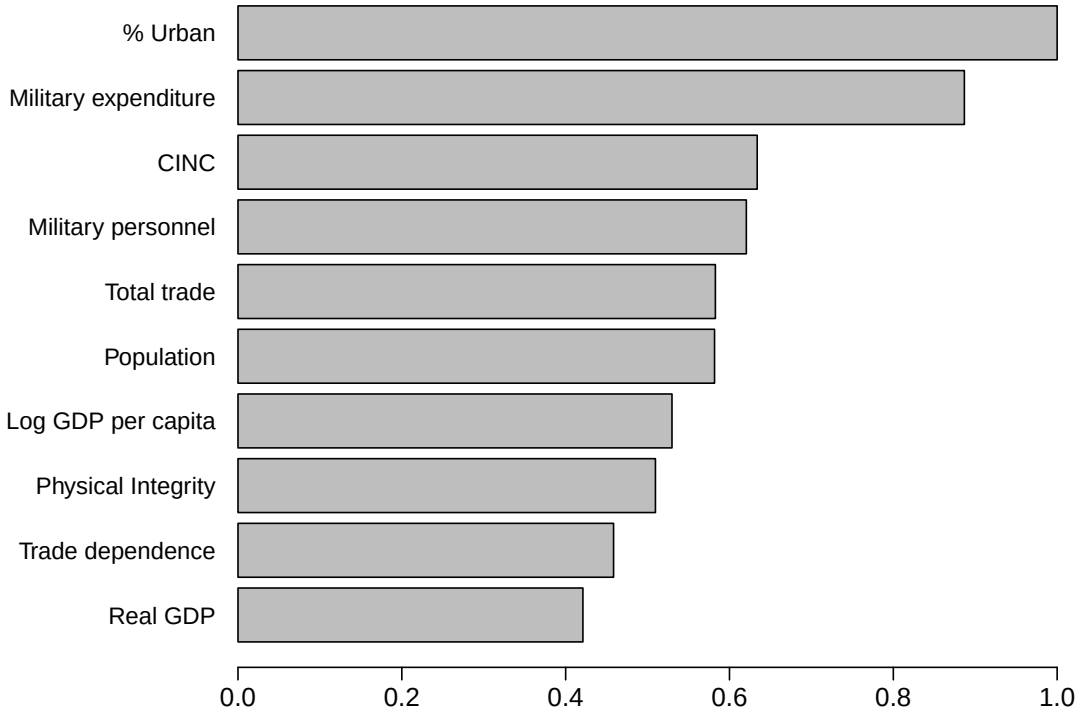

Figure 4.50: Variable Importance – Genocides and Politicides during Civil Wars (Cederman et al. Data)

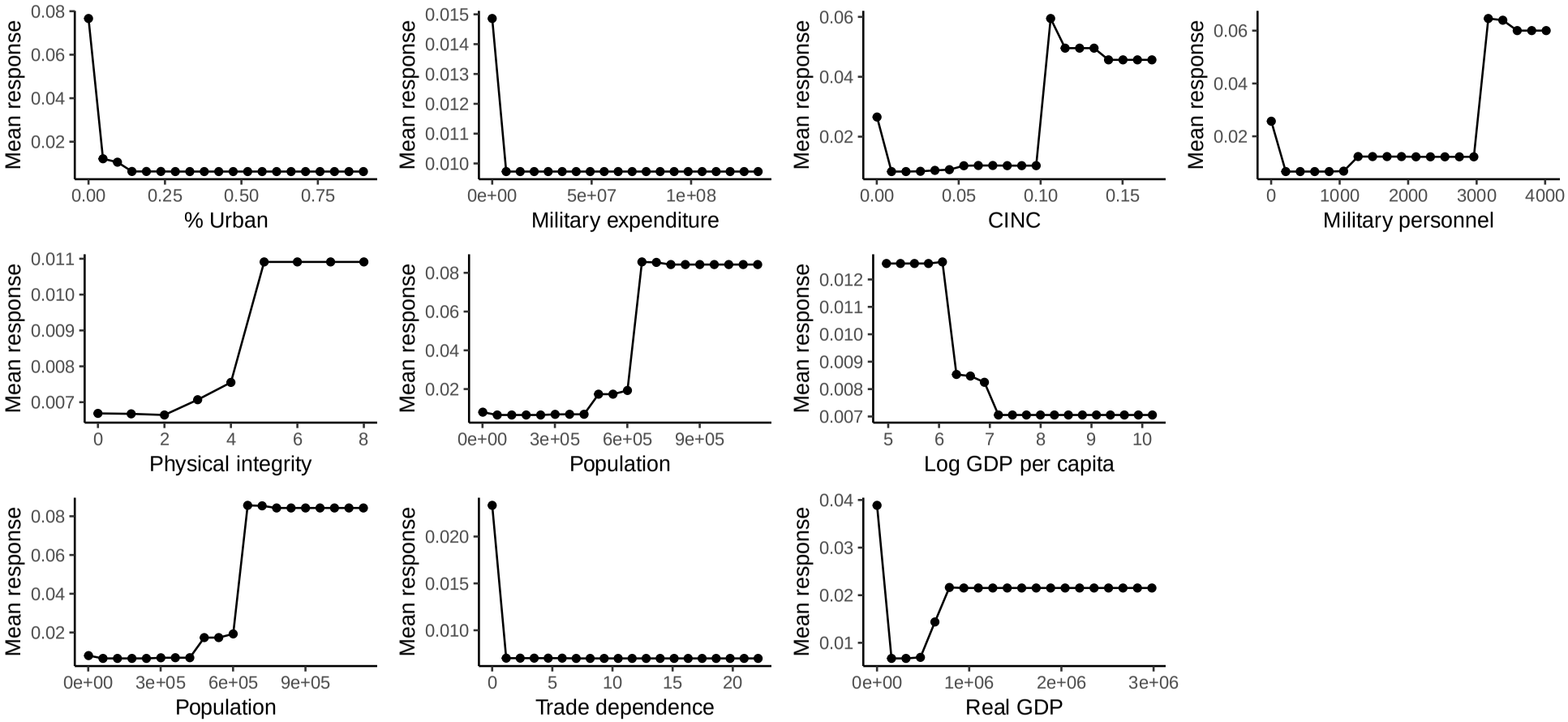

Figure 4.51: Partial Dependence Plot – Genocides and Politicides during Civil Wars (Cederman et al. Data)

4.6.7 R Code

The R code below replicates all statistical analyses and graphs included in this chapter.

```
#######################
### Data Wrangling ###
#######################

### Load required packages
if (!require("tidyverse")) {
        install.packages("tidyverse")
}
if (!require("data.table")) {
        install.packages("data.table")
}
if (!require("ExtremeBounds")) {
        install.packages("ExtremeBounds")
}
if (!require("sandwich")) {
        install.packages("sandwich")
}
if (!require("h2o")) {
  install.packages("h2o")
}
if (!require("arm")) {
        install.packages("arm")
}

### Load data
setwd("~/Documents/GitHub/mass-killings-8k/") # set the working directory
df <- haven::read_dta("data/base variables.dta") %>% setDT()

### Select and lag variables
sd.cols <- c("UCDPcivilwarstart", "UCDPcivilwarongoing", "COWcivilwarstart",
             "COWcivilwarongoing", "ethnowarstart", "ethnowarongoing",
             "assdummy", "demdummy", "elf", "lmtnest", "pop", "realgdp",
             "rgdppc", "polity2", "exclpop", "discpop", "polrqnew",
             "poltrqnew", "egiptpolrqnew", "egippolrqnew", "discrim",
             "elf2", "interstatewar", "milex", "milper", "percentpopurban",
             "postcoldwar", "coupdummy", "riotdummy", "territoryaims",
             "totaltrade", "tradedependence", "militias", "physint", "cinc",
             "totalbeaths", "guerrilladummy", "change", "sf", "regtrans")

df1 <- cbind(df, df[, shift(.SD, 1, give.names = TRUE),
                    by = ccode, .SDcols = sd.cols])

# Remove the second `ccode` variable
df1 <- as.data.frame(df1[, -c(70)])

# Add new variables
df1$logrgdppc_lag_1 <- log(df1$rgdppc_lag_1)
df1$polity2sq_lag_1 <- df1$polity2_lag_1^2
```

```
# UCDP civil war == 1
df.ucdp <- df1 %>% filter(UCDPcivilwarongoing == 1)
df.ucdp <- as.data.frame(df.ucdp[, c(1:7, 76:111)])
names(df.ucdp) <- sub("_.*","", names(df.ucdp))

# COW civil war == 1
df.cow <- df1 %>% filter(COWcivilwarongoing == 1)
df.cow <- as.data.frame(df.cow[, c(1:7, 76:111)])
names(df.cow) <- sub("_.*","", names(df.cow))

# Ethnic civil war == 1
df.eth <- df1 %>% filter(ethnowarongoing == 1)
df.eth <- as.data.frame(df.eth[, c(1:7, 75:110)])
names(df.eth) <- sub("_.*","", names(df.eth))

# Regular model
df2 <- as.data.frame(df1[, c(1:7, 70:111)])
names(df2) <- sub("_.*","", names(df2))

# Cold War period
df2.coldwar <- df2 %>% filter(year <= 1991)
df2.postcoldwar <- df2 %>% filter(year > 1991)

# Countries without civil wars
df.nowar <- df2 %>% filter(COWcivilwarongoing == 0 & UCDPcivilwarongoing == 0, ethnowarongoing == 0)

#### Same procedure with the uamkstart variable

# Preparing the dataset
df3 <- haven::read_dta("data/uamkstart.dta") %>% setDT()
sd.cols <- c("UCDPcivilwarstart", "UCDPcivilwarongoing", "COWcivilwarstart",
             "COWcivilwarongoing", "ethnowarstart", "ethnowarongoing",
             "assdummy", "demdummy", "elf", "lmtnest", "pop", "realgdp",
             "rgdppc", "polity2", "exclpop", "discpop", "polrqnew",
             "poltrqnew", "egiptpolrqnew", "egippolrqnew", "discrim",
             "elf2", "interstatewar", "milex", "milper", "percentpopurban",
             "postcoldwar", "coupdummy", "riotdummy", "territoryaims",
             "totaltrade", "tradedependence", "militias", "physint", "cinc",
             "totalbeaths", "change", "guerrilladummy", "sf", "regtrans")

df4 <- cbind(df3, df3[, shift(.SD, 1, give.names = TRUE),
                      by = ccode, .SDcols = sd.cols])

# Remove the second `ccode` variable
df4 <- as.data.frame(df4[, -c(75)])

# Add new variables
df4$logrgdppc_lag_1 <- log(df4$rgdppc_lag_1)
df4$polity2sq_lag_1 <- df4$polity2_lag_1^2

# Renaming variables
df5 <- as.data.frame(df4[, c(1:4, 72:116)])
```

```
names(df5) <- sub("_.*","", names(df5))

# UCDP civil war == 1
df.ucdp2 <- df5 %>% filter(UCDPcivilwarongoing == 1)
df.ucdp2 <- as.data.frame(df.ucdp2[, c(1:7, 14:49)])
names(df.ucdp2) <- sub("_.*","", names(df.ucdp2))

# COW civil war == 1
df.cow2 <- df5 %>% filter(COWcivilwarongoing == 1)
df.cow2 <- as.data.frame(df.cow2[, c(1:7, 14:49)])
names(df.cow2) <- sub("_.*","", names(df.cow2))

# Ethnic civil war == 1
df.eth2 <- df5 %>% filter(ethnowarongoing == 1)
df.eth2 <- as.data.frame(df.eth2[, c(1:7, 14:49)])
names(df.eth2) <- sub("_.*","", names(df.eth2))

# Cold War period
df5.coldwar <- df5 %>% filter(year <= 1991)
df5.postcoldwar <- df5 %>% filter(year > 1991)

################################
### Extreme bounds analysis  ###
################################

# Classifying a few variables as mutually exclusive variables.
# "Change" was removed because it was correlated at 0.99 with "regtrans".
# don't forget to add CINC
free.variables <- c("logrgdppc", "polity2", "mksyr")
civilwar.variables <- c("UCDPcivilwarongoing", "UCDPcivilwarstart",
                        "COWcivilwarongoing", "COWcivilwarstart",
                        "ethnowarongoing", "ethnowarstart")
doubtful.variables <- c("UCDPcivilwarongoing", "UCDPcivilwarstart",
                        "COWcivilwarongoing", "COWcivilwarstart",
                        "ethnowarongoing", "ethnowarstart", "assdummy",
                        "totaltrade", "tradedependence", "milper", "milex",
                        "pop", "totalbeaths", "guerrilladummy", "regtrans",
                        "riotdummy", "territoryaims", "militias",
                        "physint", "percentpopurban", "coupdummy",
                        "postcoldwar",  "lmtnest", "realgdp", "discrim",
                        "exclpop", "discpop", "elf",  "polrqnew",
                        "egippolrqnew", "poltrqnew", "egiptpolrqnew",
                        "polity2sq")

# Cluster-robust standard errors
se.clustered.robust <- function(model.object){
        model.fit <- vcovHC(model.object, type = "HC", cluster = "country")
        out <- sqrt(diag(model.fit))
        return(out)
}

### Models
```

```
# Main
m1 <- eba(y = "MKstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df2, vif = 7, level = 0.9,
          se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/mk.rda")

# 3 vars at a time
m1 <- eba(y = "MKstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:3,
          data = df2, vif = 7, level = 0.9, draws = 10000,
          se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/mk-3vars.rda")

# 5 vars at a time
m1 <- eba(y = "MKstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:5,
          data = df2, vif = 7, draws = 50000,
          level = 0.9, se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/mk-5vars.rda")

# Low VIF
m1 <- eba(y = "MKstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df2, vif = 2.5, level = 0.9,
          draws = 50000,
          se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/mk-low-vif.rda")

# High VIF
m1 <- eba(y = "MKstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df2, vif = 10, draws = 50000,
          level = 0.9, se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/mk-high-vif.rda")

# No VIF
m1 <- eba(y = "MKstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df2, level = 0.9, draws = 50000,
          se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/mk-no-vif.rda")

# Logit
m1 <- eba(y = "MKstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
```

```
          data = df2, level = 0.9, vif = 7, draws = 50000,
          reg.fun = bayesglm, family = binomial(link = "logit"))
save(m1, file = "~/Documents/mk/mk-logit.rda")

# Probit
m1 <- eba(y = "MKstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df2, level = 0.9, vif = 7, draws = 50000,
          reg.fun = bayesglm, family = binomial(link="probit"))
save(m1, file = "~/Documents/mk/mk-probit.rda")

# CINC
doubtful.variables <- c("UCDPcivilwarongoing", "UCDPcivilwarstart",
                        "COWcivilwarongoing", "COWcivilwarstart",
                        "ethnowarongoing", "ethnowarstart", "assdummy",
                        "totaltrade", "tradedependence", "cinc",
                        "totalbeaths", "guerrilladummy", "regtrans",
                        "riotdummy", "territoryaims", "militias",
                        "physint", "percentpopurban", "coupdummy",
                        "postcoldwar",  "lmtnest", "realgdp", "discrim",
                        "exclpop", "discpop", "elf",  "polrqnew",
                        "egippolrqnew", "poltrqnew", "egiptpolrqnew",
                        "polity2sq")

m1 <- eba(y = "MKstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df2, vif = 7, level = 0.9,
          se.fun = se.clustered.robust, draws = 50000)
save(m1, file = "~/Documents/mk/mk-cinc.rda")

# Cold War Period
civilwar.variables <- c("UCDPcivilwarstart","COWcivilwarstart","ethnowarstart")
m1 <- doubtful.variables <- c("UCDPcivilwarongoing", "UCDPcivilwarstart",
                              "COWcivilwarongoing", "COWcivilwarstart",
                              "ethnowarongoing", "ethnowarstart", "assdummy",
                              "totaltrade", "tradedependence", "cinc",
                              "totalbeaths", "guerrilladummy", "regtrans",
                              "riotdummy", "territoryaims", "militias",
                              "physint", "percentpopurban", "coupdummy",
                              "lmtnest", "realgdp", "discrim",
                              "exclpop", "discpop", "elf",  "polrqnew",
                              "egippolrqnew", "poltrqnew", "egiptpolrqnew",
                              "polity2sq")

m1 <- eba(y = "MKstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df2.coldwar, vif = 7, level = 0.9,
          se.fun = se.clustered.robust, draws = 50000)
save(m1, file = "data/mk-coldwar.rda")
```

```
# Post-Cold War
m1 <- doubtful.variables <- c("UCDPcivilwarongoing", "UCDPcivilwarstart",
                              "COWcivilwarongoing", "COWcivilwarstart",
                              "ethnowarongoing", "ethnowarstart", "assdummy",
                              "totaltrade", "tradedependence", "cinc",
                              "totalbeaths", "guerrilladummy", "regtrans",
                              "riotdummy", "territoryaims", "militias",
                              "physint", "percentpopurban", "coupdummy",
                              "lmtnest", "realgdp", "discrim",
                              "exclpop", "discpop", "elf",  "polrqnew",
                              "egippolrqnew", "poltrqnew", "egiptpolrqnew",
                              "polity2sq")

m1 <- eba(y = "MKstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df2.postcoldwar, vif = 7, level = 0.9,
          se.fun = se.clustered.robust, draws = 50000)
save(m1, file = "data/mk-postcoldwar.rda")

## Countries with no civil wars
free.variables <- c("logrgdppc", "polity2", "mksyr")
civilwar.variables <- c("UCDPcivilwarstart", "COWcivilwarstart",
                        "ethnowarstart")
m1 <- doubtful.variables <- c("UCDPcivilwarstart","COWcivilwarstart",
                              "ethnowarstart", "assdummy",
                              "totaltrade", "tradedependence", "cinc",
                              "totalbeaths","guerrilladummy",
                              "riotdummy", "territoryaims", "militias",
                              "physint", "percentpopurban", "coupdummy",
                              "postcoldwar", "lmtnest", "realgdp", "discrim",
                              "exclpop", "discpop", "elf",  "polrqnew",
                              "egippolrqnew", "poltrqnew", "egiptpolrqnew",
                              "polity2sq")

m1 <- eba(y = "MKstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df.nowar, vif = 7, level = 0.9,
          se.fun = se.clustered.robust, draws = 50000)
save(m1, file = "data/mk-nowar.rda")

### Ongoing Civil Wars

# UCDPcivilwarongoing == 1
doubtful.variables <- c("assdummy", "totaltrade", "tradedependence",
                        "milper", "milex", "pop", "totalbeaths",
                        "guerrilladummy", "regtrans", "riotdummy",
                        "territoryaims", "militias", "physint",
                        "percentpopurban", "coupdummy", "postcoldwar",
                        "lmtnest", "realgdp", "discrim", "exclpop",
                        "discpop", "elf",  "polrqnew", "egippolrqnew",
                        "poltrqnew", "egiptpolrqnew", "polity2sq")
```

```
m1 <- eba(y = "MKstart", free = free.variables,
          doubtful = doubtful.variables, k = 0:4,
          data = df.ucdp, vif = 7, draws = 50000,
          level = 0.9, se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/mk-ucdp.rda")

# COWcivilwarongoing == 1
doubtful.variables <- c("assdummy", "totaltrade", "tradedependence",
                        "milper", "milex", "pop", "totalbeaths",
                        "guerrilladummy", "regtrans", "riotdummy",
                        "territoryaims", "militias", "physint",
                        "percentpopurban", "coupdummy", "postcoldwar",
                        "lmtnest", "realgdp", "discrim", "exclpop",
                        "discpop", "elf",  "polrqnew", "egippolrqnew",
                        "poltrqnew", "egiptpolrqnew", "polity2sq")

m1 <- eba(y = "MKstart", free = free.variables,
          doubtful = doubtful.variables, k = 0:4,
          data = df.cow, vif = 7, draws = 50000,
          level = 0.9, se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/mk-cow.rda")

# Ethnic conflict == 1
doubtful.variables <- c("assdummy", "totaltrade", "tradedependence",
                        "milper", "milex", "pop", "totalbeaths",
                        "guerrilladummy", "regtrans", "riotdummy",
                        "territoryaims", "militias", "physint",
                        "percentpopurban", "coupdummy", "postcoldwar",
                        "lmtnest", "realgdp", "discrim", "exclpop",
                        "discpop", "elf",  "polrqnew", "egippolrqnew",
                        "poltrqnew", "egiptpolrqnew", "polity2sq")

m1 <- eba(y = "MKstart", free = free.variables,
          doubtful = doubtful.variables, k = 0:4,
          data = df.eth, vif = 7, draws = 50000,
          level = 0.9, se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/mk-eth.rda")

# Main
free.variables <- c("logrgdppc", "polity2", "uamkyr")
civilwar.variables <- c("UCDPcivilwarongoing", "UCDPcivilwarstart",
                        "COWcivilwarongoing", "COWcivilwarstart",
                        "ethnowarongoing", "ethnowarstart")
doubtful.variables <- c("UCDPcivilwarongoing", "UCDPcivilwarstart",
                        "COWcivilwarongoing", "COWcivilwarstart",
                        "ethnowarongoing", "ethnowarstart", "assdummy",
                        "totaltrade", "tradedependence", "milper", "milex",
                        "pop", "totalbeaths", "guerrilladummy", "regtrans",
                        "riotdummy", "territoryaims", "militias",
                        "physint", "percentpopurban", "coupdummy",
                        "postcoldwar",  "lmtnest", "realgdp", "discrim",
                        "exclpop", "discpop", "elf",  "polrqnew",
```

```
                        "egippolrqnew", "poltrqnew", "egiptpolrqnew",
                        "polity2sq")
m1 <- eba(y = "uamkstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df5, vif = 7, level = 0.9,
          se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/uamk.rda")

# 3 vars at a time
m1 <- eba(y = "uamkstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:3,
          data = df5, vif = 7, level = 0.9,
          se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/uamk-3vars.rda")

# 5 vars at a time
m1 <- eba(y = "uamkstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:5,
          data = df5, vif = 7, draws = 50000,
          level = 0.9, se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/uamk-5vars.rda")

# Low VIF
m1 <- eba(y = "uamkstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df5, vif = 2.5, level = 0.9, draws = 50000,
          se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/uamk-low-vif.rda")

# High VIF
m1 <- eba(y = "uamkstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df5, vif = 10, draws = 50000,
          level = 0.9, se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/uamk-high-vif.rda")

# No VIF
m1 <- eba(y = "uamkstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df5, level = 0.9, draws = 50000,
          se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/uamk-no-vif.rda")

# Logit
m1 <- eba(y = "uamkstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
```

```
          data = df5, level = 0.9, vif = 7, draws = 50000,
          reg.fun = bayesglm, family = binomial(link = "logit"))
save(m1, file = "~/Documents/mk/uamk-logit.rda")

# Probit
m1 <- eba(y = "uamkstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df5, level = 0.9, vif = 7, draws = 50000,
          reg.fun = bayesglm, family = binomial(link="probit"))
save(m1, file = "~/Documents/mk/uamk-probit.rda")

# CINC
doubtful.variables <- c("UCDPcivilwarongoing", "UCDPcivilwarstart",
                        "COWcivilwarongoing", "COWcivilwarstart",
                        "ethnowarongoing", "ethnowarstart", "assdummy",
                        "totaltrade", "tradedependence", "cinc",
                        "totalbeaths", "guerrilladummy", "regtrans",
                        "riotdummy", "territoryaims", "militias",
                        "physint", "percentpopurban", "coupdummy",
                        "postcoldwar",  "lmtnest", "realgdp", "discrim",
                        "exclpop", "discpop", "elf",  "polrqnew",
                        "egippolrqnew", "poltrqnew", "egiptpolrqnew",
                        "polity2sq")

m1 <- eba(y = "uamkstart", free = free.variables,
          exclusive = list(civilwar.variables),
          doubtful = doubtful.variables, k = 0:4,
          data = df5, vif = 7, level = 0.9, draws = 50000,
          se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/uamk-cinc.rda")

### Ongoing Civil Wars

# UCDPcivilwarongoing == 1
df.ucdp2 <- df5 %>% filter(UCDPcivilwarongoing == 1)
doubtful.variables <- c("assdummy", "totaltrade", "tradedependence",
                        "milper", "milex", "pop", "totalbeaths",
                        "guerrilladummy", "regtrans", "riotdummy",
                        "territoryaims", "militias", "physint",
                        "percentpopurban", "coupdummy", "postcoldwar",
                        "lmtnest", "realgdp", "discrim", "exclpop",
                        "discpop", "elf",  "polrqnew", "egippolrqnew",
                        "poltrqnew", "egiptpolrqnew", "polity2sq")

m1 <- eba(y = "uamkstart", free = free.variables,
          doubtful = doubtful.variables, k = 0:4,
          data = df.ucdp2, vif = 7, draws = 50000,
          level = 0.9, se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/uamk-ucdp.rda")

# COWcivilwarongoing == 1
df.cow2 <- df5 %>% filter(COWcivilwarongoing == 1)
```

```
doubtful.variables <- c("assdummy", "totaltrade", "tradedependence",
                        "milper", "milex", "pop", "totalbeaths",
                        "guerrilladummy", "regtrans", "riotdummy",
                        "territoryaims", "militias", "physint",
                        "percentpopurban", "coupdummy", "postcoldwar",
                        "lmtnest", "realgdp", "discrim", "exclpop",
                        "discpop", "elf",  "polrqnew", "egippolrqnew",
                        "poltrqnew", "egiptpolrqnew", "polity2sq")

m1 <- eba(y = "uamkstart", free = free.variables,
          doubtful = doubtful.variables, k = 0:4,
          data = df.cow2, vif = 7, draws = 50000,
          level = 0.9, se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/uamk-cow.rda")

# Ethnic conflict == 1
df.eth2 <- df5 %>% filter(ethnowarongoing == 1)
doubtful.variables <- c("assdummy", "totaltrade", "tradedependence",
                        "milper", "milex", "pop", "totalbeaths",
                        "guerrilladummy", "regtrans", "riotdummy",
                        "territoryaims", "militias", "physint",
                        "percentpopurban", "coupdummy", "postcoldwar",
                        "lmtnest", "realgdp", "discrim", "exclpop",
                        "discpop", "elf",  "polrqnew", "egippolrqnew",
                        "poltrqnew", "egiptpolrqnew", "polity2sq")

m1 <- eba(y = "uamkstart", free = free.variables,
          doubtful = doubtful.variables, k = 0:4,
          data = df.eth2, vif = 7, draws = 50000,
          level = 0.9, se.fun = se.clustered.robust)
save(m1, file = "~/Documents/mk/uamk-eth.rda")

######################
### Random forests ###
######################

# Load required package
library(h2o)
h2o.init(nthreads = -1, max_mem_size = "6G") # change min RAM size if necessary

df2a <- as.h2o(df2)

df2a$MKstart <- as.factor(df2a$MKstart)  #encode the binary response as a factor
h2o.levels(df2a$MKstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df2a,
                         ratios = 0.75, # train, validation
                         seed = 1234)  # reproducibility


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")
```

```
y <- "MKstart"
x <- setdiff(names(df2), c(y, "ccode", "year", "rgdppc",
                           "mksyr2", "mksyr3", "sf", "country",
                           "elf2", "polity2sq"))

#########################
### Running the models ###
#########################

rf <- h2o.grid("randomForest", x = x, y = y, training_frame = train,
               validation_frame = valid, grid_id = "grid01",
               hyper_params = list(ntrees = c(256, 512, 1024),
                                   max_depth = c(10, 20, 40),
                                   mtries = c(5, 6, 7),
                                   balance_classes = c(TRUE, FALSE),
                                   sample_rate = c(0.5, 0.632, 0.95),
                                   col_sample_rate_per_tree = c(0.5, 0.9, 1.0),
                                   histogram_type = "RoundRobin",
                                   seed = 1234))

# Saving the most accurate model
rf.grid <- h2o.getGrid(grid_id = "grid01",
                       sort_by = "auc",
                       decreasing = TRUE)

rf2 <- h2o.getModel(rf.grid@model_ids[[1]])
h2o.saveModel(rf2, path = "/Users/politicaltheory/Documents/GitHub/mass-killings-8k/data/")
summary(rf2)
h2o.varimp(rf2)
varimp <- as.data.frame(h2o.varimp(rf2))
h2o.varimp_plot(rf2)

# Second model
rf <- h2o.grid("randomForest", x = x, y = y, training_frame = train,
               validation_frame = valid, grid_id = "gridrf01b",
               hyper_params = list(ntrees = c(256, 512, 1024),
                                   max_depth = c(10, 20, 40),
                                   mtries = c(5, 6, 7),
                                   balance_classes = c(TRUE, FALSE),
                                   sample_rate = c(0.5, 0.632, 0.95),
                                   col_sample_rate_per_tree = c(0.5, 0.9, 1.0),
                                   histogram_type = "RoundRobin",
                                   seed = 4363))

# Saving the most accurate model
rf.grid <- h2o.getGrid(grid_id = "gridrf01b",
                       sort_by = "auc",
                       decreasing = TRUE)

rf2 <- h2o.getModel(rf.grid@model_ids[[1]])
h2o.saveModel(rf2, path = "/Users/politicaltheory/Documents/GitHub/mass-killings-8k/data/")
summary(rf2)
```

```
varimp <- as.data.frame(h2o.varimp(rf2))

# Third model
rf <- h2o.grid("randomForest", x = x, y = y, training_frame = train,
               validation_frame = valid, grid_id = "gridrf01c",
               hyper_params = list(ntrees = c(256, 512, 1024),
                                   max_depth = c(10, 20, 40),
                                   mtries = c(5, 6, 7),
                                   balance_classes = c(TRUE, FALSE),
                                   sample_rate = c(0.5, 0.632, 0.95),
                                   col_sample_rate_per_tree = c(0.5, 0.9, 1.0),
                                   histogram_type = "RoundRobin",
                                   seed = 7015))

# Saving the most accurate model
rf.grid <- h2o.getGrid(grid_id = "gridrf01c",
                       sort_by = "auc",
                       decreasing = TRUE)

rf2 <- h2o.getModel(rf.grid@model_ids[[1]])
h2o.saveModel(rf2, path = "/Users/politicaltheory/Documents/GitHub/mass-killings-8k/data/")
summary(rf2)
varimp <- as.data.frame(h2o.varimp(rf2))
h2o.varimp_plot(rf2)

##########################
### Ongoing civil wars ###
##########################

# UCDP == 1
df.ucdpa <- as.h2o(df.ucdp)

df.ucdpa$MKstart <- as.factor(df.ucdpa$MKstart)  #encode the binary repsonse as a factor
h2o.levels(df.ucdpa$MKstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df.ucdpa,
                         ratios = 0.75,
                         seed = 1234)


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "MKstart"
x <- setdiff(names(df.ucdp), c(y, "ccode", "year", "rgdppc",
                           "mksyr2", "mksyr3", "sf", "country",
                           "elf2", "polity2sq"))

# Running the model
rf <- h2o.grid("randomForest", x = x, y = y, training_frame = train,
               validation_frame = valid,  grid_id = "grid02",
               hyper_params = list(ntrees = c(256, 512, 1024),
```

```
                                    max_depth = c(10, 20, 40),
                                    mtries = c(5, 6, 7),
                                    balance_classes = c(TRUE, FALSE),
                                    sample_rate = c(0.5, 0.632, 0.95),
                                    col_sample_rate_per_tree = c(0.5, 0.9, 1.0),
                                    histogram_type = "RoundRobin",
                                    seed = 1234))

rf.grid <- h2o.getGrid(grid_id = "grid02",
                       sort_by = "auc",
                       decreasing = TRUE)
rf2 <- h2o.getModel(rf.grid@model_ids[[1]])
h2o.saveModel(rf2, path = "/Users/politicaltheory/Documents/GitHub/mass-killings-8k/data/")
summary(rf2)
h2o.varimp_plot(rf2)

# COW == 1
df.cowa <- as.h2o(df.cow)

df.cowa$MKstart <- as.factor(df.cowa$MKstart)  #encode the binary repsonse as a factor
h2o.levels(df.cowa$MKstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df.cowa,
                         ratios = 0.75,
                         seed = 1234)


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "MKstart"
x <- setdiff(names(df.ucdp), c(y, "ccode", "year", "rgdppc",
                               "mksyr2", "mksyr3", "sf", "country",
                               "elf2", "polity2sq"))

# Running the model
rf <- h2o.grid("randomForest", x = x, y = y, training_frame = train,
               validation_frame = valid, grid_id = "gridrf03",
               hyper_params = list(ntrees = c(256, 512, 1024),
                                    max_depth = c(10, 20, 40),
                                    mtries = c(5, 6, 7),
                                    balance_classes = c(TRUE, FALSE),
                                    sample_rate = c(0.5, 0.632, 0.95),
                                    col_sample_rate_per_tree = c(0.5, 0.9, 1.0),
                                    histogram_type = "RoundRobin",
                                    seed = 1234))

rf.grid <- h2o.getGrid(grid_id = "gridrf03",
                       sort_by = "auc",
                       decreasing = TRUE)
rf2 <- h2o.getModel(rf.grid@model_ids[[1]])
h2o.saveModel(rf2, path = "/Users/politicaltheory/Documents/GitHub/mass-killings-8k/data/")
```

```
summary(rf2)
varimp <- as.data.frame(h2o.varimp(rf2))
h2o.varimp_plot(rf2)

# Ethnic conflict == 1
df.etha <- as.h2o(df.eth)

df.etha$MKstart <- as.factor(df.etha$MKstart)  #encode the binary repsonse as a factor
h2o.levels(df.etha$MKstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df.etha,
                         ratios = 0.75,
                         seed = 1234)


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "MKstart"
x <- setdiff(names(df.eth), c(y, "ccode", "year", "rgdppc",
                              "mksyr2", "mksyr3", "sf", "country",
                              "elf2", "polity2sq"))

# Running the model
rf <- h2o.grid("randomForest", x = x, y = y, training_frame = train,
               validation_frame = valid, grid_id = "gridrf04",
               hyper_params = list(ntrees = c(256, 512, 1024),
                                   max_depth = c(10, 20, 40),
                                   mtries = c(5, 6, 7),
                                   balance_classes = c(TRUE, FALSE),
                                   sample_rate = c(0.5, 0.632, 0.95),
                                   col_sample_rate_per_tree = c(0.5, 0.9, 1.0),
                                   histogram_type = "RoundRobin",
                                   seed = 1234))

rf.grid <- h2o.getGrid(grid_id = "gridrf04",
                       sort_by = "auc",
                       decreasing = TRUE)
rf2 <- h2o.getModel(rf.grid@model_ids[[1]])
h2o.saveModel(rf2, path = "/Users/politicaltheory/Documents/GitHub/mass-killings-8k/data/")
summary(rf2)
varimp <- as.data.frame(h2o.varimp(rf2))
h2o.varimp_plot(rf2)

########################
### Cold War Period ###
########################
df2.coldwar2 <- as.h2o(df2.coldwar)

df2.coldwar2$MKstart <- as.factor(df2.coldwar2$MKstart)  #encode the binary repsonse as a factor
h2o.levels(df2.coldwar2$MKstart)
```

```
# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df2.coldwar2,
                         ratios = 0.75,
                         seed = 1234)


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "MKstart"
x <- setdiff(names(df.eth), c(y, "ccode", "year", "rgdppc",
                              "mksyr2", "mksyr3", "sf", "country",
                              "elf2", "polity2sq"))

rf <- h2o.grid("randomForest", x = x, y = y, training_frame = train,
               validation_frame = valid, grid_id = "gridrf04cw",
               hyper_params = list(ntrees = c(256, 512, 1024),
                                   max_depth = c(10, 20, 40),
                                   mtries = c(5, 6, 7),
                                   balance_classes = c(TRUE, FALSE),
                                   sample_rate = c(0.5, 0.632, 0.95),
                                   col_sample_rate_per_tree = c(0.5, 0.9, 1.0),
                                   histogram_type = "RoundRobin",
                                   seed = 1234))

rf.grid <- h2o.getGrid(grid_id = "gridrf04cw",
                       sort_by = "auc",
                       decreasing = TRUE)
rf2 <- h2o.getModel(rf.grid@model_ids[[1]])
h2o.saveModel(rf2, path = "/Users/politicaltheory/Documents/GitHub/mass-killings-8k/data/")
summary(rf2)

############################
### Post Cold War Period ###
############################
df2.postcoldwar2 <- as.h2o(df2.postcoldwar)

df2.postcoldwar2$MKstart <- as.factor(df2.postcoldwar2$MKstart)  #encode the binary repsonse as a fact
h2o.levels(df2.postcoldwar2$MKstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df2.postcoldwar2,
                         ratios = 0.75,
                         seed = 1234)


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "MKstart"
x <- setdiff(names(df.eth), c(y, "ccode", "year", "rgdppc",
                              "mksyr2", "mksyr3", "sf", "country",
                              "elf2", "polity2sq"))
```

```
rf <- h2o.grid("randomForest", x = x, y = y, training_frame = train,
               validation_frame = valid, grid_id = "gridrf04pcw",
               hyper_params = list(ntrees = c(256, 512, 1024),
                                   max_depth = c(10, 20, 40),
                                   mtries = c(5, 6, 7),
                                   balance_classes = c(TRUE, FALSE),
                                   sample_rate = c(0.5, 0.632, 0.95),
                                   col_sample_rate_per_tree = c(0.5, 0.9, 1.0),
                                   histogram_type = "RoundRobin",
                                   seed = 1234))

rf.grid <- h2o.getGrid(grid_id = "gridrf04pcw",
                       sort_by = "auc",
                       decreasing = TRUE)
rf2 <- h2o.getModel(rf.grid@model_ids[[1]])
h2o.saveModel(rf2, path = "/Users/politicaltheory/Documents/GitHub/mass-killings-8k/data/")
summary(rf2)

########################
### Only Peace Years ###
########################
df2.nowar <- as.h2o(df.nowar)

df2.nowar$MKstart <- as.factor(df2.nowar$MKstart)  #encode the binary repsonse as a factor
h2o.levels(df2.nowar$MKstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df2.nowar,
                         ratios = 0.75,
                         seed = 1234)


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "MKstart"
x <- setdiff(names(df.eth), c(y, "ccode", "year", "rgdppc",
                              "mksyr2", "mksyr3", "sf", "country",
                              "elf2", "polity2sq"))

rf <- h2o.grid("randomForest", x = x, y = y, training_frame = train,
               validation_frame = valid, grid_id = "gridrf04nowar",
               hyper_params = list(ntrees = c(256, 512, 1024),
                                   max_depth = c(10, 20, 40),
                                   mtries = c(5, 6, 7),
                                   balance_classes = c(TRUE, FALSE),
                                   sample_rate = c(0.5, 0.632, 0.95),
                                   col_sample_rate_per_tree = c(0.5, 0.9, 1.0),
                                   histogram_type = "RoundRobin",
                                   seed = 1234))

rf.grid <- h2o.getGrid(grid_id = "gridrf04nowar",
```

```
                      sort_by = "auc",
                      decreasing = TRUE)
rf2 <- h2o.getModel(rf.grid@model_ids[[1]])
h2o.saveModel(rf2, path = "/Users/politicaltheory/Documents/GitHub/mass-killings-8k/data/")
summary(rf2)

############################################################
## Same models with Genocide/Politicide variable (Harf) ###
############################################################
df5a <- as.h2o(df5)

df5a$uamkstart <- as.factor(df5a$uamkstart)  #encode the binary repsonse as a factor
h2o.levels(df5a$uamkstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df5a,
                         ratios = 0.75,
                         seed = 1234)


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "uamkstart"
x <- setdiff(names(df5), c(y, "ccode", "year", "rgdppc",
                           "uamkyr2", "uamkyr3", "sf", "country",
                           "elf2", "polity2sq"))

# Main model
rf <- h2o.grid("randomForest", x = x, y = y, training_frame = train,
               validation_frame = valid, grid_id = "gridrf05",
               hyper_params = list(ntrees = c(256, 512, 1024),
                                   max_depth = c(10, 20, 40),
                                   mtries = c(5, 6, 7),
                                   balance_classes = c(TRUE, FALSE),
                                   sample_rate = c(0.5, 0.632, 0.95),
                                   col_sample_rate_per_tree = c(0.5, 0.9, 1.0),
                                   histogram_type = "RoundRobin",
                                   seed = 1234))

# Saving the most accurate model
rf.grid <- h2o.getGrid(grid_id = "gridrf05",
                       sort_by = "auc",
                       decreasing = TRUE)

rf2 <- h2o.getModel(rf.grid@model_ids[[1]])
h2o.saveModel(rf2, path = "/Users/politicaltheory/Documents/GitHub/mass-killings-8k/data/")
summary(rf2)
varimp <- as.data.frame(h2o.varimp(rf2))
h2o.varimp_plot(rf2)

# UCDP == 1
df.ucdp2a <- as.h2o(df.ucdp2)
```

```
df.ucdp2a$uamkstart <- as.factor(df.ucdp2a$uamkstart)  #encode the binary repsonse as a factor
h2o.levels(df.ucdp2a$uamkstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df.ucdp2a,
                         ratios = 0.75,
                         seed = 1234)


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "uamkstart"
x <- setdiff(names(df.ucdp2), c(y, "ccode", "year", "rgdppc",
                           "uamkyr2", "uamkyr3", "sf", "country",
                           "elf2", "polity2sq"))

# Running the model
rf <- h2o.grid("randomForest", x = x, y = y, training_frame = train,
               validation_frame = valid, grid_id = "gridrf06",
               hyper_params = list(ntrees = c(256, 512, 1024),
                                   max_depth = c(10, 20, 40),
                                   mtries = c(5, 6, 7),
                                   balance_classes = c(TRUE, FALSE),
                                   sample_rate = c(0.5, 0.632, 0.95),
                                   col_sample_rate_per_tree = c(0.5, 0.9, 1.0),
                                   histogram_type = "RoundRobin",
                                   seed = 1234))

rf.grid <- h2o.getGrid(grid_id = "gridrf06",
                       sort_by = "auc",
                       decreasing = TRUE)
rf2 <- h2o.getModel(rf.grid@model_ids[[1]])
h2o.saveModel(rf2, path = "/Users/politicaltheory/Documents/GitHub/mass-killings-8k/data/")
summary(rf2)
varimp <- as.data.frame(h2o.varimp(rf2))
h2o.varimp_plot(rf2)

# COW == 1
df.cow2a <- as.h2o(df.cow2)

df.cow2a$uamkstart <- as.factor(df.cow2a$uamkstart)  #encode the binary repsonse as a factor
h2o.levels(df.cow2a$uamkstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df.cow2a,
                         ratios = 0.75,
                         seed = 1234)


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")
```

```
y <- "uamkstart"
x <- setdiff(names(df.cow2), c(y, "ccode", "year", "rgdppc",
                           "uamkyr2", "uamkyr3", "sf", "country",
                           "elf2", "polity2sq"))

# Running the model
rf <- h2o.grid("randomForest", x = x, y = y, training_frame = train,
               validation_frame = valid, grid_id = "gridrf07",
               hyper_params = list(ntrees = c(256, 512, 1024),
                                   max_depth = c(10, 20, 40),
                                   mtries = c(5, 6, 7),
                                   balance_classes = c(TRUE, FALSE),
                                   sample_rate = c(0.5, 0.632, 0.95),
                                   col_sample_rate_per_tree = c(0.5, 0.9, 1.0),
                                   histogram_type = "RoundRobin",
                                   seed = 1234))

rf.grid <- h2o.getGrid(grid_id = "gridrf07",
                       sort_by = "auc",
                       decreasing = TRUE)
rf2 <- h2o.getModel(rf.grid@model_ids[[1]])
h2o.saveModel(rf2, path = "/Users/politicaltheory/Documents/GitHub/mass-killings-8k/data/")
summary(rf2)
varimp <- as.data.frame(h2o.varimp(rf2))
h2o.varimp_plot(rf2)

# Ethnic conflict == 1
df.eth2a <- as.h2o(df.eth2)

df.eth2a$uamkstart <- as.factor(df.eth2a$uamkstart) #encode the binary repsonse as a factor
h2o.levels(df.eth2a$uamkstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df.eth2a,
                         ratios = 0.75,
                         seed = 1234)


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "uamkstart"
x <- setdiff(names(df.eth2), c(y, "ccode", "year", "rgdppc",
                           "uamkyr2", "uamkyr3", "sf", "country",
                           "elf2", "polity2sq"))

# Running the model
rf <- h2o.grid("randomForest", x = x, y = y, training_frame = train,
               validation_frame = valid, grid_id = "gridrf08",
               hyper_params = list(ntrees = c(256, 512, 1024),
                                   max_depth = c(10, 20, 40),
                                   mtries = c(5, 6, 7),
```

```
                                    balance_classes = c(TRUE, FALSE),
                                    sample_rate = c(0.5, 0.632, 0.95),
                                    col_sample_rate_per_tree = c(0.5, 0.9, 1.0),
                                    histogram_type = "RoundRobin",
                                    seed = 1234))

rf.grid <- h2o.getGrid(grid_id = "gridrf08",
                       sort_by = "auc",
                       decreasing = TRUE)
rf2 <- h2o.getModel(rf.grid@model_ids[[1]])
h2o.saveModel(rf2, path = "/Users/politicaltheory/Documents/GitHub/mass-killings-8k/data/")
summary(rf2)
varimp <- as.data.frame(h2o.varimp(rf2))
h2o.varimp_plot(rf2)

###############
### Graphs ###
###############

#################
# EBA Graphs ###
#################

# Main models
hist(m1, variables = c("logrgdppc", "polity2", "polity2sq", "uamkyr",
                       "UCDPcivilwarongoing",
                       "UCDPcivilwarstart", "COWcivilwarongoing",
                       "COWcivilwarstart", "ethnowarongoing", "ethnowarstart",
                       "assdummy", "totaltrade", "tradedependence", "milper",
                       "milex","pop", "totalbeaths", "guerrilladummy", "regtrans",
                       "riotdummy", "territoryaims", "militias", "physint",
                       "percentpopurban", "coupdummy", "postcoldwar",
                       "lmtnest", "realgdp", "discrim", "exclpop", "discpop",
                       "elf", "polrqnew", "egippolrqnew", "poltrqnew",
                       "egiptpolrqnew"),
     main = c("Log GDP capita", "Polity IV", "Polity IV^2", "Years last genocide",
              "UCDP ongoing", "UCDP onset", "COW ongoing", "COW onset",
              "Ethnic ongoing", "Ethnic onset", "Assassination", "Total trade",
            "Trade dependence", "Military personnel", "Military expenditure", "Population",
              "Total deaths", "Guerrilla", "Regime transition", "Riots",
              "Territory Aims", "Militias", "Physical integrity", "% Urban",
              "Coups", "Post-Cold War", "Mountainous terrain", "Real GDP",
              "Discrimination", "Excl pop", "Discrim pop", "ELF", "Groups/Eth relevant",
              "Group/Tot pop", "Inc groups/Eth relevant", "Inc groups/Tot pop"),
     density.col = "black", mu.col = "red3")

# Round
m1$coefficients$mean$beta2 <- round(as.numeric(m1$coefficients$mean$beta),4)
m1$coefficients$mean$se2 <- round(as.numeric(m1$coefficients$mean$se),4)
m1$coefficients$mean

## Models including only mass killings during civil wars
hist(m1, variables = c("logrgdppc", "polity2", "polity2sq", "uamkyr",
```

```
                        "assdummy", "totaltrade", "tradedependence", "milper",
                        "milex","pop", "totalbeaths", "guerrilladummy", "regtrans",
                        "riotdummy", "territoryaims", "militias", "physint",
                        "percentpopurban", "coupdummy", "postcoldwar",
                        "lmtnest", "realgdp", "discrim", "exclpop", "discpop",
                        "elf", "polrqnew", "egippolrqnew", "poltrqnew",
                        "egiptpolrqnew"),
     main = c("Log GDP capita", "Polity IV", "Polity IV^2", "Years last genocide",
               "Assassination", "Total trade",
             "Trade dependence", "Military personnel", "Military expenditure", "Population",
               "Total deaths", "Guerrilla", "Regime transition", "Riots",
               "Territory Aims", "Militias", "Physical integrity", "% Urban",
               "Coups", "Post-Cold War", "Mountainous terrain", "Real GDP",
               "Discrimination", "Excl pop", "Discrim pop", "ELF", "Groups/Eth relevant",
               "Groups/Tot pop", "Inc groups/Eth relevant", "Inc groups/Tot pop"),
     density.col = "black", mu.col = "red3")


# Cold War and Post-Cold War Periods
hist(m1, variables = c("logrgdppc", "polity2", "polity2sq", "UCDPcivilwarongoing",
                        "UCDPcivilwarstart",
                        "COWcivilwarongoing", "COWcivilwarstart",
                        "ethnowarongoing", "ethnowarstart", "assdummy",
                        "totaltrade", "tradedependence", "cinc",
                        "totalbeaths", "guerrilladummy", "regtrans",
                        "riotdummy", "territoryaims", "militias",
                        "physint", "percentpopurban", "coupdummy",
                        "lmtnest", "realgdp", "discrim",
                        "exclpop", "discpop", "elf",  "polrqnew",
                        "egippolrqnew", "poltrqnew", "egiptpolrqnew"),
     main = c("Log GDP capita", "Polity IV", "Polity IV^2", "Years mass killings",
               "UCDP ongoing", "UCDP onset", "COW ongoing", "COW onset",
               "Ethnic ongoing", "Ethnic onset", "Assassination", "Total trade",
               "Trade dependence", "CINC",
               "Total deaths", "Guerrilla", "Regime transition", "Riots",
               "Territory Aims", "Militias", "Physical integrity", "% Urban",
               "Coups", "Mountainous terrain", "Real GDP",
               "Discrimination", "Excl pop", "Discrim pop", "ELF", "Groups/Eth relevant",
               "Group/Tot pop", "Inc groups/Eth relevant", "Inc groups/Tot pop"),
     density.col = "black", mu.col = "red3")

#### Peacetime
hist(m1, variables = c("logrgdppc", "polity2", "polity2sq", "mksyr", "assdummy",
                        "totaltrade", "tradedependence", "milper", "milex", "pop",
                        "totalbeaths", "guerrilladummy",
                        "riotdummy", "territoryaims", "militias",
                        "physint", "percentpopurban", "coupdummy", "postcoldwar",
                        "lmtnest", "realgdp", "discrim",
                        "exclpop", "discpop", "elf",  "polrqnew",
                        "egippolrqnew", "poltrqnew", "egiptpolrqnew"),
    main = c("Log GDP capita", "Polity IV", "Polity IV^2", "Years mass killings","Assassination", "Total
            "Trade dependence", "Military Personnel", "Military Expenditure", "Population",
               "Total deaths", "Guerrilla", "Previous riots",
```

```
              "Territory Aims", "Militias", "Physical integrity", "% Urban",
              "Coups", "Post-Cold War", "Mountainous terrain", "Real GDP",
              "Discrimination", "Excl pop", "Discrim pop", "ELF", "Groups/Eth relevant",
              "Group/Tot pop", "Inc groups/Eth relevant", "Inc groups/Tot pop"),
     density.col = "black", mu.col = "red3")

######################
### Random forests ###
######################

# Main model
library(h2o)
h2o.init(nthreads = -1, max_mem_size = "6G")
a <- h2o.loadModel("grid01_model_197")
print(va <- a %>% h2o.varimp() %>% as.data.frame() %>% head(., 10))

df2a <- as.h2o(df2)

df2a$MKstart <- as.factor(df2a$MKstart)  #encode the binary repsonse as a factor
h2o.levels(df2a$MKstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df2a,
                         ratios = 0.75,  # 70%, 15%, 15%
                         seed = 1234)  # reproducibility


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "MKstart"
x <- setdiff(names(df2), c(y, "ccode", "year", "rgdppc",
                           "mksyr2", "mksyr3", "sf", "country",
                           "elf2", "polity2sq"))

# Variable Importance
par(mgp=c(2.2,0.45,0), tcl=-0.4, mar=c(2,7.5,1,1))
barplot(va$scaled_importance[10:1],
        horiz = TRUE, las = 1, cex.names=0.9,
        names.arg = c("Real GDP",
                      "Total trade",
                      "Polity IV",
                      "Population",
                      "Military expenditure",
                      "Military personnel",
                      "Trade dependence",
                      "% Urban pop.",
                      "Log GDP per capita",
                      "Years mass killing"),
        main = "")


# Partial dependence plots
```

```
mksyr <- h2o.partialPlot(object = a, data = train, cols = c("mksyr"), plot_stddev = F)
p1 <- qplot(mksyr$mksyr, mksyr$mean_response) + geom_line() + theme_classic() +
        xlab("Years mass killing") +  ylab("Mean response")

logrgdppc <- h2o.partialPlot(object = a, data = train, cols = c("logrgdppc"), plot_stddev = F)
p2 <- qplot(logrgdppc$logrgdppc, logrgdppc$mean_response) + geom_line() + theme_classic() +
  xlab("Log GDP per capita") +  ylab("Mean response")

percentpopurban <- h2o.partialPlot(object = a, data = train, cols = c("percentpopurban"), plot_stddev =
p3 <- qplot(percentpopurban$percentpopurban, percentpopurban$mean_response) + geom_line() +
        theme_classic() + xlab("% Urban pop.") + ylab("Mean response")

tradedependence <- h2o.partialPlot(object = a, data = train, cols = c("tradedependence"), plot_stddev =
p4 <- qplot(tradedependence$tradedependence, tradedependence$mean_response) + geom_line() +
        theme_classic() + xlab("Trade dependence") + ylab("Mean response")

milper <- h2o.partialPlot(object = a, data = train, cols = c("milper"), plot_stddev = F)
p5 <- qplot(milper$milper, milper$mean_response) + geom_line() + theme_classic() +
        xlab("Military personnel") + ylab("Mean response")

milex <- h2o.partialPlot(object = a, data = train, cols = c("milex"), plot_stddev = F)
p6 <- qplot(milex$milex, milex$mean_response) + geom_line() + theme_classic() +
  xlab("Military expenditure") + ylab("Mean response")

pop <- h2o.partialPlot(object = a, data = train, cols = c("pop"), plot_stddev = F)
p7 <- qplot(pop$pop, pop$mean_response) + geom_line() + theme_classic() +
        xlab("Population") + ylab("Mean response")

polity2 <- h2o.partialPlot(object = a, data = train, cols = c("polity2"), plot_stddev = F)
p8 <- qplot(polity2$polity2, polity2$mean_response) + geom_line() + theme_classic() +
  xlab("Polity IV") + ylab("Mean response")

totaltrade <- h2o.partialPlot(object = a, data = train, cols = c("totaltrade"), plot_stddev = F)
p9 <- qplot(totaltrade$totaltrade, totaltrade$mean_response) + geom_line() + theme_classic() +
  xlab("Total trade") + ylab("Mean response")

realgdp <- h2o.partialPlot(object = a, data = train, cols = c("realgdp"), plot_stddev = F)
p10 <- qplot(realgdp$realgdp, realgdp$mean_response) + geom_line() + theme_classic() +
  xlab("Real GDP") + ylab("Mean response")

# Multiplot function: http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
        library(grid)

        # Make a list from the ... arguments and plotlist
        plots <- c(list(...), plotlist)

        numPlots = length(plots)

        # If layout is NULL, then use 'cols' to determine layout
        if (is.null(layout)) {
                # Make the panel
                # ncol: Number of columns of plots
```

```
                # nrow: Number of rows needed, calculated from # of cols
                layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                                 ncol = cols, nrow = ceiling(numPlots/cols))
        }

        if (numPlots==1) {
                print(plots[[1]])

        } else {
                # Set up the page
                grid.newpage()
                pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

                # Make each plot, in the correct location
                for (i in 1:numPlots) {
                    # Get the i,j matrix positions of the regions that contain this subplot
                         matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

                         print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                                         layout.pos.col = matchidx$col))
                }
        }
}

multiplot(p1,p5,p8,p2,p6,p9,p3,p7,p10,p4, cols = 4)  # 11.1x5.14 in

#######################################
### Mass killings during civil wars ###
#######################################

# UCDP == 1
a <- h2o.loadModel("grid02_model_349")
print(va <- a %>% h2o.varimp() %>% as.data.frame() %>% head(., 10))

par(mgp=c(2.2,0.45,0), tcl=-0.4, mar=c(2,7.5,1,1))
barplot(va$scaled_importance[10:1],
        horiz = TRUE, las = 1, cex.names=0.9,
        names.arg = c("Real GDP",
                      "Military personnel",
                      "Population",
                      "Military expenditure",
                      "Total trade",
                      "Log GDP per capita",
                      "CINC",
                      "Years mass killing",
                      "Trade dependence",
                      "% Urban pop."),
        main = "")

df.ucdpa <- as.h2o(df.ucdp)

df.ucdpa$MKstart <- as.factor(df.ucdpa$MKstart)  #encode the binary repsonse as a factor
h2o.levels(df.ucdpa$MKstart)
```

```
# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df.ucdpa,
                         ratios = 0.75,  # 70%, 15%, 15%
                         seed = 1234)  # reproducibility


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "MKstart"
x <- setdiff(names(df.ucdp), c(y, "ccode", "year", "rgdppc",
                               "mksyr2", "mksyr3", "sf", "country",
                               "elf2", "polity2sq"))

percentpopurban <- h2o.partialPlot(object = a, data = train, cols = c("percentpopurban"), plot_stddev =
p1 <- qplot(percentpopurban$percentpopurban, percentpopurban$mean_response) + geom_line() +
        theme_classic() + xlab("% Urban") + ylab("Mean response")

tradedependence <- h2o.partialPlot(object = a, data = train, cols = c("tradedependence"), plot_stddev =
p2 <- qplot(tradedependence$tradedependence, tradedependence$mean_response) + geom_line() +
  theme_classic() + xlab("Trade dependence") + ylab("Mean response")

mksyr <- h2o.partialPlot(object = a, data = train, cols = c("mksyr"), plot_stddev = F)
p3 <- qplot(mksyr$mksyr, mksyr$mean_response) + geom_line() + theme_classic() +
  xlab("Years since mass killing") +  ylab("Mean response")

cinc <- h2o.partialPlot(object = a, data = train, cols = c("cinc"), plot_stddev = F)
p4 <- qplot(cinc$cinc, cinc$mean_response) + geom_line() + theme_classic() +
  xlab("CINC") +  ylab("Mean response")

logrgdppc <- h2o.partialPlot(object = a, data = train, cols = c("logrgdppc"), plot_stddev = F)
p5 <- qplot(logrgdppc$logrgdppc, logrgdppc$mean_response) + geom_line() + theme_classic() +
        xlab("Log GDP per capita") + ylab("Mean response")

totaltrade <- h2o.partialPlot(object = a, data = train, cols = c("totaltrade"), plot_stddev = F)
p6 <- qplot(totaltrade$totaltrade, totaltrade$mean_response) + geom_line() + theme_classic() +
        xlab("Total trade") + ylab("Mean response")

milex <- h2o.partialPlot(object = a, data = train, cols = c("milex"), plot_stddev = F)
p7 <- qplot(milex$milex, milex$mean_response) + geom_line() + theme_classic() +
  xlab("Military expenditure") + ylab("Mean response")

pop <- h2o.partialPlot(object = a, data = train, cols = c("pop"), plot_stddev = F)
p8 <- qplot(pop$pop, pop$mean_response) + geom_line() + theme_classic() +
  xlab("Population") + ylab("Mean response")

milper <- h2o.partialPlot(object = a, data = train, cols = c("milper"), plot_stddev = F)
p9 <- qplot(milper$milper, milper$mean_response) + geom_line() + theme_classic() +
  xlab("Military personnel") + ylab("Mean response")

realgdp <- h2o.partialPlot(object = a, data = train, cols = c("realgdp"), plot_stddev = F)
p10 <- qplot(realgdp$realgdp, realgdp$mean_response) + geom_line() + theme_classic() +
```

```
  xlab("Real GDP") + ylab("Mean response")

multiplot(p1,p5,p8,p2,p6,p9,p3,p7,p10,p4, cols = 4)  # 11.09x5.14 in


# COW == 1
a <- h2o.loadModel("gridrf03_model_41")
print(va <- a %>% h2o.varimp() %>% as.data.frame() %>% head(., 10))

par(mgp=c(2.2,0.45,0), tcl=-0.4, mar=c(2,7.5,1,1))
barplot(va$scaled_importance[10:1],
        horiz = TRUE, las = 1, cex.names=0.9,
        names.arg = c("Polarisation",
                      "Military expenditure",
                      "Military personnel",
                      "Excluded population",
                      "% Urban",
                      "Previous riots",
                      "Total battle deaths",
                      "Log GDP per capita",
                      "Years mass killing",
                      "Physical integrity"),
        main = "")

df.cowa <- as.h2o(df.cow)

df.cowa$MKstart <- as.factor(df.cowa$MKstart)  #encode the binary repsonse as a factor
h2o.levels(df.cowa$MKstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df.cowa,
                         ratios = 0.75,
                         seed = 1234)


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "MKstart"
x <- setdiff(names(df.ucdp), c(y, "ccode", "year", "rgdppc",
                               "mksyr2", "mksyr3", "sf", "country",
                               "elf2", "polity2sq"))

physint <- h2o.partialPlot(object = a, data = train, cols = c("physint"), plot_stddev = F)
p1 <- qplot(physint$physint, physint$mean_response) + geom_line() + theme_classic() +
        xlab("Physical integrity") +  ylab("Mean response")

mksyr <- h2o.partialPlot(object = a, data = train, cols = c("mksyr"), plot_stddev = F)
p2 <- qplot(mksyr$mksyr, mksyr$mean_response) + geom_line() + theme_classic() +
  xlab("Years mass killing") +  ylab("Mean response")

logrgdppc <- h2o.partialPlot(object = a, data = train, cols = c("logrgdppc"), plot_stddev = F)
p3 <- qplot(logrgdppc$logrgdppc, logrgdppc$mean_response) + geom_line() + theme_classic() +
```

```
        xlab("Log GDP per capita") + ylab("Mean response")

totalbeaths <- h2o.partialPlot(object = a, data = train, cols = c("totalbeaths"), plot_stddev = F)
p4 <- qplot(totalbeaths$totalbeaths, totalbeaths$mean_response) + geom_line() + theme_classic() +
  xlab("Total battle deaths") + ylab("Mean response")

riotdummy <- h2o.partialPlot(object = a, data = train, cols = c("riotdummy"), plot_stddev = F)
p5 <- qplot(riotdummy$riotdummy, riotdummy$mean_response) + geom_line() + theme_classic() +
        xlab("Previous riots") + ylab("Mean response")

percentpopurban <- h2o.partialPlot(object = a, data = train, cols = c("percentpopurban"), plot_stddev =
p6 <- qplot(percentpopurban$percentpopurban, percentpopurban$mean_response) + geom_line() +
  theme_classic() + xlab("% Urban") + ylab("Mean response")

exclpop <- h2o.partialPlot(object = a, data = train, cols = c("exclpop"), plot_stddev = F)
p7 <- qplot(exclpop$exclpop, exclpop$mean_response) + geom_line() +
  theme_classic() + xlab("Excluded population") + ylab("Mean response")

milper <- h2o.partialPlot(object = a, data = train, cols = c("milper"), plot_stddev = F)
p8 <- qplot(milper$milper, milper$mean_response) + geom_line() + theme_classic() +
  xlab("Military personnel") + ylab("Mean response")

milex <- h2o.partialPlot(object = a, data = train, cols = c("milex"), plot_stddev = F)
p9 <- qplot(milex$milex, milex$mean_response) + geom_line() + theme_classic() +
  xlab("Military expenditure") + ylab("Mean response")

egiptpolrqnew <- h2o.partialPlot(object = a, data = train, cols = c("egiptpolrqnew"), plot_stddev = F)
p10 <- qplot(egiptpolrqnew$egiptpolrqnew, egiptpolrqnew$mean_response) + geom_line() + theme_classic(
  xlab("Polarisation") + ylab("Mean response")

multiplot(p1,p5,p8,p2,p6,p9,p3,p7,p10,p4, cols = 4)  # 11.09x5.14 in

# Ethnic conflict == 1
a <- h2o.loadModel("gridrf04_model_52")
print(va <- a %>% h2o.varimp() %>% as.data.frame() %>% head(., 10))

par(mgp=c(2.2,0.45,0), tcl=-0.4, mar=c(2,7.5,1,1))
barplot(va$scaled_importance[10:1],
        horiz = TRUE, las = 1, cex.names=0.9,
        names.arg = c("Log GDP per capita",
                      "Democracy",
                      "Years mass killing",
                      "Polarisation",
                      "% Urban",
                      "Territorial aims",
                      "Trade dependence",
                      "Excluded population",
                      "Military personnel",
                      "Polity IV"),
        main = "")

df.etha <- as.h2o(df.eth)
```

```
df.etha$MKstart <- as.factor(df.etha$MKstart)  #encode the binary repsonse as a factor
h2o.levels(df.etha$MKstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df.etha,
                         ratios = 0.75,
                         seed = 42)


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "MKstart"
x <- setdiff(names(df.ucdp), c(y, "ccode", "year", "rgdppc",
                               "mksyr2", "mksyr3", "sf", "country",
                               "elf2", "polity2sq"))

polity2 <- h2o.partialPlot(object = a, data = train, cols = c("polity2"), plot_stddev = F)
p1 <- qplot(polity2$polity2, polity2$mean_response) + geom_line() + theme_classic() +
        xlab("Polity IV") +  ylab("Mean response")

milper <- h2o.partialPlot(object = a, data = train, cols = c("milper"), plot_stddev = F)
p2 <- qplot(milper$milper, milper$mean_response) + geom_line() + theme_classic() +
  xlab("Military personnel") + ylab("Mean response")

exclpop <- h2o.partialPlot(object = a, data = train, cols = c("exclpop"), plot_stddev = F)
p3 <- qplot(exclpop$exclpop, exclpop$mean_response) + geom_line() +
  theme_classic() + xlab("Excluded population") + ylab("Mean response")

tradedependence <- h2o.partialPlot(object = a, data = train, cols = c("tradedependence"), plot_stddev =
p4 <- qplot(tradedependence$tradedependence, tradedependence$mean_response) + geom_line() +
  theme_classic() + xlab("Trade dependence") + ylab("Mean response")

territoryaims <- h2o.partialPlot(object = a, data = train, cols = c("territoryaims"), plot_stddev = F)
p5 <- qplot(territoryaims$territoryaims, territoryaims$mean_response) + geom_line() +
  theme_classic() + xlab("Territory aims") + ylab("Mean response")

percentpopurban <- h2o.partialPlot(object = a, data = train, cols = c("percentpopurban"), plot_stddev =
p6 <- qplot(percentpopurban$percentpopurban, percentpopurban$mean_response) + geom_line() +
  theme_classic() + xlab("% Urban") + ylab("Mean response")

egiptpolrqnew <- h2o.partialPlot(object = a, data = train, cols = c("egiptpolrqnew"), plot_stddev = F)
p7 <- qplot(egiptpolrqnew$egiptpolrqnew, egiptpolrqnew$mean_response) + geom_line() + theme_classic()
  xlab("Polarisation") + ylab("Mean response")

mksyr <- h2o.partialPlot(object = a, data = train, cols = c("mksyr"), plot_stddev = F)
p8 <- qplot(mksyr$mksyr, mksyr$mean_response) + geom_line() + theme_classic() +
  xlab("Years mass killing") +  ylab("Mean response")

demdummy <- h2o.partialPlot(object = a, data = train, cols = c("demdummy"), plot_stddev = F)
p9 <- qplot(demdummy$demdummy, demdummy$mean_response) + geom_line() + theme_classic() +
  xlab("Democracy") +  ylab("Mean response")
```

```
logrgdppc <- h2o.partialPlot(object = a, data = train, cols = c("logrgdppc"), plot_stddev = F)
p10 <- qplot(logrgdppc$logrgdppc, logrgdppc$mean_response) + geom_line() + theme_classic() +
  xlab("Log GDP per capita") + ylab("Mean response")

multiplot(p1,p5,p8,p2,p6,p9,p3,p7,p10,p4, cols = 4)  # 11.09x5.14 in

########################
### Different seeds ###
########################

## Seed 4363
a <- h2o.loadModel("gridrf01b_model_73")
print(va <- a %>% h2o.varimp() %>% as.data.frame() %>% head(., 10))

df2a <- as.h2o(df2)

df2a$MKstart <- as.factor(df2a$MKstart)  #encode the binary repsonse as a factor
h2o.levels(df2a$MKstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df2a,
                         ratios = 0.75,  # 70%, 15%, 15%
                         seed = 1234)  # reproducibility


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "MKstart"
x <- setdiff(names(df2), c(y, "ccode", "year", "rgdppc",
                           "mksyr2", "mksyr3", "sf", "country",
                           "elf2", "polity2sq"))

# Variable Importance
par(mgp=c(2.2,0.45,0), tcl=-0.4, mar=c(2,7.5,1,1))
barplot(va$scaled_importance[10:1],
        horiz = TRUE, las = 1, cex.names=0.9,
        names.arg = c("Total trade",
                      "Real GDP",
                      "CINC",
                      "Population",
                      "Military personnel",
                      "Military expenditure",
                      "Years mass killing",
                      "% Urban pop.",
                      "Trade dependence",
                      "Log GDP per capita"),
        main = "")


# Partial dependence plots
logrgdppc <- h2o.partialPlot(object = a, data = train, cols = c("logrgdppc"), plot_stddev = F)
p1 <- qplot(logrgdppc$logrgdppc, logrgdppc$mean_response) + geom_line() + theme_classic() +
```

```
  xlab("Log GDP per capita") +  ylab("Mean response")

tradedependence <- h2o.partialPlot(object = a, data = train, cols = c("tradedependence"), plot_stddev =
p2 <- qplot(tradedependence$tradedependence, tradedependence$mean_response) + geom_line() +
  theme_classic() + xlab("Trade dependence") + ylab("Mean response")

percentpopurban <- h2o.partialPlot(object = a, data = train, cols = c("percentpopurban"), plot_stddev =
p3 <- qplot(percentpopurban$percentpopurban, percentpopurban$mean_response) + geom_line() +
  theme_classic() + xlab("% Urban pop.") + ylab("Mean response")

mksyr <- h2o.partialPlot(object = a, data = train, cols = c("mksyr"), plot_stddev = F)
p4 <- qplot(mksyr$mksyr, mksyr$mean_response) + geom_line() + theme_classic() +
  xlab("Years mass killing") +  ylab("Mean response")

milex <- h2o.partialPlot(object = a, data = train, cols = c("milex"), plot_stddev = F)
p5 <- qplot(milex$milex, milex$mean_response) + geom_line() + theme_classic() +
  xlab("Military expenditure") + ylab("Mean response")

milper <- h2o.partialPlot(object = a, data = train, cols = c("milper"), plot_stddev = F)
p6 <- qplot(milper$milper, milper$mean_response) + geom_line() + theme_classic() +
  xlab("Military personnel") + ylab("Mean response")

pop <- h2o.partialPlot(object = a, data = train, cols = c("pop"), plot_stddev = F)
p7 <- qplot(pop$pop, pop$mean_response) + geom_line() + theme_classic() +
  xlab("Population") + ylab("Mean response")

cinc <- h2o.partialPlot(object = a, data = train, cols = c("cinc"), plot_stddev = F)
p8 <- qplot(cinc$cinc, cinc$mean_response) + geom_line() + theme_classic() +
  xlab("CINC") + ylab("Mean response")

realgdp <- h2o.partialPlot(object = a, data = train, cols = c("realgdp"), plot_stddev = F)
p9 <- qplot(realgdp$realgdp, realgdp$mean_response) + geom_line() + theme_classic() +
  xlab("Real GDP") + ylab("Mean response")

totaltrade <- h2o.partialPlot(object = a, data = train, cols = c("totaltrade"), plot_stddev = F)
p10 <- qplot(totaltrade$totaltrade, totaltrade$mean_response) + geom_line() + theme_classic() +
  xlab("Total trade") + ylab("Mean response")

multiplot(p1,p5,p8,p2,p6,p9,p3,p7,p10,p4, cols = 4)  # 11.09x5.14 in

## Seed 7015

a <- h2o.loadModel("gridrf01c_model_409")
print(va <- a %>% h2o.varimp() %>% as.data.frame() %>% head(., 10))

df2a <- as.h2o(df2)

df2a$MKstart <- as.factor(df2a$MKstart)  #encode the binary repsonse as a factor
h2o.levels(df2a$MKstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df2a,
                         ratios = 0.75,  # 70%, 15%, 15%
```

```
                        seed = 1234)  # reproducibility


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "MKstart"
x <- setdiff(names(df2), c(y, "ccode", "year", "rgdppc",
                           "mksyr2", "mksyr3", "sf", "country",
                           "elf2", "polity2sq"))

# Variable Importance
par(mgp=c(2.2,0.45,0), tcl=-0.4, mar=c(2,7.5,1,1))
barplot(va$scaled_importance[10:1],
        horiz = TRUE, las = 1, cex.names=0.9,
        names.arg = c("Total trade",
                      "Real GDP",
                      "Military personnel",
                      "CINC",
                      "Population",
                      "Military expenditure",
                      "% Urban pop.",
                      "Years mass killing",
                      "Trade dependence",
                      "Log GDP per capita"),
        main = "")


# Partial dependence plots
logrgdppc <- h2o.partialPlot(object = a, data = train, cols = c("logrgdppc"), plot_stddev = F)
p1 <- qplot(logrgdppc$logrgdppc, logrgdppc$mean_response) + geom_line() + theme_classic() +
  xlab("Log GDP per capita") +  ylab("Mean response")

tradedependence <- h2o.partialPlot(object = a, data = train, cols = c("tradedependence"), plot_stddev =
p2 <- qplot(tradedependence$tradedependence, tradedependence$mean_response) + geom_line() +
  theme_classic() + xlab("Trade dependence") + ylab("Mean response")

mksyr <- h2o.partialPlot(object = a, data = train, cols = c("mksyr"), plot_stddev = F)
p3 <- qplot(mksyr$mksyr, mksyr$mean_response) + geom_line() + theme_classic() +
  xlab("Years mass killing") +  ylab("Mean response")

percentpopurban <- h2o.partialPlot(object = a, data = train, cols = c("percentpopurban"), plot_stddev =
p4 <- qplot(percentpopurban$percentpopurban, percentpopurban$mean_response) + geom_line() +
  theme_classic() + xlab("% Urban pop.") + ylab("Mean response")

milex <- h2o.partialPlot(object = a, data = train, cols = c("milex"), plot_stddev = F)
p5 <- qplot(milex$milex, milex$mean_response) + geom_line() + theme_classic() +
  xlab("Military expenditure") + ylab("Mean response")

pop <- h2o.partialPlot(object = a, data = train, cols = c("pop"), plot_stddev = F)
p6 <- qplot(pop$pop, pop$mean_response) + geom_line() + theme_classic() +
  xlab("Population") + ylab("Mean response")
```

```
cinc <- h2o.partialPlot(object = a, data = train, cols = c("cinc"), plot_stddev = F)
p7 <- qplot(cinc$cinc, cinc$mean_response) + geom_line() + theme_classic() +
  xlab("CINC") + ylab("Mean response")

milper <- h2o.partialPlot(object = a, data = train, cols = c("milper"), plot_stddev = F)
p8 <- qplot(milper$milper, milper$mean_response) + geom_line() + theme_classic() +
  xlab("Military personnel") + ylab("Mean response")

realgdp <- h2o.partialPlot(object = a, data = train, cols = c("realgdp"), plot_stddev = F)
p9 <- qplot(realgdp$realgdp, realgdp$mean_response) + geom_line() + theme_classic() +
  xlab("Real GDP") + ylab("Mean response")

totaltrade <- h2o.partialPlot(object = a, data = train, cols = c("totaltrade"), plot_stddev = F)
p10 <- qplot(totaltrade$totaltrade, totaltrade$mean_response) + geom_line() + theme_classic() +
  xlab("Total trade") + ylab("Mean response")

multiplot(p1,p5,p8,p2,p6,p9,p3,p7,p10,p4, cols = 4)  # 11.09x5.14 in

#### Cold War Period
# Variable Importance
a <- h2o.loadModel("data/gridrf04cw_model_100")
va <- h2o.varimp(a)

par(mgp=c(2.2,0.45,0), tcl=-0.4, mar=c(2,7.5,1,1))
barplot(va$scaled_importance[10:1],
        horiz = TRUE, las = 1, cex.names=0.9,
        names.arg = c("Population",
                      "Trade dependence",
                      "Polarisation",
                      "Previous riots",
                      "Years mass killing",
                      "Ethnic frac.",
                      "Excluded pop.",
                      "Mountainous terrain",
                      "Log GDP per capita",
                      "Polity IV"),
        main = "")


# Partial dependence plots

polity2 <- h2o.partialPlot(object = a, data = train, cols = c("polity2"), plot_stddev = F)
p1 <- qplot(polity2$polity2, polity2$mean_response) + geom_line() + theme_classic() +
  xlab("Polity IV") + ylab("Mean response")

mksyr <- h2o.partialPlot(object = a, data = train, cols = c("mksyr"), plot_stddev = F)
p6 <- qplot(mksyr$mksyr, mksyr$mean_response) + geom_line() + theme_classic() +
  xlab("Years mass killing") +  ylab("Mean response")

logrgdppc <- h2o.partialPlot(object = a, data = train, cols = c("logrgdppc"), plot_stddev = F)
p2 <- qplot(logrgdppc$logrgdppc, logrgdppc$mean_response) + geom_line() + theme_classic() +
  xlab("Log GDP per capita") +  ylab("Mean response")
```

```
lmtnest <- h2o.partialPlot(object = a, data = train, cols = c("lmtnest"), plot_stddev = F)
p3 <- qplot(lmtnest$lmtnest, lmtnest$mean_response) + geom_line() +
  theme_classic() + xlab("Mountainous terrain") + ylab("Mean response")

exclpop <- h2o.partialPlot(object = a, data = train, cols = c("exclpop"), plot_stddev = F)
p4 <- qplot(exclpop$exclpop, exclpop$mean_response) + geom_line() +
  theme_classic() + xlab("Excluded pop.") + ylab("Mean response")

elf <- h2o.partialPlot(object = a, data = train, cols = c("elf"), plot_stddev = F)
p5 <- qplot(elf$elf, elf$mean_response) + geom_line() + theme_classic() +
  xlab("Military personnel") + ylab("Mean response")

riotdummy <- h2o.partialPlot(object = a, data = train, cols = c("riotdummy"), plot_stddev = F)
p7 <- qplot(riotdummy$riotdummy, riotdummy$mean_response) + geom_line() + theme_classic() +
  xlab("Previous riots") + ylab("Mean response")

polrqnew <- h2o.partialPlot(object = a, data = train, cols = c("polrqnew"), plot_stddev = F)
p8 <- qplot(polrqnew$polrqnew, polrqnew$mean_response) + geom_line() + theme_classic() +
  xlab("Polarisation") + ylab("Mean response")

pop <- h2o.partialPlot(object = a, data = train, cols = c("pop"), plot_stddev = F)
p10 <- qplot(pop$pop, pop$mean_response) + geom_line() + theme_classic() +
  xlab("Population") + ylab("Mean response")

tradedependence <- h2o.partialPlot(object = a, data = train, cols = c("tradedependence"), plot_stddev =
p9 <- qplot(tradedependence$tradedependence, tradedependence$mean_response) + geom_line() + theme_cla
  xlab("Trade dependence") + ylab("Mean response")


multiplot(p1,p5,p8,p2,p6,p9,p3,p7,p10,p4, cols = 4)  # 11.09x5.14 in

#### Post-Cold War Period
# Variable Importance
a <- h2o.loadModel("data/gridrf04pcw_model_459")
va <- h2o.varimp(a)

par(mgp=c(2.2,0.45,0), tcl=-0.4, mar=c(2,7.5,1,1))
barplot(va$scaled_importance[10:1],
        horiz = TRUE, las = 1, cex.names=0.9,
        names.arg = c("Trade dependence",
                      "Military expenditures",
                      "Polity IV",
                      "Military personnel",
                      "Mountainous terrain",
                      "Polarisation",
                      "Log GDP per capita",
                      "Years mass killing",
                      "Total battle deaths",
                      "% Urban"),
        main = "")


# Partial dependence plots
```

```
polity2 <- h2o.partialPlot(object = a, data = train, cols = c("polity2"), plot_stddev = F)
p8 <- qplot(polity2$polity2, polity2$mean_response) + geom_line() + theme_classic() +
  xlab("Polity IV") + ylab("Mean response")

mksyr <- h2o.partialPlot(object = a, data = train, cols = c("mksyr"), plot_stddev = F)
p3 <- qplot(mksyr$mksyr, mksyr$mean_response) + geom_line() + theme_classic() +
  xlab("Years mass killing") +  ylab("Mean response")

logrgdppc <- h2o.partialPlot(object = a, data = train, cols = c("logrgdppc"), plot_stddev = F)
p4 <- qplot(logrgdppc$logrgdppc, logrgdppc$mean_response) + geom_line() + theme_classic() +
  xlab("Log GDP per capita") +  ylab("Mean response")

percentpopurban <- h2o.partialPlot(object = a, data = train, cols = c("percentpopurban"), plot_stddev =
p1 <- qplot(percentpopurban$percentpopurban, percentpopurban$mean_response) + geom_line() +
  theme_classic() + xlab("% Urban") + ylab("Mean response")

totalbeaths <- h2o.partialPlot(object = a, data = train, cols = c("totalbeaths"), plot_stddev = F)
p2 <- qplot(totalbeaths$totalbeaths,totalbeaths$mean_response) + geom_line() +
  theme_classic() + xlab("Total battle deaths") + ylab("Mean response")

egippolrqnew <- h2o.partialPlot(object = a, data = train, cols = c("egippolrqnew"), plot_stddev = F)
p5 <- qplot(egippolrqnew$egippolrqnew, egippolrqnew$mean_response) + geom_line() + theme_classic() +
  xlab("Polarisation") + ylab("Mean response")

lmtnest <- h2o.partialPlot(object = a, data = train, cols = c("lmtnest"), plot_stddev = F)
p6 <- qplot(lmtnest$lmtnest, lmtnest$mean_response) + geom_line() +
  theme_classic() + xlab("Mountainous terrain") + ylab("Mean response")

milper <- h2o.partialPlot(object = a, data = train, cols = c("milper"), plot_stddev = F)
p7 <- qplot(milper$milper, milper$mean_response) + geom_line() +
  theme_classic() + xlab("Military personnel") + ylab("Mean response")

milex <- h2o.partialPlot(object = a, data = train, cols = c("milex"), plot_stddev = F)
p9 <- qplot(milex$milex, milex$mean_response) + geom_line() +
  theme_classic() + xlab("Military expenditure") + ylab("Mean response")

tradedependence <- h2o.partialPlot(object = a, data = train, cols = c("tradedependence"), plot_stddev =
p10 <- qplot(tradedependence$tradedependence, tradedependence$mean_response) + geom_line() + theme_cl
  xlab("Trade dependence") + ylab("Mean response")


multiplot(p1,p5,p8,p2,p6,p9,p3,p7,p10,p4, cols = 4)  # 11.09x5.14 in

### Mass Killings during Peacetime

# Variable Importance
a <- h2o.loadModel("data/gridrf04nowar_model_391")
va <- h2o.varimp(a)

par(mgp=c(2.2,0.45,0), tcl=-0.4, mar=c(2,7.5,1,1))
barplot(va$scaled_importance[10:1],
        horiz = TRUE, las = 1, cex.names=0.9,
```

```
        names.arg = c("Total trade",
                      "Military personnel",
                      "CINC",
                      "Military expenditure",
                      "% Urban pop.",
                      "Real GDP",
                      "Trade dependence",
                      "Population",
                      "Years mass killing",
                      "Log GDP per capita"),
        main = "")


# Partial dependence plots
logrgdppc <- h2o.partialPlot(object = a, data = train, cols = c("logrgdppc"), plot_stddev = F)
p1 <- qplot(logrgdppc$logrgdppc, logrgdppc$mean_response) + geom_line() + theme_classic() +
  xlab("Log GDP per capita") +  ylab("Mean response")

tradedependence <- h2o.partialPlot(object = a, data = train, cols = c("tradedependence"), plot_stddev =
p4 <- qplot(tradedependence$tradedependence, tradedependence$mean_response) + geom_line() +
  theme_classic() + xlab("Trade dependence") + ylab("Mean response")

percentpopurban <- h2o.partialPlot(object = a, data = train, cols = c("percentpopurban"), plot_stddev =
p6 <- qplot(percentpopurban$percentpopurban, percentpopurban$mean_response) + geom_line() +
  theme_classic() + xlab("% Urban pop.") + ylab("Mean response")

mksyr <- h2o.partialPlot(object = a, data = train, cols = c("mksyr"), plot_stddev = F)
p2 <- qplot(mksyr$mksyr, mksyr$mean_response) + geom_line() + theme_classic() +
  xlab("Years mass killing") +  ylab("Mean response")

milex <- h2o.partialPlot(object = a, data = train, cols = c("milex"), plot_stddev = F)
p7 <- qplot(milex$milex, milex$mean_response) + geom_line() + theme_classic() +
  xlab("Military expenditure") + ylab("Mean response")

milper <- h2o.partialPlot(object = a, data = train, cols = c("milper"), plot_stddev = F)
p9 <- qplot(milper$milper, milper$mean_response) + geom_line() + theme_classic() +
  xlab("Military personnel") + ylab("Mean response")

pop <- h2o.partialPlot(object = a, data = train, cols = c("pop"), plot_stddev = F)
p3 <- qplot(pop$pop, pop$mean_response) + geom_line() + theme_classic() +
  xlab("Population") + ylab("Mean response")

cinc <- h2o.partialPlot(object = a, data = train, cols = c("cinc"), plot_stddev = F)
p8 <- qplot(cinc$cinc, cinc$mean_response) + geom_line() + theme_classic() +
  xlab("CINC") + ylab("Mean response")

realgdp <- h2o.partialPlot(object = a, data = train, cols = c("realgdp"), plot_stddev = F)
p5 <- qplot(realgdp$realgdp, realgdp$mean_response) + geom_line() + theme_classic() +
  xlab("Real GDP") + ylab("Mean response")

totaltrade <- h2o.partialPlot(object = a, data = train, cols = c("totaltrade"), plot_stddev = F)
p10 <- qplot(totaltrade$totaltrade, totaltrade$mean_response) + geom_line() + theme_classic() +
  xlab("Total trade") + ylab("Mean response")
```

```
multiplot(p1,p5,p8,p2,p6,p9,p3,p7,p10,p4, cols = 4)  # 11.09x5.14 in


#####################################################
### Same models with Genocide/Politicide (Harff) ###
#####################################################

# Main model
a <- h2o.loadModel("gridrf05_model_79")
print(va <- a %>% h2o.varimp() %>% as.data.frame() %>% head(., 10))

par(mgp=c(2.2,0.45,0), tcl=-0.4, mar=c(2,7.5,1,1))
barplot(va$scaled_importance[10:1],
        horiz = TRUE, las = 1, cex.names=0.9,
        names.arg = c("Population",
                      "Polity IV",
                      "Real GDP",
                      "Trade dependence",
                      "Log GDP per capita",
                      "Total trade",
                      "Military expenditure",
                      "Military personnel",
                      "% Urban",
                      "CINC"),
        main = "")


df5a <- as.h2o(df5)
df5a$uamkstart <- as.factor(df5a$uamkstart)  #encode the binary repsonse as a factor
h2o.levels(df5a$uamkstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df5a,
                         ratios = 0.75,
                         seed = 1234)


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "uamkstart"
x <- setdiff(names(df5), c(y, "ccode", "year", "rgdppc",
                           "uamkyr2", "uamkyr3", "sf", "country",
                           "elf2", "polity2sq"))


cinc <- h2o.partialPlot(object = a, data = train, cols = c("cinc"), plot_stddev = F)
p1 <- qplot(cinc$cinc, cinc$mean_response) + geom_line() + theme_classic() +
        xlab("CINC") +  ylab("Mean response")

percentpopurban <- h2o.partialPlot(object = a, data = train, cols = c("percentpopurban"), plot_stddev =
p2 <- qplot(percentpopurban$percentpopurban, percentpopurban$mean_response) + geom_line() +
```

```
  theme_classic() + xlab("% Urban") + ylab("Mean response")

milper <- h2o.partialPlot(object = a, data = train, cols = c("milper"), plot_stddev = F)
p3 <- qplot(milper$milper, milper$mean_response) + geom_line() +
        theme_classic() + xlab("Military personnel") + ylab("Mean response")

milex <- h2o.partialPlot(object = a, data = train, cols = c("milex"), plot_stddev = F)
p4 <- qplot(milex$milex, milex$mean_response) + geom_line() +
  theme_classic() + xlab("Military expenditure") + ylab("Mean response")

totaltrade <- h2o.partialPlot(object = a, data = train, cols = c("totaltrade"), plot_stddev = F)
p5 <- qplot(totaltrade$totaltrade, totaltrade$mean_response) + geom_line() + theme_classic() +
        xlab("Total trade") + ylab("Mean response")

logrgdppc <- h2o.partialPlot(object = a, data = train, cols = c("logrgdppc"), plot_stddev = F)
p6 <- qplot(logrgdppc$logrgdppc, logrgdppc$mean_response) + geom_line() + theme_classic() +
  xlab("Log GDP per capita") +  ylab("Mean response")

tradedependence <- h2o.partialPlot(object = a, data = train, cols = c("tradedependence"), plot_stddev =
p7 <- qplot(tradedependence$tradedependence, tradedependence$mean_response) + geom_line() +
  theme_classic() + xlab("Trade dependence") + ylab("Mean response")

realgdp <- h2o.partialPlot(object = a, data = train, cols = c("realgdp"), plot_stddev = F)
p8 <- qplot(realgdp$realgdp, realgdp$mean_response) + geom_line() +
        theme_classic() + xlab("Real GDP") + ylab("Mean response")

polity2 <- h2o.partialPlot(object = a, data = train, cols = c("polity2"), plot_stddev = F)
p9 <- qplot(polity2$polity2, polity2$mean_response) + geom_line() +
  theme_classic() + xlab("Polity IV") + ylab("Mean response")

pop <- h2o.partialPlot(object = a, data = train, cols = c("pop"), plot_stddev = F)
p10 <- qplot(pop$pop, pop$mean_response) + geom_line() +
  theme_classic() + xlab("Population") + ylab("Mean response")

multiplot(p1,p5,p8,p2,p6,p9,p3,p7,p10,p4, cols = 4)  # 11.09x5.14 in

# UCDP == 1
df.ucdp2 <- df5 %>% filter(UCDPcivilwarongoing == 1)
df.ucdp2a <- as.h2o(df.ucdp2)

df.ucdp2a$uamkstart <- as.factor(df.ucdp2a$uamkstart)  #encode the binary repsonse as a factor
h2o.levels(df.ucdp2a$uamkstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df.ucdp2a,
                         ratios = .75,
                         seed = 1234)


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "uamkstart"
```

```
x <- setdiff(names(df.ucdp2), c(y, "ccode", "year", "rgdppc",
                                "uamkyr2", "uamkyr3", "sf", "country",
                                "elf2", "polity2sq"))

a <- h2o.loadModel("gridrf06_model_275")
print(va <- a %>% h2o.varimp() %>% as.data.frame() %>% head(., 10))

par(mgp=c(2.2,0.45,0), tcl=-0.4, mar=c(2,7.5,1,1))
barplot(va$scaled_importance[10:1],
        horiz = TRUE, las = 1, cex.names=0.9,
        names.arg = c("Military expenditure",
                      "Population",
                      "Regime change",
                      "Physical integrity",
                      "Total battle deaths",
                      "Trade dependence",
                      "% Urban",
                      "Years since genocide",
                      "Log GDP per capita",
                      "Military personnel"),
        main = "")

milper <- h2o.partialPlot(object = a, data = train, cols = c("milper"), plot_stddev = F)
p1 <- qplot(milper$milper, milper$mean_response) + geom_line() +
        theme_classic() + xlab("Military personnel") + ylab("Mean response")

logrgdppc <- h2o.partialPlot(object = a, data = train, cols = c("logrgdppc"), plot_stddev = F)
p2 <- qplot(logrgdppc$logrgdppc, logrgdppc$mean_response) + geom_line() + theme_classic() +
        xlab("Log GDP per capita") + ylab("Mean response")

uamkyr <- h2o.partialPlot(object = a, data = train, cols = c("uamkyr"), plot_stddev = F)
p3 <- qplot(uamkyr$uamkyr, uamkyr$mean_response) + geom_line() + theme_classic() +
  xlab("Years since genocide") +  ylab("Mean response")

percentpopurban <- h2o.partialPlot(object = a, data = train, cols = c("percentpopurban"), plot_stddev =
p4 <- qplot(percentpopurban$percentpopurban, percentpopurban$mean_response) + geom_line() +
        theme_classic() + xlab("% Urban") + ylab("Mean response")

tradedependence <- h2o.partialPlot(object = a, data = train, cols = c("tradedependence"), plot_stddev =
p5 <- qplot(tradedependence$tradedependence, tradedependence$mean_response) + geom_line() +
        theme_classic() + xlab("Trade dependence") + ylab("Mean response")

totalbeaths <- h2o.partialPlot(object = a, data = train, cols = c("totalbeaths"), plot_stddev = F)
p6 <- qplot(totalbeaths$totalbeaths, totalbeaths$mean_response) + geom_line() +
        theme_classic() + xlab("Total battle deaths") + ylab("Mean response")

physint <- h2o.partialPlot(object = a, data = train, cols = c("physint"), plot_stddev = F)
p7 <- qplot(physint$physint, physint$mean_response) + geom_line() +
  theme_classic() + xlab("Physical integrity") + ylab("Mean response")

change <- h2o.partialPlot(object = a, data = train, cols = c("change"), plot_stddev = F)
p8 <- qplot(change$change, change$mean_response) + geom_line() +
  theme_classic() + xlab("Regime change") + ylab("Mean response")
```

```
pop <- h2o.partialPlot(object = a, data = train, cols = c("pop"), plot_stddev = F)
p9 <- qplot(pop$pop, pop$mean_response) + geom_line() +
  theme_classic() + xlab("Population") + ylab("Mean response")

milex <- h2o.partialPlot(object = a, data = train, cols = c("milex"), plot_stddev = F)
p10 <- qplot(milex$milex, milex$mean_response) + geom_line() +
  theme_classic() + xlab("Military expenditure") + ylab("Mean response")

multiplot(p1,p5,p8,p2,p6,p9,p3,p7,p10,p4, cols = 4)  # 11.09x5.14 in

# COW == 1
df.cow2 <- df5 %>% filter(COWcivilwarongoing == 1)
df.cow2a <- as.h2o(df.cow2)
df.cow2a$uamkstart <- as.factor(df.cow2a$uamkstart) #encode the binary repsonse as a factor
h2o.levels(df.cow2a$uamkstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df.cow2a,
                         ratios = .75,  # 70%, 15%, 15%
                         seed = 1234)  # reproducibility


train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "uamkstart"
x <- setdiff(names(df.cow2), c(y, "ccode", "year", "rgdppc",
                               "uamkyr2", "uamkyr3", "sf", "country",
                               "elf2", "polity2sq"))

a <- h2o.loadModel("gridrf07_model_413")
print(va <- a %>% h2o.varimp() %>% as.data.frame() %>% head(., 10))

par(mgp=c(2.2,0.45,0), tcl=-0.4, mar=c(2,7.5,1,1))
barplot(va$scaled_importance[10:1],
        horiz = TRUE, las = 1, cex.names=0.9,
        names.arg = c("Trade dependence",
                      "Real GDP",
                      "Population",
                      "CINC",
                      "Total trade",
                      "Total battle deaths",
                      "Log GDP per capita",
                      "% Urban",
                      "Years since genocide",
                      "Military personnel"),
        main = "")

milper <- h2o.partialPlot(object = a, data = train, cols = c("milper"), plot_stddev = F)
p1 <- qplot(milper$milper, milper$mean_response) + geom_line() +
  theme_classic() + xlab("Military personnel") + ylab("Mean response")
```

```
uamkyr <- h2o.partialPlot(object = a, data = train, cols = c("uamkyr"), plot_stddev = F)
p2 <- qplot(uamkyr$uamkyr, uamkyr$mean_response) + geom_line() + theme_classic() +
  xlab("Years since genocide") +  ylab("Mean response")

percentpopurban <- h2o.partialPlot(object = a, data = train, cols = c("percentpopurban"), plot_stddev =
p3 <- qplot(percentpopurban$percentpopurban, percentpopurban$mean_response) + geom_line() +
  theme_classic() + xlab("% Urban") + ylab("Mean response")

logrgdppc <- h2o.partialPlot(object = a, data = train, cols = c("logrgdppc"), plot_stddev = F)
p4 <- qplot(logrgdppc$logrgdppc, logrgdppc$mean_response) + geom_line() + theme_classic() +
  xlab("Log GDP per capita") + ylab("Mean response")

totalbeaths <- h2o.partialPlot(object = a, data = train, cols = c("totalbeaths"), plot_stddev = F)
p5 <- qplot(totalbeaths$totalbeaths, totalbeaths$mean_response) + geom_line() +
  theme_classic() + xlab("Total battle deaths") + ylab("Mean response")

totaltrade <- h2o.partialPlot(object = a, data = train, cols = c("totaltrade"), plot_stddev = F)
p6 <- qplot(totaltrade$totaltrade, totaltrade$mean_response) + geom_line() +
  theme_classic() + xlab("% Urban") + ylab("Mean response")

cinc <- h2o.partialPlot(object = a, data = train, cols = c("cinc"), plot_stddev = F)
p7 <- qplot(cinc$cinc, cinc$mean_response) + geom_line() +
  theme_classic() + xlab("CINC") + ylab("Mean response")

pop <- h2o.partialPlot(object = a, data = train, cols = c("pop"), plot_stddev = F)
p8 <- qplot(pop$pop, pop$mean_response) + geom_line() +
  theme_classic() + xlab("Population") + ylab("Mean response")

realgdp <- h2o.partialPlot(object = a, data = train, cols = c("realgdp"), plot_stddev = F)
p9 <- qplot(realgdp$realgdp, realgdp$mean_response) + geom_line() +
  theme_classic() + xlab("Real GDP") + ylab("Mean response")

tradedependence <- h2o.partialPlot(object = a, data = train, cols = c("tradedependence"), plot_stddev =
p10 <- qplot(tradedependence$tradedependence, tradedependence$mean_response) + geom_line() +
  theme_classic() + xlab("Trade dependence") + ylab("Mean response")

multiplot(p1,p5,p8,p2,p6,p9,p3,p7,p10,p4, cols = 4)  # 11.09x5.14 in


# Ethnic conflict == 1
df.eth2 <- df5 %>% filter(ethnowarongoing == 1)

df.eth2a <- as.h2o(df.eth2)

df.eth2a$uamkstart <- as.factor(df.eth2a$uamkstart)  #encode the binary repsonse as a factor
h2o.levels(df.eth2a$uamkstart)

# Partition the data into training, validation and test sets
splits <- h2o.splitFrame(data = df.eth2a,
                         ratios = .75,  # 70%, 15%, 15%
                         seed = 1234)  # reproducibility
```

```
train <- h2o.assign(splits[[1]], "train.hex")
valid <- h2o.assign(splits[[2]], "valid.hex")

y <- "uamkstart"
x <- setdiff(names(df.eth2), c(y, "ccode", "year", "rgdppc",
                               "uamkyr2", "uamkyr3", "sf", "country",
                               "elf2", "polity2sq"))


a <- h2o.loadModel("gridrf08_model_173")
print(va <- a %>% h2o.varimp() %>% as.data.frame() %>% head(., 10))

par(mgp=c(2.2,0.45,0), tcl=-0.4, mar=c(2,7.5,1,1))
barplot(va$scaled_importance[10:1],
        horiz = TRUE, las = 1, cex.names=0.9,
        names.arg = c("Real GDP",
                      "Trade dependence",
                      "Physical Integrity",
                      "Log GDP per capita",
                      "Population",
                      "Total trade",
                      "Military personnel",
                      "CINC",
                      "Military expenditure",
                      "% Urban"),
        main = "")

percentpopurban <- h2o.partialPlot(object = a, data = train, cols = c("percentpopurban"), plot_stddev =
p1 <- qplot(percentpopurban$percentpopurban, percentpopurban$mean_response) + geom_line() +
  theme_classic() + xlab("% Urban") + ylab("Mean response")

milex <- h2o.partialPlot(object = a, data = train, cols = c("milex"), plot_stddev = F)
p2 <- qplot(milex$milex, milex$mean_response) + geom_line() +
  theme_classic() + xlab("Military expenditure") + ylab("Mean response")

cinc <- h2o.partialPlot(object = a, data = train, cols = c("cinc"), plot_stddev = F)
p3 <- qplot(cinc$cinc, cinc$mean_response) + geom_line() +
  theme_classic() + xlab("CINC") + ylab("Mean response")

milper <- h2o.partialPlot(object = a, data = train, cols = c("milper"), plot_stddev = F)
p4 <- qplot(milper$milper, milper$mean_response) + geom_line() +
  theme_classic() + xlab("Military personnel") + ylab("Mean response")

totaltrade <- h2o.partialPlot(object = a, data = train, cols = c("totaltrade"), plot_stddev = F)
p5 <- qplot(totaltrade$totaltrade, totaltrade$mean_response) + geom_line() +
  theme_classic() + xlab("% Urban") + ylab("Mean response")

pop <- h2o.partialPlot(object = a, data = train, cols = c("pop"), plot_stddev = F)
p6 <- qplot(pop$pop, pop$mean_response) + geom_line() +
  theme_classic() + xlab("Population") + ylab("Mean response")

logrgdppc <- h2o.partialPlot(object = a, data = train, cols = c("logrgdppc"), plot_stddev = F)
p7 <- qplot(logrgdppc$logrgdppc, logrgdppc$mean_response) + geom_line() + theme_classic() +
```

```
  xlab("Log GDP per capita") + ylab("Mean response")

physint <- h2o.partialPlot(object = a, data = train, cols = c("physint"), plot_stddev = F)
p5 <- qplot(physint$physint, physint$mean_response) + geom_line() +
  theme_classic() + xlab("Physical integrity") + ylab("Mean response")

tradedependence <- h2o.partialPlot(object = a, data = train, cols = c("tradedependence"), plot_stddev =
p9 <- qplot(tradedependence$tradedependence, tradedependence$mean_response) + geom_line() +
  theme_classic() + xlab("Trade dependence") + ylab("Mean response")

realgdp <- h2o.partialPlot(object = a, data = train, cols = c("realgdp"), plot_stddev = F)
p10 <- qplot(realgdp$realgdp, realgdp$mean_response) + geom_line() +
  theme_classic() + xlab("Real GDP") + ylab("Mean response")

multiplot(p1,p5,p8,p2,p6,p9,p3,p7,p10,p4, cols = 4)  # 11.09x5.14 in
```

Bibliography

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies*, 72(1):1–19. Cited on page 16.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490). Cited on pages 2, 7, 16, 25, 26 and 30.

Abadie, A., Diamond, A., and Hainmueller, J. (2011). Synth: An R Package for Synthetic Control Methods in Comparative Case Studies. *Journal of Statistical Software*, 42(13). Cited on page 16.

Abadie, A., Diamond, A., and Hainmueller, J. (2014). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*. Cited on pages 2, 7, 17, 18 and 23.

Abadie, A. and Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *The American Economic Review*, pages 113–132. Cited on pages 2, 7, 16, 17 and 30.

Achen, C. H. (1992). Social Psychology, Demographic Variables, and Linear Regression: Breaking the Iron Triangle in Voting Research. *Political Behavior*, 14(3):195–211. Cited on page 18.

Achen, C. H. (2002). Toward a New Political Methodology: Microfoundations and ART. *Annual Review of Political Science*, 5(1):423–450. Cited on page 18.

Agência Brasil (2012). Jogo do Bicho e Exploração de Caça-Níquel Entram na Lista de Crimes de Lavagem de Dinheiro. https://bit.ly/2D5AeX6. Access: 2018-10-16. Cited on page 56.

Ahnen, R. (2003). Between Tyranny of the Majority and Liberty: The Persistence of Human Rights Violations under Democracy in Brazil. *Bulletin of Latin American Research*, 22(3):319–339. Cited on page 5.

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, 84(3):488–500. Cited on pages 50 and 58.

Allansson, M., Melander, E., and Themnér, L. (2017). Organized violence, 1989–2016. *Journal of Peace Research*, 54(4):574–587. Cited on pages 78, 87, 91 and 122.

Ames, B. (1995). Electoral Strategy under Open-List Proportional Representation. *American Journal of Political Science*, 39(2):406–433. Cited on page 63.

Anderton, C. H. and Carter, J. R. (2015). A new look at weak state conditions and genocide risk. *Peace Economics, Peace Science and Public Policy*, 21(1):1–36. Cited on page 86.

Andrews, G. R. (1991). *Blacks & Whites in São Paulo, Brazil, 1888–1988*. Madison: University of Wisconsin Press. Cited on page 51.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press. Cited on pages 2, 7, 67 and 69.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1):17–21. Cited on page 71.

Arguello, K. (2012). Criminalização dos Jogos de Azar: Contradição Entre Lei e Realidade Social. *Revista da EMERJ*, 15(60):239–250. Cited on page 60.

Arias, E. D. (2006). The Dynamics of Criminal Governance: Networks and Social Order in Rio de Janeiro. *Journal of Latin American Studies*, 38(02):293–325. Cited on page 62.

Arias, E. D. (2009). *Drugs and Democracy in Rio de Janeiro: Trafficking, Social Networks, and Public Security*. Chapel Hill: University of North Carolina Press. Cited on page 13.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books. Cited on page 57.

Axelrod, R. and Keohane, R. O. (1985). Achieving Cooperation under Anarchy: Strategies and Institutions. *World Politics*, 38(1):226–254. Cited on page 57.

Balcells, L. (2010). Rivalry and Revenge: Violence against Civilians in Conventional Civil Wars. *International Studies Quarterly*, 54(2):291–313. Cited on page 86.

Balcells, L. (2011). Continuation of politics by two means: Direct and indirect violence in civil war. *Journal of Conflict Resolution*, 55(3):397–422. Cited on page 86.

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. Cited on page 87.

Banks, A. S. (1999). *Cross-National Time-Series Data Archive User's Manual.* Center for Social Analysis, State University of New York at Binghamton. Cited on page 87.

Barata, R. B. and Ribeiro, M. C. S. d. A. (2000). Relação entre Homicídios e Indicadores Econômicos em São Paulo, Brasil. *Panamerican Journal of Public Health*, 7(2):118–24. Cited on pages 6 and 8.

Barbarino, A. and Mastrobuoni, G. (2014). The Incapacitation Effect of Incarceration: Evidence from Several Italian Collective Pardons. *American Economic Journal: Economic Policy*, 6(1):1–37. Cited on page 11.

Baser, O. (2006). Too Much Ado about Propensity Score Models? Comparing Methods of Propensity Score Matching. *Value in Health*, 9(6):377–385. Cited on page 18.

BBC (2012). Brazil's Illegal Numbers Game Under Pressure. https://bbc.in/2PVgxnJ. Access: 2017-03-07. Cited on page 62.

BBC (2016). PCC Não Derrubou Homicídios Sozinho em SP, Dizem Pesquisadores. Cited on page 13.

Beattie, S. and Mole, A. (2007). Police Resources in Canada. *Canadian Centre for Justice Statistics*. Cited on page 11.

Becker, G. S. (1968). Crime and Punishment: An Economic Approach. In *The Economic Dimensions of Crime*, pages 13–68. Springer. Cited on pages 6 and 9.

Bell, M. S. (2015). Examining explanations for nuclear proliferation. *International Studies Quarterly*, 60(3):520–529. Cited on pages 67, 69 and 71.

Bell, M. S. and Miller, N. L. (2015). Questioning the effect of nuclear weapons on conflict. *Journal of Conflict Resolution*, 59(1):74–92. Cited on page 84.

Belém, E. d. F. (2015). Livro Diz que Capitão do Exército Modernizou O Jogo Do Bicho e Produziu Uma Máfia Tropical. https://bit.ly/2ETTI2S. Access: 2018-10-14. Cited on page 54.

Benatte, A. P. (2002). *Dos Jogos que Especulam Com o Acaso: Contribuição À História do Jogo de Azar no Brasil (1890–1950)*. PhD thesis, University of Campinas. Cited on page 48.

Bersch, K., Praça, S., and Taylor, M. M. (2017). State Capacity, Bureaucratic Politicization, and Corruption in the Brazilian State. *Governance*, 30(1):105–124. Cited on page 64.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *Quarterly Journal of Economics*, 119(1):249–275. Cited on pages 2 and 7.

Besançon, M. L. (2005). Relative resources: Inequality in ethnic wars, revolutions, and genocides. *Journal of Peace Research*, 42(4):393–415. Cited on pages 75 and 86.

Bethell, L. (2008). Politics in Brazil under Vargas 1930–45. In Bethell, L., editor, *The Cambridge History of Latin America, Volume IX: Brazil since 1930*, volume 9, chapter 1, pages 3–86. Cambridge: Cambridge University Press, 1 edition. Cited on page 60.

Bezerra, L. A. (2009). O Mecenato do Jogo do Bicho e a Ascensão da Beija-Flor no Carnaval Carioca. *Textos Escolhidos de Cultura e Artes Populares*, 6(1):139–150. Cited on pages 3, 59 and 61.

Biderman, C., Lima, R. S. D., and Mello, J. M. P. D. (2016). Pax Monopolista and Crime: The Case of the Emergence of the Primeiro Comando da Capital in São Paulo. Cited on page 13.

Billmeier, A. and Nannicini, T. (2013). Assessing Economic Liberalization Episodes: A Synthetic Control Approach. *Review of Economics and Statistics*, 95(3):983–1001. Cited on page 7.

Biondi, K. (2010). *Junto e Misturado: Uma Etnografia do PCC*. São Paulo: Editora Terceiro Nome. Cited on pages 12 and 13.

Boettke, P. (2001). *Calculation and Coordination: Essays on Socialism and Transitional Political Economy*. London & New York: Routledge. Cited on page 48.

Brasil de Fato (2013). Com Maior População Carcerária do Brasil, São Paulo Registra 15 mil Prisões em Um Ano. Cited on page 9.

Breiman, L. (2001a). Random forests. *Machine Learning*, 45(1):5–32. Cited on page 71.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231. Cited on pages 72 and 73.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. (2015). Inferring Causal Impact Using Bayesian Structural Time-Series Models. *The Annals of Applied Statistics*, 9(1):247–274. Cited on page 26.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable Selection for Propensity Score Models. *American Journal of Epidemiology*, 163(12):1149–1156. Cited on page 18.

Buchanan, J. M. (1965). An Economic Theory of Clubs. *Economica*, 32(125):1–14. Cited on page 54.

Bueno, S. (2014). Letalidade na Ação Policial: Os Desafios para a Consolidação de uma Agenda de Políticas Públicas no Estado de São Paulo. *Administração Pública e Gestão Social*, 7(1):9–15. Cited on page 8.

Bueno De Mesquita, E. (2005). The Quality of Terror. *American Journal of Political Science*, 49(3):515–530. Cited on page 1.

Bueno de Mesquita, E. and Dickson, E. S. (2007). The Propaganda of the Deed: Terrorism, Counterterrorism, and Mobilization. *American Journal of Political Science*, 51(2):364–381. Cited on page 1.

Bulutgil, H. Z. (2015). Social cleavages, wartime experience, and ethnic cleansing in europe. *Journal of Peace Research*, 52(5):577–590. Cited on page 86.

Bundervoet, T. (2009). Livestock, land and political power: The 1993 killings in burundi. *Journal of Peace Research*, 46(3):357–376. Cited on page 86.

Buonanno, P. and Raphael, S. (2013). Incarceration and Incapacitation: Evidence from the 2006 Italian Collective Pardon. *The American Economic Review*, 103(6):2437–2465. Cited on page 11.

Cabral, M. V. d. F. (2016a). *Avaliação do Impacto do Infocrim sobre as Taxas de Homicídio nos Municípios Paulistas: Uma Aplicação do Método de Diferenças em Diferenças Espaciais*. PhD thesis, Federal University of Juiz de Fora. Cited on page 10.

Cabral, S. (2016b). *Escolas de Samba do Rio de Janeiro*. Rio de Janeiro: Editora Lazuli. Cited on page 61.

Cahuc, P. and Dormont, B. (1997). Profit-sharing: Does It Increase Productivity and Employment? A Theoretical Model and Empirical Evidence on French Micro Data. *Labour Economics*, 4(3):293–319. Cited on page 62.

Camargo, A. B. M. (2007). Mortes por Causas Violentas no Estado de São Paulo. *São Paulo em Perspectiva*, 21(1):31–45. Cited on page 6.

Cardia, N., Adorno, S., and Poleto, F. Z. (2003). Homicide Rates and Human Rights Violations in São Paulo, Brazil: 1990 to 2002. *Health and Human Rights*, pages 14–33. Cited on page 6.

Carey, S. C., Mitchell, N. J., and Lowe, W. (2013). States, the security sector, and the monopoly of violence: A new database on pro-government militias. *Journal of Peace Research*, 50(2):249–258. Cited on page 87.

Carvalho, S. d. and Freire, C. R. (2005). O Regime Disciplinar Diferenciado: Notas Críticas À Reforma do Sistema Punitivo Brasileiro. *Revista Transdisciplinar de Ciências Penitenciárias*, 4(1):7–26. Cited on pages 6 and 9.

Cavalcanti, M. L. V. d. C. (2006). *Carnaval Carioca: Dos Bastidores ao Desfile*. Rio de Janeiro: Editora UFRJ. Cited on pages 59 and 61.

Cederman, L.-E., Wimmer, A., and Min, B. (2010). Why do ethnic groups rebel? new data and analysis. *World Politics*, 62(1):87–119. Cited on pages 75, 78, 83, 87, 91 and 122.

Cerqueira, D. (2013). Mapa de Homicídios Ocultos no Brasil. Cited on page 5.

Cerqueira, D. and Mello, J. M. P. d. (2013). Evaluating a National Anti-Firearm Law and Estimating the Causal Effect of Guns on Crime. *PUC, Rio de janeiro. Departamento de Economia. Texto para Discussão*, (607). Cited on page 9.

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6. Cited on page 74.

Chazkel, A. (2007). Beyond Law and Order: The Origins of the Jogo do Bicho in Republican Rio de Janeiro. *Journal of Latin American Studies*, 39(03):535–565. Cited on pages 47, 52, 53 and 59.

Chazkel, A. (2011). *Laws of Chance: Brazil's Clandestine Lottery and the Making of Urban Public Life*. Durham: Duke University Press. Cited on pages 3, 48, 51, 52, 53 and 59.

Cheadle, D. and Prendergast, J. (2007). *Not on Our Watch: The Mission to End Genocide in Darfur and Beyond*. Dublin: Hachette Books. Cited on page 66.

Chen, M. K. and Shapiro, J. M. (2007). Do Harsher Prison Conditions Reduce Recidivism? A Discontinuity-Based Approach. *American Law and Economics Review*, 9(1):1–29. Cited on page 11.

Chinelli, F. and Machado, L. A. (1993). O Vazio da Ordem: Relações Políticas e Organizacionais entre as Escolas de Samba e o Jogo do Bicho. *Revista do Rio de Janeiro*, 1(1):42–52. Cited on pages 54, 61 and 62.

Cingranelli, D. L. and Richards, D. L. (2010). The cingranelli and richards (ciri) human rights data project. *Human Rights Quarterly*, 32(2):401–424. Cited on page 87.

Clarke, K. A. (2005). The Phantom Menace: Omitted Variable Bias in Econometric Research. *Conflict Management and Peace Science*, 22(4):341–352. Cited on pages 18 and 69.

Clarke, K. A. (2009). Return of the Phantom Menace Omitted Variable Bias in Political Research. *Conflict Management and Peace Science*, 26(1):46–66. Cited on page 18.

Clayton, G. and Thomson, A. (2016). Civilianizing civil conflict: Civilian defense militias and the logic of violence in intrastate conflict. *International Studies Perspectives*, 60(3):499–510. Cited on page 86.

Coffman, M. and Noy, I. (2012). Hurricane Iniki: Measuring the Long-Term Economic Impact of a Natural Disaster using Synthetic Control. *Environment and Development Economics*, 17(02):187–205. Cited on page 7.

Cohen, A. and Siegelman, P. (2010). Testing for Adverse Selection in Insurance Markets. *Journal of Risk and Insurance*, 77(1):39–84. Cited on page 58.

Cohen, D. K. (2013). Explaining Rape During Civil War: Cross-National Evidence (1980–2009). *American Political Science Review*, 107(03):461–477. Cited on page 1.

Colaresi, M. and Carey, S. (2008). To kill or to protect: Security forces, domestic institutions, and genocide. *Journal of Conflict Resolution*, 52(1):39–67. Cited on pages 68 and 86.

Collier, P. (2003). *Breaking the Conflict Trap: Civil War and Development Policy*. Washington, DC: World Bank Publications. Cited on page 85.

Collier, P. and Hoeffler, A. (2004). Greed and Grievance in Civil War. *Oxford Economic Papers*, 56(4):563–595. Cited on page 1.

Congresso em Foco (2007). Políticos do Rio Receberam Dinheiro de Bicheiros, diz PF. https://bit.ly/2qcut1y. Access: 2017-02-07. Cited on pages 60 and 64.

Consejo Ciudadano para la Securidad Pública y Justicia Penal (2014). *The 50 Most Violent Cities in the World 2014*. Mexico City. Cited on pages 2 and 5.

Cook, D. (2017). *Practical Machine Learning with H2O*. Sebastopol: O'Reilly. Cited on page 73.

Cornish, D. B. and Clarke, R. V. (2014). *The Reasoning Criminal: Rational Choice Perspectives on Offending*. London: Transaction Publishers. Cited on page 6.

Cross, J. C. and Peña, S. (2006). Risk and Regulation In Informal and Illegal Markets. In Fernández-Kelly, P. and Shefner, J., editors, *Out of the Shadows: Political Action and the Informal Economy in Latin America*, chapter 02, pages 49–80. Penn State University Press, University Park. Cited on page 52.

Dal Bó, P. (2005). Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games. *The American Economic Review*, 95(5):1591–1604. Cited on page 54.

DaMatta, R. and Soárez, E. (1999). *Águias, Burros e Borboletas: Um Estudo Antropológico do Jogo do Bicho*. Rio de Janeiro: Rocco. Cited on pages 3, 47, 48, 52, 53 and 59.

Dehejia, R. H. and Wahba, S. (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economics and statistics*, 84(1):151–161. Cited on page 7.

Del Río, S., López, V., Benítez, J. M., and Herrera, F. (2014). On the use of mapreduce for imbalanced big data using random forest. *Information Sciences*, 285:112–137. Cited on page 74.

Dias, C. C. N. (2009a). Da Guerra À Gestão: Trajetória do Primeiro Comando da Capital (PCC) nas Prisões de São Paulo. *Revista Percurso*, pages 79–96. Cited on pages 12 and 13.

Dias, C. C. N. (2009b). Ocupando as Brechas do Direito Formal: O PCC como Instância Alternativa de Resolução de Conflitos. *Dilemas*, 2(4):83–105. Cited on page 12.

Dias, C. C. N. (2011). *Da Pulverização ao Monopólio da Violência: Expansão e Consolidação do Primeiro Comando da Capital (PCC) no Sistema Carcerário Paulista.* PhD thesis, Universidade de São Paulo. Cited on pages 1, 12, 13 and 15.

Dietterich, T. G., Hild, H., and Bakiri, G. (1995). A comparison of id3 and backpropagation for english text-to-speech mapping. *Machine Learning*, 18(1):51–80. Cited on page 72.

Downes, A. B. (2006). Desperate times, desperate measures: The causes of civilian victimization in war. *International Security*, 30(4):152–195. Cited on page 86.

Downes, A. B. (2007). Restraint or propellant? democracy and civilian fatalities in interstate wars. *Journal of Conflict Resolution*, 51(6):872–904. Cited on page 86.

Doyle, M. W. and Sambanis, N. (2006). *Making War and Building Peace: United Nations Peace Operations.* Princeton, NJ: Princeton University Press. Cited on page 69.

Easterly, W., Gatti, R., and Kurlat, S. (2006). Development, democracy, and mass killings. *Journal of Economic Growth*, 11(2):129–156. Cited on pages 75 and 86.

Eck, J. E. and Maguire, E. R. (2006). Have Changes in Policing Reduced Violent Crime? An Assessment of the Evidence. In Blumstein, A. and Wallman, J., editors, *The Crime Drop in America*, pages 207–265. Cambridge: Cambridge University Press. Cited on page 11.

Eck, K. and Hultman, L. (2007). One-sided violence against civilians in war: Insights from new fatality data. *Journal of Peace Research*, 44(2):233–246. Cited on pages 82 and 86.

Ehrlich, I. (1973). Participation in Illegitimate Activities: A Theoretical and Empirical Investigation. *The Journal of Political Economy*, pages 521–565. Cited on page 9.

Elwert, F. and Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40:31–53. Cited on page 69.

Estado de São Paulo (2006). Bicheiro Dono de Máquinas Caça-Níqueis É Preso em Belém. https://bit.ly/2qbBdfS. Access: 2016-11-14. Cited on page 55.

Estado de São Paulo (2011). Justiça do Rio Condena 11 por Máfia dos Caça-Níqueis. https://bit.ly/2RflGae. Access: 2018-10-16. Cited on page 56.

Esteban, J., Morelli, M., and Rohner, D. (2015). Strategic mass killings. *Journal of Political Economy*, 123(5):1087–1132. Cited on pages 75 and 86.

Evans, P. B. (1995). *Embedded Autonomy: States and Industrial Transformation*. Princeton: Princeton University Press. Cited on page 64.

Exame (2016). As 15 Maiores Empresas da Cidade do Rio de Janeiro. https://abr.ai/2qdG4NO. Access: 2017-02-01. Cited on page 48.

Farias, E. S. (2013). A Afirmaćão de uma Situaćão Sociocomunicativa: Desfile de Carnaval e Tramas da Cultura Popular Urbana Carioca. *Caderno CRH*, 26(67):157–178. Cited on page 61.

Fausto, B. (1972). *A Revolução de 1930: história e historiografia*. São Paulo: Brasiliense. Cited on page 60.

Fazal, T. M. and Greene, B. C. (2015). A particular difference: European identity and civilian targeting. *British Journal of Political Science*, 45(4):829–851. Cited on page 86.

Fearon, J. D. and Laitin, D. D. (2003). Ethnicity, Insurgency, and Civil War. *American Political Science Review*, 97(01):75–90. Cited on pages 1 and 87.

Feltran, G. d. S. (2010). Crime e Castigo na Cidade: Os Repertórios da Justiça e a Questão do Homicídio nas Periferias de São Paulo. *Caderno CRH*, 23(58). Cited on page 12.

Feltran, G. d. S. (2012a). Governo que Produz Crime, Crime que Produz Governo: O Dispositivo de Gestão do Homicídio em São Paulo (1992–2011). *Revista Brasileira de Segurança Pública*, 6(2):232–255. Cited on pages 9, 12 and 13.

Feltran, G. d. S. (2012b). Manter A Ordem nas Periferias de São Paulo: Coexistência de Dispositivos Normativos na "Era PCC". Cited on page 13.

Figueiredo, A. C. and Limongi, F. (2000). Presidential Power, Legislative Organization, and Party Behavior in Brazil. *Comparative Politics*, pages 151–170. Cited on page 63.

Finkel, E. and Straus, S. (2012). Macro, meso, and micro research on genocide: Gains, shortcomings, and future areas of inquiry. *Genocide Studies and Prevention*, 7(1):56–67. Cited on page 66.

FitzRoy, F. R. and Kraft, K. (1987). Cooperation, Productivity, and Profit Sharing. *The Quarterly Journal of Economics*, 102(1):23–35. Cited on page 62.

Fjelde, H. and Hultman, L. (2014). Weakening the enemy: A disaggregated study of violence against civilians in africa. *Journal of Conflict Resolution*, 58(7):1230–1257. Cited on page 86.

Fligstein, N. (1996). Markets as Politics: A Political-Cultural Approach to Market Institutions. *American Sociological Review*, 61(4):656–673. Cited on page 56.

Folha de São Paulo (2001). Estatuto do PCC Prevê Rebeliões Integradas. Cited on page 12.

Folha de São Paulo (2006). Bicheiro Dá Garantia Contra Apostas Altas. https://bit.ly/2PWtXjp. Access: 2016-12-19. Cited on page 55.

Folha de São Paulo (2016). Ministros de Temer Querem a Legalização de Jogos de Azar. https://bit.ly/1OxrIuU. Access: 2016-05-17. Cited on page 52.

Franco, G. (1987). Reforma Monetária e Instabilidade Durante a Transição Republicana. Master's thesis, Catholic University of Rio de Janeiro. Cited on page 51.

Franzese, R. J., Covey, H. C., and Menard, S. (2016). *Youth Gangs*. Springfield: Charles C Thomas Publisher. Cited on page 1.

Freire, D. (2014). Entering the Underworld: Prison Gang Recruitment in São Paulo's Primeiro Comando da Capital. Master's thesis, The Graduate Institute, Geneva. Cited on pages 1 and 15.

Freyre, G. (1933). *Casa-Granda & Senzala: Formação da Família Brasileira Sob o Regime de Economia Patriarcal*. Rio de Janeiro: José Olympio. Cited on page 61.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232. Cited on page 73.

Gambetta, D. (1996). *The Sicilian Mafia: The Business of Private Protection*. Cambridge: Harvard University Press. Cited on page 49.

Gambetta, D. (2009). *Codes of the Underworld: How Criminals Communicate*. Princeton: Princeton University Press. Cited on pages 50 and 57.

Gaspari, E. (2002). *A Ditadura Escancarada*. Sã Paulo: Companhia das Letras. Cited on page 62.

Gassebner, M., Gutmann, J., and Voigt, S. (2016). When to expect a coup d'état? an extreme bounds analysis of coup determinants. *Public Choice*, 169(3):293–313. Cited on pages 69, 70 and 71.

Gassebner, M., Lamla, M. J., and Vreeland, J. R. (2013). Extreme bounds of democracy. *Journal of Conflict Resolution*, 57(2):171–197. Cited on page 69.

Geddes, B. and Neto, A. R. (1992). Institutional Sources of Corruption in Brazil. *Third World Quarterly*, 13(4):641–661. Cited on page 63.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, pages 1360–1383. Cited on pages 84 and 105.

Gelman, A. and Su, Y.-S. (2016). `arm`*: Data Analysis Using Regression and Multilevel/Hierarchical Models*. R package version 1.9-3. Cited on page 105.

Genuer, R., Poggi, J.-M., and Tuleau, C. (2008). Random forests: Some methodological insights. *arXiv*. Cited on page 73.

Glaeser, E. L., Sacerdote, B., and Scheinkman, J. A. (1996). Crime and Social Interactions. *The Quarterly Journal of Economics*, 111(2):507–548. Cited on page 11.

Gleditsch, K. S. (2002). Expanded trade and gdp data. *Journal of Conflict Resolution*, 46(5):712–724. Cited on page 87.

Gleditsch, N. P., Wallensteen, P., Eriksson, M., Sollenberg, M., and Strand, H. (2002). Armed conflict 1946–2001: A new dataset. *Journal of peace research*, 39(5):615–637. Cited on pages 78, 87, 91 and 122.

Goertzel, T. and Kahn, T. (2009). The Great São Paulo Homicide Drop. *Homicide Studies*, 13(4):398–410. Cited on pages 2, 6 and 9.

Goldman, R. (2017). Assad's history of chemical attacks, and other atrocities. Cited on page 66.

Goldsmith, B. E., Butcher, C. R., Semenovich, D., and Sowmya, A. (2013). Forecasting the onset of genocide and politicide: Annual out-of-sample forecasts on a global dataset, 1988–2003. *Journal of Peace Research*, 50(4):437–452. Cited on pages 68 and 86.

Goldstein, B., Hubbard, A., Cutler, A., and Barcellos, L. (2010). An application of random forests to a genome-wide association dataset: Methodological considerations & new findings. *BMC Genetics*, 11(1):49. Cited on page 73.

Goldstein, D. (2013). *Laughter Out of Place: Race, Class, Violence, and Sexuality in a Rio Shantytown*. Berkeley: University of California Press. Cited on page 62.

Gurr, T. R. (2000). *Peoples Versus States: Minorities at Risk in the New Century*. Washington, DC: US Institute of Peace Press. Cited on page 76.

Hafner-Burton, E. M. (2005). Right or robust? the sensitive nature of repression to globalization. *Journal of Peace Research*, 42(6):679–698. Cited on page 69.

Hall, M. M. (1969). *The Origins of Mass Immigration in Brazil, 1871–1914*. New York: Columbia University Press. Cited on page 51.

Harff, B. (2003). No lessons learned from the holocaust? assessing risks of genocide and political mass murder since 1955. *American Political Science Review*, 97(1):57–73. Cited on pages 68, 76, 83, 86, 114 and 138.

Harff, B. and Gurr, T. R. (1988). Toward Empirical Theory of Genocides and Politicides: Identification and Measurement of Cases Since 1945. *International Studies Quarterly*, 32(3):359–371. Cited on pages 1 and 66.

Hegre, H., Ellingsen, T., Gates, S., and Gleditsch, N. P. (2001). Toward a democratic civil peace? democracy, political change, and civil war, 1816–1992. *American political science review*, 95(1):33–48. Cited on page 75.

Hegre, H. and Sambanis, N. (2006). Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution*, 50(4):508–535. Cited on pages 67, 69 and 70.

Heim, B. T. and Lurie, I. Z. (2014). Does Health Reform Affect Self-Employment? Evidence from Massachusetts. *Small Business Economics*, 43(4):917–930. Cited on page 7.

Hertzman, M. A. (2013). *Making Samba: A New History of Race and Music in Brazil*. Durham: Duke University Press. Cited on page 61.

Hill, D. W. and Jones, Z. (2014). An empirical evaluation of explanations for state repression. *American Political Science Review*, 108(3):661–687. Cited on pages 67 and 73.

Hinrichs, P. (2012). The Effects of Affirmative Action Bans on College Enrollment, Educational Attainment, and the Demographic Composition of Universities. *Review of Economics and Statistics*, 94(3):712–722. Cited on page 7.

Hlavac, M. (2016). ExtremeBounds: Extreme bounds analysis in R. *Journal of Statistical Software*, 72(9):1–22. Cited on page 71.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, 15(3):199–236. Cited on pages 7 and 16.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844. Cited on page 72.

Höglund, K. (2009). Electoral Violence in Conflict-Ridden Societies: Concepts, Causes, and Consequences. *Terrorism and political violence*, 21(3):412–427. Cited on page 1.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960. Cited on pages 7 and 29.

Holmes, J. S., Piñeres, S. A. G., and Curtin, K. M. (2006). Drugs, Violence, and Development in Colombia: A Department-Level Analysis. *Latin American Politics and Society*, 48(3):157–184. Cited on page 1.

HuffPost Brasil (2015). Caixa Estima que Legalização de Jogos de Azar Pode Quintuplicar Arrecadação do Brasil. https://bit.ly/2SVrrf3. Access: 2018-10-25. Cited on page 48.

Hughes, P. J. A. (2004). Segregação Socioespacial e Violência na Cidade de São Paulo: Referências para a Formulação de Políticas Públicas. *São Paulo em Perspectiva*, 18(4):93–102. Cited on pages 8 and 13.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press. Cited on pages 7 and 67.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1):5–86. Cited on page 3.

Isaac, R. M., Walker, J. M., and Thomas, S. H. (1984). Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations. *Public Choice*, 43(2):113–149. Cited on page 57.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449. Cited on page 74.

Jinjarak, Y., Noy, I., and Zheng, H. (2013). Capital controls in Brazil–Stemming a Tide with a Signal? *Journal of Banking & Finance*, 37(8):2938–2952. Cited on page 7.

Jones, G. (2009). *Youth Violence in Latin America: Gangs and Juvenile Justice in Perspective*. Berlin: Springer. Cited on page 1.

Jones, Z. and Linder, F. (2015). Exploratory data analysis using random forests. pages 1–31. Cited on pages 72 and 73.

Jones, Z. M. and Lupu, Y. (2018). Is There More Violence in the Middle? *American Journal of Political Science*, 62(3):652–667. Cited on pages 67 and 72.

Jornal do Brasil (2011). Jogo do bicho e Política: Influência vem de Longa Data. https://bit.ly/2D9irhM. Access: 2017-02-07. Cited on page 60.

Joshi, M. and Quinn, J. M. (2017). Who kills whom? the micro-dynamics of civilian targeting in civil war. *Social Science Research*, 63:227–241. Cited on page 86.

Jupiara, A. and Otavio, C. (2015). *Os Porões da Contravenção*. Rio de Janeiro: Editora Record. Cited on pages 48, 50, 54, 62 and 63.

Kahn, T. and Zanetic, A. (2005). O Papel dos Municípios na Segurança Pública. *Estudos Criminológicos*, 4:1–68. Cited on pages 2 and 9.

Kalyvas, S. N. (2006). *The Logic of Violence in Civil War*. Cambridge: Cambridge University Press. Cited on pages 1 and 66.

Kaufmann, C. (1996). Possible and Impossible Solutions to Ethnic Civil Wars. *International security*, 20(4):136–175. Cited on page 1.

Kim, D. (2010). What makes state leaders brutal? examining grievances and mass killing during civil war. *Civil Wars*, 12(3):237–260. Cited on page 86.

Kim, N. K. (2016). Revolutionary leaders and mass killing. *Journal of Conflict Resolution*, page 0022002716653658. Cited on page 86.

Kimbrough, E. O., Rubin, J., Sheremeta, R. M., and Shields, T. W. (2015). Commitment Problems in Conflict Resolution. *Journal of Economic Behavior & Organization*, 112:33–45. Cited on page 50.

King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, pages 666–687. Cited on page 71.

King, G., Keohane, R. O., and Verba, S. (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press. Cited on page 67.

Kisangani, E. and Wayne Nafziger, E. (2007). The political economy of state terror. *Defence and Peace Economics*, 18(5):405–414. Cited on page 86.

Koren, O. (2017). Means to an end: Pro-government militias as a predictive indicator of strategic mass killing. *Conflict Management and Peace Science*, 34(5):461–484. Cited on pages 78 and 86.

Krain, M. (1997). State-Sponsored Mass Murder: The Onset and Severity of Genocides and Politicides. *Journal of Conflict Resolution*, 41(3):331–360. Cited on pages 1, 68, 76 and 86.

Krain, M. (2005). International intervention and the severity of genocides and politicides. *International Studies Quarterly*, 49(3):363–387. Cited on page 1.

Kruse, D. L. (1992). Profit Sharing and Productivity: Microeconomic Evidence from the United States. *The Economic Journal*, 102(410):24–36. Cited on page 62.

Labronici, R. B. (2012). Para Todos Vale o Escrito: Uma Etnografia do Jogo do Bicho. Master's thesis, Universidade Federal Fluminense. Cited on pages 3, 53, 54, 55, 58, 59 and 61.

Labronici, R. B. (2014). Sorteio de Bicho: Uma Análise do Lazer para Fora da Lei. *Recorde: Revista de História do Esporte*, 7(2). Cited on pages 48 and 53.

Lacina, B. and Gleditsch, N. P. (2005). Monitoring trends in global combat: A new dataset of battle deaths. *European Journal of Population/Revue Européenne de Démographie*, 21(2-3):145–166. Cited on page 87.

LaFree, G. (1999). A Summary and Review of Cross-National Comparative Studies of Homicide. In Smith, M. D. and Zahn, M. A., editors, *Homicide: A Sourcebook of Social Research*. London. Cited on page 19.

Lall, R. (2016). How multiple imputation makes a difference. *Political Analysis*, 24(4):414–433. Cited on pages 71 and 73.

Lauerhass, L. (1972). *Getúlio Vargas and the Triumph of Brazilian Nationalism: A Study on the Rise of the Nationalist Generation of 1930*. PhD thesis, University of California, Los Angeles. Cited on page 60.

Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review*, 75(3):308–313. Cited on page 69.

Leeson, P. T. (2009). *The Invisible Hook: The Hidden Economics of Pirates*. Princeton University Press. Cited on page 49.

Leeson, P. T. (2010). Pi*rational* Choice: The Economics of Infamous Pirate Practices. *Journal of Economic Behavior & Organization*, 76(3):497–510. Cited on pages 49 and 57.

Lesser, J. (2013). *Immigration, Ethnicity, and National Identity in Brazil, 1808 to the Present*. Cambridge: Cambridge University Press. Cited on page 51.

Lessing, B. (2015). Logics of Violence in Criminal War. *Journal of Conflict Resolution*, 59(8):1486–1516. Cited on page 1.

Levin, J. (2001). Information and the Market for Lemons. *RAND Journal of Economics*, 32(4):657–666. Cited on page 58.

Levine, R. and Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *The American Economic Review*, 82(4):942–963. Cited on page 69.

Levine, R. M. (1992). *Vale of Tears: Revisiting the Canudos Massacre in Northeastern Brazil, 1893–1897*. Berkeley: University of California Press. Cited on page 70.

Levitt, S. D. (1996). The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Litigation. *The Quarterly Journal of Economics*, 111(2):319–351. Cited on pages 9 and 11.

Levitt, S. D. (1997). Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime. *American Economic Review*, 87(3):270–90. Cited on page 9.

Levitt, S. D. (2004). Understanding Why Crime Fell in the 1990s: Four Factors that Explain the Decline and Six that Do Not. *The Journal of Economic Perspectives*, 18(1):163–190. Cited on page 11.

Lobo, E. M. L. (2001). *Imigração Portuguesa no Brasil*, volume 43. São Paulo: Editora Hucitec. Cited on page 51.

Magalhães, F. S. (2005). *Ganhou Leva ... Do Vale o Impresso ao Vale o Escrito: Uma História Social do Jogo do Bicho no Rio de Janeiro (1890–1960)*. PhD thesis, Universidade Federal do Rio de Janeiro. Cited on pages 3, 48, 51, 55, 57, 58 and 59.

Magalhães, M. D. B. d. (1997). A Lógica da Suspeição: Sobre os Aparelhos Repressivos À Época da Ditadura Militar no Brasil. *Revista Brasileira de História*, 17(34):203–220. Cited on page 62.

Mamdani, M. (2014). *When Victims Become Killers: Colonialism, Nativism, and the Genocide in Rwanda*. Princeton: Princeton University Press. Cited on page 1.

Manekin, D. (2013). Violence against civilians in the second intifada: The moderating effect of armed group structure on opportunistic violence. *Comparative Political Studies*, 46(10):1273–1300. Cited on page 86.

Manso, B. P. (2012). *Crescimento e Queda dos Homicídios em SP entre 1960 e 2010 – Uma Análise dos Mecanismos da Escolha Homicida e das Carreiras no Crime*. PhD thesis. Cited on page 11.

Manso, B. P. and Godoy, M. (2014). 20 Anos de PCC – o Efeito Colateral da Política de Segurança Pública. *Interesse Nacional*, 24(6):26–35. Cited on page 15.

Marsh, J. L., Hutton, J. L., and Binks, K. (2002). Removal of Radiation Dose Response Effects: An Example of Over-Matching. *BMJ*, 325(7359):327–330. Cited on page 18.

Marshall, M. G., Gurr, T. R., and Harff, B. (2017). Pitf state failure problem set, 1955-2016. Cited on pages 66, 68 and 87.

Mattos, M. B. (1991). Vadios, Jogadores, Mendigos e Bêbados na Cidade do Rio de Janeiro do Início do Século. Master's thesis, Fluminense Federal University. Cited on page 51.

McCann, B. (2004). *Hello, Hello Brazil: Popular Music in the Making of Modern Brazil*. Durham: Duke University Press. Cited on page 60.

McDoom, O. S. (2013). Who killed in rwanda's genocide? micro-space, social influence and individual participation in intergroup violence. *Journal of Peace Research*, 50(4):453–467. Cited on page 86.

McDoom, O. S. (2014). Predicting violence within genocide: A model of elite competition and ethnic segregation from rwanda. *Political Geography*, 42:34–45. Cited on page 86.

Melander, E., Öberg, M., and Hall, J. (2009). Are 'new wars' more atrocious? battle severity, civilians killed and forced migration before and after the end of the cold war. *European Journal of International Relations*, 15(3):505–536. Cited on page 86.

Mello, J. M. P. d. and Schneider, A. (2010). Mudança Demográfica e a Dinâmica dos Homicídios no Estado de São Paulo. *São Paulo em Perspectiva*, 21(1):19–30. Cited on pages 8 and 10.

Mello, M. P. d. (1989). A História Social dos Jogos de Azar no Rio de Janeiro (1808–1946). Master's thesis, Instituto Universitário de Pesquisas do Rio de Janeiro. Cited on page 59.

Miraglia, P. (2015). Drugs and Drug Trafficking in Brazil: Trends and Policies. *Center for 21st Century Security and Intelligence Latin America Initiative*, pages 1–16. Cited on page 65.

Misse, M. (2007). Mercados Ilegais, Redes de Proteção e Organização Local do Crime no Rio de Janeiro. *Estudos Avançados*, 21(61):139–157. Cited on pages 50, 53 and 61.

Misse, M. (2009). Sobre A Acumulação Social da Violência no Rio de Janeiro. *Civitas-Revista de Ciências Sociais*, 8(3):371–385. Cited on page 62.

Misse, M. (2011). Crime Organizado e Crime Comum no Rio de Janeiro. *Revista de Sociologia e Política*, 19(40):13–25. Cited on pages 60, 61, 62, 63 and 65.

Montalvo, J. G. (2011). Voting after the Bombings: A Natural Experiment on the Effect of Terrorist Attacks on Democratic Elections. *Review of Economics and Statistics*, 93(4):1146–1154. Cited on page 7.

Montalvo, J. G. and Reynal-Querol, M. (2005). Ethnic Polarization, Potential Conflict, and Civil Wars. *The American Economic Review*, 95(3):796–816. Cited on page 1.

Montalvo, J. G. and Reynal-Querol, M. (2008). Discrete polarisation with an application to the determinants of genocides. *The Economic Journal*, 118(533):1835–1865. Cited on pages 68 and 86.

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press. Cited on page 7.

Muchlinski, D. (2014). Grievances and opportunities: Religious violence across political regimes. *Politics and Religion*, 7(4):684–705. Cited on page 75.

Muchlinski, D., Siroky, D., He, J., and Kocher, M. (2015). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, 24(1):87–103. Cited on page 72.

Nagin, D. S. (2007). Moving Choice to Center Stage in Criminological Research and Theory. *Criminology*, 45(2):259–272. Cited on page 10.

Nichols, M. (2017). South sudan's government using food as weapon of war - u.n. report. Cited on page 66.

Nivette, A. E. (2011). Cross-National Predictors of Crime: A Meta-Analysis. *Homicide Studies*, 15(2):103–131. Cited on page 19.

Nyseth Brehm, H. (2017). Re-examining risk factors of genocide. *Journal of Genocide Research*, 19(1):61–87. Cited on page 76.

O Dia (2016). 'O Carnaval Está em Pé Graças À Contravenção', diz Anísio Abraão David. https://bit.ly/2SixnhS. Access: 2017-03-08. Cited on page 60.

O Estado de São Paulo (2012). O 'Gene' do Jogo do Bicho dos Cachoeira. https://bit.ly/2PoG0Jd. Access: 2018-10-25. Cited on page 63.

O Globo (2012a). Bicheiro Recebia Políticos em sua Casa. https://glo.bo/2PT9pby. Access: 2017-03-07. Cited on pages 64 and 65.

O Globo (2012b). PF: País Foi Fatiado pelas Quadrilhas de Contraventores. https://glo.bo/2O6hR5t. Access: 2016-09-18. Cited on page 53.

O Globo (2015). Caça-Níqueis: Bicheiros São Condenados a 25 Anos de Prisão. https://glo.bo/2D6ihYm. Access: 2016-10-03. Cited on pages 55 and 65.

O Globo (2017). Castor de Andrade, Chefão do Bicho, Cria Império À Base de Corrupćão. https://glo.bo/2OQ3sjb. Access: 2018-10-15. Cited on pages 54 and 55.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5):673–690. Cited on pages 71 and 99.

O'Donnell, G. (1993). On the State, Democratization and Some Conceptual Problems: A Latin American View with Glances at Some Postcommunist Countries. *World Development*, 21(8):1355–1369. Cited on page 62.

Oliveira, L. L. (2001). *O Brasil dos Imigrantes*. Rio de Janeiro: Zahar. Cited on page 51.

Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. (2012). How many trees in a random forest? In *Machine Learning and Data Mining in Pattern Recognition*, pages 154–168, Berlin, Germany. Springer. Cited on page 73.

Ostrom, E. (1990). *Governing the Commons*. Cambridge: Cambridge University Press. Cited on page 57.

Owens, E. G. (2009). More Time, Less Crime? Estimating the Incapacitative Effect of Sentence Enhancements. *Journal of Law and Economics*, 52(3):6. Cited on page 11.

Pacheco, R. J. C. (1957). *Antologia do Jôgo de Bicho*. Rio de Janeiro: Organização Simões. Cited on page 48.

Pape, R. A. (2003). The Strategic Logic of Suicide Terrorism. *American Political Science Review*, 97(03):343–361. Cited on page 1.

Paternoster, R. (2010). How Much Do We Really Know about Criminal Deterrence? *The Journal of Criminal Law and Criminology*, pages 765–824. Cited on page 9.

Pearl, J. (2001). Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc. Cited on page 15.

Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press. Cited on page 18.

Peres, M. F. T., Vicentin, D., Nery, M. B., de Lima, R. S., de Souza, E. R., Cerda, M., Cardia, N., and Adorno, S. (2011). Queda dos Homicídios em São Paulo, Brasil: Uma Análise Descritiva. *Revista Panamericana de Salud Publica*, 29(1):17. Cited on pages 11 and 14.

Pilster, U., Böhmelt, T., and Tago, A. (2016). The differentiation of security forces and the onset of genocidal violence. *Armed Forces & Society*, 42(1):26–50. Cited on page 86.

Pinheiro, P. S. (2000). Democratic Governance, Violence, and the (Un)Rule of Law. *Daedalus*, pages 119–143. Cited on pages 5, 13 and 62.

Pinheiro, P. S. (2001). The Paradox of Democracy in Brazil. *Brown Journal of World Affairs*, 8:113. Cited on page 5.

Pinker, S. (2011). *The Better Angels of Our Nature: The Decline of Violence in History and Its Causes*. London: Penguin UK. Cited on page 75.

Power, S. (2013). *"A Problem From Hell": America and the Age of Genocide*. New York: Basic Books. Cited on page 1.

Queiroz, M. I. P. d. (1992). *Carnaval Brasileiro: O Vivido e o Mito*. São Paulo: Brasiliense. Cited on pages 50, 59 and 61.

Querido, C. M. (2009). State-sponsored mass killing in african wars–greed or grievance? *International Advances in Economic Research*, 15(3):351. Cited on page 86.

Raleigh, C. (2012). Violence against civilians: A disaggregated analysis. *International Interactions*, 38(4):462–481. Cited on page 86.

Reiter, B. and Mitchell, G. (2009). The New Politics of Race in Brazil. In Reiter, B. and Mitchell, G., editors, *Brazil's New Racial Politics*, pages 1–18. Boulder: Lynne Rienner, 1 edition. Cited on page 61.

Richani, N. (2013). *Systems of Violence: The Political Economy of War and Peace in Colombia*. New York: Suny Press. Cited on page 1.

Risso, M. (2014). Intentional Homicides in São Paulo City: A New Perspective. *Stability: International Journal of Security and Development*, 3(1). Cited on page 10.

Rodgers, D. (2006). Living in the Shadow of Death: Gangs, Violence and Social Order in Urban Nicaragua, 1996–2002. *Journal of Latin American Studies*, 38(2):267–292. Cited on page 1.

Rose-Ackerman, S. (1978). *Corruption: A Study in Political Economy*. New York: Academic Press. Cited on page 50.

Rost, N. (2013). Will it happen again? on the possibility of forecasting the risk of genocide. *Journal of Genocide Research*, 15(1):41–67. Cited on pages 75 and 86.

Roth, A. and Murnighan, J. K. (1978). Equilibrium Behavior and Repeated Play of the Prisoner's Dilemma. *Journal of Mathematical Psychology*, 17(2):189–198. Cited on page 54.

Roth, M. G. and Skarbek, D. (2014). Prison Gangs and the Community Responsibility System. *Review of Behavioral Economics*, 1(3):223–243. Cited on page 51.

Rubin, D. B. (1973). Matching to Remove Bias in Observational Studies. *Biometrics*, pages 159–183. Cited on pages 7 and 16.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press. Cited on page 16.

Rummel, R. J. (1995). Democracy, power, genocide, and mass murder. *Journal of Conflict Resolution*, 39(1):3–26. Cited on pages 66, 69, 75 and 86.

Sala-i-Martin, X. (1997). I just ran two million regressions. *The American Economic Review*, 87(2):178–183. Cited on pages 69, 70 and 75.

Sala-i-Martin, X., Doppelhofer, G., and Miller, R. I. (2004). Determinants of long-term growth: A bayesian averaging of classical estimates (bace) approach. *American Economic Review*, 94(4):813–835. Cited on pages 84 and 95.

Salla, F. (2007). De Montoro a Lembo: As Políticas Penitenciárias em São Paulo. *Revista Brasileira de Segurança Pública*, 1(1):72–90. Cited on pages 6 and 9.

Sambanis, N. (2001). Do Ethnic and Nonethnic Civil Wars Have the Same Causes? A Theoretical and Empirical Inquiry (Part 1). *Journal of Conflict Resolution*, 45(3):259–282. Cited on page 1.

Samuels, D. (2000). Ambition and Competition: Explaining Legislative Turnover in Brazil. *Legislative Studies Quarterly*, pages 481–497. Cited on page 63.

Santos, F. F. S. d. (2008). *Um Partido, Três Agendas?: Política de Segurança Pública no Estado de São Paulo: 1995–2006.* PhD thesis. Cited on page 9.

Sarkees, M. R. and Wayman, F. W. (2010). *Resort to War.* Washington DC: CQ Press. Cited on pages 71, 78, 91 and 122.

Schelling, T. (1960). *The Strategy of Conflict.* Cambridge: Harvard University Press. Cited on page 50.

Schneider, G. and Bussmann, M. (2013). Accounting for the dynamics of one-sided violence: Introducing kosved. *Journal of Peace Research*, 50(5):635–644. Cited on page 86.

Schneider, R. M. (1996). *Brazil: Culture and Politics in a New Industrial Powerhouse.* Boulder: Westview Press. Cited on page 48.

Schulz, J. (2008). *The Financial Crisis of Abolition.* New Haven: Yale University Press. Cited on page 51.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression. *Center for Bioinformatics & Molecular Biostatistics*, pages 1–14. Cited on page 73.

Shirk, D. A. (2010). Drug Violence in Mexico: Data and Analysis from 2001–2009. *Trends in Organized Crime*, 13(2-3):167–174. Cited on page 1.

Shleifer, A. and Vishny, R. W. (2002). *The Grabbing Hand: Government Pathologies and Their Cures.* Cambridge: Harvard University Press. Cited on page 50.

Singer, J. D. (1988). Reconstructing the correlates of war dataset on material capabilities of states, 1816–1985. *International Interactions*, 14(2):115–132. Cited on page 87.

Singer, J. D., Bremer, S., and Stuckey, J. (1972). Capability distribution, uncertainty, and major power war, 1820-1965. In Russett, B., editor, *Peace, War, and Numbers*, pages 19–48. Beverly Hills: Sage. Cited on page 87.

Siroky, D. and Dzutsati, V. (2015). The empire strikes back: Ethnicity, terrain, and indiscriminate violence in counterinsurgencies. *Social Science Quarterly*, 96(3):807–829. Cited on page 86.

Skarbek, D. (2011a). Governance and Prison Gangs. *American Political Science Review*, 105(04):702–716. Cited on pages 1, 49 and 57.

Skarbek, D. (2011b). Governance and Prison Gangs. *American Political Science Review*, 105(04):702–716. Cited on page 15.

Skarbek, D. (2012). Prison Gangs, Norms, and Organizations. *Journal of Economic Behavior & Organization*, 82(1):96–109. Cited on pages 1, 49 and 57.

Skarbek, D. (2014). *The Social Order of the Underworld: How Prison Gangs Govern the American Penal System*. Oxford: Oxford University Press. Cited on pages 1 and 49.

Skidmore, T. E. (1967). *Politics in Brazil, 1930–1964: An Experiment in Democracy*. New York: Oxford University Press. Cited on page 60.

Skidmore, T. E. (1993). *Black Into White: Race and Nationality in Brazilian Thought*. Durham: Duke University Press. Cited on page 51.

Smith, J. M. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press. Cited on page 57.

Soares, S. S. F. (1993). *O Jogo do Bicho: A Saga de um Fato Social Brasileiro*. Rio de Janeiro: Editora Bertrand. Cited on page 3.

Sobel, M. E. (1987). Direct and Indirect Effects in Linear Structural Equation Models. *Sociological Methods & Research*, 16(1):155–176. Cited on pages 1 and 15.

Soihet, R. (1998). *A Subversão Pelo Riso: Estudos sobre o Carnaval Carioca da Belle Époque ao Tempo de Vargas*. São Paulo: Fundação Getúlio Vargas Editora. Cited on page 61.

Spector, P. E. and Brannick, M. T. (2011). Methodological urban legends: The misuse of statistical control variables. *Organizational Research Methods*, 14(2):287–305. Cited on page 69.

Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics*, 87(3):355–374. Cited on pages 50 and 58.

Stanton, J. (2015). Regulating militias: Governments, militias, and civilian targeting in civil war. *Journal of Conflict Resolution*, 59(5):899–923. Cited on pages 66, 68 and 86.

Stanton, J. A. (2013). Terrorism in the Context of Civil War. *The Journal of Politics*, 75(4):1009–1022. Cited on page 68.

Stiglitz, J. E. and Weiss, A. (1981). Credit Rationing in Markets with Imperfect Information. *The American Economic Review*, 71(3):393–410. Cited on page 58.

Stockler, J. S. (2011). The Invention of Samba and National Identity in Brazil. Working Papers in Nationalism Studies. . Access: February 2017. Cited on page 60.

Straus, S. (2007). Second-generation comparative research on genocide. *World Politics*, 59(3):476–501. Cited on pages 66, 68 and 85.

Straus, S. (2012a). "destroy them to save us": Theories of genocide and the logics of political violence. *Terrorism and Political Violence*, 24(4):544–560. Cited on page 85.

Straus, S. (2012b). Wars Do End! Changing Patterns of Political Violence in Sub-Saharan Africa. *African Affairs*, 111(443):179–201. Cited on page 75.

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25(1):1–21. Cited on page 7.

Sturm, J.-E. and de Haan, J. (2002). How robust is sala-i-martin's robustness analysis? Technical report, University of Groningen: Mimeo. Cited on page 71.

Sullivan, C. M. (2012). Blood in the village: A local-level investigation of state massacres. *Conflict Management and Peace Science*, 29(4):373–396. Cited on page 86.

Terra (2011). Rio: Fábrica de Caça-Níquel de Bicheiro É Descoberta pela Polícia. https://bit.ly/2JhZV7g. Access: 2016-10-05. Cited on page 55.

The H2O.ai Team (2017). h2o: *R Interface for H2O*. R package version 3.14.0.3. Cited on pages 72 and 120.

Time Magazine (1966). Brazil: The Animal Game. https://ti.me/2D8gj94. Access: 2018-10-25. Cited on page 48.

Tir, J. and Jasinski, M. (2008). Domestic-level diversionary theory of war: Targeting ethnic minorities. *Journal of Conflict Resolution*, 52(5):641–664. Cited on page 86.

Tollison, R. D. (1982). Rent Seeking: A Survey. *Kyklos*, 35(4):575–602. Cited on page 50.

Topik, S. (2014). *The Political Economy of the Brazilian State, 1889–1930*. Austin: University of Texas Press. Cited on page 51.

Torcato, C. E. M. (2011). *A Repressão Oficial ao Jogo do Bicho: Uma História dos Jogos de Azar em Porto Alegre (1885–1917)*. PhD thesis, Federal University of Rio Grande do Sul. Cited on page 52.

Trent, C. L. and Pridemore, W. A. (2012). A Review of the Cross-National Empirical Literature on Social Structure and Homicide. In *Handbook of European Homicide Research*, pages 111–135. Cited on page 19.

Trento, A. (1989). *Do Outro Lado do Atlântico: Um Século de Imigração Italiana no Brasil*. São Paulo: Studio Nobel. Cited on page 51.

Triner, G. D. and Wandschneider, K. (2005). The Baring Crisis and the Brazilian Encilhamento, 1889–1891: An Early Example of Contagion among Emerging Capital Markets. *Financial History Review*, 12(02):199–225. Cited on page 51.

Ulfelder, J. (2012). Forecasting onsets of mass killing. Technical report. Accessed: January 2018. Cited on pages 68 and 86.

Ulfelder, J. and Valentino, B. (2008). Assessing risks of state-sponsored mass killing. Technical report. Accessed: January 2018. Cited on pages 66, 70, 75 and 86.

United Nations Office on Drugs and Crime (2013). *Global Study on Homicide 2013: Trends, Contexts, Data*. Cited on pages 2 and 5.

UOL (2016). Sorteios da Caixa São Confiáveis? Veja Como é o Processo de Auditoria. https://bit.ly/2zD2Kv2. Access: 2017-03-08. Cited on page 58.

Uzonyi, G. (2014). Unpacking the Effects of Genocide and Politicide on Forced Migration. *Conflict Management and Peace Science*, 31(3):225–243. Cited on page 1.

Uzonyi, G. (2015). Civil war victory and the onset of genocide and politicide. *International Interactions*, 41(2):365–391. Cited on page 86.

Uzonyi, G. (2016). Domestic unrest, genocide and politicide. *Political Studies*, 64(2):315–334. Cited on pages 68 and 86.

Valentino, B. (2014). Why we kill: The political science of political violence against civilians. *Annual Review of Political Science*, 17:89–103. Cited on pages 66 and 75.

Valentino, B., Huth, P., and Balch-Lindsay, D. (2004). "draining the sea": Mass killing and guerrilla warfare. *International Organization*, 58(2):375–407. Cited on pages 66, 68 and 86.

Valentino, B., Huth, P., and Croco, S. (2006). Covenants without the sword international law and the protection of civilians in times of war. *World Politics*, 58(3):339–377. Cited on page 86.

Varese, F. (2001). *The Russian Mafia: Private Protection In A New Market Economy*. Oxford: Oxford University Press. Cited on page 57.

Verpoorten, M. (2012). Leave none to claim the land: A malthusian catastrophe in rwanda? *Journal of Peace Research*, 49(4):547–563. Cited on page 86.

Vianna, H. (1995). *O Mistério do Samba*. Rio de Janeiro: Jorge Zahar/Ed. UFRJ. Cited on page 61.

Villar, J. L. M. (2008). *Contravenção e a Cultura da Ascensão Social*. São Paulo: Blucher Acadêmico. Cited on page 51.

Waiselfisz, J. J. (2011). Mapa da Violência 2011: Os Jovens do Brasil. Cited on page 6.

Waiselfisz, J. J. (2014). Mapa da Violência 2014: Os Jovens do Brasil. Cited on pages 2 and 5.

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. Cited on page 73.

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. Cited on page 73.

Wayman, F. W. and Tago, A. (2010). Explaining the onset of mass killing, 1949–87. *Journal of Peace Research*, 47(1):3–13. Cited on pages 68, 82 and 86.

Western, B., Kling, J. R., and Weiman, D. F. (2001). The Labor Market Consequences of Incarceration. *Crime & Delinquency*, 47(3):410–427. Cited on page 11.

Wig, T. and Tollefsen, A. F. (2016). Local institutional quality and conflict violence in africa. *Political geography*, 53:30–42. Cited on page 86.

Wilkinson, S. I. (2006). *Votes and Violence: Electoral Competition and Ethnic Riots in India*. Cambridge: Cambridge University Press. Cited on page 1.

Williams, D. (2001). *Culture Wars in Brazil: The First Vargas Regime, 1930–1945*. Durham: Duke University Press. Cited on page 60.

Willis, G. D. (2014). Antagonistic Authorities and the Civil Police in São Paulo, Brazil. *Latin American Research Review*, 49(1):3–22. Cited on page 13.

Willis, G. D. (2015). *The Killing Consensus: Police, Organized Crime, and the Regulation of Life and Death in Urban Brazil*. Berkeley: University of California Press. Cited on pages 12 and 13.

Wood, E. J. (2006). Variation in Sexual Violence During War. *Politics & Society*, 34(3):307–342. Cited on page 1.

Wood, E. J. (2009). Armed Groups and Sexual Violence: When is Wartime Rape Rare? *Politics & Society*, 37(1):131–161. Cited on page 1.

World Health Organization (2015). Social Determinants of Health. Cited on page 6.

Yanagizawa-Drott, D. (2014). Propaganda and conflict: Evidence from the rwandan genocide. *The Quarterly Journal of Economics*, 129(4):1947–1994. Cited on page 86.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis*, 13(2):157–170.

Cited on page 84.