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

Danilo Freire Gary Uzonyi

18 July 2018

Abstract

The literature on state-sponsored violence has grown significantly over the last decades. 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. We 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 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.

Keywords: extreme bounds analysis, genocide, mass killings, random forest, state-sponsored violence

JEL Classification: C52, C53, D74, H56, K10

*Freire: PhD candidate, Department of Political Economy, King's College London. Email address: danilofreire@gmail.com. Uzonyi: Assistant Professor, Department of Political Science; Research Fellow, Howard H. Baker Jr. Center for Public Policy, University of Tennessee. Email address: guzonyi@utk.edu. We thank Robert McDonnell and David Skarbek for helpful suggestions and comments, and Mark S. Bell for sharing `R` code to estimate the penalised-likelihood models. All replication materials are available at https://github.com/danilofreire/mass-killings.

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).

Ultimately, effective prevention requires us to understand why these atrocities occur. In this vein, the academic community has laboured tremendously to establish evidence-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 we draw from these findings.

To overcome these limitations and provide a better understanding of government atrocity, we employ extreme bounds analysis and random forests to identify the most robust determ-

[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., Finkel and Straus 2012; Straus 2012a; Valentino et al. 2004), others claim genocide and politicide follow a different logic from other forms of government violence (Kalyvas 2006; Stanton 2015).

inants of state-sponsored atrocities. Our approach is similar to Hegre and Sambanis (2006) seminal analysis on the causes of civil war onset, but we provide additional tests to check which variables are able to predict out-of-sample cases of mass violence as suggested by Hill and Jones (2014), Muchlinski et al. (2015), and Ward et al. (2010). In conducting this analysis, we 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 advocates, 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 war (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-year (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 Uzonyi 2016; Wayman and Tago 2010).

Our analysis tests the sensitivity of 40 variables on a sample of 177 countries from 1945 to 2013. Our findings partially confirm previous research – 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, we 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, we 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 our conclusion is that unstable countries are violent, then preventing atrocity likely requires significant investments of time and resources in state-building, which is often politically and practically unfeasible (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.

2 Empirical Methods

To conduct our analysis, we 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, we identified 45 articles which employed logit or probit models of mass killing onset in a global sample. We then included all variables that appeared in at least two of these papers in our 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 we considered and a complete list of the variables we included in our models. Next, we estimated an extreme bounds analysis to determine which variables were the most robust in explaining the onset of government atrocity. Then, we estimated a distributed random forest analysis to see which of the variables best predicted the onset of these atrocities. In this section, we provide more detail on each of these estimation procedures before turning to the results of both analyses in the next section.

2.1 Extreme Bounds Analysis

Th first method we 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, we use Sala-i-Martin's 1997 more flexible version of EBA and consider the whole range of values of $CDF(0)$. We choose to use the whole distribution because the aggregate $CDF(0)$ allows us to move away from a binary indicator of robustness and present the estimations with their appropriate degrees of confidence. Our focus is the percentage of the variable's cumulative distribution function that is smaller or greater than zero. We do not assume that the CDFs are normally distributed and use Sala-i-Martin's generic model instead.[2] We specify our models as follows:

$$Mass\ Killing\ Onset_{it} = \beta_M M_{it} + \beta_F F_{it} + \beta_Z Z_{it} + \upsilon_{it} \tag{1}$$

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

Our 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 our 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 we are 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. We thus only use this notation to help clarify the connection of our 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), we lagged the independent variables one year to reduce the risk of endogeneity.

Although our dependent variable is dichotomous, we use linear probability models in our 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, we also use cluster-robust standard errors.

As a precaution against collinearity, we 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 we use a moderately conservative VIF of 7. As robustness tests, we 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). We have decided to omit democracy because of its evident correlation with the Polity measure and interstate war due to

its correlation with our 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, we 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, we agree with 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, our data set has a number of missing observations, so model comparison measures could be misleading (Lall 2016).

2.2 Distributed Random Forest

We also make use of distributed random forest (DRF) (Breiman 2001a; The H2O.ai Team 2017) to measure the predictive power of our set of independent variables. Random 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 the risk of 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

[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 our core variables, we 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.

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; Strobl et al. 2007).

In this paper we use distributed random forest (DRF) to model our 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 our purposes. 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. When building decision trees, the DRF treats the missing observations as a separate category that can go either left or right. 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, we adopt an automated procedure to select the model parameters. We perform a grid search where the algorithm starts with a random combination of parameters, jumps to another randomly-chosen set, and then stops after it reaches a certain threshold (Cook 2017, 123). We follow the literature on predictive political science and use the area under the ROC curve (AUC) as our model evaluation metric (e.g., Hill and Jones 2014; Ward et al. 2010,

[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).

2013). We set the metric as follows: if five random models had not increased the AUC by at least in 0.1% comparing to the previous ones, the algorithm considers the result to be optimal.

We 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." We 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, we let the algorithm decide among using 10, 20, or 40 levels for each tree.

We test whether having balanced classes of our dependent variable (mass killing onset) affects the predictive ability of the model. Since the response measure is heavily imbalanced, oversampling the positive responses could potentially improve our results (Chawla et al. 2004; Del Río et al. 2014; Japkowicz and Stephen 2002).

We 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 we have 40 covariates of interest, we 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 we include the options of using 50% and 100% of the data. Similarly, we 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, we use four 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 offers four choices of histogram selection and we include all of them in our estimations.[5]

3 Results

We endeavour to answer three questions in our 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 1 summarises our 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.[6] 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.[7] This is our main statistic of interest, and we consider a covariate to be robust if it has a $CDF(0)$ of 0.9 or higher (Sala-i-Martin 1997, 181). Lastly, we 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 1: Extreme Bounds Analysis – Mass Killings (Robust Variables Only)

[5] The methods are described at https://goo.gl/jGvX2e (access: June 2018).

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

[7] We 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.

Seven variables pass our EBA criteria 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 we employ the UCDP measures of violent conflict. We 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 our 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). Our 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 we implement below and the sturdiness of this finding remains open to question.

Overall, the EBA indicates two patterns in answer to our 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 1 presents the six most important predictors of state-sponsored violence in out-of-sample tests. Overall, the random forest machine learning estimations have a good fit, with an

AUC of about 0.8 in the test samples. The random forest models confirm some of the main findings of EBA, yet they also show some interesting prediction patterns. Only democracy and state capacity appear to both be robust explanators of mass killing and strongly predict the likelihood of atrocity in the future. These patterns further support the refrain that stable states tend to stay stable states. Interestingly, several variables that are not robust explanators of mass killing in the EBA, have strong predictive power in the machine learning estimates. Parametrisation and interactions may explain this difference. The linear model imposes a parametric structure to the covariate, and the relationship between the 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.

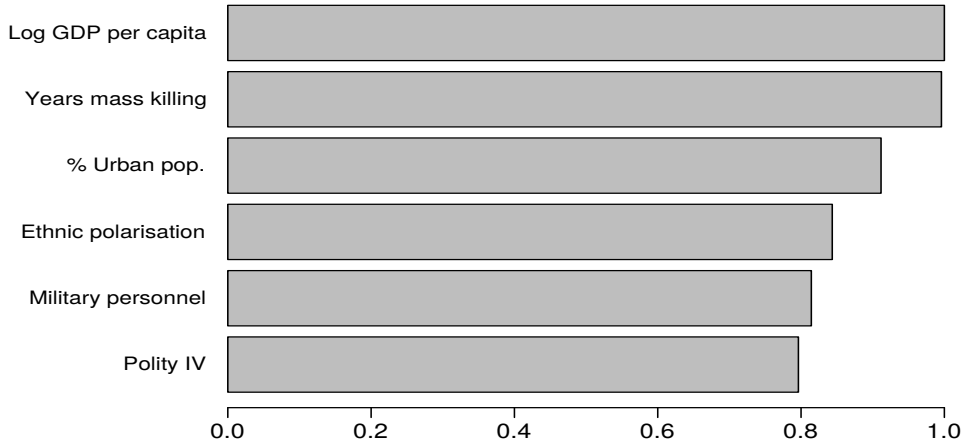

Figure 1: Distributed Random Forest – Variable Importance (Scaled)

Figure 2 displays the partial dependence plots for the six variables that the distributed random forests highlight as highly predictive 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, we are able to see that the effect of Military Personnel takes a significant jump in predictive power when the army's size reaches roughly 4,000 members.

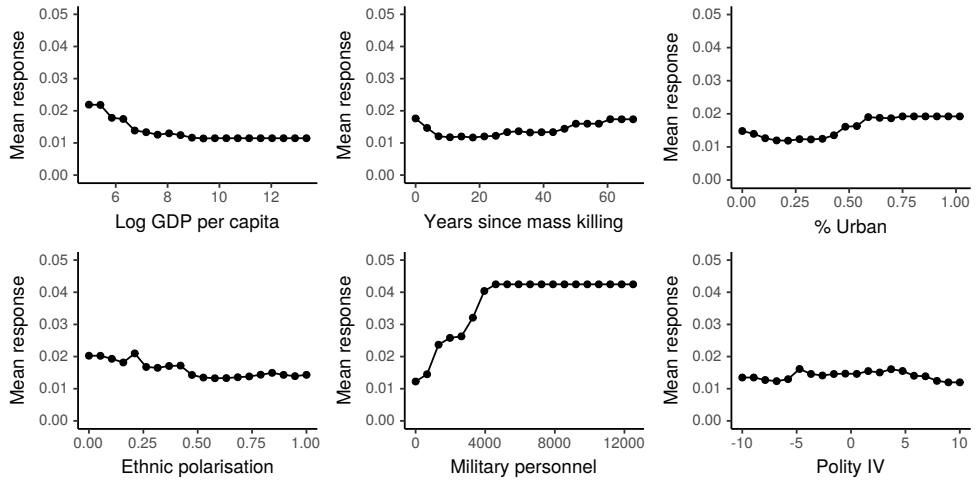

Figure 2: Distributed Random Forest – Partial Dependence Plots

Table 2 presents our results when we restrict our analysis to only civil war years to answer Question 2. We 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). We 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 we code civil war. Across our three codings, 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 we 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 2: EBA – Mass Killings during Civil Wars (Robust Variables Only)

When we analyse the predictive power of the variables across the three codings of civil war using our random forest analysis, we 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 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 our correlation-based models do not produce robust findings across different civil war data sets, our predictive 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 3–5 display the partial dependence plots for the variables with the highest predictive power in each of the three civil war data sets.

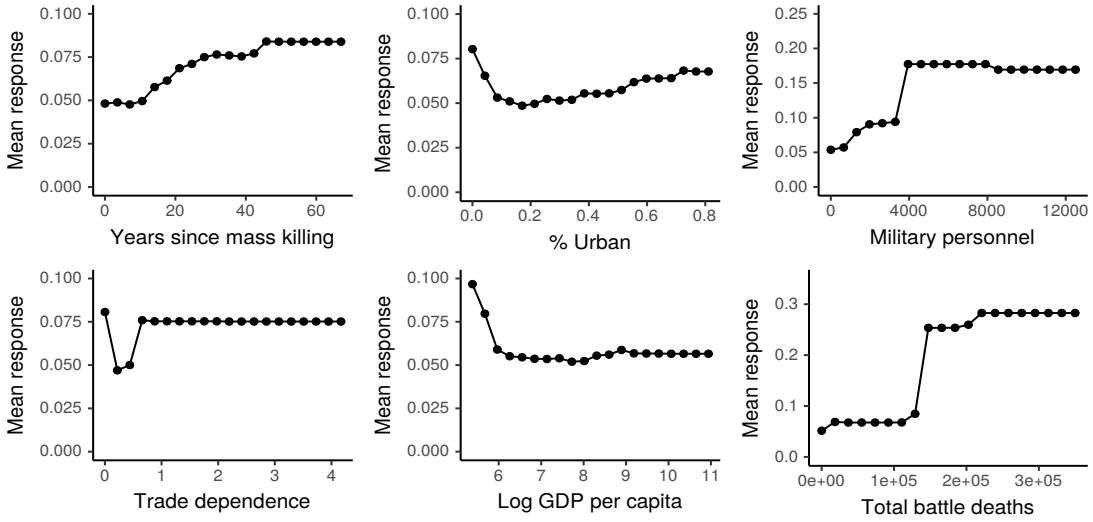

Figure 3: Partial Dependence Plots – Mass Killings during Civil Wars (UCDP Data)

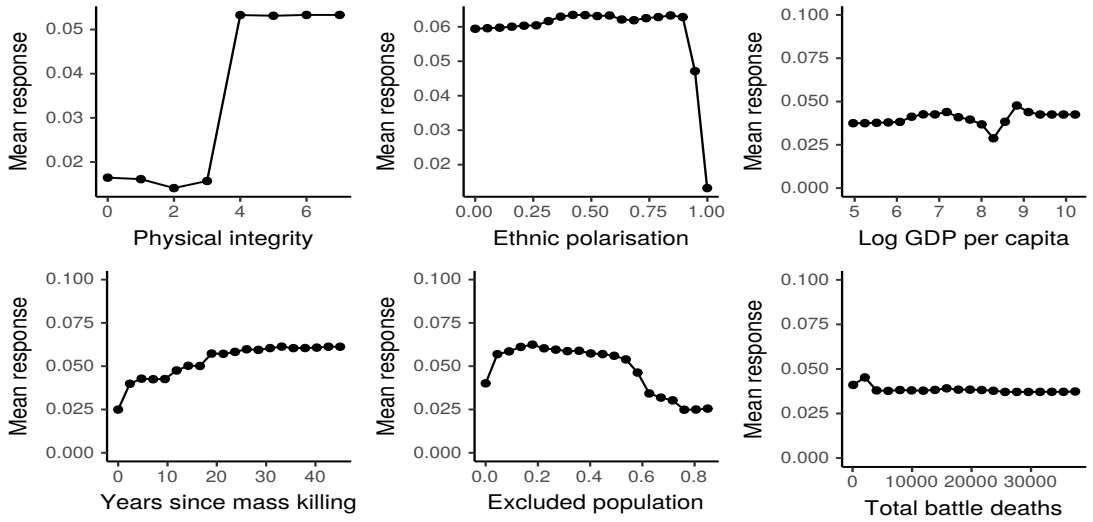

Figure 4: Partial Dependence Plots – Mass Killings during Civil Wars (COW Data)

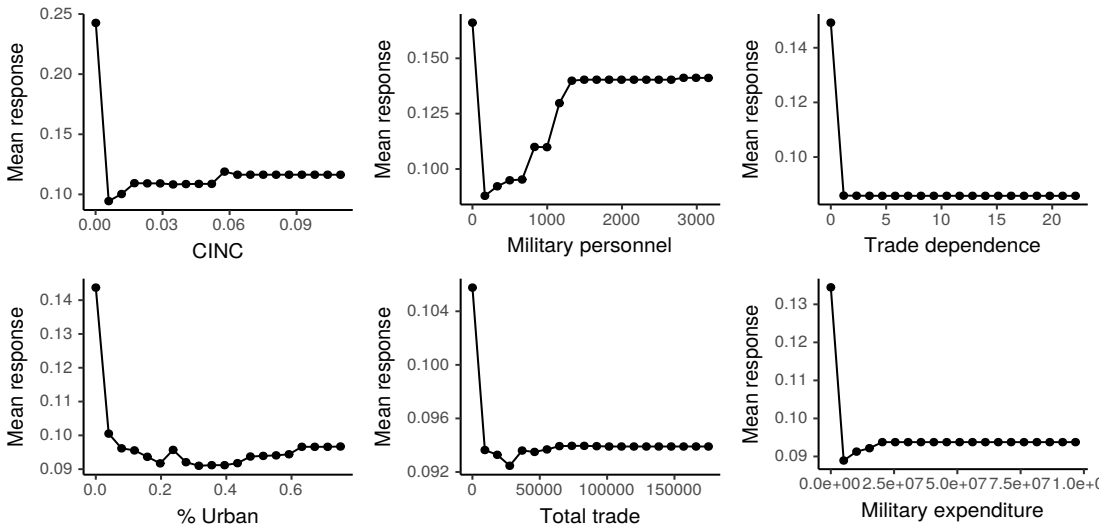

Figure 5: Partial Dependence Plots – Mass Killings during Civil Wars (Cederman et al. Data)

To answer Question 3, we estimate the same regressions using Harff's (2003) indicator of genocide and politicide. We find that no variable appears significant in our EBA models for genocide or politicide onset in the full data set. When we limit our 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, our dummy variable for political assassinations has a negative impact on the onset of genocides. Overall, from these EBA analyses, we 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 we 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 Additional Tests

We estimate a set of additional regressions to assess the robustness of our main findings. In regard to EBA, we include 10 variants of our original model. They largely confirm our prior results. First, we 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, we place different restrictions on the variance inflation factor (VIF) to test whether multicollinearity is driving our 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 our threshold by a very small margin.

We 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) we 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.

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, we performed a series of grid searches using three different seeds obtained from `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 online appendix.

5 Conclusion

In this paper, we 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. We find strong evidence that mass killings are unlikely to happen in rich, stable countries. Nevertheless, there is considerable heterogeneity in some of our 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 we see this diversity of outcomes under a positive light. Our results suggest new avenues for research, and we believe 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, 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).

References

Allansson, M., Melander, E., and Themnér, L. (2017). Organized violence, 1989–2016. *Journal of Peace Research*, 54(4):574–587.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

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

Bell, M. S. (2015). Examining explanations for nuclear proliferation. *International Studies Quarterly*, 60(3):520–529.

Bell, M. S. and Miller, N. L. (2015). Questioning the effect of nuclear weapons on conflict. *Journal of Conflict Resolution*, 59(1):74–92.

Besançon, M. L. (2005). Relative resources: Inequality in ethnic wars, revolutions, and genocides. *Journal of Peace Research*, 42(4):393–415.

Breiman, L. (2001a). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231.

Cederman, L.-E., Wimmer, A., and Min, B. (2010). Why do ethnic groups rebel? new data and analysis. *World Politics*, 62(1):87–119.

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.

Cheadle, D. and Prendergast, J. (2007). *Not on Our Watch: The Mission to End Genocide in Darfur and Beyond*. Dublin: Hachette Books.

Clarke, K. A. (2005). The Phantom Menace: Omitted Variable Bias in Econometric Research. *Conflict Management and Peace Science*, 22(4):341–352.

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.

Collier, P. (2003). *Breaking the Conflict Trap: Civil War and Development Policy*. Washington, DC: World Bank Publications.

Cook, D. (2017). *Practical Machine Learning with H2O*. Sebastopol: O'Reilly.

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.

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.

Doyle, M. W. and Sambanis, N. (2006). *Making War and Building Peace: United Nations Peace Operations*. Princeton, NJ: Princeton University Press.

Easterly, W., Gatti, R., and Kurlat, S. (2006). Development, democracy, and mass killings. *Journal of Economic Growth*, 11(2):129–156.

Elwert, F. and Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40:31–53.

Esteban, J., Morelli, M., and Rohner, D. (2015). Strategic mass killings. *Journal of Political Economy*, 123(5):1087–1132.

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.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

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.

Gassebner, M., Lamla, M. J., and Vreeland, J. R. (2013). Extreme bounds of democracy. *Journal of Conflict Resolution*, 57(2):171–197.

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.

Genuer, R., Poggi, J.-M., and Tuleau, C. (2008). Random forests: Some methodological insights. *arXiv*.

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.

Goldman, R. (2017). Assad's history of chemical attacks, and other atrocities.

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.

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.

Gurr, T. R. (2000). *Peoples Versus States: Minorities at Risk in the New Century*. Washington, DC: US Institute of Peace Press.

Hafner-Burton, E. M. (2005). Right or robust? the sensitive nature of repression to globalization. *Journal of Peace Research*, 42(6):679–698.

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.

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.

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.

Hegre, H. and Sambanis, N. (2006). Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution*, 50(4):508–535.

Hill, D. W. and Jones, Z. (2014). An empirical evaluation of explanations for state repression. *American Political Science Review*, 108(3):661–687.

Hlavac, M. (2016). ExtremeBounds: Extreme bounds analysis in R. *Journal of Statistical Software*, 72(9):1–22.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449.

Jones, Z. and Linder, F. (2015). Exploratory data analysis using random forests. pages 1–31.

Kalyvas, S. N. (2006). *The Logic of Violence in Civil War*. Cambridge: Cambridge University Press.

King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, pages 666–687.

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.

Krain, M. (1997). State-Sponsored Mass Murder: The Onset and Severity of Genocides and Politicides. *Journal of Conflict Resolution*, 41(3):331–360.

Lall, R. (2016). How multiple imputation makes a difference. *Political Analysis*, 24(4):414–433.

Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review*, 75(3):308–313.

Levine, R. and Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *The American Economic Review*, 82(4):942–963.

Levine, R. M. (1992). *Vale of Tears: Revisiting the Canudos Massacre in Northeastern Brazil, 1893–1897*. Berkeley: University of California Press.

Marshall, M. G., Gurr, T. R., and Harff, B. (2017). Pitf state failure problem set, 1955-2016.

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.

Muchlinski, D. (2014). Grievances and opportunities: Religious violence across political regimes. *Politics and Religion*, 7(4):684–705.

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.

Nichols, M. (2017). South sudan's government using food as weapon of war - u.n. report.

Nyseth Brehm, H. (2017). Re-examining risk factors of genocide. *Journal of Genocide Research*, 19(1):61–87.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5):673–690.

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.

Pinker, S. (2011). *The Better Angels of Our Nature: The Decline of Violence in History and Its Causes*. London: Penguin UK.

Rost, N. (2013). Will it happen again? on the possibility of forecasting the risk of genocide. *Journal of Genocide Research*, 15(1):41–67.

Rummel, R. J. (1995). Democracy, power, genocide, and mass murder. *Journal of Conflict Resolution*, 39(1):3–26.

Sala-i-Martin, X. (1997). I just ran two million regressions. *The American Economic Review*, 87(2):178–183.

Sarkees, M. R. and Wayman, F. W. (2010). *Resort to War*. Washington DC: CQ Press.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression. *Center for Bioinformatics & Molecular Biostatistics*, pages 1–14.

Spector, P. E. and Brannick, M. T. (2011). Methodological urban legends: The misuse of statistical control variables. *Organizational Research Methods*, 14(2):287–305.

Stanton, J. (2015). Regulating militias: Governments, militias, and civilian targeting in civil war. *Journal of Conflict Resolution*, 59(5):899–923.

Stanton, J. A. (2013). Terrorism in the Context of Civil War. *The Journal of Politics*, 75(4):1009–1022.

Straus, S. (2007). Second-generation comparative research on genocide. *World Politics*, 59(3):476–501.

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.

Straus, S. (2012b). Wars Do End! Changing Patterns of Political Violence in Sub-Saharan Africa. *African Affairs*, 111(443):179–201.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25.

Sturm, J.-E. and de Haan, J. (2002). How robust is sala-i-martin's robustness analysis? Technical report, University of Groningen: Mimeo.

The H2O.ai Team (2017). `h2o`*: R Interface for H2O*. R package version 3.14.0.3.

Ulfelder, J. (2012). Forecasting onsets of mass killing. Technical report. Accessed: January 2018.

Ulfelder, J. and Valentino, B. (2008). Assessing risks of state-sponsored mass killing. Technical report. Accessed: January 2018.

Uzonyi, G. (2016). Domestic unrest, genocide and politicide. *Political Studies*, 64(2):315–334.

Valentino, B. (2014). Why we kill: The political science of political violence against civilians. *Annual Review of Political Science*, 17:89–103.

Valentino, B., Huth, P., and Balch-Lindsay, D. (2004). "draining the sea": Mass killing and guerrilla warfare. *International Organization*, 58(2):375–407.

Ward, M. D., Greenhill, B. D., and Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375.

Ward, M. D., Metternich, N. W., Dorff, C. L., Gallop, M., Hollenbach, F. M., Schultz, A., and Weschle, S. (2013). Learning from the past and stepping into the future: Toward a new generation of conflict prediction. *International Studies Review*, 15(4):473–490.

Wayman, F. W. and Tago, A. (2010). Explaining the onset of mass killing, 1949–87. *Journal of Peace Research*, 47(1):3–13.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis*, 13(2):157–170.