77th Economic Policy Panel Meeting

20-21 April 2023

Power Mismatch and Civil Conflict: An Empirical Investigation

Laura Ogliari, Massimo Morelli, Long Hong

Power Mismatch and Civil Conflict: An Empirical Investigation*

Massimo Morelli [†] Laura Ogliari [‡] Long Hong [§]

February 28, 2023

Abstract

This paper empirically shows that the imbalance between an ethnic group's political and military power is crucial to understand the likelihood that a group engages in a conflict. We develop a novel measure of a group's military power by combining machine learning techniques with rich data on ethnic group characteristics and outcomes of civil conflicts in Africa and the Middle East. We couple this measure with available indicators of ethnic groups' political power as well as with a novel proxy based on information about the ethnicity of cabinet members. We find that groups characterized by a higher mismatch between military and political power are approximately 30% more likely to engage in a conflict against their government. We also find that the effects of power mismatch are nonlinear, which is in agreement with the predictions of a simple model that accounts for the cost of conflict. Moreover, our results suggest that high-mismatched groups are typically involved in larger and centrist conflicts. The policy implication is that power sharing recommendations and institutional design policies for peace should consider primarily the reduction of power mismatches between relevant groups, rather than focusing exclusively on equalizing military or political power in isolation.

Keywords: Civil War, Military Power, Political Power, Mismatch, Machine Learning

*Morelli wishes to thank the European Research Council, ERC Advanced Grant 694583. We wish to thank Giulia Buccione, Jacopo Bellabio, Matteo Bertoli, and especially Mohamed Badaoui for indispensable and amazing research assistance and Yannick Pengl, Dominic Rohner, Austin Wright, and Ekaterina Zhuravskaya for very useful comments and suggestions. We also thank seminar and conference participants at the Political Economy of Power Relation conference, the CEPR ITRE seminar series, Zurich Political Economy Seminar Series (Z-PESS) and Sciences Po. The usual disclaimer applies.

†Bocconi University, IGIER, Baffi and CEPR; massimo.morelli@unibocconi.it

‡University of Bergamo & CEPR; laura.ogliari@unibg.it

§University of Wisconsin - Madison, long.hong@wisc.edu

1 Introduction

When thinking about the causes of wars, the concept of *power* should come to mind. International relations scholars, however, almost exclusively refer to military power when studying conflicts. In this paper, we provide new empirical evidence highlighting the importance of simultaneously accounting for multiple dimensions of power. Herrera et al. (n.d.) establish theoretically that when bargaining fails – for example, for commitment or asymmetric information problems –, the probability of conflict depends on the *power mismatch* between the disputant groups, where the mismatch is defined as the difference between the relative military strength and the relative political-economic power of the disputant groups. If the two types of power are unequally but similarly distributed (e.g., the player that has greater military power also has greater political-economic power), no war should ensue. Conflict can instead arise when one of the players is relatively stronger in one of the two dimensions of power.[1] We contribute to the literature by providing the first empirical evidence that power mismatch matters for conflict. No systematic empirical evidence on the role of power mismatches exists, and such evidence could be crucial in order to assess the potential effectiveness of alternative policies for peace.

To progress in this direction, we build a novel measure of military power of ethnic groups, exploiting a wealth of ethnic group-level data and machine learning techniques. We combine this new measure with detailed information on civil conflict events, ethnic groups' characteristics, and measures of political power. We obtain a novel dataset, covering virtually the universe of politically relevant ethnic groups in Africa and the Middle East for the period 1992-2012. By using these new measures of the relative military and political power of ethnic groups, we can calculate the groups' power mismatch and use it to appraise, at the intra-state level, the role it plays in the decision to engage in conflicts.[2]

Using different specifications that account for many observed and unobserved confounding factors, we find that within the same country-year, high-mismatched groups have a much higher probability of partaking in a conflict than groups characterized by low mismatch.[3] The results are extremely robust, and the use of country$\times$year and ethnic group fixed effects should be reassuring in terms of identification.

When using a continuous measure of the mismatch, we also observe a convex relationship between power mismatch and conflict: when the mismatch is small, a marginal increase in the mismatch is not associated to the probability of conflict. On the other hand, a further increase in the mismatch raises the likelihood of conflict participation more than proportionally for high-mismatch groups. Lastly, we examine the relationship between mismatch and conflict characteristics. We show that mismatched groups have a higher chance to be involved in conflicts that are bigger (in terms of fatalities) and that concern

[1] The result is shown to hold even when allowing for bargaining because the evolution of military and political powers and their future use in case of an indecisive war cannot be contracted ex-ante. This is one of the main reasons why the power mismatch giving rise also to the Russian attack incentives with Ukraine is difficult to re-balance at a negotiation table.

[2] At the inter-state level Herrera et al. (n.d.) find some preliminary supporting evidence, using GDP ratios as a very rudimentary measure of economic power.

[3] As shown in section 5, a mismatched group is 50% more likely to partake in a conflict against the government than a similar group that is not mismatched.

power-sharing at the central level (centrist conflicts). The latter result is also consistent with additional theoretical findings in Esteban et al. (2022): it is precisely when a dispute is about power distribution in a country (centrist dispute rather than a dispute about autonomy or potential secession) that the mismatch of powers is salient, and the bigger the stakes, the bigger the conflict.

The international relations literature debates the pros and cons of balance versus preponderance of power, focusing on military power alone. Even theorists emphasizing commitment problems as the primary cause of war always refer to the difficulty to commit not to use military power, and the comparison with political economic power is mostly ignored (see e.g., the seminal work of Fearon, 1995). A notable exception is found in the body of work by Cederman (see e.g., Cederman et al., 2013 for a recent detailed analysis of the different grievances that might lead to civil war). Cederman's research stresses the role of economic inequality and political exclusion as a trigger for civil conflict. In the same spirit, in this paper, we use the concept of "relative political power" to capture the advantage conferred to a player by the existing political institutions—for example, the relative control of the political bodies governing the allocation of resources in peace.

Our paper speaks to the debate within the literature on power sharing. As discussed e.g. in Hartzell and Hoddie (2003) and Reilly (2012), the two opposed theories on how power sharing should be advocated to resolve or avoid conflicts are consociationalism and centripetalism: the former relates to proportional access to political power by all relevant groups, like in Lebanon; the latter refers to attempts to create multi-ethnic parties competing for power, like in Kenya, Indonesia and Nigeria. Both theories, in any case, focus on the pros and cons of different types of distributions of political power alone. The whole debate must shift, we argue, in the direction of considering simultaneously all the relevant dimensions of power, not just political power. The most successful case of power sharing agreement, the Good Friday Agreement of 1998, which was proposed to end conflict in northern Ireland, is a case in which indeed the power mismatch has been addressed, since deposition of weapons and access to political power and public sector jobs have been part of the multidimensional deal – see e.g., O'Leary (2001). As far as the parallel literature on the usefulness of proportional representation as a peace-inducing electoral mechanism is concerned, – see e.g., Horowitz (1990, 2000, 2003, 2005) – the example of the UNITA rebel group in Angola is telling. In the rhetoric of UNITA, civil war against the MPLA government in Luanda was justified by the group's exclusion from power at the time of independence in 1975. However, after signing a peace agreement in 1991 and losing the winner-take-all type of elections in 1992, UNITA returned to war until it was finally induced to sign a power-sharing agreement in 1994. This shows that elections are not a panacea unless the electoral system and proportion of voters can determine an implicit commitment to power sharing. However, even power sharing often fails since forms of credible power-sharing agreement do not necessarily reflect the desirable elimination of a mismatch: for example, a pure democracy with a proportional electoral system could guarantee a group with 30 percent of ethnic group voters 30 percent of political power, but if such a group has a probability of victory against the majority group much higher or much lower than 30 percent the mismatch is not eliminated.[4] Given that simply using elections (when they are fair) does not guarantee the credible elimination of a mismatch, democracy has to be supplemented by inventive institutional designs, for example, creating commitments in terms of public jobs, political roles, and military quotas in exchange for deposition of weapons, as in the peace treaty that led to the demilitarization of the IRA. Similarly, also in Colombia, the demilitarization of the FARC had to go hand-in-hand with the concession of a political role.

The paper is organized as follows: in section 2 we recall the baseline model that formalizes the mismatch theory of conflict; in section 3 we describe in detail the data collection efforts on all fronts, discussing the relevant previous literature; In section 4 we zoom in particular on the description of the machine learning procedures used to create our main novel military power measure. Section 5 presents the empirical results, and section 6 discusses the potential causal interpretation of the results. Finally, section 7 concludes. The appendix contains all the technical aspects and additional figures and tables.

2 A Simple Theoretical Framework

Consider a government controlling group G and an ethnic group E that has to decide whether to rebel or not.[5] Let $p \equiv \frac{p_E}{p_E + p_G}$ denote the relative political power of E.[6] Finally, let m denote the probability of winning of E in case of war against G and let c be the cost of war for each player.

Denote by S the divisible surplus and consider first a case in which $m > p$. If E decides not to challenge the status quo, her payoff is $U_E = pS$; on the other hand, in case of conflict, the payoff for E is, with the standard costly lottery assumptions, $mS - c_E$. Thus, conflict is initiated by E in a one-shot game iff $c_E < (m - p)S$. Given any ex ante uncertainty on c_E, represented by a distribution $F(\cdot)$ on the domain $[0, \infty)$, E rebels with probability $F((m - p)S)$, and hence the incentive to rebel is increasing in $(m - p)$, which represents the *mismatch*. It is also clearly increasing in the size of the divisible surplus.

Whenever the players make decisions on the basis of an expected cost (rather than knowing their own cost of war), then the mismatches below such an expected cost do not lead to war, while the ones above the expected cost do. For this reason, we expect to find in the data an intuitive form of non linearity: if a given increase in mismatch happens starting from a status quo with low mismatch, the impact on risk of war should be low, because such a marginal increase makes the new mismatch still likely to remain below

[4] See Spears (2000) for a comprehensive discussion on power-sharing agreements in Africa.

[5] In Herrera et al (2022) the focus is on showing that whichever the two players involved in a bilateral dispute are, the mismatch matters for war and duration. But most of the disputes that can lead to a conflict involve governments (or at least one group of those holding power). Given that in the empirical analysis of this paper we focus on the bilateral conflicts involving a government group, and given that this type of disputes are those ending in conflict with incomparably higher frequency, we limit attention to such pairs also in the brief sketch of the theory that we present in this section.

[6] In a parliamentary system one measure of this in the status quo could be the relative number of seats in the parliament or the relative number of ministries in the government. But in our sample, the regimes and meaning of political power are quite different from country to country, and the construction of an appropriate measure of relative political power is one of the contributions of the paper.

the expected cost. On the other hand, when the same marginal increase in the mismatch happens in a status quo with an already significant mismatch, the likelihood that the new mismatch is considered higher than the expected cost is higher. Hence the impact of a marginal increase in the mismatch should be expected to be higher in situations with an already high level of mismatch.

In the rare cases in which $p > m$, i.e., when an ethnic group has political power but is very weak militarily, the government may have an incentive to start a (repression) conflict if $c_G < [(1-m)-(1-p)]S$.[7] Conflict exist with the corresponding probability that c_G is less than $G\left((p-m)S\right)$, where $G(\cdot)$ denotes the cumulative probability distribution of the possible realizations of c_G.

The conflicts are usually initiated by ethnic groups that rebel against a status quo where they are given too little political-economic power. However, the data do not allow to distinguish initiation, and, moreover, both cases actually say the same thing in terms of the role of the mismatch. In sum:
Main prediction: Conflict is more likely to happen when $|m-p|$ is high.

The more general model, allowing for dynamics, bargaining and stalemates, can be found in Herrera et al. (n.d.).

3 Data description

To test the validity of the mismatch theory for civil conflicts, we need at least three pieces of information: (i) what are the relevant groups that may be tempted to participate in conflicts; (ii) the group's political power; (iii) the group's military strength. To test whether more mismatched groups are more likely to enter into conflicts, we need to know both military and political power, even for the groups that never participated in a conflict. To the best of our knowledge, none of these pieces of information is readily available in existing datasets. This section provides a detailed description of how we identify the ethnic groups in conflict and how we measure group-level political and military power.

3.1 Ethnic conflicts

Throughout the analysis, we focus on conflicts involving the government group (represented by at least one dominant ethnic group) and one or more ethnic opposition groups over the period 1992-2012. We restrict attention to conflicts involving a government related group because for such congflicts we can build meaningful measures of relative political and military power for the two opposing sides.[8] We restrict our attention to conflicts occurring in Africa and the Middle East to make sure that ethnicity represents a

[7]Powell (2012,2013) and Debs and Monteiro (2014) argue that an additional reason for these types of wars could be the fear of future power shifts that would make the group currently weaker militarily a stronger one to repress in the future. We do not need to invoke these considerations on expectations for the future for our simple goal to establish the relevance of the mismatch as a reason for war.

[8]On the contrary, in conflicts against civilians one of the side is not determined, while in conflicts between two rebel groups it is not clear what is the relevant measure of political power.

salient cleavage.[9] We construct a dataset that records ethnic groups and conflict information by linking the UCDP Georeferenced Event Dataset (UCDP-GED) and Ethnic Power Relations (EPR). In what follows, we describe in detail the procedure used to link the two databases.

UCDP-GED We use the UCDP Georeferenced Event Dataset (UCDP-GED) Global version 5.0,[10] which contains information on conflict events from 1989 to 2015. The database has four key features that make it particularly suitable for our purposes. First, it classifies the type of violence, allowing us to identify conflicts in which one of the actors is a government and the other is an organized rebel group. It also has separate coding for civil vs interstate conflicts – where one of the actors belong to a different country. Second, it provides detailed information on the precise location of conflict events, which will help us linking actors from different databases. Third, UCDP-GED reports all the incidents that result in at least one direct associated death, which allows us to extend the analysis to include also small-scale conflicts. Finally, the database also provides estimates of fatalities separately for civilian and each side involved in the conflict. This feature—not available to the best of our knowledge in any other civil conflict database—is essential to estimate the military power measure, as discussed in Section 4.

Ethnic Power Relation We use the list of ethnic groups provided in the Ethnic Power Relations (EPR) Core dataset (version 2018.1.1),[11] which identifies 817 politically relevant ethnic groups worldwide for the period 1946 - 2018. EPR defines ethnic groups according to the ethnic categories most salient for national politics in each country. An ethnic group is considered politically relevant if one or more significant political actors claim to represent their interests in the national political arena, or if the group's members are subjected to systematic and intentional discrimination in the realm of public politics. This feature of the database ensures that the groups analyzed are politically relevant and likely to represent organized actors who could make the decision to engage in a conflict.[12] It is worth noting that, due to the dynamic political environment of the country, the politically relevant groups are time-varying.

Government ethnicity EPR provides rich information on political power. Specifically, the database contains a variable that ranks each ethnic group's political power from 1

[9]Table B.1 in Appendix A reports the list of countries in our sample.

[10]The database is introduced by Sundberg and Melander (2013). See Eck (2012) for a detailed discussion on the strengths and weaknesses of the database.

[11]The database was first introduced by Cederman et al. (2010) and further developed by Vogt et al. (2015).

[12]Alternatively, one could use the Geographic Representation of Ethnic Groups dataset (GREG) which digitally represents settlement patterns of ethnic groups worldwide coming from a version of the Atlas Narodov Mira (ANM) (Bruk and Apenchenko, 1964): a series of maps collected by Soviet ethnographers charting ethnic groups across space. However, besides being potentially outdated, the main limitation of this dataset is that it focuses exclusively on the list of ethnic groups given by the ANM authors, even if the linguistic differences on which the ANM focuses do not correspond to ethnic cleavages that are politically relevant (see e.g., Posner, 2004; Cederman and Girardin, 2007; Chandra and Wilkinson, 2008; Wucherpfennig et al., 2011)

to 7.[13] In particular, if a group rules alone, the group is either "monopolist" (rank 7) or "dominant" (rank 6). If groups share powers, they could be either "senior partners" (rank 5) or "junior partners" (rank 4). If a group is excluded from power, the group is either "self-excluded" (rank 3), or "powerless" (rank 2), or "discriminated" (rank 1). Table B.2 shows the share of the observations we have in each power rank category.

We follow different steps to find the ethnicity of the government group in a given year. First, if a group rules alone, we label it as government group. Second, if more than one group has the highest power rank in a given year—this occurs for "senior partners"—we consult the EPR Atlas, containing details of how the political rank was created, to determine whether the groups are allied. Appendix B provides a detailed discussion about our manual checks. In case they are, we consider them both as government groups and treat them as a single entity. If they are not allied, we try to determine who enjoys a larger advantage using external sources. If we are unsuccessful, we exclude them from the analysis (3.9% of the sample).

Rebel group ethnicity Our objective is to assign conflict against the government to ethnic groups in the EPR dataset. This requires linking actors of the GED dataset to Ethnic groups in EPR. There is no direct correspondence between actors in UCDP-GED and ethnic groups in EPR so, we apply a multi-step procedure to assign ethnicity to the UCDP-GED rebel groups. As a first step, we use the ACD2EPR (Version 2021) conversion table developed by Wucherpfennig et al. (2011) and Vogt et al. (2015), which integrates UCDP/PRIO Armed Conflict Dataset (Version 17.1) with EPR.[14] Out of the 369 rebel groups present in our sample of countries, 78 can be matched directly through ACD2EPR (this amounts to 30.2% of the conflicts in our sample).

For the remaining rebel groups, we exploit the location of the conflicts they have been involved in to build a link with ethnic groups. We take all the conflict events (both against the government and against other rebel groups) which involve one rebel group.

We first build a link between rebel and ethnic groups, then, we assign the conflicts against the government the rebel group has been involved in to the corresponding ethnicity. As a first step, we overlay conflict events (UCDP-GED) on the geoEPR ethnic group's homeland polygons. Appendix Figure A.2 shows an illustration of this step. We start with 27,117 events in our main country list, reported in Table B.1. We specifically exclude all the conflicts where the exact location of the event is not known or coded (45.8%), drop all conflicts that are against civilians and foreign governments (25.3%), and discard all the conflict events outside the boundaries of any ethnic groups (2.1%). We further exclude the events that happen in the homeland of government and irrelevant groups (6.3%) as they do not contain useful information to locate the homelands of rebel groups.[15] The

[13] There is another category called "political irrelevance", which we exclude from the analysis. In a few cases, the country is in a state of collapse. We exclude the country-year observations from our primary analysis.

[14] UCDP/PRIO Armed Conflict Dataset, first introduced by Gleditsch et al. (2002), is an old version of the UCDP-GED. The conversion table contains a smaller set of rebel groups than that used in UCDP-GED because UCDP/PRIO only records large conflict events where the number of involved casualties is at least 25.

[15] In EPR all groups that are considered irrelevant do not have information on the ethnic homeland poly-

sample restrictions leave us with 2,975 events. This leaves us with approximately 5,000 ethnicity-event observations which we use to match rebel to ethnic groups.

Second, we count the number of times a rebel group has a conflict event in the homeland of a particular ethnic group. Finally, we assign the ethnicity with the highest count to the rebel group. In some cases, there is a tie among the counts. We choose the homeland of the ethnic group with the highest fatalities occurred. In rare cases, the fatalities are also in a tie. We break the tie by randomly choosing one.[16] Finally, having matched rebel groups to ethnic groups, we assign all the conflicts a rebel group participated in—including those occurring in the government homeland and in the homeland of irrelevant groups—to the corresponding ethnic group.[17]

To have a sense of the performance of this procedure, for the conflicts that are assigned directly through ACD2EPR we can compare the correct rebel-to-ethnic-group match with that generated by the geomatching procedure. If we do so, geomatching can identify correctly 59 of the 78 matches (76%). Zooming in on the "bad matches", the procedures fails when there are few conflict events on which to base the geomatch. For this reason, we exclude from the sample matches that are based on fewer than 3 events. We further provide robustness checks where results are obtained using a sample that excludes cases where ethnicities are identified by fewer than 5 events, as well as results based only on the ACD2EPR matches.[18]

To give a simple but concrete example, Figure 1 shows a map of Liberia with all its ethnic groups in EPR, which is geo-matched with all conflict events associated with the Liberians United for Reconciliation and Democracy (LURD) and Independent National Patriotic Front of Liberia (INPFL), respectively. Take LURD as an illustration of our geomatching procedures. We first exclude all the events in the dominant group's homeland and the events associated with the irrelevant group. Using the remaining conflict events, we assign LURD to Mandingo because most of their events took place in the ethnic group's homeland.[19] In the case of INPFL, most of the conflict events happen in the dominant group's homeland. Using only one event for identification, we would incorrectly infer its ethnic group as Mandingo, while ACD2EPR has identified it as Gio.

gons. Similarly, conflicts falling in the government homeland do not give useful information on the ethnicity of the rebel group, as we cannot assign to it the government ethnicity.

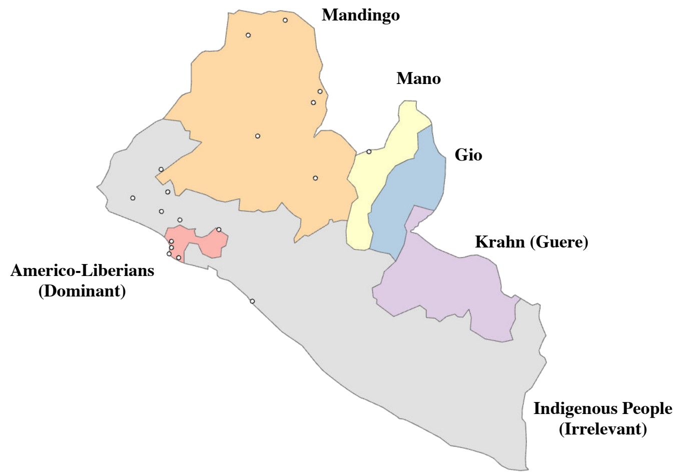
[16] We have also conducted a manual check to identify the ethnicity of rebel groups through online resources. Among the rebel groups that we could manually identify, around 70% are in line with the results from our method.

[17] Michalopoulos and Papaioannou (2016) and Moscona et al. (2020) use direct matching of battles with the ethnic group whose homeland contains the battle. Employing this direct match of battles to ethnicities in our event dataset would lead to inconsistent matches.

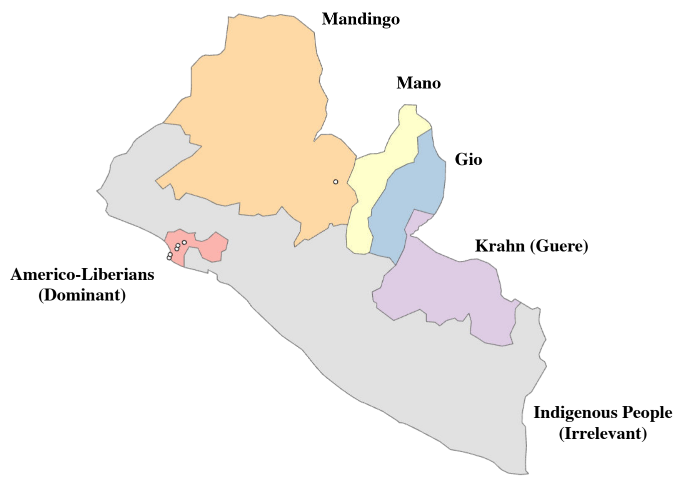
[18] As shown in Table 7 below, the results are close to our baseline results.

[19] We have also included a similar map of the National Patriotic Front of Liberia (NPFL) in Appendix Figure A.3. We have correctly identified the rebel's ethnic group as Gio.

Figure 1: Rebel group coding: LURD and INPFL in Liberia as an example



(a) LURD conflict events



(b) INPFL conflict events

Notes: This is a map of Liberia, where each colored polygon represents an ethnic group listed in EPR. The dots represent all conflict events in UCDP-GED associated with the rebel group LURD or INPFL.

Key dependent variables Our primary goal is the analysis of civil conflicts between incumbent governments and (ethnic) rebel groups. Using the information on involved parties in UCDP-GED, we restrict our attention to conflict events where one of the actors is the government and the other is an organized rebel group. We aggregate the events at the ethnic group-year level using each rebel group's ethnicity. We use conflict *incidence* as dependent variable, an indicator that equals one if the ethnic group is involved in a (at least one) conflict event in a given year.[20]

[20]We use all the conflict events – regardless of the location precision and whether it happens in the homeland of the government – but conflicts against civilians and foreign governments are excluded from our

3.2 Relative political power

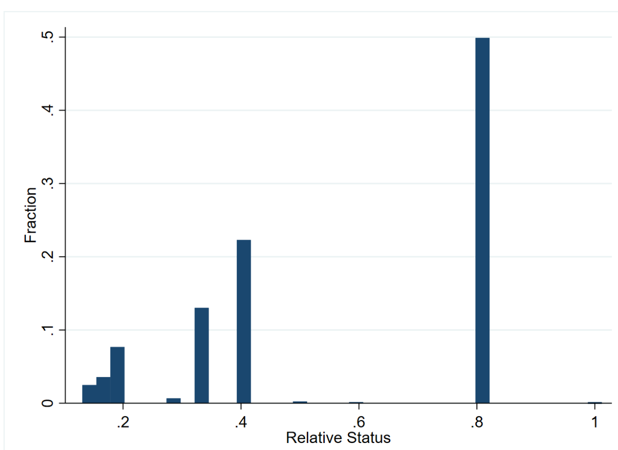
Building a measure of the political power of ethnic groups with respect to the government presents challenges. The literature used relative group size and the ethnicity of the leader or relied on experts' opinions. We build two proxies of groups' political power. The first approach directly uses the discrete power rank index provided in EPR. The advantage of the measure is that it is available for all the politically relevant ethnic groups in our sample. However, since it is a discrete index (i) it does not present a lot of variation, and (ii) the index is an ordinal variable, and its value does not represent any "real" quantification of power. In our second approach, we restrict our attention to a sub-sample of African countries and exploit the ethnicity of cabinet members to build a continuous measure of ethnic groups' political power. We discuss the two approaches in detail below.

Discrete measure We assign EPR's power rank as a measure of the political power of each ethnic group and define the relative political power, p_{eg}^{PR}, of ethnic group e to government g as follows:

$$p_{eg}^{PR} = \frac{C_e^{PR}}{C_g^{PR}} \in (0,1),$$

where C is the discrete power rank. Figure 2 shows the distribution of the measure. As expected, the distribution of p_{eg}^{PR} contains few points. Groups with high political power are basically those with $p_{eg}^{PR} = 0.8$, that is, groups that are considered junior partners in the government (i.e. their rank is 4 and the government group has rank 5). Groups with low political power are groups that are considered either powerless or that are discriminated against.

Figure 2: Histogram of the relative political power p_{eg}^{PR}



analysis.

Continuous measure We collect data about cabinet membership in 14 countries in Sub-Saharan Africa over 21 years (from 1992 to 2012).[21] We choose these fourteen countries for two reasons. First, their location is in Saharan Africa, where conflicts are most likely to be ethnic related. Second, as shown in the previous works by Francois et al. (2015); Rainer and Trebbi (2016), these countries, while not democracies, are organized with some form of power sharing configuration, and the share of cabinet members can be considered as a valid measure of how political power is distributed among ethnic groups.[22][23]

To identify the ethnicity of cabinet members we follow a procedure that entails different steps. First, we obtain all the cabinet membership information from the CIA's "Chiefs of State and Cabinet Members of Foreign Governments".[24] We precisely extract all the incumbent cabinet members at the end of each calendar year. We manually check the cabinet members' names using various online sources to identify their ethnic affiliation and use that information to match the ethnic list in EPR. If direct evidence is not available, we turn to two alternatives: the ethnicity of the minister's parents and the birthplace. Specifically, using the coordinates of the birthplace and the geo-referenced EPR map, we assign the ethnic homeland of the birthplace of the cabinet member as the ethnicity. When the birthplace is not available (e.g., some minister is born in a foreign country), we use the location of his primary school or the location of the district of her first election. With this procedure, we identify 82 politically relevant ethnic groups.[25]

The share of the cabinet seat held by an ethnicity should represent the share of the political power of the ethnic group insofar as cabinet members decide the allocation of the resources of a country. In fact, if cabinet membership contained information on political power we would expect it to monotonically decrease with the EPR Power rank measure. Table 1 confirms such a relationship: dominant groups on average hold more than 50% of the cabinet seats. On the other hand, the powerless or discriminated groups tend to have negligible shares.

We define the relative political power $p_{i,c}$ of group i in country c as the ratio of the cabinet seats of the group in a given year relative to the seats held by the government's ethnicity, g:

$$\hat{p}_{i,g,c} := \frac{\#\big(\text{Cabinet seats belonging to ethnicity } i\big)}{\#\big(\text{Cabinet seats belonging to ethnicity } g\big)}$$

Figure 3 plots the distribution of $\hat{p}_{i,g,c}$. Unlike the discrete measure, the relative political measure now is quite smooth. It also contains a mass point at zero, which indicates

[21] They are: Benin, Cameroon, Republic of Congo, Democratic Republic of Congo, Ivory Coast, Gabon, Ghana, Guinea, Kenya, Liberia, Nigeria, Sierra Leone, Tanzania, and Uganda.

[22] Indeed, if you look the EPR power rank index for the government group of these countries, only 4 of them (Guinea, Nigeria, Sierra Leone, and Tanzania) have periods where the government group ruled alone (rank 6 - dominant), in all other cases the government group has rank 5 (senior partner). This is in line with Francois et al. (2015) claim that in these countries some form of power sharing is in place.

[23] Francois et al. (2015) and Rainer and Trebbi (2016) use a different, much finer, categorization of ethnicity from EPR.

[24] For detailed information, please refer to the CIA website: `https://www.cia.gov/resources/world-leaders`.

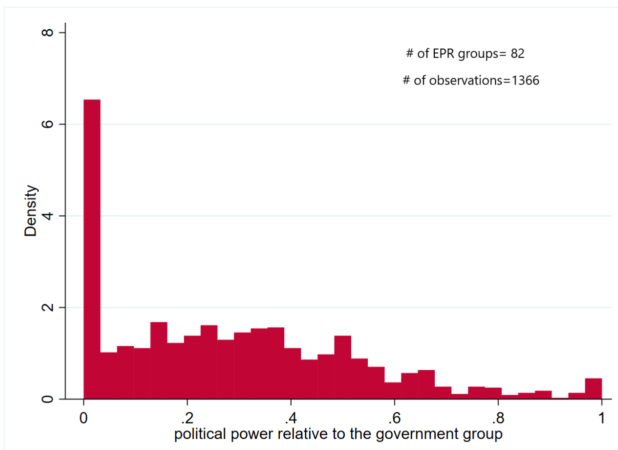
[25] There are 2,557 raw events in the 14 countries of interest from 1992 to 2012. Among these relevant groups, there are 1,866 conflicts.

Table 1: A comparison between EPR power rank and Cabinet member shares

		Political power (cabinet member shares)	
		Mean	Standard Dev.
Power Rank	Dominant	0.53	0.188
	Senior partner	0.261	0.141
	Junior partner	0.143	0.125
	Powerless	0.092	0.098
	Discriminated	0.068	0.063

that many groups are not represented in the cabinet, implying a certain degree of political power inequality in the restricted sample.

Figure 3: Histogram of relative political power



3.3 Relative military power

Finding measures of military power at the ethnic-group level is extremely challenging. Traditional information that is available at the country level – military expenditure, military personnel, trade in arms – does not exist at the group level. This is why the literature often uses the group's population size or GDP per capita as proxies of military power. In a recent paper, however, Carroll and Kenkel (2019) shows that these two variables perform no better than random guesses when used to predict the probability of winning a conflict in the context of inter-state conflict.

To be consistent with the theory described in Section 2, we define relative military strength as the probability of winning a conflict against the government. Estimating such a probability poses one difficulty: we need to compute the probability of winning a conflict for those groups that have never participated in a conflict. Moreover, even for groups that did participate in conflicts, inferring the probability of winning during the peace years requires a non-trivial technique.

We rely on some insights in Carroll and Kenkel (2019), who propose a machine-learning technique to overcome the challenges. We modify their algorithm and use an extended sample (described below) of conflicts in Asia and Africa combined with a rich set of ethnic group-level and country-level variables to infer the probability of victory for all potential conflicts between every ethnic (rebel) group and the government. The details of the machine learning procedure are described in Section 4.

3.4 Mismatch measure

With the measures of political and military power in hand, we can construct our primary independent variable: the empirical power mismatch measure $M_{e,g}$. We propose two measures of power mismatch at the group level: the Mismatch Dummy $M^D_{e,g}$, and a continuous mismatch variable. The mismatch dummy is an indicator that takes value one if political power is low and military power is high or vice versa. Specifically, we define the mismatch dummy as follows:

$$M^D_{e,g} = \begin{cases} 1 \;\; if \;\; (p^{PR}_{eg} \leq \bar{p}^{PR}_{p50} \wedge m_{eg} > \bar{m}_{p66}) \vee (p^{PR}_{eg} > \bar{p}^{PR}_{p50} \wedge m_{eg} \leq \bar{m}_{p33}) \\ 0 \;\; otherwise \end{cases} , \tag{1}$$

Where $\bar{p}^{PR}_{p50}$ is the median of the distribution of relative political power computed using the EPR index, and $\bar{m}_{p33}$ and $\bar{m}_{p66}$ are the values of the first and third terciles of the military power distribution. In other words, $M^D_{e,g}$ captures the presence of a high imbalance between the relative political power of a group and its relative military strength. Figure 4 panel (a) shows the share of years ethnic groups are defined as mismatched according to this indicator.

The continuous measure is defined as

$$M_{e,g} := |m_{e,g} - p_{e,g}| \tag{2}$$

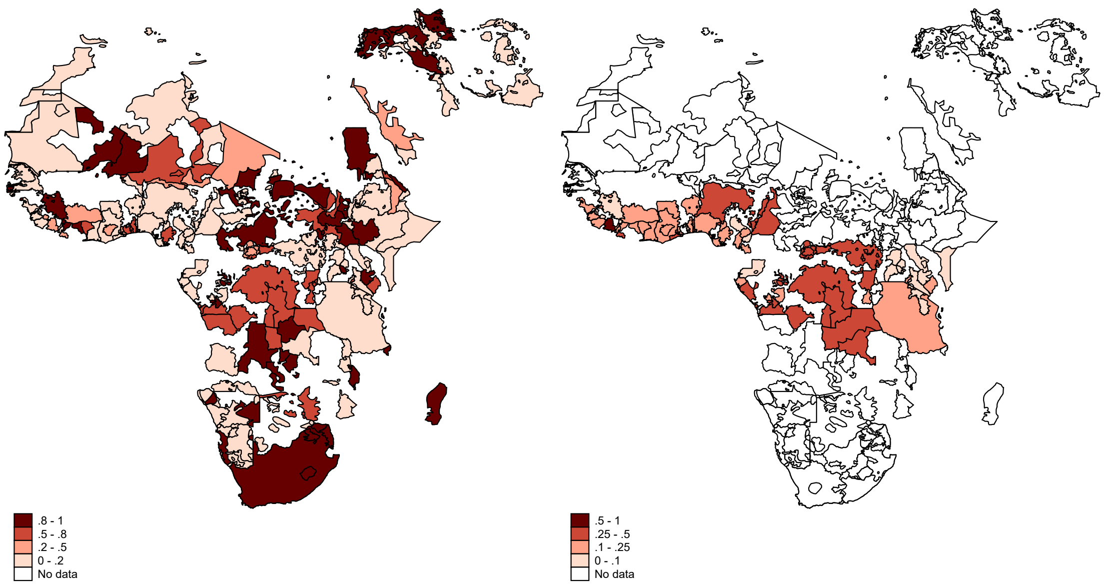
where $m_{e,g}$ is the predicted probability of winning a conflict against the government, and $p_{e,g}$ is the relative political power measured by using the ethnicity of the cabinet members. Note that since both $m_{e,g}$ and $p_{e,g}$ change over time, our mismatch definition is also time-varying. Figure 4 panel (b) shows the spatial distribution of the average of the continuous measure of mismatch for the restricted sample. Figure A.1 in the appendix plots the within group standard deviation for the two variables.

3.5 Control variables

In our analysis, we control for an extensive set of ethnic-level variables mainly constructed from GROWup (2019) and GRID-PRIO (v.2.0).[26] Control variables are grouped into five categories. These should account for possible determinants of war highlighted by the literature on conflict. First, since groups that have experienced a recent conflict are more likely to participate in another conflict, we include information on the number

[26] GROWup is developed by Girardin et al. (2015) and GRID-PRIO is introduced by Tollefsen et al. (2012)

Figure 4: Mismatch Measures



(a) Mismatch Dummy
Share of years with Mismatch Dummy=1

(b) Continous measure of Mismatch
Average value of the ethnic group

of peace years of each ethnic group. Second, as the fates of co-ethnic groups may affect both conflict likelihood and power mismatch, we control for information about kinship relationship. Third, we include geographic characteristic such as land area, elevation, distance to capital and country borders. Forth, as natural resources may be linked both to conflict events and the political power of ethnic groups, we include information on gold and diamond veins, diamonds and gem mines and active oil production. Finally, in some specifications we include socio-economic controls: share of land devoted to agriculture, share of land devoted to pasture, population, urban population, nightlight density, and group inequality.

4 Predicting military power via machine learning

Our goal is to estimate the probability of winning for all ethnic groups, each against their corresponding government. However, when an ethnic group has never experienced a conflict, traditional methods cannot be used for estimation. According to Carroll and Kenkel (2019), the standard approach of using a linear probability model to predict the winning probability of a conflict may result in accuracy no better than a random guess. Therefore, we use a machine learning algorithm, adapted from theirs and incorporating a rich set of observed ethnic group-level variables, to compute the probability of winning for all ethnic rebel groups against their government in our sample. In the following sections, we describe our *training dataset* used to train our machine learning algorithm, and discuss the algorithm and its performance.

4.1 Training set

In order to train our algorithm, we require a training set that includes plausible conflict outcomes – which side wins the conflict. However, this variable is typically not reported in standard databases such as GED. Therefore, we use a novel approach to determine the winning side of a conflict. We leverage the idea that the winning party should have a lower number of fatalities (relative to the population) than the losing party, and thus use the fatalities ratio as a measure of whether a group has won the battle. This approach is supported by the strong correlation between winning and having a smaller number of deaths (Sarkees and Wayman, 2010).

Formally, let f_g^t, f_e^t be government g and group e's fatalities (normalized by the corresponding group's population) in year t, respectively. We define the binary outcome Y, which is equal to 1 if the government wins and 0 otherwise:

$$Y := \mathbb{I}\{F_{eg}^t = \frac{f_g^t}{f_g^t + f_e^t} \geq c\} \tag{3}$$

That is, the ethnic group wins the conflict at time t if its share of fatalities is less than a threshold c.[27]

Figure 5 displays the distribution of F_{eg}^t in red. The graph shows two focal points at the two endpoints, indicating that the majority of fatality ratios are concentrated at zero or one. The remaining ratios are distributed (almost) uniformly. Based on this data pattern, we choose the threshold $c = 0.5$. However, the results remain robust even if we choose a different threshold.[28]

We use a rich set of predictors to train the model. In particular, we include ethnic-group level demographic, geographic, and meteorological information, as well as their external support and ethnic kinship information, extracted from the GROWup database. We include year dummies, region dummies, and the longitude and latitude of countries to capture aggregate trends. A detailed list is discussed in Appendix C.1. We adopt an advanced learning algorithm, which conducts both variable selection and cross-validation testing (more details in Appendix C.3).

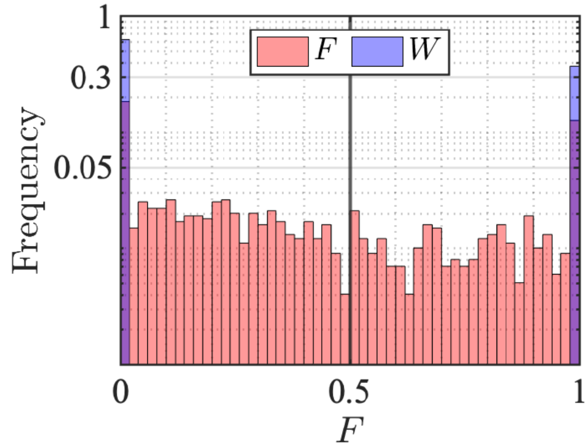
Since machine learning methods are powerful when data points are dense, to train the algorithm we use augmented sample of African and Asian conflicts from 1989 to 2016. Moreover, since we want the fatality ratio to be a sensible measure, we only consider conflicts with 25 or more fatalities in our training dataset.[29] In the final training sample, we have 726 observations with 198 predictors, where the share of outcomes where government wins is around 64%.

[27] We do not consider civilians deaths in the computation of the fatality ratio.

[28] An alternative dataset for documenting inter-state conflict outcomes is ACLED, where battle outcomes can be inferred by exploring whether control of territory has changed. However, using this method, we would categorize as as stalemates more than 90% of the conflicts. Correlates of War (COW) also reports conflict outcomes, although the number of conflicts that match our sample is very limited. Using the matched conflicts, if we attribute victory to the side with less fatalities we guess correctly 82% of the conflict outcomes.

[29] Results remain qualitatively similar when we restrict the sample to countries in the extended sample or if we consider also small conflicts, but the predictive power of the algorithm is reduced.

Figure 5: Training outcome distribution and fatalities ratio



Note: In red, the distribution of the fatality ratio F in our training set. In blue, the outcome, i.e. the *binarisation* of the ratio according to eq. 3. The frequency is represented on a logarithmic scale for better visualisation.

Algorithm Formally, we use a binary learning model $\mathcal{M}$ based on the training data (Y, X), where X contains all the ethnic level characteristics as described in Section 4.1 and Y is a binary outcome variable defined in equation (3). The optimal trained model $\mathcal{M}^*_{(Y,X)}$ is obtained by minimizing a pre-defined loss metric $\mathcal{L}$ given our dataset (Y, X), as shown below.

$$\mathcal{M}^*_{(Y,X)} = \arg\min_{\mathcal{M}\in\mathbb{M}} \mathcal{L}\big(Y - \mathbb{I}\{\mathcal{M}(X) > 0.5\}\big). \tag{4}$$

Intuitively, we look for the model with the highest prediction level given the training data. We also allow cross-validation techniques to avoid the over-fitting problem. Using the trained model, we can predict the winning probability in all situations.

It is important to point out that what we label military power, is the predicted probability that a group wins a conflict against the government. This implies: (i) the military power estimate is a "dyadic" measure, so it would be best to compare it within country; (ii) given the predictors we feed the ML algorithm, what we are estimating is "potential" or structural military power, which differs from actual military personnel, military expenditure, size of the armaments, etc.

Performance The performance of a binary classification model is commonly measured by the logarithmic loss metric. Given the classification model $\mathcal{M}^*_{(Y,X)}$ obtained from equation (4), the log-loss is calculated as:

$$\mathrm{L}\big(\mathcal{M}^*_{(Y,X)}\big) = -\frac{1}{N}\sum_i Y_i \log\big(\mathcal{M}^*_{(Y,X)}(X_i)\big) + (1 - Y_i)\log\big(1 - \mathcal{M}^*_{(Y,X)}(X_i)\big).$$

The smaller it is, the closer the likelihood of correct classification. We are interested in the cross-validated log-loss for the sake of maximal prediction accuracy. Following Carroll and Kenkel (2019), we calculate the Proportional Reduction Loss (PRL) that gives the predictive power accuracy of our model relative to a null model i.e. a classification scheme that assigns the label by majority rule (hence accurate at the 64.2%). The higher the PRL, the stronger is our prediction compared to the null model. The PRL writes:

$$\mathrm{PRL}\big(\mathcal{M}^*_{(Y,X)}\big) = \frac{\mathrm{L}_{null} - \mathrm{L}\big(\mathcal{M}^*_{(Y,X)}\big)}{\mathrm{L}_{null}}.$$

The loss L_{null} from a null model is always higher than the trained model's loss $\mathrm{L}\big(\mathcal{M}^*_{(Y,X)}\big)$. The higher the PRL, the bigger the performance normalized difference of $\mathcal{M}^*_{(Y,X)}$ with respect to the null model, and the better is our model.

Our reported PRL is comparable with the one obtained by CK for inter-state conflict. They obtain a PRL of 23%, only slightly higher than ours. Considering that we predict the outcome of intra-state conflicts, on which the data is very limited, the performance is quite good. Among the predictors that CK use, there are important metrics such as Iron and Steel production, Military expenditure and personnel, and Primary Energy consumption. Such data is absent for ethnic groups. The variables used to build our model are much harder to calculate, and most of them rely on satellite images.

Compared with the conventional military power proxies used in the literature, where population and night light have been commonly used as the main proxy for relative military power in intra-state conflicts (e.g., Esteban et al. (2012) use population). Table 2 shows that the population ratio or the night light ratio can predict the outcomes only marginally better than random guessing. Our algorithm, on the other hand, performs 15% better than random guessing based on the PRL, which is a considerable improvement on state of the art.[30]

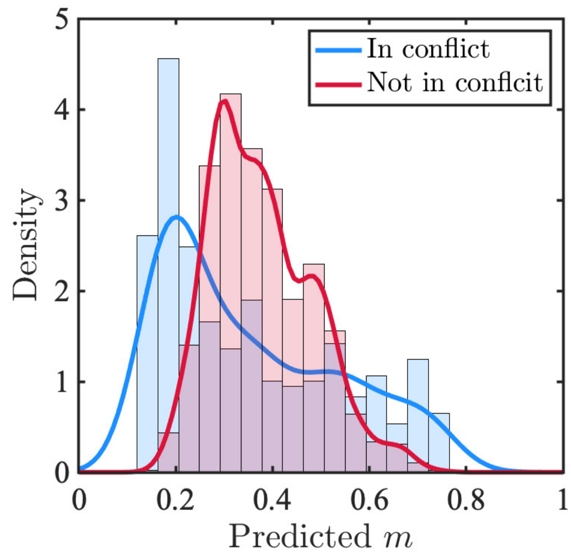
Table 2: Algorithm's predictive power.

	CV Log-loss	PRL	Accuracy	Δ_{Null}(Accuracy)
Full model	0.554	15%	70.2%	9.1%
Population ratio	0.650	0.2%	64.4%	0%
Night light ratio	0.646	0.8%	65.1%	1.1%

Since groups that experience conflict are a selected sample, one concern is that the predicted military power of the groups that are in conflict (and that we use to estimate the parameters) is systematically different from that of groups that do not experience conflict. Given that in the empirical analysis we use (a transformation of) military power to predict conflict participation, we want the distribution of military power for the groups in the training sample and for the groups that do not experience conflict to have common support. Figure 6 shows the distribution of predicted military power for groups who are not in conflict (in red) and those who are in conflict (in blue). The figure shows that groups that are in conflict have a more dispersed distribution, but the two distributions have common support.

[30]Notice that, even if the algorithm does not correctly predict who is more likely to win, by looking at fatalities the exercise might still be able to approximate how threatening the rebel group is for the government. The explanatory power of the variable could be even higher than what the validation would suggest.

Figure 6: Distribution of the predicted military power



Finally, Table 3 reports the correlation of the predicted probability of winning a conflict, which we will use as our baseline measure of military power *m1*, with alternative predicted probabilities based on different proxies for the winning outcome. In particular, in model *m2*, we change the threshold we use to assign the outcome of the conflict based on the fatality ratio. To do so, we exploit the Correlates of War database, which is, to the best of our knowledge, the only database that reports both the outcome of a conflict and the fatalities borne by each party. We restrict the sample to intra-state conflicts against the government with a win/lose outcome, and compute the fatality ratio that maximizes the correct classification of the conflict outcome. We find this ratio to be 0.58, which yields a correct classification in 82% of the cases.[31] Model *m2* assigns a loss if the fatality ratio is greater or equal to 0.58, and the prediction of this model is extremely correlated with the baseline prediction. In model *m3*, instead, we include small conflicts (i.e., those with total fatalities lower than 25) in the training sample. Including small conflicts adds a bit of noise to the outcome variable, this reduces the performance of the model and yields a prediction of military power that is slightly less correlated with the baseline one. In model *m4*, instead, the fatality ratio is built using cumulative fatalities (fatalities over the entire conflict duration) instead of the yearly fatalities, keeping the threshold fixed at 0.5. Again, using cumulative fatalities yield predictions that are very similar to those of the baseline model. Appendix C.2, provides further results that show the robustness of the predictions to changes in the machine learning parameters.

[31] Using our preferred threshold of 0.5 we correctly classify the outcome of the conflicts in the COW database 77% of the times.

Table 3: Correlation of military power predictions

	$m1$ Baseline	$m2$ Fatality threshold .58	$m3$ Small conflicts	$m4$ Cumulative deaths
$m1$ Baseline	1.00			
$m2$ Fatality threshold .58	0.952	1.00		
$m3$ Small conflicts	0.722	0.649	1.00	
$m4$ Cumulative deaths	0.924	0.858	0.796	1.00

4.2 Variables' relevance and summary statistics

We can explore the question of which predictors might be the most important in predicting military power. However, due to the nature of the algorithm, it is somewhat challenging to tell how much a single predictor affects military power. We explore the importance of a particular set of predictor using the following algorithm (following CK). We remove a predictor of interest, rerun the entire algorithm, and compare the resulting PRL to the original PRL. Effectively, the larger the difference between the resulting PRL and the original one (or PRL loss), the more important the predictor is. Instead of having an extensive list of all variables, we report in Table 4 some of the important predictors, which are external support variables (Ext), geographic variables i.e., border distance, capital distance, and travel time (Geo), population and demographic growth variables (Pop), peace years and war history (Py Wh), affiliated ethnic groups (Tek), land characteristics (Land), country-level variables, i.e., latitude, longitude and region dummy (Country).

Table 4: Variables' predictive power.

	Ext	Geo	Pop	PyWh	Tek	Land	Country
PRL loss (%)	3.1	1.9	2.3	5.6	1	1.6	0.5

As expected, characteristics related to conflicts (war histories, ability to have external support from other national or international powers) are important to explain conflict outcomes. But also the size af the group and the geographic characteristics of the ethnic homeland seem to play an important role. Finally, Table 5 provides summary statistics for the measure of the estimated military power. Observing the number we see that the majority of the variation of the variable comes from differences across different ethnic groups of the same country-year, while the within group variation is quite low. For this reason, we always report in the analysis results exploiting both the within group variation and variation within country-year.

5 Empirical Analysis

This section is devoted to the exploration of the main prediction by Herrera et al. (n.d.), namely that conflict is more likely when groups are mismatched. We start by analyzing

Table 5: Military power

	Obs.	N. of groups	Mean	Median	Max	Min	sd overall	sd within country-year	sd within group
military power	4260	238	0.282	0.217	0.866	0.076	0.184	0.114	0.048

the relationship between the government-ethnic-group power mismatch and conflict incidence using the indicator of mismatch in an extended sample of countries in Africa and the Middle East. [32] Then, we restrict the focus of our analysis to the sample of 14 African countries for which we could build the continuous measure of power mismatch. Finally, we conclude the section by providing additional results on mismatch and conflict type.

5.1 Evidence using the Mismatch Dummy

Table 6: Power Mismatch and Conflict Incidence

Dependent variable: conflict incidence

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mismatch dummy	0.0854***	0.0633***	0.0714***	0.0609***	0.0624***	0.0546***	0.0803***	0.0461**
	(0.0185)	(0.0154)	(0.0155)	(0.0144)	(0.0139)	(0.0171)	(0.0217)	(0.0227)
Observations	4,260	4,260	4,260	4,260	3,612	2,674	4,260	4,260
R-squared	0.335	0.561	0.566	0.574	0.588	0.660	0.596	0.708
Controls								
Peace years		✓	✓	✓	✓	✓	✓	✓
Family			✓	✓	✓	✓	✓	✓
Natural resources				✓	✓	✓	✓	✓
Geographic					✓	✓		
Socio-ecomic						✓		
Fixed effects								
Country × year	✓	✓	✓	✓	✓	✓		✓
Ethnic group							✓	✓
Year							✓	

Note: The dependent variable is conflict incidence. Control variables are defined as follows. *Peace years* is the number of years since the last time the group participate in a conflict; *Family controls* are: the population of kin groups that is in power (log) and dummy variables that indicate whether groups in the family had an upgrade or a downgrade in their power rank over the previous two/ten years. *Natural resources controls* are: dummy variables for the presence of gold, diamonds, and other precious gems; a dummy variable that indicates active oil extraction. *Geographic controls* are: (log) distance from the capital, (log) distance from the closest border, (log) area of the group homeland, and (log) mean and standard deviation of the elevation. *Socio-economic controls* are: 1990 share of group's homeland used for agriculture, 1990 share used for pasture, 1990 share of group homeland that is urbanized, 1990 population (log), 1990 nighttime luminosity (log), and group inequality measured as $ln[(group\ nightlight\ per\ capita)/(government\ nightlight\ per\ capita)]^2$. In all specifications standard errors are clustered at the country-year level. *** p<0.01, ** p<0.05, * p<0.1.

We start by studying whether ethnic groups that have a high power mismatch against

[32] The list of countries in the sample and the number of ethnic groups for each country is reported in Table B.1 in the Appendix.

their government are more likely to participate in a conflict against the government. For this purpose, we use an extended sample of 44 countries in Africa and the Middle East. Our main independent variable is the mismatch dummy, as defined in (1).

We investigate the relationship between mismatch and conflict estimating the following model:

$$Conflict\,incidence_{ect} = \alpha_{ct} + \beta Mismatch\,dummy_{ect} + X_{ect}\gamma + \delta_e + \varepsilon_{ect}, \tag{5}$$

where $Conflict\,incidence$ is an indicator that takes value 1 if ethnic group e in country c in year t is taking part in a conflict; $Mismatch\,dummy$ is our variable of interest and takes value 1 if group e in country c in year t has imbalance between military and political power; X_{ect} is a matrix of ethnic-group-level controls; $\alpha_c t$ is a full set of country $\times$ year fixed effects and δ_e are ethnic group fixed effects. Results are collected in Table 6. Moving across columns of Table 6, we add different sets of group-level controls to the specifications. In column (1), we report the result of a simple OLS regression of Conflict incidence on the mismatch dummy only controlling for country$\times$year fixed effects. This set of fixed effects absorbs all time-varying country-level variables and country-level shocks that may simultaneously affect the probability of being in conflict and the military/political power of a group. Crucially, it also takes into account the fact that mismatch is always defined relative to the dominant group of a specific country.[33] Hence, the mismatch-dummy coefficient is identified only by variation across ethnic groups in the same country and year. In column (2) we add a variable that measures the number of years since the last conflict. While, in column 3, we include controls for the group's cross-border relationship. Specifically, using the Ethnic Power Relations Transborder Ethnic Kin Dataset (Vogt et al., 2015), we build a measure of the population of co-ethnic groups that are in power in other countries. The idea is that the size of groups of the same ethnicity in power in different countries could impact the political and military power as they may put pressure on the government or provide logistic/material support during a conflict. In the same spirit, we add dummy variables that signal whether groups in the family had an upgrade or a downgrade in their power rank over the previous ten years or in the previous 2 years.[34]

Natural resources are arguably an important determinant of civil conflicts and could also affect the degree of political/economic power of the group sitting on the resources. To take into account these factors, we augment the specification by adding controls for the availability of natural resources at the group level (column 4). In particular, we add dummies for the presence of productive gold veins, diamond mines and other precious gems, and petroleum. Other factors that may impact both the decision of participating in a conflict and the allocation of power might be related to geographic characteristics (take as an example secessionist conflicts, or rebellions for regional autonomy as in Esteban et al., 2022). Hence in column (5) we repeat the exercise by adding geographic controls (log distance from the capital, log distance to the closest border, log area of the group homeland,

[33]Indeed, both the military power and the political power measures are relative with respect to the government power.

[34]We control for changes in the political power of kin groups in different temporal span as it may take time for a group that seized power settle and provide help and resources to a kin group in a different country. Changing the temporal span does not have any impact on our results.

and the mean and standard deviation of the elevation of the group homeland). When building the mismatch variable, we primarily focus on the military and political dimensions of power. However, economic power could well be an important determinant of a group's opportunity-cost of conflict. Accordingly, in column (6), we control for economic conditions in the pre-sample period. Specifically, we add log population and log nightlight luminosity as proxies for a group GDP, and various controls for land use[35]. Moreover, we build a time-variying variable that captures the group economic inequality vis à vis the dominant group as in Cederman et al. (2010).[36] Even if economic inequality is a "bad control" (Angrist and Pischke, 2009, 2014), adding it to the specification is important as it shows that power mismatch as an effect on conflict that is independent of that of economic inequality.

In all specifications, the mismatch dummy is positively related to the probability of being in conflict, and the relationship is always statistically significant. This means that, within the same country-year, high-mismatched groups have a higher probability of partaking in a conflict against the government than groups characterized by low mismatch. A comparison of the mismatch-dummy coefficients across column (1)-(6) suggests that group-level observable characteristics explain about one-third of the correlation between mismatch and conflict. The magnitude of the effect is sizable: the probability of being in conflict is approximately 6 percentage points higher for high-mismatch groups compared to low-mismatch ones, indicating that a mismatched group is 50% more likely to partake in a conflict against the government than a similar group that is not mismatched.[37]

In columns (7) of Table 6, we leverage the panel dimension of the dataset and substitute country$\times$year fixed effects with ethnic group and year fixed effects, thereby using only within-group variation for identification of the mismatch coefficient. While ethnic group fixed effects absorb all time-invariant characteristics of the groups, including them reduces effective number of observations used to identify the mismatch-dummy coefficient as a group needs to have variation both in the mismatch dummy and in the incidence variable to contribute to identification.[38] [39] Finally, column (8) of Table 6 reports the results of an extremely demanding specification in which we include both country$\times$year fixed effects and ethnic group fixed effects. This specification is akin to difference in difference model where we use for identification variation of the mismatch variable from the individual average in deviation from the country-year average. The mismatch-dummy coefficient remains positive and significantly different from zero. The size of the coefficient indicates that a groups that become mismatched has a probability of entering a conflict that is 4,6 percentage point higher (approximately 36%) compared to a group in is country and year whose mismatch does not change.

[35]Share of the group land devoted to agriculture, pasture, the share of urban land, all computed in 1990.

[36]The group inequality vis-à-vis the government group is defined as $[\ln(nightlight_{e,t}/nightlight_{g,t})]^2$.

[37]The average value of conflict incidence for low-mismatched groups is 0.12.

[38]This means that groups that are always/never in conflict and groups that always have high/low mismatch are not used in the estimation of the coefficient of interest.

[39]In our sample, the definition of politically relevant ethnic groups is time-varying. Indeed one-third of the ethnic group codes in the sample are observed for less than the whole sample period. Including ethnic group fixed effect helps taking into account the fact that some groups might become irrelevant or might merge with other groups during the sample period.

Table 7: Robustness checks on the geomatching procedure

Dependent variable: conflict incidence

	Mismatch Dummy (se)	R^2	Obs.	Mismatch Dummy (se)	R^2	Obs.
	(1)	(2)	(3)	(4)	(5)	(6)
Less than 50% of matches identifyed by 3 (or less) events	0.0727*** (0.0154)	0.601	3,740	0.0425* (0.0232)	0.734	3,740
Less than 50% of matches identifyed by 5 (or less) events	0.0835*** (0.0169)	0.612	3,544	0.0429* (0.0253)	0.744	3,544
No match identifyed by 5 (or less) events	0.0846*** (0.0179)	0.544	3,327	0.0541** (0.0273)	0.683	3,327
No geomatching (ACD2EPR only)	0.0392*** (0.0116)	0.530	3,612	0.0343* (0.0201)	0.693	3,612
No geomatching - EPR incidence	0.0410*** (0.0101)	0.500	3,453	0.0384** (0.0175)	0.726	3,453
Model fixed effects	Country × Year			Ethnic group & Country ×Year		

Note: The dependent variable is conflict incidence. Column (1)-(3) report specifications with country×year fixed effects, column (4)-(6) those with country×year and ethnic group fixed effects. All specifications include *peace years*, *family*, *natural resources* controls (see notes to Table 6). Standard errors are clustered at the country-year level in column. *** p<0.01, ** p<0.05, * p<0.1.

While the correlation between mismatch dummy and conflict incidence is robust to the inclusion of different sets of controls and fixed effects, the credibility of the estimates depends on the robustness of the matching procedure between ethnic and rebel groups. Table 7 collects the results of the estimation of model 5 using samples where we eliminate potential ”bad matches”. Columns (1)-(3) reports the coefficients, standard errors, R-squared, and number of observation for the model estimated in column (4) of Table 6 which includes country × year fixed effects, peace years as well as family and natural resources controls. In columns (4)-(6) are collected the results of the model in column (8) of Table 6, which adds to the specification ethnic-group fixed effects. The condition that define a ”bad match” are reported in the row headings. In general, we think of a match as potentially problematic if it is determined by few events. Specifically, in row (a) we exclude all the ethnic groups that have more than half of the matches identified by 3 or fewer events. In row (b) and (c) we define a match bad if it is based on 5 or fewer events. So we respectively exclude the ethnic groups that have more than half of the matches identified by 5 or fewer events, and all those group where which contain at least one bad match. In row (d) we keep only ethnic groups that are matched to rebel groups through the ACD2EPR conversion table, thus doing away with the geomatching procedure altogether. Finally, in row (e), we directly use as dependent variable the measure of conflict incidence provided in the Ethnic Power relations Core Dataset. This measure not only does not use the geomatching procedure, but also includes only conflicts with more than 25 fatalities. The coefficient of the mismatch dummy is always positive and statistically significant. The size of the coefficient seems to decrease in samples built without the geomatching procedure (row d-e), but the quantification of the effect is remarkably similar to that of the baseline sample. In fact, when we do not employ geomatching, the number of groups in conflict is lower and the average value of conflict incidence is 0.042 for the sample in row (d) and 0.035 for the sample in row (e); this implies that, within a country-year, groups that are mismatched are about 50% more likely to participate in conflict than their non mismatch counterparts. An effect that is extremely close in magnitude to that found in Table 6.

5.2 Results with a continuous measure of mismatch

Thus far we used a dummy proxy of power mismatch. We now exploit the ethnicity of cabinet members for a selected sample of 14 African countries to build a continuous measure of relative political power and construct an empirical counterpart of the theoretical definition of mismatch, $M = |m - p|$. This allows us to revisit the results presented in the previous section by analyzing the impact of marginal changes in the imbalance between the two dimensions of power. Moreover, it allows us to test whether the relationship between power mismatch and conflict is linear.

In conducting this analysis we are making two changes with respect to the previous Section: (i) the main explanatory variable is different—now continuous—, (ii) the sample we are using for the estimation is different. To understand whether sample selection is driving the results, for each model we report both a specification where *Mismatch dummy* is our main explanatory variable, and a specification where we replace the indicator with the continuous variable *Mismatch cont.*

We start by repeating the specifications of Table 6 Table 8 reports the results. In columns (1)-(4) we use *Mismatch dummy* as our main explanatory variable, while in columns (5)-(8) we repeat the analysis using the continuous measure of mismatch. Moving across columns we increase the number of control variables included in the model. Specifically, columns (1) and (5) only control for country$\times$year fixed effects; in columns (2) and (6) we introduce controls for peace years, ethnic relations and natural resources; in column (3)-(7) we add controls for geographic and socio-economic characteristics. The coefficients and the magnitude of the effect are similar to those found using the extended sample. This suggests that countries that are in the restricted sample behave similarly to those in the extended sample and indicates that sample selection is not affecting the results. Findings in columns (5)-(8), which exploit the continuous measure of mismatch, confirm the positive relationship between mismatch and conflict participation. Using the most demanding specification, column (8), results indicate that a one-standard-deviation increase in conflict mismatch (0.17) is associated with an increase of the probability of participating in a conflict by 2.7 percentage points (an increase that approximately corresponds to 33% of the sample mean).

Our simple theoretical framework implies that, for any realization of costs around the mean, only sufficiently high mismatches determine a rational incentive to attack. Then, a second question we can ask, taking advantage of the continuous measure of mismatch, is whether the effect of mismatch on conflict is linear. To answer this question, we start by splitting the sample using the median of the mismatch distribution as cutoff value, and estimate the empirical model separately for groups whose mismatch is below the median value and for those whose mismatch exceeds it. Results are collected in Table 9. Columns (1) and (2) report results for the sample below the median while columns (3) and (4) those for the sample above the median. Odd numbered columns contains specifications that include the baseline set of controls (peace years, family and natural resources

Table 8: Power Mismatch - Continuous Measure

Dependent variable: conflict incidence.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mismatch dummy	0.0794***	0.0553**	0.0534**	0.0705**				
	(0.0255)	(0.0223)	(0.0254)	(0.0313)				
Mismatch cont.					0.253***	0.158**	0.161**	0.166**
					(0.0687)	(0.0661)	(0.0771)	(0.079)
Observations	1,247	1,247	995	1,247	1,247	1,247	995	1,247
R-squared	0.232	0.481	0.571	0.615	0.234	0.481	0.571	0.615
Controls								
Peace years		✓	✓	✓		✓	✓	✓
Family		✓	✓	✓		✓	✓	✓
Natural resources		✓	✓	✓		✓	✓	✓
Geographic			✓				✓	
Socio-ecomic			✓				✓	
Fixed effects								
Country × year	✓	✓	✓	✓	✓	✓	✓	✓
Ethnic group				✓				✓

Note: The dependent variable is conflict incidence. See notes to Table 6 for the definitions of the groups of controls. Standard errors are clustered at the country-year level. *** p<0.01, ** p<0.05, * p<0.1.

controls) and country×year fixed effects. Specifications in even numbered columns are augmented with ethnic-group fixed effects. The results show that the positive correlation between power mismatch and conflict is detectable only in the sample above the median. This suggests that the relationship between mismatch and conflict is non-linear and marginally increasing the mismatch increases conflict probability only for those groups characterized by a relatively high asymmetry between the two dimension of power. In particular, a one-standard deviation increase in mismatch (0.16) is associated with a 6.8-percentage-point increase in the likelihood of conflict participation, corresponding to a 60% raise of the probability compared to the sample average, 0.108. We corroborate this finding by adding a quadratic term in the baseline specification, *Mismatch squared*. Column (7) of Table 9 reports the result of the estimation of the model with group-level controls and country×year fixed effects, while column (8) those obtained by adding ethnic-group fixed effects. The coefficients on the two variables indicate that the relationship between mismatch and conflict is convex: the coefficient on the linear term is negative and statistically significant, while the coefficient on the square of the mismatch is positive and statistically significant. The point estimates suggest that the relationship between mismatch and conflict is positive for values of the mismatch higher than 0.2 (average of the mismatch in this sample is 0.199).

We further probe into the non-linearity of the effect using non-parametric regressions. Figure7 visualizes the results of non-parametric regressions that uses third-order B-spline as the basis. Panel (a) reports the result of a simple non-parametric regression of conflict incidence on the mismatch variable. In panel (b) we condition the regression on

Table 9: Non-linearity of the effect

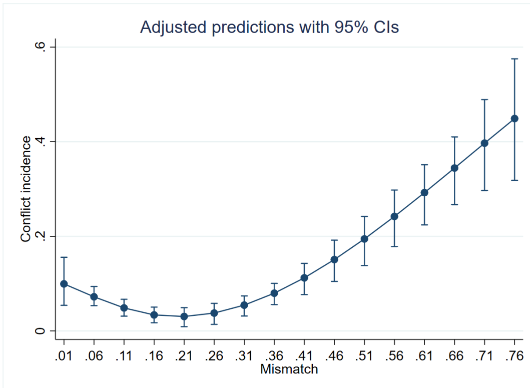
Dependent variable: conflict incidence

	Below median		Above median		Whole sample	
	(1)	(2)	(3)	(4)	(5)	(6)
Mismatch cont.	-0.487*	-0.460	0.434***	0.406**	-0.356**	-0.487**
	(0.262)	(0.281)	(0.128)	(0.180)	(0.147)	(0.191)
Mismatch squared					0.846***	1.137***
					(0.272)	(0.331)
Observations	541	530	564	550	1,226	1,225
R-squared	0.558	0.785	0.539	0.636	0.490	0.623
Fixed effects						
Country × year	✓	✓	✓	✓	✓	✓
Ethnic group		✓		✓		✓

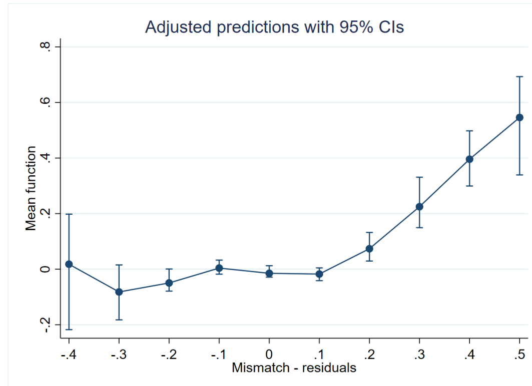
Note: The dependent variable is conflict incidence. Odd-numbered columns reports specification with country×year fixed effects, and even-numbered columns the specification with country×year and ethnic group fixed effects. All specifications include peace years, family controls, and natural resources controls (see notes to Table 6). Standard errors are clustered at the country-year level. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

country×year fixed effect, in panel (c), instead, we add ethnic group fixed effects. The plots clearly confirm that the relationship between power mismatch and conflict is non-linear and that an increase in mismatch raises the likelihood of conflict only for values of mismatch above a certain threshold. This result is immediate if one returns to the intuition of the simple theoretical framework: for any realization of costs around the mean, only sufficiently high mismatches determine a rational incentive to attack, i.e., when the mismatch is small, the costs can outweigh, in expectation, the benefits of conflict.

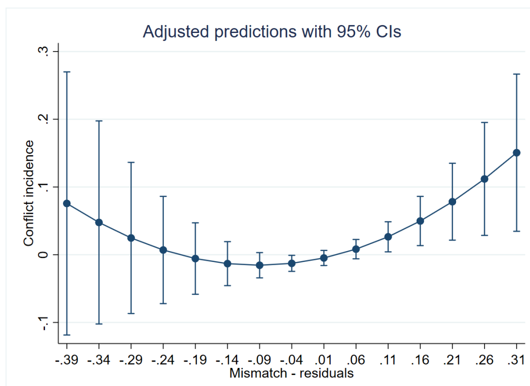
Figure 7: Non-Parametric regressions



(a) No controls



(b) Residuals after controlling for country×year fixed effects



(c) Residuals after controlling for ethnic-group and country×year fixed effects

5.3 Further results

In the previous sections, we tested the main aspect of the theory of power wars, showing a positive (and convex) relationship between power mismatch and the probability of participating in a conflict. In this section, we first examine whether groups with high military power and low political power are affected differently by mismatch than groups with high political power and low military power. We then investigate whether power mismatch is also associated with different conflict characteristics, particularly conflict size and grievance type.

To investigate whether the effect of power mismatch is symmetric, regardless of whether military power is larger or smaller than political power, we split the mismatch measures into two separate variables. Specifically, for the indicator of mismatch we create two dummy variables: *Mismatched dummy Military* equals one if military power is above the median of the distribution and political power is in the first tercile; *Mismatched dummy Political* equals one if political power is in the 3^rd tercile of the distribution and military power is below the median. In a similar spirit, *Mismatch cont. m¿p* contains the value of the continuous measure of mismatch if military power is grater than political power, while *Mismatch cont. p¿m* contains the value of the continuous measure of mismatch if military power is smaller than political power. Table 10 collects the results of this analysis. As in previous tables odd columns report the specifications with country×year fixed effects and even columns those with both country×year and ethnic-group fixed effects. All specification include the usual ethnic group level controls. Column (1) and (2) reports the results for the extended sample using the indicator variables of power mismatch; in Columns (3) and (4) we restrict the sample to the 14 Sub-Saharan countries and still use the mismatch indicators as our main explanatory variable; in Columns (5) and (6) we use the continuous measure of mismatch for the restricted sample. Results are consistent across samples, measures of power mismatch and sources of variation used for identification, and they indicate that the effect of power mismatch is asymmetric. Indeed, coefficients of the variables indicating a mismatch driven by a political power higher than military power are small and generally not significant. On the contrary, coefficients on the variables indicating a mismatch driven by high military power are always positive and statistically significant. This suggests that the likelihood that a group partakes in a conflict against the government rises when its military power increases without being balanced by an increase in political power.[40]

Table 10: Signed Mismatch

Dependent variable: conflict incidence

	(1)	(2)	(3)	(4)	(5)	(6)
Mismatch dummy military	0.0733***	0.0787**	0.0787**	0.0952*		
	(0.0239)	(0.0337)	(0.0329)	(0.0511)		
Mismatch dummy political	0.0480***	0.00311	0.0334	0.0466		
	(0.0140)	(0.0260)	(0.0203)	(0.0355)		
Mismatch cont., m>p					0.248***	0.220**
					(0.0786)	(0.0872)
Mismatch cont., p>m					-0.0222	0.0663
					(0.0885)	(0.128)
Observations	4,260	4,260	1,247	1,247	1,247	1,247
R-squared	0.574	0.708	0.482	0.615	0.489	0.616
Fixed effects						
Country × year	✓	✓	✓	✓	✓	✓
Ethnic group		✓		✓		✓
Sample	Full	Full	Restricted	Restricted	Restricted	Restricted

Note: The dependent variable is conflict incidence. All specifications include *peace years*, *family*, *natural resources* controls (see notes to Table 6). The sample used in the specification is reported at the bottom of the table. Standard errors are clustered at the country-year level. *** p<0.01, ** p<0.05, * p<0.1.

These results on the particular relevance of a mismatch when $m > p$ is consistent with

[40]The effect is not driven by the fact that there is little within-group variation in the *Mismatch Dummy Political* variable. In fact, for both dummy variables, approximately 10% of the sample switches from low to high mismatch.

our general intuition on the conflict initiation incentives: even though conflict initiation data does not exist—and would be anyway unreliable—, $m > p$ cases of mismatch are those where it is clearly the militarily strong but politically weaker group that has the maximum incentive to challenge the status quo. The fewer cases of the other type of mismatch where $p > m$, may end up in conflict because of the power shift logic (Powell, 2012,2013), but it is intuitively a more indirect type of incentive, and hence, intuitively, less likely to show up in the data.

We now ask whether we can associate power mismatch to some conflict characteristics. First, we divide conflicts based on the type of incompatibility underlying them. The GED dataset contains a variable that categorizes civil conflicts in two types: *territorial* (i.e., the incompatibility concerns the status of a territory, secession, or autonomy) and *centrist* (i.e., the incompatibility concerns the type of political system, the replacement of the central government, or the change of its composition). We use this information to build two new measures of conflict incidence that we use as the dependent variable in the analysis. Second, we leverage the completeness of the GED dataset—not limited to civil wars, but also containing smaller conflicts— and ask whether power mismatch has different implications for major and minor conflicts. We define a conflict as *big* if the yearly average number of casualties is above 25.[41] Again, we use as dependent variable conflict incidence for big and small conflicts separately.[42]

Table 11: Centrist vs territorial conflicts

Conflict incidence:	Centrist (1)	Terrirorial (2)	Centrist (3)	Terrirorial (4)	Centrist (5)	Territorial (6)	Centrist (7)	Territorial (8)
Mismatch dummy	0.0583***	0.0151	0.0535**	-0.0136				
	(0.0139)	(0.00965)	(0.0209)	(0.0148)				
Mismatch cont.					0.151**	0.0210	0.147*	0.00147
					(0.0672)	(0.0222)	(0.0771)	(0.0247)
Observations	4,110	3,914	4,110	3,914	1,235	1,157	1,235	1,157
R-squared	0.546	0.449	0.700	0.752	0.451	0.351	0.614	0.433
Fixed effects								
Country × year	✓	✓	✓	✓	✓	✓	✓	✓
Ethnic group			✓	✓			✓	✓

Note: The dependent variable is in the column headings. All specifications include *peace years*, *family*, *natural resources* controls (see notes to Table 6). Standard errors are clustered at the country-year level. *** p<0.01, ** p<0.05, * p<0.1.

Results are reported in Tables 11 and 12. In both tables, columns (1)-(2) and (5)-(6) present the results of the specifications with country×year fixed effects, columns (3)-(4) and (7)-(8) those with both country×year and ethnic-group fixed effects. In columns (1)-(4) we report the results for the extended sample and the dummy variable for mismatch,

[41] For instance, assume that a conflict lasts 3 years. In the first year, there are 10 casualties reported, in the second 20 casualties, and in the third 100 casualties, then the average number of casualties is 43.3 and the conflict is considered “big”. Following the convention of UCDP/PRIO, we chose 25 as a cutoff.

[42] The sample used in the regressions for “big” conflicts is different from that used in the regressions for “small” conflicts. This is because groups that experience a big conflict are not considered in the small conflict analysis and vice versa.

while columns (5)-(8) collect the results for the restricted sample and the continuous measure of mismatch. All the specifications include the usual group-level controls.

The results of Table 11 show that power mismatch is positively correlated to conflict incidence only when considering centrist conflicts, regardless of the sample, the measure of mismatch, and the variation used to identify the coefficients. This finding seems reasonable when considering the dimension of political power captured by the mismatch variable. Both when using the EPR index and the continuous mismatch variable, political power is intended as participation in the decision-making process at the central level. These results square with the theory by Esteban et al. (2022), which highlights how dimensions such as cultural and religious group identity increase the group preferences for autonomy and may lead to territorial conflict even in the absence of a substantial power mismatch. Turning to 12, the results on the relationship between mismatch and the size of conflicts are less sharp. The coefficient of the mismatch variable for small conflict is generally small and imprecisely estimated. On the other hand, power mismatch is more important for big conflicts. This suggests that, if the mismatch is high and cannot be bargained away, resolving the grievance will entail big and probably long conflicts.

Table 12: Big vs small conflicts

Conflict incidence	Big (1)	Small (2)	Big (3)	Small (4)	Big (5)	Small (6)	Big (7)	Small (8)
Mismatch dummy	0.0516***	0.0175*	0.0444**	0.0107				
	(0.0143)	(0.00909)	(0.0222)	(0.0142)				
Mismatch cont.					0.131*	0.0527	0.0622	0.133
					(0.0676)	(0.0419)	(0.0578)	(0.0853)
Observations	4,150	3,883	4,150	3,883	1,220	1,174	1,174	1,220
R-squared	0.534	0.416	0.698	0.550	0.440	0.327	0.444	0.606
Fixed effects								
Country × year	✓	✓	✓	✓	✓	✓	✓	✓
Ethnic group			✓	✓			✓	✓

Note: The dependent variable is in the column headings. All specifications include *peace years*, *family*, *natural resources* controls (see notes to Table 6). Standard errors are clustered at the country-year level. *** p<0.01, ** p<0.05, * p<0.1.

Overall, the findings in this section paint a consistent picture and suggest that power mismatches are associated with higher risk of conflict when the military power of ethnic group excluded from power is high. Moreover, groups with a high power mismatch with respect to the government tend to participate in conflicts that are bigger and whose underlying grievance concerns the division of central political power.

6 Discussion

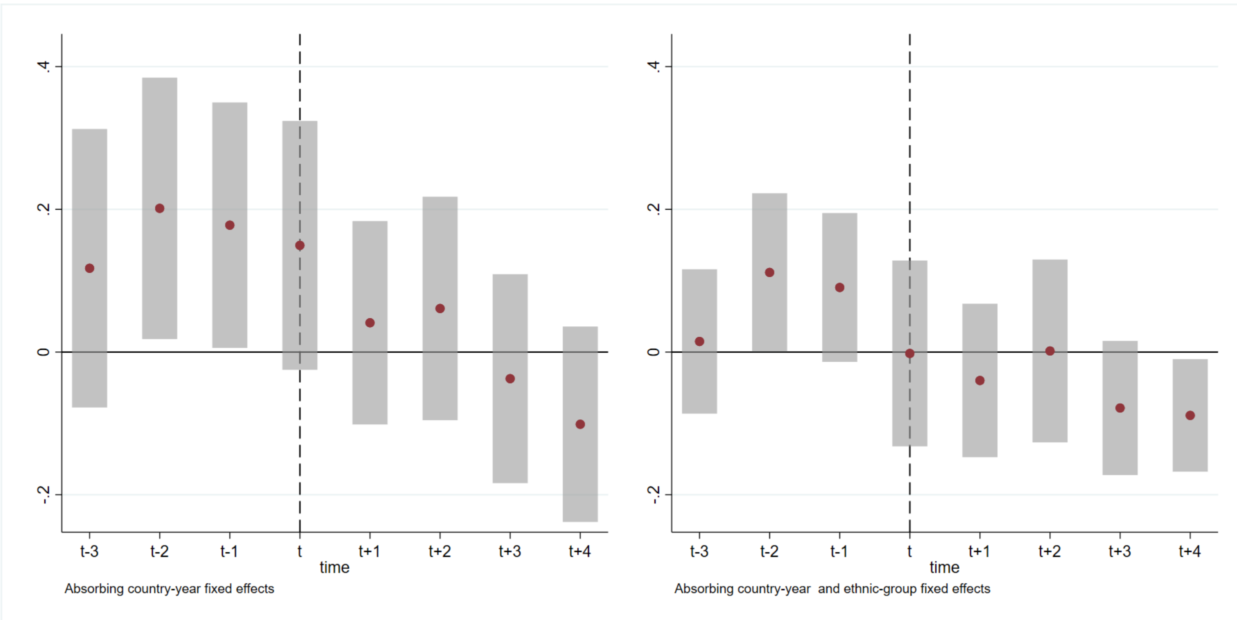
The objective of this paper is mainly descriptive. Nonetheless, it is important to discuss the potential sources of endogeneity. While we control for many of the determinants of conflicts proposed by the literature, and country-level shocks are absorbed by country$\times$year fixed effect, we cannot claim the causality of the relationship between mismatch and conflict. A first concern is reverse causality. Groups that are excluded from power enter into conflict and precisely because they are experiencing conflict have higher military power than groups that are not. While we cannot completely rule out this channel, we can test whether mismatch or conflict comes first. If conflict participation was causing mismatch we would expect the latter to rise during/after the conflict. Figure 8 shows the dynamics of mismatch for groups that experienced conflict in our sample. Panel (a) and (b) report the coefficients and 90% confidence intervals of a regression of mismatch on lags and leads of conflict onset using respectively the discrete and the continuous measure of power mismatch. Looking at the evolution of mismatch around the onset of a conflict we see that, on average, it increases before the conflict—reaching its peak usually in the year before the conflict—and then it decreases. This suggests that our mismatch variables moves before the conflict occurs and not vice versa.

Still, one may be worried that groups with low political power increase their military capabilities precisely because they want to enter into a conflict against their government in order to gain political power. Hence, since some group excluded from power expect conflict in the future, they start accumulating military power in the years before the conflict breaks out. However, it is very unlikely that our mismatch variable is capturing this forward-looking behavior. In fact, the bulk of the variation of the estimated military power occurs across groups: 88% (77%) of the variance of the military power measure is explained by differences across ethnic groups in the extended (restricted) sample. Moreover, estimated military power is extremely persistent over time. Fitting an AR(1) process on the military power variable, we obtain a ρ coefficient that ranges between 0.68 and 0.97. This should not come as a surprise, since the majority of the predictors we use in the machine learning procedure are either time invariant or vary little over time. Hence, our military power measure captures the "structural" or "potential" military power of an ethnic group and, as such, it does not respond to the forward-looking behavior of actors.

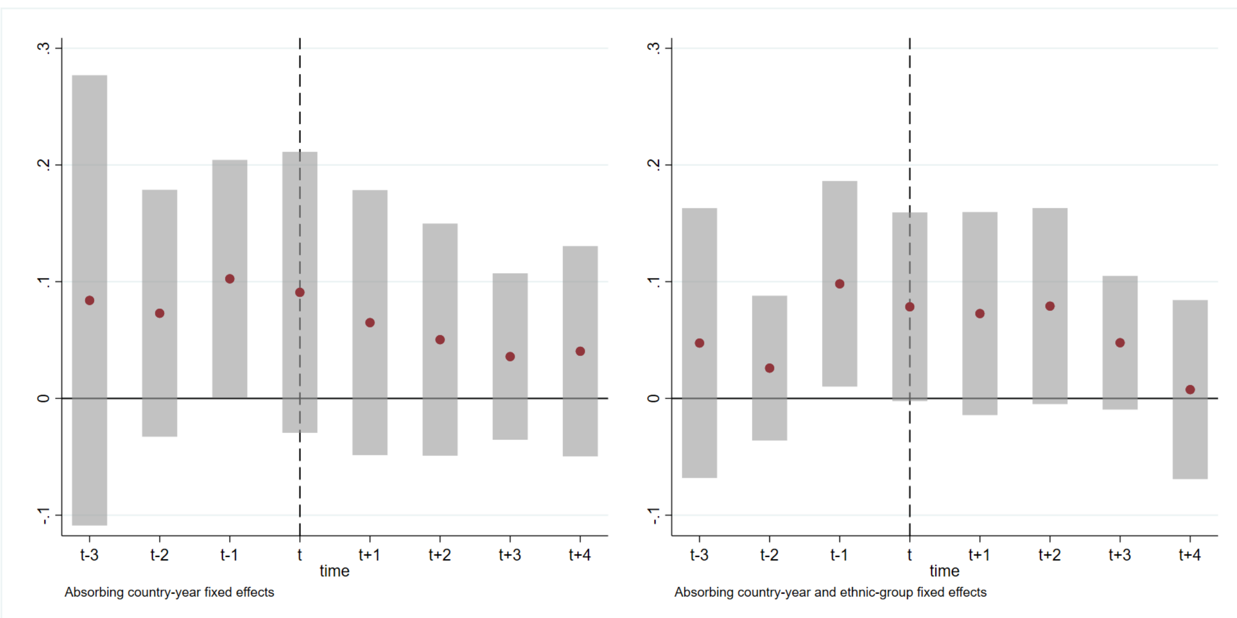
A final worry could arise about political power. A recent body of work (e.g., Deiwiks et al., 2012; Wimmer et al., 2009) emphasizes the role of political exclusion as a trigger for conflict. In our setting, this might be problematic if the mismatch variable is just capturing exclusion from political power. Indeed, we find that conflict participation is more likely when group political power is low and group military power is high. Suppose that a group has average military power and average political power, suppose also that for some reason it loses political power, this would translate in an increase of the mismatch variable. If we observe conflict after the group has been "downgraded", then we would attribute it to the increase of the mismatch but, in reality, the trigger for the conflict was just the exclusion from power. Even though we cannot completely rule out this possibility, we can learn something looking at episodes of political power downgrading in our data. In the extended sample there are 28 groups (30 cases) that have a reduction in the power rank measure.[43] For each of these events, we compute whether the group participated in a conflict against the government in the year of the power downgrading or in the following 5 years. We report the descriptive statistics in Table 13. Only 10 out of the

[43] We drop 2 cases where the downgrade happens at the end of the sample period.

Figure 8: Power Mismatch Evolution



(a) Mismatch dummy



(b) Continous Mismatch

Table 13: Political power downgrading

	Frequency	Rank before Downgrade		Δ Rank		Military Power	
		mean	median	mean	median	mean	median
Conflict	10	3.8	4	-2	-2	0.638	0.742
No Conflict	20	3.95	4	-2	-2	0.393	0.275

30 groups that experienced a reduction in political power have a conflict in the following 5 years. The characteristics in terms of political power enjoyed before the downgrade and the magnitude of the downgrade are extremely similar across groups that enter into a conflict and those who do not. Indeed, on average, these groups switch from a political rank of 4 (i.e., junior partner in the government) to a political rank of 2 (powerless). The only significant difference is their military power: groups that experience conflict after a political downgrade have an estimated military power that is much higher than the one of groups that do not experience conflict. To help gauge the size of the difference, downgraded groups that do not experience conflict have a median military power equal to 0.27, a value that is close to the median of the whole sample (for comparison, the average value of military power in the extended sample is 0.28 and the median value 0.22). the median of military power for groups that are downgraded and do experience conflict is 0.74, higher than the value of the 95th percentile of the military power distribution in the whole sample (0.68). This simple exercise suggests that being excluded from power is not enough to predict conflict, and that our mismatch variable is not just picking up the effect of power downgrading, but an actual asymmetry between political and military power.

7 Conclusions

This paper makes two contributions to the literature and for future research use: it provides new measures for the study of ethnic conflict, filling an important gap, especially on the military power dimension, and it provides the first empirical analysis of the mismatch theory of power wars. About the first contribution, the use of machine learning techniques allowed us to provide a new measure of the relative military power of each ethnic group in the dyadic confrontation with the corresponding government controlling group, varying over time and context. This new measure performs well in predicting victory and allows us to improve on traditional proxies of military power used in the empirical literature on civil conflict.

As for the second contribution, theoretical work already suggested that absolute measures of military power are not enough to explain conflict participation and that researchers should focus on imbalances of different dimensions of power. In this paper, we take up the challenge to empirically show that power mismatch matters for conflict. Armed with the new measure of ethnic-group military power, we show the existence of a relationship between the likelihood of conflicts and the imbalance between relative military strength and relative political power. Exploiting data on civil (ethnic) conflicts, we find evidence that high-mismatch groups are approximately 50% more likely to take part in a conflict against their government. Moreover, these conflicts tend to be more deadly than those where low-mismatch groups are involved.

Our findings, albeit robust to the inclusion of different sets of fixed effects and to the addition of many group-level controls, are mainly descriptive. However, they paint a picture that is extremely consistent with the theory of power wars (Herrera et al., n.d.). Hence, we believe that our evidence on the key role of power mismatch should encourage further research, both in the direction of precise identification and forecasting of future conflicts. In addition, our evidence bears an important policy implication: when trying to understand and prevent conflict outbreaks, one needs to pay attention to the imbalance between different dimensions of (relative) power. Focusing just on military strength or economic or political power may be misleading: militarily strong groups may not be those who start a war if they have enough political power; similarly, groups that are discriminated against may not pose a threat if they are militarily weak.

From a policy perspective, it could also be interesting to dig deeper into the origin of power mismatch and into the causes of its persistence. While a full-fledged analysis is beyond the scope of this paper, if we correlate power mismatch with country characteristics, we find that more democratic (polity 2 index), and ethnically and culturally more homogeneous countries, tend to have a larger share of high-mismatch groups. Contrary to popular belief, it seems that groups at high conflict risk reside in countries that are traditionally believed not to be at risk of conflict. This could be due to the winner-takes-all aspect of some electoral democracies, where a small difference in votes can be enough to completely exclude some (strong) groups from political power. In such cases, elections may not guarantee the credible elimination of a mismatch, and democracy may need to be supplemented by inventive institutional designs such as commitments in terms of public jobs, political roles, and military quotas, which favor power-sharing.

References

Angrist, Joshua D and Jörn-Steffen Pischke, *Mostly harmless econometrics: An empiricist's companion*, Princeton university press, 2009.

_ and _ , *Mastering'metrics: The path from cause to effect*, Princeton university press, 2014.

Breiman, Leo, "Stacked regressions," *Machine learning*, 1996, *24* (1), 49–64.

Bruk, Solomon I and VS Apenchenko, "Atlas narodov mira (Atlas of the World's Peoples)," *Moscow: Glavnoye Upravleniye Geodezii i Kartografii*, 1964.

Carroll, Robert J and Brenton Kenkel, "Prediction, proxies, and power," *American Journal of Political Science*, 2019, *63* (3), 577–593.

Cederman, Lars-Erik and Luc Girardin, "Beyond fractionalization: Mapping ethnicity onto nationalist insurgencies," *American Political science review*, 2007, *101* (1), 173–185.

_ , Andreas Wimmer, and Brian Min, "Why do ethnic groups rebel? New data and analysis," *World politics*, 2010, *62* (1), 87–119.

_ , Kristian Skrede Gleditsch, and Halvard Buhaug, *Inequality, grievances, and civil war*, Cambridge University Press, 2013.

Chandra, Kanchan and Steven Wilkinson, "Measuring the effect of "ethnicity"," *Comparative Political Studies*, 2008, *41* (4-5), 515–563.

Deiwiks, Christa, Lars-Erik Cederman, and Kristian Skrede Gleditsch, "Inequality and conflict in federations," *Journal of Peace Research*, 2012, *49* (2), 289–304.

der Laan, Mark J Van, Eric C Polley, and Alan E Hubbard, "Super learner," *Statistical applications in genetics and molecular biology*, 2007, *6* (1).

Eck, Kristine, "In data we trust? A comparison of UCDP GED and ACLED conflict events datasets," *Cooperation and Conflict*, 2012, *47* (1), 124–141.

Esteban, Joan, Laura Mayoral, and Debraj Ray, "Ethnicity and conflict: An empirical study," *American Economic Review*, 2012, *102* (4), 1310–42.

_ , Sabine Flamand, Massimo Morelli, and Dominic Rohner, "Stay, Split or Strike: Theory and Evidence on Secessionist vs Centrist Conflict," 2022.

Fearon, James D, "Rationalist explanations for war," *International organization*, 1995, *49* (3), 379–414.

Francois, Patrick, Ilia Rainer, and Francesco Trebbi, "How is power shared in Africa?," *Econometrica*, 2015, *83* (2), 465–503.

Girardin, Luc, Philipp Hunziker, Lars-Erik Cederman, Nils-Christian Bormann, and Manuel Vogt, "Growup-geographical research on war, unified platform," *ETH Zurich. http://growup. ethz. ch*, 2015.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand, "Armed conflict 1946-2001: A new dataset," *Journal of peace research*, 2002, *39* (5), 615–637.

Hartzell, Caroline and Matthew Hoddie, "Institutionalizing peace: power sharing and post-civil war conflict management," *American Journal of Political Science*, 2003, *47* (2), 318–332.

Herrera, Helios, Massimo Morelli, and Salvatore Nunnari, "A theory of power wars," *Quarterly Journal of Political Science*, *17* (1), 1–30.

Horowitz, Donald L, "Ethnic conflict management for policymakers," *Conflict and peacemaking in multiethnic societies*, 1990, *115.*

_ , *Ethnic groups in conflict, updated edition with a new preface*, Univ of California Press, 2000.

_ , "Electoral systems and their goals," *Journal of Democracy*, 2003, *14* (4), 115–127.

Horowitz, Donald L., *Power Sharing: New Challenges For Divided Societies*, Pluto Press, 2005.

Michalopoulos, Stelios and Elias Papaioannou, "The long-run effects of the scramble for Africa," *American Economic Review*, 2016, *106* (7), 1802–48.

Moscona, Jacob, Nathan Nunn, and James A Robinson, "Segmentary Lineage Organization and Conflict in Sub-Saharan Africa," *Econometrica*, 2020, *88* (5), 1999–2036.

O'Leary, Brendan, "The character of the 1998 agreement: results and prospects," *Aspects of the Belfast Agreement*, 2001, pp. 49–83.

Posner, Daniel N, "The political salience of cultural difference: Why Chewas and Tumbukas are allies in Zambia and adversaries in Malawi," *American Political Science Review*, 2004, *98* (4), 529–545.

Rainer, Ilia and Francesco Trebbi, *5. New Tools for the Analysis of Political Power in Africa*, Chicago: University of Chicago Press,

Reilly, Benjamin, "Institutional designs for diverse democracies: Consociationalism, centripetalism and communalism compared," *European Political Science*, 2012, *11*, 259–270.

Sarkees, Meredith Reid and Frank Wayman, *Resort to war: 1816-2007*, Cq Press, 2010.

Spears, Ian S, "Understanding inclusive peace agreements in Africa: the problems of sharing power," *Third World Quarterly*, 2000, *21* (1), 105–118.

Sundberg, Ralph and Erik Melander, "Introducing the UCDP georeferenced event dataset," *Journal of Peace Research*, 2013, *50* (4), 523–532.

Tibshirani, Robert and Jerome H Friedman, *The elements of statistical learning [electronic resource]: data mining, inference, and prediction: with 200 full-color illustrations*, Springer, 2001.

Tollefsen, Andreas Forø, Håvard Strand, and Halvard Buhaug, "PRIO-GRID: A unified spatial data structure," *Journal of Peace Research*, 2012, *49* (2), 363–374.

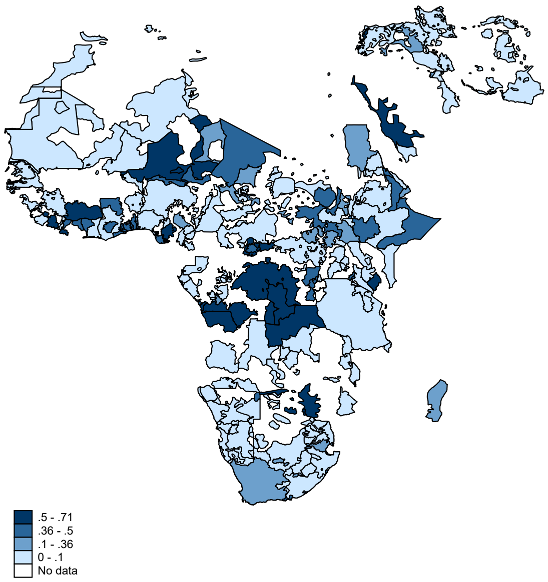
Vogt, Manuel, Nils-Christian Bormann, Seraina Rüegger, Lars-Erik Cederman, Philipp Hunziker, and Luc Girardin, "Integrating data on ethnicity, geography, and conflict: The ethnic power relations data set family," *Journal of Conflict Resolution*, 2015, *59* (7), 1327–1342.

Wimmer, Andreas, Lars-Erik Cederman, and Brian Min, "Ethnic politics and armed conflict: A configurational analysis of a new global data set," *American sociological review*, 2009, *74* (2), 316–337.

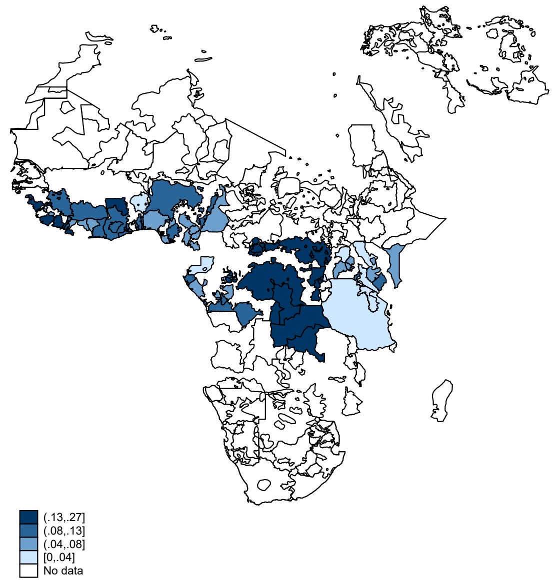
Wucherpfennig, Julian, Nils B Weidmann, Luc Girardin, Lars-Erik Cederman, and Andreas Wimmer, "Politically relevant ethnic groups across space and time: Introducing the GeoEPR dataset," *Conflict Management and Peace Science*, 2011, *28* (5), 423–437.

A Additional Figures

Figure A.1: Mismatch Measures – Within-Group Standard Deviation

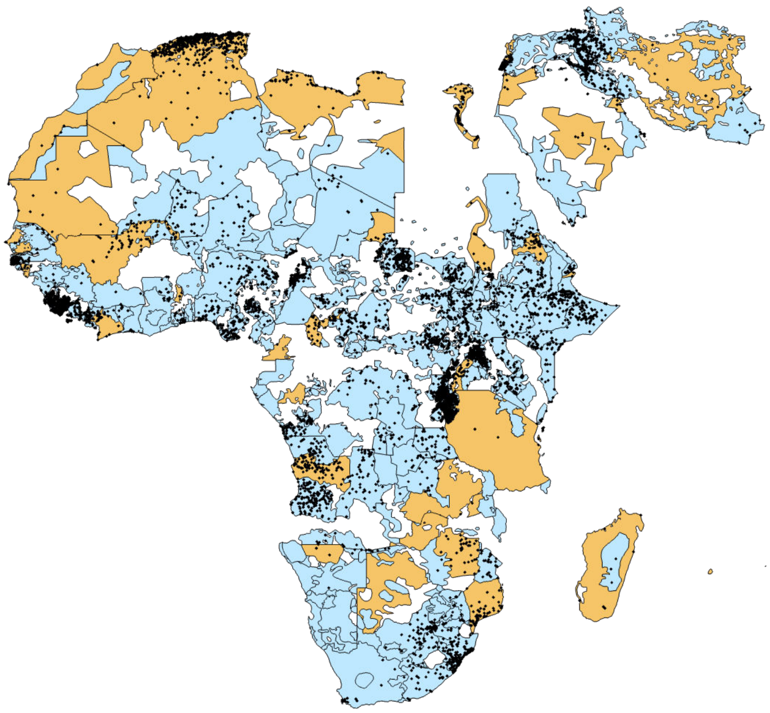


(a) Mismatch Dummy



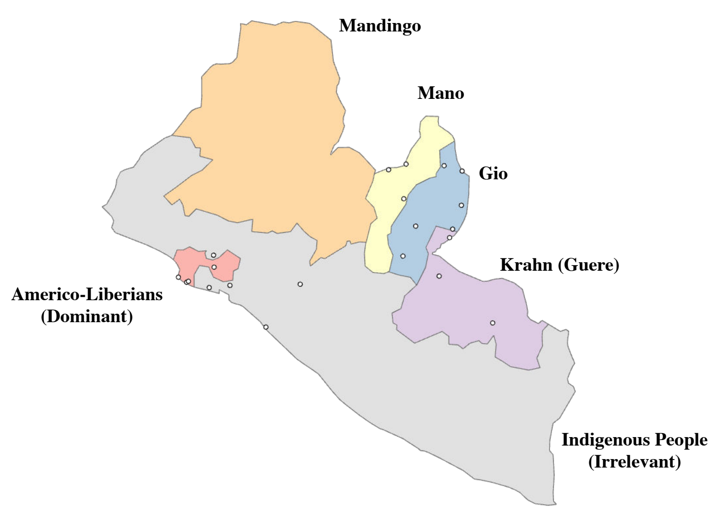
(b) Continous measure of Mismatch

Figure A.2: Geo-matching between GED and GeoEPR in Africa and Middle East



Notes. The polygons are the homeland of each politically relevant ethnic group in Africa and the Middle East as coded in GeoEPR. Specifically, the golden-colored polygons are the homeland of the government ethnic groups, and the blue-colored ones are the ethnic homeland of the rebel groups. The irrelevant ethnic groups are not geo-coded in GeoEPR, so they are left blank on the map. The black dots are the conflict events recorded in GED.

Figure A.3: Rebel group coding: NPFL in Liberia as an example



Notes: This is a map of Liberia, where each colored polygon represents an ethnic group listed in EPR. The dots represent all conflict events in UCDP-GED associated with the rebel group NPFL.

B Sample

This section provides the list of countries in our extended and restricted sample, the distribution of the observations in the different EPR power rank categories and a description of the detailed manual checks on multi-ethnic governments.

Country list For each country Table B.1 reports the average number of ethnic groups in EPR over the sample period and whether the country also belongs to the restricted sample.

Table B.1: Country list

Country	Avg # of groups	Restricted sample	Country	Avg # of groups	Restricted sample
Algeria	1		Lebanon	8	
Angola	4		Liberia	4	✓
Bahrain	1		Libya	3	
Benin	3	✓	Madagascar	1	
Botswana	9		Malawi	2	
Cameroon	5	✓	Mali	2	
Central African Repub.	3.8		Mauritania	2	
Chad	5		Mauritius	5.2	
Comoros	2		Morocco	2	
Congo	5	✓	Mozambique	2	
Congo, DRC	11.6	✓	Namibia	11	
Cote d'Ivoire	4	✓	Niger	3.9	
Djibouti	1		Nigeria	5	✓
Egypt	1.6		Saudi Arabia	3	
Equatorial Guinea	4		Senegal	4	
Eritrea	3		Sierra Leone	3.7	✓
Ethiopia	8.2		South Africa	12.5	
Gabon	3.2	✓	Sudan	14.8	
Ghana	4	✓	Syria	4	
Guinea	2		Tanzania	3.18	✓
Guinea-Bissau	3.9		The Gambia	4	
Iran	10		Togo	1	
Iraq	4		Uganda	8	✓
Jordan	2		Zambia	4	
Kenya	6.8	✓	Zimbabwe	2	
Kuwait	2				

Power rank in EPR The definition of the Power rank index in the EPR core dataset are The Power rank database is provided by in the EPR Core dataset and is defined as:

- The group rules alone:
  - Monopoly: Elite members hold monopoly power in the executive to the exclusion of members of all other ethnic groups
  - Dominance: Elite members of the group hold dominant power in the executive but there is some limited inclusion of ”token” members of other groups who however do not have real influence on decision making.
- The group shares power:
  - Senior Partner: Representatives of the group participate as senior partners in a formal or informal power-sharing arrangement. By power sharing, we mean any arrangement that divides executive power among leaders who claim to represent particular ethnic groups and who have real inuence on political decision making.
  - Junior Partner: Representatives participate as junior partners in government
- The group is excluded from power:
  - Self-exclusion: The special category of self-exclusion applies to groups that have excluded themselves from central state power, in the sense that they control a particular territory of the state which they have declared independent from the central government
  - Powerless: Elite representatives hold no political power (or do not have influence on decision making) at the national level of executive power, without being explicitly discriminated against.
  - Discrimination: Group members are subjected to active, intentional, and targeted discrimination by the state, with the intent of excluding them from political power. Such active discrimination can be either formal or informal, but always refers to the domain of public politics (excluding discrimination in the socio-economic sphere).

Table B.2 shows the political power distribution of in the extended and restricted samples.

Table B.2: Power rank in EPR

	Government groups			Other groups			
Status	Monopoly	Dominant	Senior Partner	Junior Partner	Self-Excluded	Powerless	Discriminated
Rank	7	6	5	4	3	2	1
Share of obs. extended	4.79%	15.73%	79.48%	47.04%	0.55%	34.07%	18.31%
Share of obs. restricted	0	8.58%	91.52%	56.4%	0	30.69%	12.84%

Detailed manual checks on government's ethnicity

Guinea 2009 The military coup by Capt. Camara, a Kpelle, seized control of the government after the death of the previous president. Kpelle is not a politically relevant ethnic group. EPR documentation states that ”Generally, ethnicity still matters for national politics and the major political parties have easily identifiable ethnic bases”. In this specific period, however, ”all major ethnic groups are included in the cabinet leadership” (Malinke, Susu and Peul). It is, however, possible that the government was heavily influenced by the Malinke ethnic group since the two top positions, the military junta's second man and the prime minister, were Malinke. EPR classifies the three main ethnic groups as Senior Partners. However, further references [1] report attacks of the military against the Peul. Malinke and Peul have been historically clashing. It appears right to depart from EPR classification and assume that the Malinke were dominating the government in 2009. Further support for this is the Malinke dominating after 2009.

Liberia 1990-6: the First Liberian Civil War EPR classifies the country is in state collapse. The EPR documentation states that ”there is no central authority anymore and the state is unable to perform any of its *empirical functions* outside Monrovia”. They further state that ”there is no functioning central political power during this time and the country is ruled by different rebel groups, warlords, criminal gangs etc., the term *access to state power* becomes completely meaningless”. The EPR documentation does not describe the events. A brief account of the events can be found on Wikipedia (citing academic books and BBC articles) and is summarised below.

1. 1990: Invasion by Taylor (NPFL) overthrew president Doe. Taylor controlled most of the country (80%). NPFL fighters mainly from Gio and Mano ethnic groups of northern Liberia who were persecuted under Doe's regime.

2. NPFL efforts to capture the capital city of Monrovia were thwarted by the arrival of the Economic Community of West African States (ECOWAS) cease-fire monitoring group the Economic Community of West African States Monitoring Group (ECOMOG).

3. The NPFL instead set up in 1991 an alternative national administration away from the capital (the National Patriotic Reconstruction Assembly Government - NPRAG).

4. Johnson broke away from the NPFL and founded his own party INPFL. Johnson captured quickly the capital Monrovia. NPFL's power declined however after 1992.

5. in 1992 ECOMOG declared an Interim Government of National Unity (IGNU) with Amos Sawyer as their president, with the broad support of Johnson.

6. October 1992: Taylor launches an assault on Monrovia but was pushed back.

7. The interim government lasted until the next 1997 elections won by Taylor.

[1] https://www.refworld.org/docid/537db96b4.html

Liberia 2004-5 EPR states that all countries relevant ethnic groups are Senior Partners. In the aftermath of the Second Liberian Civil War 1999-2003, the country started its democratic transition under the National Transitional Government of Liberia until the 2005 democratic Election won by Ellen Johnson Sirleaf. EPR documentation states ″the cabinet posts and National Transitional Legislative Assembly seats were equally divided between the civil society (and other neutral political forces) and the warring factions (i.e. the corresponding ethnic groups)″ and that ″Chairman of this power-sharing government was Gyude Bryant, a neutral politician″.

Sierra-Leone 1993-6 EPR states that the country is in state collapse. Military coup by Cap. Cap. Valentine Strasser in 1993 who was a Creole, a non-politically relevant ethnic group. According to EPR documentation ″the constitution is suspended, and all political parties and activities are banned″. There is civil war since 1991 with the major opposing rebel group being the Revolutionary United Front (RUF). Sources disagree on whether Strasser favored an ethnic group: there is information that Strasser favoured the Mende. EPR documentation doubts this based on other sources. It is also important to remark that according to many sources, RUF did not fight for a certain group or region (with claims for social justice and pan-Africanism). It seems that it would be hard to attribute a correct ethnic group to both the rebels and the government. For this reason, no ethnicity is attributed to the government following EPR.

Sierra Leone 1998-2002 EPR states that the country is in state collapse. New military coup against Kabbah (from the Mende group) and The Armed Forces Revolutionary Council (AFRC), allied with the RUF, is established as the new government. ″The government even lacks an official army after 1998, being completely dependent on foreign peacekeepers and local militias″. Kabbah (in power in 1997)i s reinstalled by ECOMOG forces in March 1998 but the civil war continues, with rebels reentering the capital Freetown in January 1999. EPR documentation states ″The dramatic increase of warring parties and the shifting alliances completely blur the picture of who holds political power″ and ″the virtual loss of control by the central government and the totally nebulous situation regarding political power during this period of intensified civil war, this period is again coded as *state collapse*″.

Congo DRC 2004-6 Ethnicity constantly played a role in Congo DRC politics. Joseph Kabila, the new president is more moderate than the previous one: had 4 vice-presidents each one representing a different faction. Kabila is supported by two related ethnic groups which are both coded as Senior Partners: Lunda-Yeke and Luba Shaba which have are related and form Kabila main power base. Here we consider them as allies.

Congo DRC 2007-12 Joseph Kabila is re-elected with less inclusive politics than his first term. Moreover, his power-base is unchanged and both Lunda-Yeke and Luba Shaba beeing Senior Partners and considered allies.

Kenya 2008-11 Ethnicity has been dominant in Kenyan politics: political parties are organised along ethnic identities. In December 2007, ethnic divisions turned violent following suspected fraud by the president during elections. An agreement, brokered by former U.N. Secretary Kofi Annan, was reached parties reached an agreement (brokered by former U.N. Secretary Kofi Annan) on a coalition government with equally shared cabinet reflecting Kenya's ethnic diversity. The two Senior partners are therefore considered allies.

Kenya 2012 The government did not change composition. Kenyan politics are however marked by a fight against the Somali ethnic group suspected to support the terrorist organisation Al-Shabaab.

Burundi 2002-5 The two politically relevant ethnic groups, Tutsi and Hutu, share government power "50-50" according to EPR documentation The Hutu hold a larger share in government posts which is compensated by the Tutsi's hold over the army. This equal sharing, after a historical hostility between the two ethnic groups, is the result of the Arusha Agreements in 2001 (not the same as the Arusha Agreements in Rwanda): a transitional peace treaty which brought the Burundian Civil War to an end.

Zambia 1966-2012 Zambian politics didn't involve ethnical conflicts and parties have been historically inclusive. From the EPR documentation: "leaders of Zambian parties have always attempted to appoint to significant position, members of diverse ethnic group, in the hope of increasing their share of national votes".

Madagascar 2002-12 According to the EPR documentation "from the literature surveyed there was no evidence that ethnicity is the basis for political discrimination". There is however "at least one interest group claiming to represent the interest of an ethnic group" until 2001. There is also evidence that ethnic tensions played a major role in "each of the major political transitions" until 2001. However, the political parties cannot be identified as representing a specific group. In 2002, Marc Ravalomanana is elected and his "ethnicity was eclipsed by his sense of nationalism and his call for a united Malagasy people". The two political ethnicities Côtiers and Highlanders are therefore coded as Irrelevant.

Comoros 2002-12 (included since they had conflicts) The inhabitants of each island are understood as different ethnic group. The three ethnicities corresponding to the three islands are Senior Partners in 2002. Each island has a president and they are vice-presidents in the Comorian Union government. The office of the president of the Union rotates among the three islands in 4-year terms.

C Military measure

C.1 Training set

This section describes the variables that we include in the training dataset. We collect all the ethnic level characteristics from the GrowUp Dataset (Girardin et al., 2015). First, we contain land characteristics, which include the ethnic territory's land area, the percentage area of the ethnic territory covered by mountainous, forest, barren lands, shrubland, urban area, water, grassland, and pasture land, as well as the area equipped for irrigation in the ethnic territory. We also contain a set of rich geographic characteristics, which include the spherical distances and travel time from the ethnic territory centroid to the border of the nearest land-contiguous neighboring country, and to the border of the nearest neighboring country (regardless of whether the nearest country is located across international waters), and to the border of the territorial outline of the country it belongs to, as well as to the national capital city in the corresponding country. Moreover, we contain rich socio-demographic characteristics, which include the population size of the ethnic group, population growth in the last 10 years, infant mortality rate, and prevalence of child malnutrition. Climate information is also captured by controlling the yearly total amount of precipitation and mean temperature. Besides, we include the number of affiliated ethnic groups outside the country considered and the number of affiliated ethnic groups outside the country considered in a position of power to capture kinship characteristics. Finally, we include external support information such as whether the group has ever received foreign support, whether the group has received support before the year of observation, and whether the group is currently receiving support. We also include a set of aggregate level variables, which include year, country latitude and longitude, and regional indicators.

C.2 Robustness checks on the parameters of the algorithm

We vary arbitrarily the parameters of the algorithm and show that the prediction is highly robust. We report both the relative difference in PRL and the correlation of the baseline military measure and the military measure using a different set S_i of parameters.

- S_1: the number of cross-validation folds is changed from 6 to 10;
- S_2: #(CV-folds) from 6 to 10; Random Forests' depth and span is increased; Random Forests' column sampling is also increased;
- S_3: #(CV-folds) from 6 to 10, Boosted Decision Tree's learning rate increased;
- S_4: Random Forests' column sampling is decreased;
- S_5: Random Forests' column sampling is increased; Boosted Decision Tree's learning rate increased; Boosted Decision Tree's total number of allowed trees decreased;
- S_6: in both Random Forests and Boosted Decision Tree the row sampling rate is increased.

Table C.3: Robustness check: algorithm's parameters.

Models	S_1	S_2	S_3	S_4	S_5	S_6
$\text{corr}(m_{S_0}, m_{S_i})$	> 0.99	> 0.99	> 0.99	> 0.99	0.98	0.97
$\Delta_{\text{PRL}}^{S_o,S_i}$	0.01	0.02	0.03	0.03	0.04	0.05

C.3 Military measure: machine learning algorithm

In this section, we describe in detail our machine learning algorithm and how to use it to predict the probability of winning for each ethnic group against the government.

We follow CK and train a machine learning algorithm to infer the dyadic probability of winning. We use a stacked ensemble learner which is a method that combines "multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms"[2].

Stacking, or Super Learning, is a procedure that aims to find the optimal combination of prediction algorithms. Concretely, Stacking solves a main issue in the inference problem: when not knowing the exact form of the underlying distribution, stacking allows to combine several possible forms. Stacking uses cross-validation i.e. a random partition of the training set into n subsets of equal size. The procedure consists of generating n models, each one based on a different $\frac{n-1}{n}th$ fraction of the training set and the models are then used for prediction on the $\frac{1}{n}th$ fraction left. We can then obtain the cross-validated prediction for each of the N observations in out training set. Generally, the cross-validated error of the learner is simply the average error made on each of the N predictions. It is typically used for model selection among the considered learners when the available data is scarce enough to not be able to afford a proper testing set. The Stacking algorithm instead, uses the cross-validated prediction for each observation and for each learning algorithm.

Precisely a base learner is defined as a mapping from an N-observations dataset (Y, X) (predictors X and responses Y) to a function of the predictors. Given L base learners, we obtain by cross-validation a matrix $\mathcal{Z} \in \mathbb{M}^{N \times L}$ storing the predictions. The matrix $\mathcal{Z}$ and the original response vector $Y \in \mathbb{R}^N$ is the so-called "level-one data".

The meta-learning algorithm, then, combines the L base learning models to build a single learner. The ensemble model consists of the L base learning models and the meta-learning model. In practice, the true response vector Y is regressed on the meta-predictors $\mathcal{Z}$. The form of the meta-regression function $\Psi(\mathcal{Z}) = \mathbb{E}(Y|Z)$ is left as input by the user and the parameters of Ψ are chosen as to minimise a loss function: typically the cross-validated risk $\Psi = \arg\min_{\psi \in \mathcal{P}} \sum_i \left(Y_i - \psi(Z_i)\right)^2$ where $\mathcal{P}$ is the set of functions considered. If $\mathcal{P}$ allows a great deal of complexity, penalisation or further cross-validation can be used to avoid over-fitting on the cross-validated risk criteria[3]. A commonly used meta-regression function is a linear one with $\beta_l \geq 0$ and $\|\beta\|_1 = 1$, that is a weighted average of the predictions, the weights (to each model) being determined to minimise the cross-validated risk.

[2] h2o documentation accessible at https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/stacked-ensembles.html.

[3] for example, the method of the Lasso can be used if $\mathcal{P}$ is the space of multi-linear functions in X.

To wrap up, the stacked ensemble learner of Y for a value X is obtained by evaluating the meta-regression function Ψ on the L cross-validated predictions at X of the L base learning models. Van der Laan et al. (2007) show that the super learner will perform asymptotically as well as the best learner among all base learners. Therefore, it is clear that with an appropriate meta-regression function (eg. a simple linear function) the stacked ensemble learner outperforms each of the base learners. The functional form of the meta-learner that we use is a generalised linear model, the recommended one by the h2o library. Empirical studies such as Breiman (1996) show that, generally, the more diversified are the base learners forms, the better the super learner performs. For this reason, our base models consist of three different families: *random forests*, *gradient boosting machines* and *generalised linear models* that we proceed to describe. More details on each single method can be found in Tibshirani and Friedman (2001).

C.3.1 Random Forests

A random forest is a tree based learning algorithm. A simple decision tree performs well in terms of training error with a low bias but it typically suffers from high variance i.e., it tends to over-fit the training set. Random forests consist in applying bagging, or bootstrap aggregation, to decision trees. The method relies on a simple insight from statistics: averaging a set of i.i.d. observations reduces the variance. To reduce variance and therefore increase the prediction accuracy, the algorithm generates multiple decision trees and the final prediction is an average of the predictions of all the decision trees. To effectively reduce the variance, the decision trees have to be sufficiently decorrelated. Two features of a random forest algorithm realise this task: (1) each decision tree is based on a different bootstrapped training dataset which is the core of bagging; (2) each time a split is considered in a tree (branching in the decision-tree building algorithm), only a random sample of m predictors is considered among all the predictors. In particular, (2) compromises the training error and increases the bias but greatly reduces the final variance by reducing the correlation between the trees.

It is recommended to use value of m that is close to $\sqrt{p}$ where p is the total number of predictors. A smaller value of m is used when the data set is made of a large number of correlated predictors. In our case, it is not the case and we tend to span m slightly above the recommended value. The number of trees in a Random forest algorithm is not a sensible parameter: if we allow for several trees, over-fitting does not occur. It is therefore recommended to use a sufficient number of trees such that the error rate stabilises. Another important parameter in Random forests, and as a matter of fact, in any tree based algorithm, is the depth of the tree. The depth is determined by the number of branchings allowed. Allowing a high depth naturally increases the complexity of each tree and of the resulting random forest. We do not allow for high depth to not over-fit the data since the number of observations in our training dataset is not very high.

C.3.2 Gradient Boosting Machine

A gradient boosting machine, like random forest, is a tree based learning algorithm. GBMs attempt to solve the same issue as Random Forests but in a radically different way.

Instead of bagging, GBMs are based on boosting: trees are generated sequentially, each tree using the information provided by the previous ones. In that sense, GBM belongs to the forward leaning ensemble methods in that information flows only in the input-to-output direction. The boosting approach is based on the belief that learning slowly is preferable to straightforwardly fitting the data. The algorithm sequentially adds trees to the existing tree to reduce the residuals and better fit the data. Once a new tree is calculated in order to reduce the residuals, it is added in a "shrunken" version that is with an attenuating coefficient called the shrinkage parameter or learning rate. The residuals are then recalculated for the updated boosted model i.e. the shrunk sum of all trees. The smaller the shrinkage parameter is, the slower will the boosted model evolve and the lower will be the learning rate. The intuition behind the a slow learning rate is that it allows to slowly improve the boosted model in regions where it relatively mis-performs without perturbing dramatically the model. In that sense it is "safer" and generally leads to better results. A learning rate that is too small may however require many trees to converge. It is usual to set a lower maximal depth than in random forests since GBM is additive and the new tree growth takes into account the other trees that have already been built. In our case, we sacrifice computational time and use a small learning rate to avoid overfitting that we compensate with a high number of trees. The maximal depth of the trees allowed has been spanned from 1 to 5.

C.3.3 Generalised linear models

Generalised linear models are very commonly used and allow a flexible extension of linear models for responses that are not necessarily normally distributed. Given that we have a binary classification problem, we use a binomial model with a logit link function. To prevent over-fitting and increase the prediction accuracy of the statistical model, we span the Elastic net regularisation from pure Lasso to pure Ridge penalisation, which are described below.

- *Lasso* imposes a constraint on the coefficients of the form $\|\beta\|_1 \leq t$ for some t and consequently performs both regularisation and variable selection. It limits the magnitude of the coefficients multiplying the predictors such that only the most important predictors are kept in the model, the others being put to exactly 0. While the variable selection feature of the Lasso is an attracting property for interpretation, it can be a drawback in prediction: some variables having possibly an impact, albeit small, on the response are nonetheless rejected.

- *Ridge regularisation* imposes a constraint of the form $\|\beta\|_2 \leq t$ for some t and it is therefore very unlikely (with probability 0) that a coefficient is put to 0. Consequently ridge regression does not perform variable selection but only shrinkage.

Because each of the two types of constraints has advantages and drawbacks, it is interesting to use different mixtures of l^1 and l^2 penalisation. The resulting Elastic net regularisation still allows for variable selection but it is less "sharp" as the l^2 dominates the l^1 penalisation. In all cases, we let the algorithm search for the optimal magnitude of penalisation which is determined through cross-validation. The trade-off is evident: a low penalisation will cause high variance and decrease the prediction accuracy while a high penalisation will induce a high bias and decrease the training accuracy.