CESifo WORKING PAPERS

10609
2023

August 2023

# Cultural Distance and Ethnic Civil Conflict

*Eleonora Guarnieri*

CESifo

## Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10609*

# Cultural Distance and Ethnic Civil Conflict

## Abstract

Ethnically diverse countries are more prone to conflict, but why do some groups engage in conflict while others do not? I show that civil conflict is explained by ethnic groups' cultural distance to the central government: an increase in cultural distance, proxied by linguistic distance, increases an ethnicity's propensity to fight over government power. To identify this effect, I leverage within-ethnicity variation in linguistic distance resulting from power transitions between ethnic groups over time. I provide evidence that the effects can be attributed to differences in preferences over both the allocation and the type of public goods.

JEL-Codes: D740, Z100, O550.

Keywords: ethnic civil war, culture, linguistic distance, Africa, Bantu expansion.

*Eleonora Guarnieri*
*University of Exeter / United Kingdom*
*e.guarnieri@exeter.ac.uk*

August 8, 2023
I thank Eren Arbatli, Miriam Artiles, Vojtěch Bartoš, Anke Becker, Graziella Bertocchi, Mathias Bühler, Filipe Campante, Davide Cantoni, Damian Clarke, Klaus Desmet, Quoc-Anh Do, Marc Fabel, Christian Fons-Rosen, Ariel Gomez, Andrea Guariso, Soeren Henn, Roland Hodler, Paul Hufe, Andreas Kotsadam, Panu Poutvaara, Helmut Rainer, Patrick Reich, Thorsten Rogall, Wayne Sandholtz, Paul Schaudt, Augustin Tapsoba, Ana Tur-Prats, Romain Wacziarg, Fabian Waldinger, and seminar participants at LMU Munich, the NBER Political Economy Program Meeting 2022, NEUDC 2020, ASSA Meetings 2023, NICEP 2023 Conference, EDGE Jamboree, PUC Chile, University of St Gallen, University of Glasgow, Queen's University Belfast, Lund University, University of Milan Bicocca, ECARES, CUNEF, EI University, University of Alicante, IESEG, and University of Exeter for helpful suggestions. All remaining errors are my own.

## 1 Introduction

Since the end of World War II, civil conflicts have been the most common form of war around the world. Most civil conflicts have taken place in Africa, and the deadliest have been fought along ethnic lines.[1] Globally, civil conflict has caused five times as many deaths as inter-state war, and has led to long-lasting economic and social disruptions (Fearon and Laitin, 2003). Yet, what triggers ethnic civil conflict remains largely debated. A well-established literature has analyzed the role of ethnic-diversity measures at the country level using ethnic fractionalization (Fearon and Laitin, 2003; Esteban et al., 2012) and polarization indices (Montalvo and Reynal-Querol, 2005; Esteban et al., 2012), and has found that more ethnically diverse countries tend to experience more conflict. However, given that all ethnicities in a country face the same aggregate level of diversity, why do some rebel against the central government while others do not? And why do they rebel sometimes and not at other times?

To answer these questions, I move the study of diversity and conflict from the country level to the ethnicity level and explore how ethnic groups' cultural *distance* to the central government affects their decision to rebel against it. In line with existing theories of conflict (Esteban and Ray, 2011; Spolaore and Wacziarg, 2017), I focus on disputes over government power (i.e., public goods).[2] Every group in a country is subjected to public goods and policies provided by the government. Different preferences over public goods—either their *type*, their *allocation*, or both—between an ethnic group and the government may give rise to disagreement and ultimately conflict. I test the hypothesis that an increase in cultural distance increases an ethnicity's propensity to fight over government power.

[1] Ethnic civil conflicts, the focus of this paper, are "armed conflicts between the government of a state and one or more internal opposition groups that cause at least 25 battle-related deaths within a year in which armed groups (i) explicitly pursue ethno-nationalistic aims, motivations, and interests; and (ii) recruit fighters and forge alliances on the basis of ethnic affiliation" (Wimmer et al., 2009).

[2] Between 1961 and 2017, 66 percent of the whole prevalence of ethnic conflict in Africa was accounted for by conflict over power (Sundberg and Melander, 2013). The remaining conflicts were fought over territory.

While prominent in the theoretical literature on diversity and conflict (Esteban and Ray, 2011; Caselli and Coleman, 2013; Spolaore and Wacziarg, 2017), this hypothesis has never been empirically tested at the ethnic-group level in the context of civil war. I generate a dataset that measures how culturally distant each ethnic group is from the central government at a given point in time. I measure cultural distance using linguistic distance, as is standard in the literature.[3] I follow all politically relevant ethnic groups over time and construct an indicator for whether they rebel over government power in a specific year or not. The resulting dataset is a panel of 265 distinct ethnic groups in 44 African countries observed over a period of 57 years (1961-2017).[4]

Using changes in the ethnic identity of the government as a source of within-group time variation in cultural distance, I find support for my hypothesis. After a power transition, the prevalence of conflict over power increases among ethnic groups that become culturally more distant from the government, but not among those who become culturally closer or whose cultural distance remains unchanged. This effect cannot be attributed to differential trends in conflict involvement prior to a government transition between these subgroups of ethnicities. The reaction to cultural distance occurs immediately after a change in leadership and is large in magnitude: a one standard deviation increase in cultural distance increases the prevalence of conflict over government power by 0.4 standard deviations.[5]

My results hold after conditioning on a rich set of fixed effects. The data allows me to include ethnicity fixed effects, thus isolating the effect of cultural distance

[3]See, for example, Fearon and Laitin (2003), Desmet et al. (2009), Desmet et al. (2012), Esteban et al. (2012). The use of linguistic distance is motivated by the notion that language is a salient dimension of culture transmitted through generations (Spolaore and Wacziarg, 2016a). Different languages are the result of horizontal separations between populations, and these separations are likely to go hand in hand with cultural divergence.

[4]I focus on Africa for two reasons. First, it constitutes a unique setting due to the high degree of ethnic diversity within countries. Second, for the purpose of identification, Africa offers two unique natural experiments I exploit in my analysis: the Bantu expansion and the random allocation of country borders.

[5]For comparison, this corresponds to more than double the effect of an increase in resource inequality uncovered by Guariso and Rogall (2017). They show that a one standard deviation increase in rainfall inequality between an ethnicity and the government increases the likelihood of conflict by 0.15 standard deviations.

from any time-invariant ethnic-specific characteristic potentially associated with conflict: cultural norms, social structure, ancestral arrangements and geographic conditions, whether the group was split by country borders, or a general propensity to rebel. The inclusion of country-by-year fixed effects ensures that the coefficients are not confounded by government characteristics or by country-level time shocks affecting all ethnicities within a country, including the overall effect of a power transition. Country-by-year fixed effects also allow me to keep constant the aggregate level of diversity and focus on the effect of cultural distance between groups.

Reassuringly, results hold after conducting a large set of robustness tests and employing alternative specifications. For instance, estimates are similar when I restrict the sample to those ethnicities that were artificially partitioned across countries during the Scramble for Africa, which allows me to add ethnicity-year fixed effects to my baseline specification.[6]

While I interpret my findings as the effect of cultural distance on conflict, I consider alternative explanations. First, linguistic distance may be correlated with income differences between groups. In line with the existing literature (Morelli and Rohner, 2015; Guariso and Rogall, 2017), I find that income differences are associated with conflict, but the effect of linguistic distance on conflict occurs over and above the effect of income differences. Second, the effects are not confounded by political exclusion. When ethnic groups lose power after a political transition, their linguistic distance to the government mechanically increases. Therefore, results could be consistent with ethnicities rebelling after their members are excluded from the coalition for reasons not necessarily linked to cultural considerations. While political exclusion is directly correlated with a group's propensity to rebel, I do not find that political exclusion, once controlled for in the main specification, explains away the effect of linguistic distance on conflict. Taken together, these additional findings show that the estimates reflect the effect of cultural distance, and not of other confounding factors.

[6]This approach is analogous to the one employed by Dickens (2018). Michalopoulos and Papaioannou (2014) were the first exploiting the partition of African ethnicities in multiple countries as a source of within-ethnicity variation.

A remaining potential concern with the analysis is the presence of other omitted factors affecting both cultural distance and conflict. For instance, historical conflict between two ethnicities might have a direct impact on both the probability of current animosities between them and their contemporary linguistic distance. I address these endogeneity concerns using a novel instrumental variables (IV) approach. To instrument for cultural distance, I exploit the Bantu expansion, a massive migration that occurred 5,000 years ago, which changed language and culture in some parts of Central and sub-equatorial Africa but not others. To my knowledge, this is the first paper exploiting this cultural shock as a source of exogenous variation in cultural distance between today's African ethnicities.

I construct and assign to each ethnicity a *Bantu index*, which captures the extent to which ethnic homelands were exposed to the route of Bantu migration. I then instrument for cultural distance between two ethnic groups using the absolute difference in their Bantu index. The idea behind this instrument is the following. Groups whose homelands were highly exposed to the Bantu migration route inherited Bantu culture, and should be culturally distant to those whose culture remained unaffected. Conversely, two groups with a similarly high or similarly low Bantu exposure should be culturally close to each other, because they either both inherited Bantu culture, or they both kept their pre-existing one. Consistent with this idea, the first stage documents a strong positive association between the absolute distance in the Bantu index and cultural distance. In the second stage, the effect of cultural distance on conflict over power becomes larger in magnitude when compared to the OLS estimates.

In the remainder of the paper, I explore factors that may explain why cultural distance triggers conflict over the control of the central government. I find evidence consistent with conflict arising due to cultural disagreements over both the *type* and the *allocation* of public goods. Theory predicts that if linguistic distance operates through a disagreement over the public policies that all groups in a country must share, then we should expect linguistic distance to explain only conflict over public goods, and not over rival goods like territory and resources (Spolaore and

Wacziarg, 2016b, 2017; Esteban and Ray, 2011). In line with these predictions, I do not find a positive association between linguistic distance to the government and rebellions over territorial control.

To more directly test whether linguistic distances between groups are associated with diverging preferences over public policies, I exploit individual-level data from seven rounds of the Afrobarometer survey for more than 110,000 respondents in 28 African countries. First, I find that respondents are more likely to oppose a wide range of current government policies if they identify with ethnicities that are culturally distant from the ethnic groups in power at the time of the survey. Second, for each ethnic group in the sample, I construct an index capturing group preferences over public policies. The index is based on respondents' views about the most important public matter that the government should address among health, infrastructure, services, food security, governance, or the economy. I then generate a dyadic dataset in which I pair every ethnic group to all other groups residing in the same country. Using a gravity specification, I find that the larger the linguistic distance between a pair of ethnicities, the more likely they are to differ in their preferences over public policies.

In a final test, I investigate whether my findings can be explained by disagreement over the allocation of public goods, their type, or both. Established work has documented the presence of ethnic favouritism in Africa (Burgess et al., 2015; De Luca et al., 2018), which extends to non-coethnic but linguistically similar groups (Dickens, 2018). Is conflict over government power triggered solely by favouritism, i.e., discontent with the unequal *allocation* of public resources, or do diverging preferences over the *type* of public good provided also play a role? To shed light on this question, I exploit the V-Dem dataset, which provides information at the country-year level on (i) whether the national budget is mostly spent on "private" or "public" goods and (ii) whether public services are equally distributed across social groups or not. If only disagreement over the allocation of public goods was driving the effects, then linguistic distance should not trigger conflict in settings where the national budget is mostly spent on equally-distributed

public goods. On the contrary, I find that the effect of linguistic distance on conflict, albeit smaller in magnitude, is still positive and significant in such settings. I take this as evidence highlighting the role of cultural divergences over the preferred type of public goods in explaining ethnic conflict over government power.

This paper contributes to several strands of literature. First, it adds to the vast literature on civil war (see Blattman and Miguel (2010) for a review). Among studies focusing on ethnic conflict, some have highlighted time-invariant factors that make certain ethnicities more likely to experience conflict, like segmentary lineage organization (Moscona et al., 2020), inter-personal diversity (Arbatlı et al., 2020), and ethnic partitioning (Michalopoulos and Papaioannou, 2016). However, even after acknowledging these different propensities to experience conflict, what drives ethnic groups to fight against other specific ethnicities at specific points in time remains unclear. By focusing on cultural distance between ethnic groups and governments, I relate ethnicities to their potential opponents and show that the prevalence of conflict is a function of the characteristics of both sides involved, and not only of the characteristics of one side. Two studies adopt a similar approach. Guarnieri and Tur-Prats (2023) focus on the intensive margin of violence, and show that cultural distance in gender norms between ethnic belligerents increase the use of conflict-related sexual violence. Guariso and Rogall (2017) show that income inequality between an ethnic group and the leading group in a country increases the likelihood of ethnic conflict between them. Whereas Guariso and Rogall's (2017) focus is on the economic drivers of conflict, my paper explores the role of deep-rooted cultural determinants.

This study is closely related to theoretical and empirical work on ethnic diversity and conflict, summarized in the next section. Empirically, this literature has mostly focused on the relationship between diversity indices and conflict at the country level (Esteban and Ray, 2011; Esteban et al., 2012) or at the district level (Amodio and Chiovelli, 2018). This aggregate-level approach, however, remains silent on the group-specific heterogeneity in the decision to fight. By moving the analysis to the ethnicity level, this paper unpacks the country-level associations

and identifies precisely which ethnicities engage in conflict, when they do so, and why. More broadly, this paper adds to the literature on the consequences of ethnic diversity. In addition to studies focusing on conflict, others have analyzed a large range of economic outcomes (see Alesina and La Ferrara (2005) for a review), adopting as unit of analysis either countries, cities, or grid cells (see, for instance, Montalvo and Reynal-Querol (2017)). My study is part of a new line of work that keeps the aggregate level of diversity constant and focuses on distances between entities within aggregate units. While this paper studies ethnicities within countries, Gomes (2020) examines individuals within geographical radii, and finds that children of mothers who are linguistically distant from their neighbors have worse health outcomes.

Finally, this paper contributes to the literature exploring the determinants of ethnolinguistic diversity. Michalopoulos (2012) finds that contemporary ethnic diversity is rooted into geographic variability. Work by Ashraf and Galor (2013) shows that genetic diversity, determined during the prehistoric migration of humans out of Africa, is a fundamental determinant of ethnic heterogeneity within countries. Galor et al. (2018) explore the geographic roots of specific linguistic traits like the structure of the future tense, sex-based grammatical gender, and politeness distinctions. A recent contribution by Dickens (2020) studies linguistic distance between neighboring groups and finds that ethnicities separated across geographic regions with high variation in land productivity are more similar than those separated across more homogeneous regions due to historical trade. My study speaks to this literature by linking the prehistoric Bantu migration to cultural distance between ethnic groups today.

## 2 Conceptual Framework

The link between diversity and conflict has been debated by a large strand of interdisciplinary literature. This debate originates from the so-called primordialist

view, according to which dissimilarities between groups spur conflict.[7] An anthropological formulation of this view posits that a society formed by culturally divergent groups can only be sustained through a political order in which one of the cultural groups politically dominates the others (Smith, 1965). In turn, this political structure inevitably generates cultural dissensus, given that all groups within a jurisdiction are bound to the public policies provided by a culturally distant group. Yet, cultural differences might also impede conflict, by "focusing the ambitions of various groups on alternative sources of gratification, thereby preventing them from impinging on each other" (Horowitz, 2000, p. 138). Taken together, these opposing views hint at a potentially ambiguous relationship between cultural differences and conflict.

Literature in economics has recently shed further theoretical and empirical light on these issues and has highlighted the importance of distinguishing between conflict over private goods and conflict over public goods. In their study of inter-state conflict, Spolaore and Wacziarg (2016b) provide empirical evidence in support of Horowitz's (2000) conjecture. They find that inter-state conflict is more likely between similar populations with similar preferences over the same type of resources or territories—the typical bone of contention in international conflict. Esteban and Ray (2011) and Esteban et al. (2012) focus on intra-state conflict and analyze the relationship between a country's level of diversity and the equilibrium level of conflict. They show that conflict can be approximated by a weighted average of three measures of diversity: a Gini coefficient, an index of ethnic fractionalization, and a measure of ethnic polarization. The weights of each of these measures depend on a country-level measure of "publicness of the prize." When groups fight for the control of an excludable private good, inter-group distance—best captured by a measure of polarization—is not a powerful explanation for conflict. The only explanatory factor will be a country's distribution of group sizes, best captured by an index of fractionalization. At the other extreme, when a prize is fully public,

[7]This view is often juxtaposed to the so-called instrumentalist view, which sees ethnic conflict as a result of grievances and inequality between groups, rather than the result of different identities.

differences in group preferences over public goods will emerge and conflict will be better explained by indices of polarization and Gini coefficients, which capture the degree of inter-group distances within a country.

Along similar lines, Spolaore and Wacziarg's (2017) theoretical framework highlights that the impact of cultural distance on conflict depends on whether groups are fighting over the control of public goods or rival goods. Conflict over the control of public goods—e.g., public policies that all must share within a jurisdiction—is more likely between culturally distant groups that hold different preferences, consistently with what highlighted in Esteban and Ray's (2011) theory. I bring this hypothesis to the data and, for the first time, test this prediction at the ethnic group level in the context of ethnic civil conflict.

## 3 Data and Descriptive Statistics

### *3.1 Main Dataset Construction*

To examine the relationship between cultural distance and conflict, I assemble a new dataset measuring how culturally distant an ethnic group is from the government of a country in a specific year and which groups engage in violent conflict at a certain point in time. In this section, I describe the procedure I adopted to construct the dataset.[8]

**Ethnic groups.** The first step consists of retrieving information on ethnic groups. To this end, I exploit the Ethnic Power Relations (EPR) Dataset Family. For each country of the world, EPR lists all politically relevant ethnic groups between 1946-2017 and their access to government power. A group is defined as *politically relevant* if a political actor claims to represent the interests of a group in the national political arena, or if group members are systematically and intentionally discriminated against. EPR codes the degree to which each ethnic group's representatives hold executive-level state power in each year, which ranges from total control of the government to political discrimination.[9] Based on this data, I

[8]See Appendix B for additional details on data sources and variables.

[9]Whenever government changes happen in the same year as a conflict, EPR reports the power

construct a panel that follows each ethnic group over time.

**Ethnic civil conflicts.** To construct the dependent variable, I use the UCDP-PRIO Armed Conflict Dataset, which includes information on civil conflicts and the actors involved in them. The UCDP/PRIO dataset defines civil conflicts as "armed conflicts between the government of a state and one or more internal opposition groups that cause at least 25 battle-related deaths within a year." I focus on *ethnic* civil conflicts in Africa fought between 1961 and 2017, those where armed groups "explicitly pursue ethno-nationalistic aims, motivations, and interests; and recruit fighters and forge alliances on the basis of ethnic affiliation" (Wimmer et al., 2009).

UCDP/PRIO provides information on the issue over which rebel groups fight. First, a conflict can be about *government power*, i.e., about the type of political system, the replacement of the central government, or the change of its composition. Second, a conflict can be about *territory*, i.e., about the exclusive control of a specific region for own settlement, local resource use, or, in extreme cases, secession. Examples of struggles over state power include the case of the Liberians United for Reconciliation and Democracy (LURD) (2000-2003) against the government led by Charles Taylor, or the Tigray People's Liberation Front (TPLF) in Ethiopia (1976-1986) fighting against the hegemony of the Amharan government. Conflicts fought over territory include the one between the Movement of Democratic Forces of Casamance (MFDC) and the government of Senegal for the control of the Casamance region (1990-2011). This category also includes cases where rebel groups seek control over specific resources linked to a certain territory, such as the Niger Delta People's Volunteer Force (NDPVF) fighting for controlling the petroleum resources of the delta region in Nigeria (2004).

**Rebels' ethnic identity.** To assign an ethnic identity to rebel groups, I exploit the ACD2EPR dataset. This dataset maps each rebel group to one or more ethnic groups in the EPR dataset. To continue some of the examples above, members of the Mandingo and Krahn ethnic groups formed the LURD rebel group in Liberia;

relations that were in place *before* the conflict outbreak.

the Tigry ethnic group formed the TPLF rebel group in Ethiopia; and the Ijaw group formed the NDPVF in Nigeria. I merge this information to the panel of ethnicities, each of which constitutes a *potential rebel*. I construct a binary variable that is equal to 1 if a potential rebel is involved in a conflict against the central government in a given year and zero otherwise.[10]

**Governments' ethnic identity.** To assign an ethnic identity to the government, I exploit information provided by the EPR dataset, which indicates which ethnic groups hold executive government power in a country at a certain point in time.[11] In some cases, government power is held exclusively by one ethnic group, that EPR classifies as either *monopolist* or *dominant*. In other cases, the government results from a coalition of ethnic groups, which can take either the role of *junior* or *senior partner*.[12]

**Linguistic distance.** To measure cultural distance between each potential rebel and the government, I use linguistic distance, as is standard in the literature.[13] The source of information on languages is the EPR Ethnic Dimensions (EPR-ED) dataset. EPR-ED assigns up to three languages to each EPR ethnic group, indicating the three largest language segments spoken by group members in descending order and their relative size.[14] The advantage of this dataset is that it provides a nuanced description of the linguistic (and therefore, cultural) identity

[10]All groups within a country can be potential rebels, apart from monopolist or dominant ones. By definition, these groups cannot rebel against themselves, because they have the exclusive control of the government. Indeed, there are no instances in which a rebel group involved in a conflict is associated with a dominant or monopolist ethnicity. Whenever an ethnicity becomes dominant or monopolist, it drops out of the panel. Note that these cases are quite rare. In a robustness test, I re-run the analysis excluding these ethnicities.

[11]EPR takes into consideration where executive power is exercised when coding ethnicities' access to it. This includes the presidency, the cabinet, and senior posts in the administration, including the army, or the ruling party leadership in one-party states.

[12]Whether a group is classified as senior or junior partner depends on the group's absolute influence in the executive, measured by the number and importance of the positions controlled by group members.

[13]An advantage of using linguistic distance instead of measures of cultural distance based on survey data is that the latter may be affected by contemporary conflict, while linguistic distance is pre-determined and arguably more exogenous. Reassuringly, linguistic distance is positively correlated with a Facebook-based measure of cultural distance developed by Obradovich et al. (2022) (see the paper's Figure 2A and Appendix Figure B1).

[14]These refer to indigenous African languages, not those imported by colonizers.

of an ethnicity also when the latter is not homogeneous. Therefore, my measure of cultural distance incorporates this intra-group heterogeneity. I merge languages to linguistic trees in the Ethnologue database. For each language, the Ethnologue provides a classification starting with the language family (e.g. Afro Asiatic, Nilo-Saharan, Creole), followed by "nodes", i.e., the branching points of the linguistic tree, and ending with the language itself.

To compute linguistic distance between each potential rebel and the government, I employ a three-steps procedure. First, I calculate the distance between each pair of languages ($x$ and $y$) as follows:

$$d_{xy} = 1 - \left( \frac{\#\text{ of common nodes between } x \text{ and } y}{\frac{1}{2}(\#\text{ of nodes of language } x + \#\text{ of nodes of language } y)} \right)^{\lambda}$$

This measure, called cladistic distance, is the most frequently used in the literature (Fearon and Laitin, 2003). Languages originating from different families have no nodes in common, and their distance will be equal to 1. The parameter $\lambda$ ranges between 0 and 1, and is used to attribute higher weight to earlier common nodes, as early separations in the language tree are likely to signify larger cultural divergence on average than later separations. I follow Fearon and Laitin (2003) and assign to $\lambda$ a value of 0.5.[15] Second, I calculate linguistic distance between potential rebels ($r$) and each ethnic group forming the government ($g_i$):

$$LD_{rg_i} = \sum_{x=1}^{3} \sum_{y=1}^{3} (w_{rx} \times w_{g_i y} \times d_{xy}) \tag{1}$$

where $w_{rx}$ and $w_{g_i y}$ are the fraction of population speaking language $x$ in group $r$ and language $y$ in group $g_i$, respectively. Third, I compute linguistic distance between each potential rebel and the government:

$$LD^{W} = \sum_{i=1}^{N} p_{g_i} \times LD_{rg_i} \tag{2}$$

[15] In a robustness test, I alter the value of this parameter.

where $N$ is the total number of ethnicities forming the government, and $p_{g_i}$ is a weight reflecting the position of power of group $g_i$ in the government. When the government is composed only by one ethnic group, $LD^W$ equals $LD_{rg_i}$. In the case of a government coalition, I assign a higher weight to linguistic distance between each potential rebel and the senior partner.[16] Senior partners will receive double the weight of each junior partner.[17] Alternatively, I use an unweighted version of the just described linguistic distance measure.[18]

**Exposure to the Bantu expansion.** To estimate the exposure of each ethnic group to the Bantu expansion, I digitize the Bantu expansion route reconstructed by Grollemund et al. (2015) (see Figure A-3, left). I then overlay the path onto the Murdock map.[19] I leverage ethnic groups' ancestral homelands to capture the exposure to the Bantu expansion for the following reasons. First, EPR provides ethnic settlements for contemporary ethnic groups, which are likely the result of recent phenomena endogenous to conflict, as well as of modern-day country borders imposed artificially by colonial powers. Second, the geographic locations of ethnic groups in EPR capture groups' regional presence and contemporary settlements, rather than homelands as in the Ethnographic Atlas.

**Geographic controls.** The GeoEPR dataset provides polygons describing each ethnic group's geographic location, allowing overlaps between different groups' settlements.[20] To add information on geographic characteristics to each settlement, I combine GeoEPR with data on elevation (from which I compute ruggedness) at

[16]Rebels may be intentionally fighting against one specific group in the coalition, or against the government as a whole. Since I do not have systematic information on this issue, I consider all ethnic groups forming the government.

[17]In the example illustrated in Table A-1, in the distance between Lari/Bakongo and the government, $p_{g_i}$=0.5 when $g_i$=Mbochi (the senior partner), and $p_{g_i}$=0.25 when $g_i$=Bateke or Kouyou (the junior partners).

[18]I compute the following: $LD^{UW} = \sum_{i=1}^{N} \frac{LD_{rg_i}}{N}$.

[19]The Murdock Map records the location of ethnic groups' ancestral homelands prior to European contact as recorded by Murdock. I merge contemporary ethnic groups to pre-colonial ethnic settlements through the Linking Ethnic Data for Africa (LEDA) R-package. For additional details on the methodology, see Müller-Crepon et al. (2020) and Section C-2 in the Appendix.

[20]GeoEPR does not provide the location of groups classified as living exclusively in urban areas. To these, I assign the coordinates of the country's capital city. These constitute 2% of ethnicities in my sample.

the grid level provided by Worldclim (Fick and Hijmans, 2017). Furthermore, I use the Caloric Suitability Index (CSI) (pre-1500 average) from Galor and Özak (2016) to add information on potential agricultural output.

Figure A-1 and Table A-1 illustrate the structure of my data and the identifying variation with an example from Congo. There are six politically-active ethnic groups: Lari/Bakongo, Kouyou, Mbochi, Bateke, Bemba, Vili. I follow each of these "potential rebels" over time, and track both their involvement in conflict and their cultural distance to the government. The within-ethnicity variation in cultural distance is driven by power transitions, like the one in Congo in 1998.

### *3.2 Summary Statistics*

The resulting dataset includes 265 ethnic groups in 44 African countries over a period of 57 years (1961-2017) (N=9,990).[21] The source of within-ethnicity variation in cultural distance stems from 138 changes in the ethnic composition of the government. Table C-1 reports the number of changes experienced by each country in the sample. Ten countries did not experience any change in the ethnic composition of the government in the period considered. Large variation comes from west-central Africa, where countries such as Niger and Nigeria experienced the highest number of power transitions. Table C-2 reports descriptive statistics of the main variables used in the analysis. 22 countries experienced at least one conflict over government power during the period considered, with a prevalence of 0.048. Of the 265 potential rebels in the sample, 64 became rebels at least once. Within countries, the average linguistic distance between potential rebels and governments is 0.44. Potential rebels and governments are culturally close in Malawi (0.09 on average) and culturally very distant in Liberia (0.77 on average). The Liberian example speaks to the accuracy of linguistic distance in capturing cultural differences. The Americo-Liberians were the group in power over almost the whole period considered. This group originates from free-born and formerly

[21] I exclude countries where there is only one ethnic group and countries that are not part of the EPR dataset family.

enslaved African Americans who emigrated in the 19th century and became the founders of Liberia. They imported African-American and Caribbean culture, and are therefore culturally different from the other Liberian groups.

## 4 Empirical Strategy

The objective of the empirical analysis is to investigate whether changes in cultural distance to the government have an effect on ethnic groups' involvement in ethnic civil conflict. To this end, I estimate the following linear probability model:

$$\text{Conflict}_{rct} = \lambda_{ct} + \zeta_{rc} + \theta_{rc}t + \beta \text{LD}_{rct} + \gamma \text{G}_{rct} + \epsilon_{rct} \tag{3}$$

where the dependent variable is an indicator that takes value 1 if there is a rebel group of ethnicity $r$ fighting the government of country $c$ in year $t$. $\text{LD}_{rct}$ indicates linguistic distance between ethnic group $r$ and the government of country $c$ in year $t$. $\lambda_{ct}$ denotes a full set of country-year fixed effects. These capture time-invariant characteristics at the country-level—e.g., colonial history or geography—that might make some countries overall more or less prone to conflict than others. They also account for time-specific shocks common to all ethnic groups in a country and for government characteristics, such as whether the government represents a large versus a small share of the population; whether it is a coalition of groups or formed by a single dominant one; whether it is a bellicose or peaceful government; the quality and type of its public policies; the strength of the army; or whether a certain government is established as a result of a conflict and thus is more vulnerable to retaliation from opposing groups. Crucially, through country-year fixed effects I can isolate the effect of cultural distance from that of a country's aggregate level of fractionalization and polarization.

My dataset allows me to add a full set of ethnicity-by-country fixed effects ($\zeta_{rc}$), as well as ethnicity-by-country year trends ($\theta_{rc}t$). By adding ethnicity fixed effects, I control for any time-invariant characteristics specific to an ethnic group that might affect its propensity to experience conflict. These include all ethnic traits that have been associated with conflict by the literature like social structure

(Moscona et al., 2020), within-group inter-personal diversity (Arbatlı et al., 2020), or ethnic partitioning (Michalopoulos and Papaioannou, 2016). By controlling for an ethnicity-specific year trend, I can furthermore account for dimensions at the ethnicity level that change linearly over time. Finally, to isolate the effect of cultural distance from the effect of geography, I control for a vector of differences between potential rebels and the government ($G_{rct}$): geodesic distance, distance in elevation, ruggedness, and the caloric suitability index (CSI).[22] I cluster standard errors two-way by country and year.[23]

## 5 Main Results

Table 1 reports estimates from the linear probability model described in equation 3. Column 1 shows that a one standard deviation increase in cultural distance increases conflict over government power by 2.5 percentage points. The coefficient remains positive and significant and increases in magnitude when adding ethnicity year trends and geographic controls in columns 2 and 3. In my preferred specification in column 3, a one standard deviation increase in cultural distance increases conflict over power by 8.6 percentage points (0.4 standard deviations). Results are similar when adopting an unweighted measure of linguistic distance that does not account for the power structure in the government coalition (columns 4-6).

### *5.1 Effect Dynamics*

Having established that an increase in cultural distance to the government increases an ethnicity's propensity to rebel, I turn to examining the dynamics of the effect through an event study design. The event of interest is a government transition experienced by all ethnic groups in a country and year. I start by conducting a canonical event study to detect the effect that a government transition

[22]Given the dynamic nature of the GeoEPR dataset, ethnic settlements may vary over time, making geographic characteristics also time-variant. However, since ethnicity fixed effects absorb the variation almost entirely, I do not control for potential rebels' elevation, ruggedness, and CSI to avoid collinearity issues.

[23]I provide alternative clustering in Table A-2.

TABLE 1: *Cultural Distance and Ethnic Civil Conflict over Power*

| | Dependent variable: Ethnic conflict over power | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Linguistic distance$^{W}$ | 0.025*** | 0.075*** | 0.086*** | | | |
| | (0.008) | (0.016) | (0.019) | | | |
| Linguistic distance$^{UW}$ | | | | 0.021** | 0.071*** | 0.081*** |
| | | | | (0.009) | (0.016) | (0.017) |
| Mean of dep. var. | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 |
| Country-year FE | yes | yes | yes | yes | yes | yes |
| Ethnicity-country FE | yes | yes | yes | yes | yes | yes |
| Ethnicity-country trend | | yes | yes | | yes | yes |
| Geographic controls | | | yes | | | yes |
| Observations | 9,990 | 9,990 | 9,990 | 9,990 | 9,990 | 9,990 |
| Adjusted R-squared | 0.447 | 0.506 | 0.507 | 0.447 | 0.506 | 0.507 |

*Notes*: The unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government and 0 otherwise. *Linguistic distance*$^{W}$ captures linguistic distance between each potential rebel and the ethnic groups at the government, weighted by the position of power of each ethnic group in case the government is a coalition. *Linguistic distance*$^{UW}$ is an unweighted average of linguistic distance between each potential rebel and each ethnic group at the government. The linguistic distance measures are standardized. Geographic controls include the logged geodesic distance between each ethnic group and the government, absolute distance in elevation, absolute distance in ruggedness, and absolute distance in the caloric suitability index (CSI). The sample includes 265 distinct ethnic groups in 44 African countries over 57 years (1961-2017). Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

has on the *overall* level of conflict in a country. This exercise allows me to rule out that government transitions—my source of variation—systematically arise due to abnormal conflict dynamics. I estimate the following:

$$\text{Conflict}_{rct} = \sum_{\substack{k=-5,\\ k\neq -1}}^{5} \alpha_k \text{T}_{rck} + \lambda_{ct} + \zeta_{rc} + \theta_{rc} t + \gamma \text{G}_{rct} + \epsilon_{rct} \tag{4}$$

where $\text{T}_{rck}$ is an indicator for $k$ years before (or after) a government transition experienced by ethnicity $r$ in country $c$.[24] Panel A of Figure 1 shows that the overall level of conflict is fairly stable around a power transition, which is reassuring

[24]All other elements are equivalent to those in equation 3. I estimate this on a sample of 210 ethnic groups in 34 countries where at least one government transition took place between 1961 and 2017. I focus on a 5-year window around the event because the median time period between government transitions is 5 years in my sample.

in that it suggests that government changes do not seem to systematically emerge from (or generate) peculiar conflict dynamics.

Next, I test whether the results described in Section 5 are confirmed in the event-study setting and explore how long the effect of cultural distance persists after a transition. To this end, I add an interaction term to examine whether there is a differential response for ethnic groups that become culturally more distant after the event compared to groups whose distance decreases or remains unchanged. I estimate the following:

$$\text{Confl}_{rct} = \sum_{\substack{k=-5,\\ k\neq -1}}^{5} \alpha_k \text{T}_{rck} + \sum_{\substack{k=-5,\\ k\neq -1}}^{5} \tau_k \text{T}_{rck} \times \text{MoreDist}_{rc} + \lambda_{ct} + \zeta_{rc} + \theta_{rc} t + \gamma \text{G}_{rct} + \epsilon_{rct} \quad (5)$$

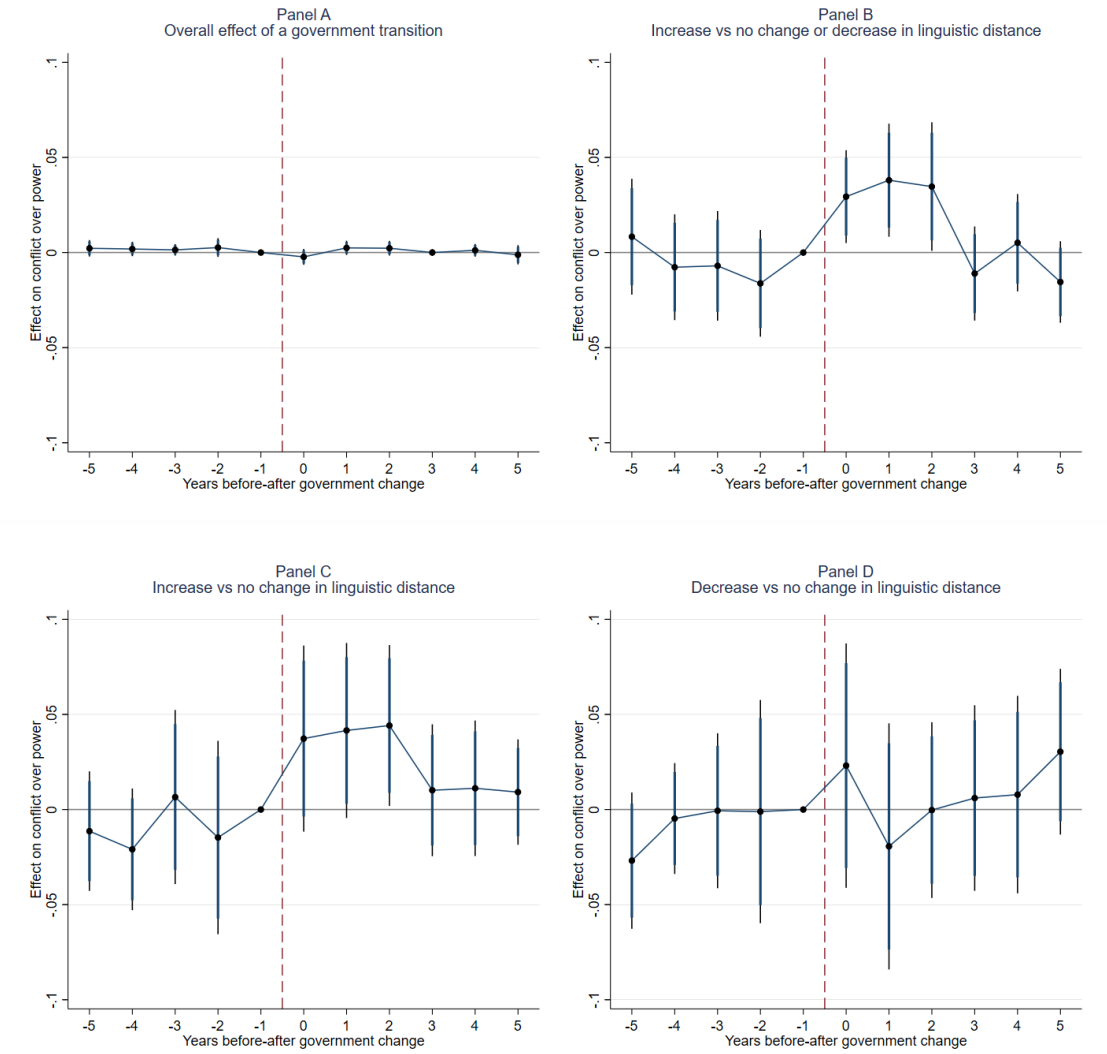
where $\text{MoreDist}_{rc}$ is an indicator that equals 1 if ethnic group $r$ becomes more culturally distant following the government transition of interest in country $c$ and 0 if an ethnic group's cultural distance either remains unaffected or decreases. Panel B of Figure 1 presents estimates of $\tau_k$ and shows that, while there is no differential involvement in conflict prior to a transition—a key identifying assumption—, groups whose cultural distance increases become more likely to rebel afterwards. This reaction persists for three consecutive periods after the new government takes power and, as shown in Panels C and D, is driven by an increase in conflict among groups becoming culturally more distant to the government, and not by a decrease in conflict among groups becoming culturally closer.[25]

### *5.2 Ruling out Alternative Explanations*

To attribute the uncovered effects to cultural distance, I assess the role of two potential confounders: political representation and exogenous income differences between groups.

[25]To tease this out, I proceed as follows. In panel C, I estimate equation 5 removing from the sample ethnic groups whose linguistic distance decreases after a transition. I thus compare the reaction of groups becoming culturally more distant only to that of groups who face the transition but whose linguistic distance remains unchanged. In panel D, I estimate an equivalent of equation 5, but replacing $\text{More Dist}_{rct}$ with an indicator that equals 1 if a group becomes culturally *closer* following the transition and 0 if an ethnic group's linguistic distance remains unchanged. Ethnic groups whose linguistic distance increases are removed from this sample.

FIGURE 1: *Ethnic Conflict around a Government Transition*




*Notes*: The figures plot coefficient estimates of event studies in equation 4 (Panel A) and 5 (Panels B, C, D). Sample sizes: 9,304 in panels A and B (groups whose linguistic distance increases account for 42 percent of the sample); 5,166 in panel C (groups whose linguistic distance increases account for 74 percent of the sample); 5,277 in panel D (groups whose linguistic distance decreases account for 74 percent of the sample). Bold blue vertical bars and thin black vertical bars denote 90 and 95 confidence intervals respectively. Standard errors are clustered at the ethnicity-by-country level.

**Political representation.** My results could be confounded by ethnic groups entering and exiting power. An ethnicity's cultural distance to the government mechanically decreases whenever one or more of its members enter a coalition. This could prevent an ethnic group from rebelling for reasons not necessarily related to cultural closeness (e.g., the desire not to hurt the newly-elected group

members). Conversely, an increase in cultural distance might coincide with an ethnicity losing power. This could trigger conflict not much because of cultural differences, but due to the desire to regain the just-lost power position. To delve into the role of this confounder, in Table 2 I run a horse-race between cultural distance and a binary variable that equals 1 if a potential rebel is represented in the government coalition and zero otherwise. As expected, this indicator is negatively correlated with conflict: an ethnicity is less likely to rebel when holding a position in the government coalition (column 1). However, while the magnitude of the cultural distance estimates slightly decreases when controlling for this indicator, the coefficient on linguistic distance remains large and significant (see columns 2-3 and columns 4-5). Taken together, these results indicate that the effect of cultural distance on conflict occurs over and above the effect of political representation.

**Income differences.** Another factor that may potentially confound my results is income distance between groups. If income distance is systematically correlated with linguistic distance, my estimates could pick up the effect of economic grievances and relative deprivation, factors that the literature has linked to conflict in my setting (Guariso and Rogall, 2017). To rule out this possibility, I employ two exogenous measures of income differences between groups introduced by previous literature: distance in rainfall during the crop-growing season (Guariso and Rogall, 2017) and distance in the presence of oil fields (Morelli and Rohner, 2015).[26] I split the income distance measures into two continuous components: one capturing instances where the potential rebel holds *more* income than the groups in power and one capturing cases where the potential rebel holds *less* income. In Table 3, I start by replicating the association between income differences and conflict uncovered by Guariso and Rogall (2017): whenever potential rebels are relatively poorer than the groups in power, they become more likely to rebel (columns 1 and 4). However, I do not find that the effect of linguistic distance dissipates once

[26]See Appendix C-1 for more details on the income measures. Note that this exercise is not intended to capture the role of ethnic favouritism (i.e., income *changes* resulting from discrimination in the allocation of public resources), but of exogenous or pre-existing income differences. I delve into the role of ethnic favouritism as a potential mechanism in section 7.

TABLE 2: *Cultural Distance, Ethnic Civil Conflict and Political Representation*

| | Dependent variable: Ethnic conflict over power | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Linguistic distance$^{W}$ | | 0.086*** | 0.055** | | |
| | | (0.019) | (0.021) | | |
| Linguistic distance$^{UW}$ | | | | 0.076*** | 0.050** |
| | | | | (0.018) | (0.023) |
| In government | -0.110*** | | -0.092** | | -0.093** |
| | (0.041) | | (0.041) | | (0.043) |
| Mean of dep. var. | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 |
| Country-year FE | yes | yes | yes | yes | yes |
| Ethnicity-country FE | yes | yes | yes | yes | yes |
| Ethnicity-country trend | yes | yes | yes | yes | yes |
| Geographic controls | yes | yes | yes | yes | yes |
| Observations | 9,990 | 9,990 | 9,990 | 9,990 | 9,990 |
| Adjusted R-squared | 0.509 | 0.507 | 0.511 | 0.507 | 0.511 |

*Notes*: The unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government for gaining power. *In government* is equal to 1 if an ethnicity is represented in the government coalition in a certain year and 0 otherwise. For a description of all explanatory variables, refer to notes in Table 1. The sample includes 44 African countries, 57 years (1961-2017), and 265 distinct ethnic groups. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

controlling for these measures of income differences (see columns 2-3 and columns 5-6). This indicates that the effect of linguistic distance is orthogonal to that of income distance between groups.

### *5.3 Other Robustness Tests*

Table A-2 shows that results are robust to a tighter specification that includes ethnicity-by-year fixed effects and only exploits variation coming from ethnic groups partitioned across countries (columns 1-2).[27] Columns 3 to 5 show that the significance of the coefficients is not affected by the choice of clustering of standard

[27]Section B-1 in the Appendix describes this alternative estimation strategy in detail.

TABLE 3: *Cultural Distance, Income Distance, and Ethnic Civil Conflict*

| Income measure: | Dependent variable: Ethnic conflict over power | | | | | |
|---|---|---|---|---|---|---|
| | *Rainfall in growing season Guariso & Rogall (2021)* | | | *Oil fields Morelli & Rohner (2015)* | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Linguistic distance | | 0.079*** | 0.098*** | | 0.061*** | 0.073*** |
| | | (0.016) | (0.019) | | (0.015) | (0.017) |
| More income | 0.015 | 0.006 | 0.011 | 0.036 | 0.034 | 0.034 |
| | (0.010) | (0.009) | (0.009) | (0.031) | (0.030) | (0.030) |
| Less income | 0.022* | 0.010 | 0.015 | 0.369* | 0.323 | 0.336 |
| | (0.013) | (0.011) | (0.012) | (0.208) | (0.213) | (0.214) |
| Mean of dep. var. | 0.051 | 0.051 | 0.051 | 0.048 | 0.048 | 0.048 |
| Country-year FE | yes | yes | yes | yes | yes | yes |
| Ethn.-country FE | yes | yes | yes | yes | yes | yes |
| Ethn.-country trend | yes | yes | yes | yes | yes | yes |
| Geographic controls | | | yes | | | yes |
| Observations | 8,947 | 8,947 | 8,947 | 9,990 | 9,990 | 9,990 |
| Adjusted R-squared | 0.497 | 0.501 | 0.502 | 0.509 | 0.511 | 0.512 |

*Notes*: For a description of all explanatory variables, refer to notes in Table 1. The sample in columns 1-3 includes 44 African countries, 57 years (1961-2017), and 238 ethnic groups. The sample in columns 4-6 includes 44 African countries, 57 years (1961-2017), and 265 ethnic groups. See Appendix C-1 for details about the income measures. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

errors. Results also hold when adding climatic controls to the baseline specification (column 6).[28] Columns 7 to 9 restrict the sample to countries that experienced at least one conflict between 1961 and 2017, to those that experienced at least one power transition, and to a balanced panel of ethnicities. Finally, column 10 shows robustness to altering the $\lambda$ parameter in the linguistic distance computation.[29]

[28]I include a set of differences in precipitation and mean temperature and time-variant climatic variables (temperature and precipitation) specific to each potential rebel. These account for the potential confounding effect of climatic shocks, which are important determinants of conflict in the African continent according to Harari and La Ferrara (2018).

[29]I conduct a series of additional robustness tests. Section B-2 shows that results hold when employing an alternative linguistic distance measure that relies on Francois et al.'s (2015) data on African cabinet ministers, only available for a subsample of 15 countries. In Figure A-2, I check for outliers by re-running my estimations dropping one country at a time. Table A-3

## 6 Instrumental Variable Approach

A remaining potential concern with my strategy is that the measure of cultural distance might be endogenous to conflict. If there are omitted factors that affect linguistic distance between groups and contemporary conflict, then the estimates presented in the previous section are biased. An example is historical conflict. Old antagonisms between groups might have affected linguistic distance and, at the same time, can have a direct effect on conflict today.[30] Since not all members of an ethnic group speak the same language[31], a potential source of endogeneity could impact the relative importance of a language within an ethnicity's population.

The direction of the potential bias is ex-ante unclear. For instance, suppose past conflict between two ethnicities is positively associated with current conflict. A history of violence might also have reduced two ethnicities' linguistic distance over time, e.g., due to genocides eliminating linguistically distant subgroups. If this was the case, then the main coefficients on conflict over government power would be biased downwards. On the other hand, if past conflict widened linguistic distance by reinforcing linguistic enclaves or ethnic boundaries, then my coefficients would be biased upwards.

### *6.1 The Instrument*

To instrument for cultural distance, I exploit the Bantu expansion, a natural experiment unique to African history. 5,000 years ago, a climatic shock generated a temporary loss of rainforest in Central Africa. Through increased seasonality of the monsoon, a lowering of the sea surface temperature in the Guinean Gulf, and

shows that the main coefficient remains positive and significant when controlling for a binary variable that is equal to 1 if a potential rebel was involved in a conflict over the previous year and zero otherwise. Table A-3 also shows that results hold when employing a lagged measure of linguistic distance and when controlling for a lead measure of linguistic distance. The small and insignificant coefficient on the latter rules out the presence of anticipatory effects, consistently with what illustrated in Figure 1.

[30]For evidence on how past conflict is correlated with current conflict in Africa, see Besley and Reynal-Querol (2014).

[31]Recall that the linguistic distance measure is a weighted average of all pairwise linguistic distances, where weights reflect the percentage of group members speaking each language.

less rainfall, the shock favored the emergence of savanna corridors (Bostoen et al., 2015). After the climatic crisis ended, the new savanna environments disappeared, supplanted by rainforest.

Archaeological, anthropological, and linguistic evidence suggests that the temporary opening up of parts of the—beforehand impenetrable—rainforest facilitated the expansion of the Bantu, a tribe residing in Cameroon, throughout the subcontinent (see Bostoen et al. (2013) and Bostoen et al. (2015) for a summary of the interdisciplinary literature). The exact reason for why Bantu people started migrating is largely debated. One hypothesis is that the climate-driven opening of the forest gave hunters access to "naive" animal populations that were previously trapped in the forest and then became suddenly available (Bostoen et al., 2013). However, there is by now consensus among scholars on how the *path* of the Bantu migration was exogenously shaped by climatic events, a crucial feature for my instrumental variables analysis (Grollemund et al., 2015; Bostoen et al., 2015).

How is this massive prehistoric migration related to contemporary cultural distances between ethnic groups? The Bantu expansion has been defined as the most important linguistic, cultural and demographic process in Late Holocene Africa (Robbeets and Savelyev, 2017). Bantu farmers, in gradually expanding from their homeland in Northwestern region of Cameroon, spread their new language and culture assimilating or displacing earlier inhabitants of the regions they crossed, i.e., Pygmy and Khoisan hunter-gatherers (Diamond, 1997). Today, one out of three Africans is fluent in at least one of the approximately 500 existing languages belonging to the Bantu family. Crucially, Bantu people did not move and settle everywhere, and this is exactly the idea behind my instrument. Within today's countries, some territories happened to be crossed by the route of the Bantu expansion, while others did not. This is demonstrated by the existence of pre-Bantu hunter-gatherer populations that remain today culturally and linguistically distinct such as the Pygmies in Central Africa, the Hadza in Tanzania, and Khoisan people in southern Africa.

I exploit this unique event in African prehistory to construct a novel instrument

for cultural distance between ethnic groups. Groups whose homelands were highly exposed to the Bantu migration route and have inherited Bantu culture should be culturally distant to those that remained unaffected. Instead, groups with a similarly high or low exposure to the Bantu expansion should be culturally close to each other, because they either both inherited Bantu culture, or kept their pre-existing one.[32] Based on this idea, I construct a Bantu index ranging between 0 and 1, which captures the extent to which a group's ancestral homeland was exposed to the route of the Bantu expansion. Section B-3 in the Appendix reports a detailed description of how I construct this index. I then use the absolute difference in this index as an instrument for cultural distance. Figure 2 displays the geographic variation of the index.[33]

### *6.2 Empirical Strategy*

I use a two-stage least-square (2SLS) procedure to estimate equation 3. In the first stage, I estimate the effect of the distance in the Bantu index ($\text{BD}_{rct}$) between potential rebels and governments on linguistic distance:

$$\text{LD}_{rct} = \lambda_{ct} + \zeta_{rc} + \theta_{rc}t + \beta\text{BD}_{rct} + \gamma\text{G}_{rct} + u_{rct} \tag{6}$$

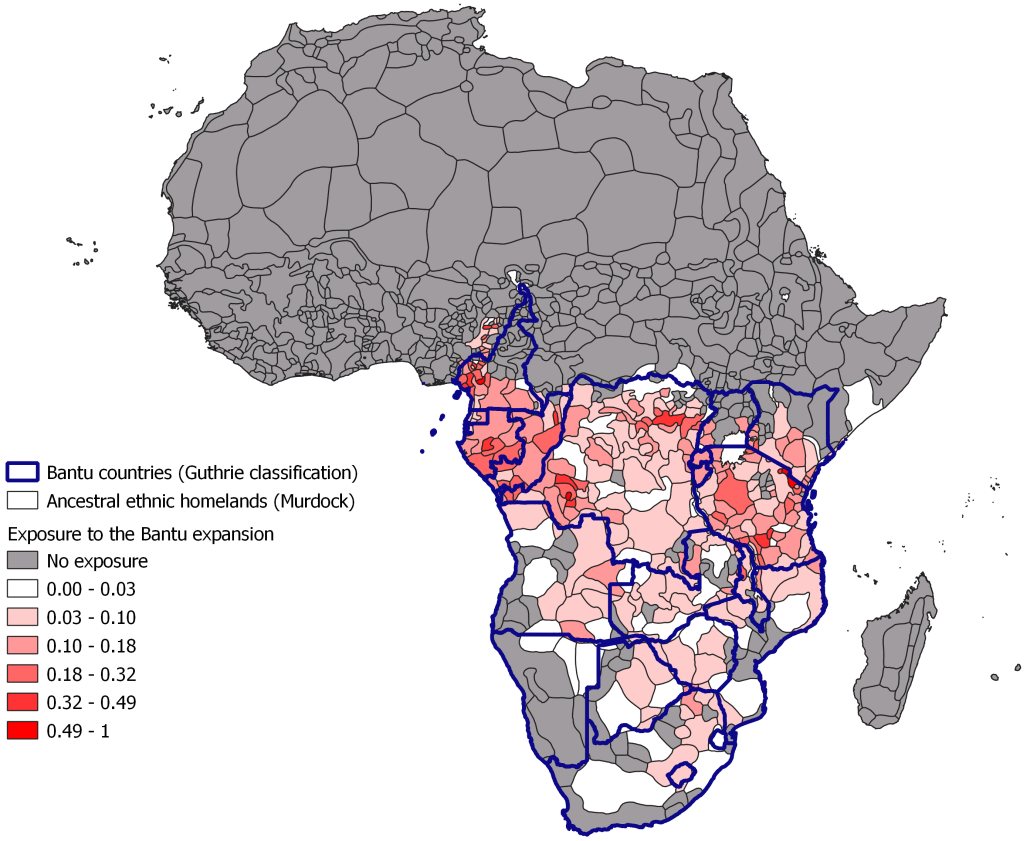
I run the IV analysis for the subset of 17 sub-Saharan countries where the Bantu tribe migrated, and that are classified as Bantu countries according to the Guthrie classification of languages (see notes in Figure 2). Table C-3 in the Appendix reports summary statistics for this subsample of 95 ethnic groups (N=3,775).[34]

[32]The assumption here is that, *within contemporary countries*, pre-Bantu groups were on average culturally similar to each other, and culturally dissimilar to the Bantu. This is partly supported by archaeological evidence showing that pre-Bantu societies shared the same mode of subsistence, i.e., hunting and gathering, while Bantu people were predominantly farmers (Diamond, 1997).

[33]Blouin (2021) exploits the Bantu migration to test Diamond's (1997) axis-orientation hypothesis and distinguishes between the South and the East Bantu migration. Among other findings, he shows that descendants of southern migrants are more geographically and culturally isolated than descendants of eastern migrants. Since my analysis exploits variation *across ethnicities within countries*, Blouin's (2021) findings do not pose a threat to the validity of my instrument.

[34]The identifying variation stems from 40 government changes. In the Bantu region, the prevalence of conflict over power is 0.05 and the average linguistic distance between potential

FIGURE 2: *The Bantu Index and Countries in the Bantu Region*




*Notes*: The figure displays the index of exposure to the Bantu expansion based on the historical migration route reconstructed by Grollemund et al. (2015), as well as contemporaneous country borders of the 17 countries classified as belonging to the Bantu region according to the Guthrie classification: Angola, Botswana, Burundi, Cameroon, Congo, Congo DRC, Equatorial Guinea, Gabon, Kenya, Malawi, Mozambique, Rwanda, South Africa, Tanzania, Uganda, Zambia, Zimbabwe.

### *6.3 Results*

Table 4 reports results from the instrumental variables analysis, which confirm the linear probability model estimates. In Panel A, columns 1 and 2 report OLS estimates for the sub-sample of ethnicities in the Bantu region. Results are consistent with the estimates in the full African continent. Panel C reports the first-stage results of equation 6 and shows how a larger difference in the exposure to the

rebels and governments is 0.31.

Bantu expansion is positively associated with cultural distance, measured through linguistic distance.[35] In Panel B, I find that there is a positive reduced-form effect of Bantu distance on conflict over government power, which is reassuring for the validity of the instrument. A one standard deviation increase in Bantu distance increases conflict prevalence by 6.1-6.9 percentage points.

Columns 3 and 4 in Panel A report the 2SLS result, which are consistent with the OLS estimates of columns 1 and 2. In all specifications, both for the weighted and unweighted measures of cultural distance, the magnitude of the coefficients increases significantly compared to the OLS specification, indicating that the original coefficients were biased downwards. This is consistent, for instance, with past conflict between two ethnicities being positively associated with contemporary conflict, and, at the same time, reducing two groups' linguistic distance (e.g., through genocides eliminating linguistically distant subgroups). Section B-3.1 in the Appendix discusses potential violations of the exclusion restriction and provides an empirical test supporting the validity of instrument.

## 7 Channels

Why are ethnic groups more likely to fight over power when the government is culturally distant from them? In this section, I find evidence consistent with conflict arising due to cultural disagreements over both the *allocation* and the *type* of public goods.

**Cultural distance and conflict over territory.** The finding that cultural distance triggers conflict fought over the control of the central government is consistent with Spolaore and Wacziarg's (2017) and Esteban and Ray's (2011) theoretical frameworks. As described in Section 2, a key prediction of these models is that, if cultural distance affects conflict by generating diverging preferences over public policies that everyone in a country must share, then we should not expect

[35] A recent contribution by Lee et al. (2022) proposes a new test to assess the validity of the second-stage t-ratio inference when the first-stage F statistic is smaller than 104.7, as in this case. Table A-4 in the Appendix reports second-stage results with standard errors based on Lee et al.'s (2022) $tF$ procedure and shows that inference considerations remain unchanged.

TABLE 4: *IV: Bantu Exposure, Cultural Distance and Ethnic Civil Conflict*

| *Panel A:* | Dependent variable: Ethnic conflict over power | | | |
|---|---|---|---|---|
| | *OLS* | | *2SLS* | |
| | (1) | (2) | (3) | (4) |
| Linguistic distance | 0.073*** | 0.076*** | 0.200*** | 0.149*** |
| | (0.013) | (0.009) | (0.037) | (0.029) |
| Mean of dep. var | 0.050 | 0.050 | 0.050 | 0.050 |
| *Panel B:* | | | *Reduced form* | |
| Bantu distance | | | 0.069*** | 0.061*** |
| | | | (0.016) | (0.018) |
| Adjusted R-squared | | | 0.534 | 0.533 |
| *Panel C:* | | | *First stage* | |
| | | | Linguistic distance | |
| Bantu distance | | | 0.346*** | 0.405*** |
| | | | (0.084) | (0.070) |
| Kleibergen-Paap F-statistic | | | 12.75 | 25.47 |
| Distance type | w | uw | w | uw |
| Country-year FE | yes | yes | yes | yes |
| Ethnicity-country FE | yes | yes | yes | yes |
| Ethnicity-country trends | yes | yes | yes | yes |
| Geographic controls | yes | yes | yes | yes |
| Observations | 3,775 | 3,775 | 3,775 | 3,775 |

*Notes*: the unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting for government power and 0 otherwise. *Linguistic distance* captures linguistic distance between each potential rebel and the ethnic groups at the government, either weighted or unweighted. *Bantu distance* denotes the absolute difference in the exposure to the Bantu expansion between potential rebels and the government. Linguistic distance and Bantu distance are standardized. For a description of geographic controls, refer to the notes in Table 1. The sample includes 17 African countries in the Bantu region, 57 years (1961-2017), and 95 distinct ethnic groups. Two-way clustered standard errors by year and country are reported in parenthesis, and are adjusted for the low number of clusters using the number of countries. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

cultural distance to generate conflict over rival goods such as territory and resources. This prediction allows me to take a first step in empirically assessing the role of diverging preferences as a mechanism. I re-run my main specification and the IV design replacing the dependent variable with an indicator that equals 1 if an ethnic group fights over territory and zero otherwise.[36] As shown in Table 5, cultural distance is not associated with conflict over territory. If anything, the relationship reverses, suggesting that disputes over rival goods are more likely to occur among similar populations, as shown by Spolaore and Wacziarg (2016b).

TABLE 5: *Cultural Distance and Ethnic Civil Conflict over Territory*

| | Dep. var.: ethnic conflict over territory | | | |
|---|---|---|---|---|
| | *Baseline* | | *IV* | |
| | (1) | (2) | (3) | (4) |
| Linguistic distance | -0.015 | -0.021 | -0.008*** | -0.010*** |
| | (0.018) | (0.022) | (0.003) | (0.003) |
| Mean of dep. var. | 0.024 | 0.024 | 0.008 | 0.008 |
| Distance type | w | uw | w | uw |
| Country-year FE | yes | yes | yes | yes |
| Ethnicity-country FE | yes | yes | yes | yes |
| Ethnicity-country trend | yes | yes | yes | yes |
| Geographic controls | yes | yes | yes | yes |
| Observations | 9,990 | 9,990 | 3,775 | 3,775 |

*Notes*: For a description of all explanatory variables, refer to notes in Table 1. The sample includes 44 African countries, 57 years (1961-2017), and 265 ethnic groups (columns 1 and 2) and 17 African countries, 57 years (1961-2017), and 95 ethnic groups (columns 3 and 4). Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

**Cultural distance and dissatisfaction with government performance.** As a next step, I turn to survey data to more directly test whether, as existing theory and my results so far suggest, linguistic distance to the groups forming the central government is associated with disagreement over the mix of public policies that

[36]See Appendix C-1 for details on this variable.

the government provides. I use 7 waves of the Afrobarometer Survey, where respondents are asked to elicit their opinions about the performance of the current government in various policy domains including the economy, education, health, infrastructure, minorities, national security, and other social issues. Using PCA, I group these opinions into a single index ranging between 0-1, with 1 denoting the highest degree of disagreement.[37] I manually link respondents' self-reported linguistic affiliation to the Ethnologue and compute linguistic distance to the current government as described in Section 3.

To establish whether cultural distance to the government comes with higher disagreement over public policies, I run the following specification:

$$\mathrm{GP}_{irct} = \lambda_{ct} + \zeta_{rc} + \theta_{rc}t + \beta \mathrm{LD}_{rct} + \gamma \mathrm{G}_{rct} + \Phi \mathrm{I}_i + \epsilon_{irct} \tag{7}$$

where $\mathrm{GP}_{irct}$ is a continuous variable that ranges between 0 and 1 and captures the extent to which individual $i$ belonging to ethnic group $r$ thinks that the government of country $c$ in year $t$ is performing fairly badly or very badly. $\mathrm{LD}_{rct}$ indicates linguistic distance between individual $i$'s ethnic group $r$ and the government in country $c$ and year $t$. In addition to the set of fixed effects and ethnicity-level controls included in the baseline specification (see equation 3), I also add a set of individual-level controls $\mathrm{I}_i$ (a dummy for female, age, age squared, and a dummy for urban residents).

Table 6 shows that respondents are more likely to disagree with a wide range of public policies implemented by the government when they are more linguistically distant to the ethnic groups in power. A one standard deviation increase in an individual's linguistic distance to the government increases the degree of opposition to government policies by 6-9 percentage points, which corresponds to a 10-15 percent effect compared to the average degree of disagreement in the sample.[38]

[37] Table A-6 reports the results of the PCA for each survey round. The first principal component explains 34-61% of the common variance of the domains. Each domain always loads positively on the first principal component, suggesting that individuals who dislike government performance tend to do so along the whole range of listed public policies.

[38] Table A-5 shows that results are similar when constructing the dependent variable by taking

TABLE 6: *Cultural Distance and Dissatisfaction with Government Performance*

| | Dependent variable: Thinks that the government is performing badly | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Linguistic distance | 0.090*** | 0.090*** | 0.060*** |
| | (0.024) | (0.024) | (0.023) |
| Mean of dep. var. | 0.583 | 0.583 | 0.583 |
| Country-year FE | yes | yes | yes |
| Ethnicity-country FE | yes | yes | yes |
| Ethnicity-country trends | yes | yes | yes |
| Individual controls | | yes | yes |
| Geographic controls | | | yes |
| Observations | 117,012 | 117,012 | 117,012 |
| Adjusted R-squared | 0.175 | 0.176 | 0.176 |

*Notes*: The unit of observation is an individual belonging to one of 904 distinct ethnolinguistic groups in 28 African countries interviewed between 1999 and 2018 in 7 rounds of the Afrobarometer survey. The dependent variable is a binary variable capturing the extent to which the person thinks the government is performing very badly or fairly badly in handling a variety of policy matters (see Table A-6 for a complete list and Appendix C-1 for details). *Linguistic distance* captures the weighted linguistic distance between a respondent and the ethnic groups in power and is standardized. For a description of geographic controls, refer to the notes in Table 1. Individual controls include age, age squared, a dummy for female and a dummy for residence in a rural area. Standard errors clustered at the ethnicity level are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

**Cultural distance and differences in preferences over policy.** While informative about the presence of dissatisfaction with the government whenever cultural differences are more pronounced, the just-shown results could be picking up hatred or animosity towards groups in power, and not just diverging preferences. In a next step, I therefore seek to establish whether linguistically distant groups hold different views about which public policies should be implemented, irrespective of who is in power at a given point in time.

I construct an index at the ethnic group level proxying preferences over public goods. The index is based on group members' views about the most important

the average disagreement across all policy domains instead of using PCA.

public matter that the government should address among health, infrastructure, services, food security, governance, the economy, and others. For each ethnic group, I compute the share of respondents mentioning a given domain as the most important one and then group these shares into a single 0-1 index using PCA.[39] I then generate a dyadic dataset in which I pair every ethnic group to all other groups surveyed in year $t$ and residing in the same country $c$. I compute linguistic distance ($\text{LD}_{ij}$) and the absolute distance in the preferences index ($\text{PD}_{ijct}$) between groups $i$ and $j$ forming each pair. To assess whether linguistic distance generates differences in preferences over policy, I estimate the following gravity specification:

$$\text{PD}_{ijct} = \zeta_{ict} + \phi_{jct} + \alpha \text{LD}_{ij} + \gamma \text{G}_{ij} + \epsilon_{ijct} \tag{8}$$

$\zeta_{ict}$ and $\phi_{jct}$ denote country- and year-specific fixed effects for ethnicity $i$ and $j$ in the pair, respectively.[40] $\text{G}_{ij}$ controls for the same set of geographical differences between ethnic groups adopted in the baseline specification. Table 7 displays the coefficients estimates of $\alpha$. Results corroborate what the literature on diversity and conflict has often assumed but never directly tested empirically: the larger the linguistic distance between a pair of ethnicities, the more likely they are to differ in their preferences over public policies. In column 4, a one standard deviation increase in linguistic distance increases distance in preferences by 0.017, equivalent to a 8.8 percent effect compared to the average distance in preferences between dyads in the sample.

**Cultural distance and preferences over the allocation and type of public goods.** Having established that culturally distant groups hold diverging preferences over public policies, what is left to understand is whether what triggers conflict are differences in preferences over the type of public goods to be provided, their allocation, or both. Established work has documented the presence

[39] Table A-7 reports the results of the PCA for each survey round. The first principal component explains 20-23% of the common variance.

[40] Alternatively, I include country-specific ethnicity $i$ and $j$ fixed effects and country-by-year fixed effects. These set of country- and year-specific fixed effects allow me to isolate the role of linguistic distance between group $i$ and group $j$ from the potentially confounding effect of a specific government being in power at the time of the survey.

TABLE 7: *Cultural Distance and Preferences over Policy*

| | Dep. var.: Distance in preferences over policy | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Linguistic distance | 0.019*** | 0.015*** | 0.020*** | 0.017*** |
| | (0.005) | (0.005) | (0.005) | (0.005) |
| Mean of dep. var. | 0.194 | 0.194 | 0.194 | 0.194 |
| Country-year FE | yes | yes | | |
| Ethnicity $i$-country FE | yes | yes | | |
| Ethnicity $j$-country FE | yes | yes | | |
| Ethnicity $i$-country-year FE | | | yes | yes |
| Ethnicity $j$-country-year FE | | | yes | yes |
| Geographic distances | | yes | | yes |
| Observations | 11,239 | 11,239 | 11,239 | 11,239 |
| Adjusted R-squared | 0.523 | 0.524 | 0.603 | 0.605 |

*Notes*: The unit of observation is a pair of ethnic groups in a country and year. The sample includes 662 distinct ethnic groups. The dependent variable is the absolute difference between ethnicity $i$ and ethnicity $j$ in a 0-1 index capturing group-level preferences on the most important problems the government should address. The index is constructed through principal component analysis (see Table A-7 for details). *Linguistic distance* is standardized. Standard errors clustered two-way at the ethnicity $i$-country and ethnicity $j$-country level are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

of ethnic favouritism in Africa (Burgess et al., 2015; De Luca et al., 2018), which extends to non-coethnic but linguistically similar groups (Dickens, 2018). Caselli and Coleman (2013) theorize that language is an ethnic marker that facilitates discrimination in the allocation of public goods and resources and that this may, in turn, generate conflict over the control of these resources. Given this literature, I aim to understand whether conflict over government power is triggered solely by discontent with the unequal *allocation* of public resources (i.e., ethnic favouritism), or whether diverging preferences over the *type* of public good provided also play a role.

To shed light on this, I exploit the V-Dem dataset, which provides information at the country-year level on (i) whether the national budget is mostly spent on

"private" (i.e., particularistic) or "public" goods and (ii) whether public services are equally distributed across social groups or not.[41] If only disagreement over the allocation of public resources was driving the effects, then linguistic distance should not trigger conflict in settings where the national budget is mostly spent on equally-distributed public goods. To test this conjecture, I estimate the following:

$$\text{Conflict}_{rct} = \lambda_{ct} + \zeta_{rc} + \theta_{rc}t + \beta \text{LD}_{rct} + \rho \text{VDEM}_{ct} \times \text{LD}_{rct} + \gamma \text{G}_{rct} + \epsilon_{rct} \quad (9)$$

where $\text{VDEM}_{ct}$ is an indicator that equals 1 if, in a given country $c$ and year $t$, the national budget is mostly spent on public goods, and 0 if it is mostly spent on particularistic goods. Alternatively, $\text{VDEM}_{ct}$ equals 1 if public services are equally distributed across social groups and zero otherwise. Table 8 reports the results. The negative estimates of $\rho$ in columns 2 and 3 demonstrate that the reaction to cultural distance is smaller if the national budget is mostly spent on public goods and if the latter are equally distributed across social groups.[42] Hence, favouritism in the allocation of public resources seems to explain part of the effects I uncover. Yet, the reaction to cultural distance is still present in settings where ethnic favouritism is less prevalent: when restricting the sample to countries and years in which the two just-described V-Dem indicators both equal 1, I find a smaller but still positive and significant effect of linguistic distance on conflict.[43] I take this as evidence highlighting the role of cultural divergences over the preferred type of public goods in explaining ethnic conflict over government power.

## 8 Conclusion

Conflict—and in particular, civil conflict fought along ethnic lines—is more prevalent in ethnolinguistically diverse societies. Yet, despite facing the same aggregate level of ethnolinguistic diversity, some ethnic groups rebel and others do not. This

[41] See details in Appendix C-1 on how these variables are constructed and for V-Dem's definition of public vs particularistic goods.

[42] Notice, however, that the former estimate is not statistically significant at conventional levels.

[43] In the baseline specification in column 1, a one standard deviation increase in linguistic distance increases conflict over power by 0.49 standard deviations; in column 4, a one standard deviation increase in linguistic distance increases conflict over power by 0.33 standard deviations.

TABLE 8: *Disagreement over the Allocation and Type of Public Goods*

| | Conflict over government power | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | | | | Countries and years with public goods & equal distrib. |
| Linguistic distance | 0.087*** | 0.097*** | 0.103*** | 0.027** |
| | (0.019) | (0.020) | (0.018) | (0.012) |
| Linguistic distance × Public goods | | -0.023 | | |
| | | (0.014) | | |
| Linguistic distance × Equal distribution | | | -0.068*** | |
| | | | (0.020) | |
| Mean of dep. var. | 0.049 | 0.049 | 0.049 | 0.006 |
| Country-year FE | yes | yes | yes | yes |
| Ethnicity-country FE | yes | yes | yes | yes |
| Ethnicity-country trend | yes | yes | yes | yes |
| Geographic controls | yes | yes | yes | yes |
| Observations | 9,694 | 9,694 | 9,694 | 2,786 |
| Adjusted R-squared | 0.514 | 0.514 | 0.514 | 0.472 |

*Notes*: The unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government for gaining power and 0 otherwise. *Linguistic distance* captures weighted linguistic distance between a potential rebel and the ethnic groups at the government. The sample includes 43 African countries, 57 years (1961-2017), and 256 ethnic groups. *Public goods* and *Equal distribution* are indicator variables at the country-year level taken from V-Dem (see Appendix C-1 for details) capturing whether the national budget is mostly spent on public goods and whether these are equally distributed across social groups, respectively. The sample in column 4 includes 18 countries and 95 ethnic groups in which the variables *Public goods* and *Equal distribution* take a value of 1. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

paper sheds light on why this is the case by uncovering that a group's involvement in conflict increases with its cultural distance to the ethnic groups in power.

My paper relies on a key innovation: moving the analysis of diversity and conflict from the country level to the ethnicity level. Besides unpacking the country-level associations and uncovering which groups are more likely to rebel at a given point in time, the analysis at the ethnicity level allows me to delve into the mechanisms through which cultural distance generates conflict over government power. I document that the reaction to cultural distance occurs over and above the effect of

exclusion from power or income differences, and that ethnic favouritism explains only part of the effect. Culturally distant groups disagree over both the allocation and the type of public goods, and thus rebel to wrest control of the central government. Identifying which groups are more likely to rebel, when, and why they do so is essential for effectively targeting conflict-prevention efforts. The results in this paper suggest that investing public resources in public goods and ensuring that they are equally distributed across social groups can attenuate the effect of cultural distance on conflict, but might not be enough: in multicultural societies, providing a mix of public goods accommodating diverging preferences may also be key.

This paper focuses on ethnic civil conflict and, as a result, on a group's cultural distance to the ethnic groups in power. Civil conflict, while largely prevalent, is only one type of the many possible manifestations of inter-group violence. A promising avenue for future research is to test the role of cultural distance in explaining non-civil conflict, i.e., inter-group violence not involving the government, so far largely overlooked by the literature.[44] Another promising area for future research is improving the measurement of cultural distance. Linguistic distance is by now a well-established measure, but it does not allow to test which exact dimension of culture triggers conflict. Efforts to shed light on this issue are underway,[45] but exploring new metrics that allow to separately examine different cultural components is a promising area for future lines of inquiry.

[44]A notable exception is the work by Depetris-Chauvin and Özak (2020).

[45]See, for example, Guarnieri and Tur-Prats (2023), who focus on distance in gender norms and investigate its impact on the intensive margin of violence.

## References


**Afrobarometer Data**, 1999-2017.

**Alesina, Alberto and Eliana La Ferrara**, "Ethnic diversity and economic performance," *Journal of economic literature*, 2005, *43* (3), 762–800.

**Allansson, Marie, Erik Melander, and Lotta Themnér**, "Organized violence, 1989–2016," *Journal of Peace Research*, 2017, *54* (4), 574–587.

**Amodio, Francesco and Giorgio Chiovelli**, "Ethnicity and violence during democratic transitions: Evidence from south africa," *Journal of the European Economic Association*, 2018, *16* (4), 1234–1280.

**Arbatlı, Cemal Eren, Quamrul H Ashraf, Oded Galor, and Marc Klemp**, "Diversity and conflict," *Econometrica*, 2020, *88* (2), 727–797.

**Ashraf, Quamrul and Oded Galor**, "Genetic diversity and the origins of cultural fragmentation," *American Economic Review*, 2013, *103* (3), 528–33.

**Besley, Timothy and Marta Reynal-Querol**, "The legacy of historical conflict: Evidence from Africa," *American Political Science Review*, 2014, pp. 319–336.

**Blattman, Christopher and Edward Miguel**, "Civil war," *Journal of Economic Literature*, 2010, *48* (1), 3–57.

**Blouin, Arthur**, "Axis-orientation and knowledge transmission: evidence from the Bantu expansion," *Journal of Economic Growth*, 2021, *26* (4), 359–384.

**Bostoen, Koen, Bernard Clist, Charles Doumenge, Rebecca Grollemund, Jean-Marie Hombert, Joseph Koni Muluwa, Jean Maley, Roger Blench, Pierpaolo Di Carlo, Jeff Good et al.**, "Middle to late Holocene Paleoclimatic change and the early Bantu expansion in the rain forests of Western Central Africa," *Current Anthropology*, 2015, *56* (3), 367–368.

**_ , Rebecca Grollemund, and Joseph Koni Muluwa**, "Climate-induced vegetation dynamics and the Bantu Expansion: Evidence from Bantu names for pioneer trees (Elaeis guineensis, Canarium schweinfurthii, and Musanga cecropioides)," *Comptes Rendus Geoscience*, 2013, *345* (7-8), 336–349.

**Burgess, Robin, Remi Jedwab, Edward Miguel, Ameet Morjaria, and Gerard Padró i Miquel**, "The value of democracy: evidence from road building in Kenya," *American Economic Review*, 2015, *105* (6), 1817–1851.

**Caselli, Francesco and Wilbur John Coleman**, "On the theory of ethnic conflict," *Journal of the European Economic Association*, 2013, *11* (suppl_1), 161–192.

**Coppedge, Michael and Gerring, John and Knutsen, Carl Henrik and Lindberg, Staffan I. and others** , "V-Dem Dataset v13," 2023.

**Depetris-Chauvin, Emilio and Ömer Özak**, "Borderline Disorder:(De facto) Historical Ethnic Borders and Contemporary Conflict in Africa," 2020.

**Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg**, "The political economy of linguistic cleavages," *Journal of development Economics*, 2012, *97* (2), 322–338.

**_ , Shlomo Weber, and Ignacio Ortuño-Ortín**, "Linguistic diversity and redistribution," *Journal of the European Economic Association*, 2009, *7* (6), 1291–1318.

**Diamond, Jared**, "Guns, Germs, and Steel: The Fates of Human Societies," 1997.

**Dickens, Andrew**, "Ethnolinguistic favoritism in african politics," *American Economic Journal: Applied Economics*, 2018, *10* (3), 370–402.

_ , "Understanding Ethnic Differences: The Roles of Geography and Trade," 2020.

**Esteban, Joan and Debraj Ray**, "Linking conflict to inequality and polarization," *American Economic Review*, 2011, *101* (4), 1345–74.

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

**Fearon, James D and David D Laitin**, "Ethnicity, insurgency, and civil war," *American political science review*, 2003, *97* (1), 75–90.

**Fick, Stephen E and Robert J Hijmans**, "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas," *International journal of climatology*, 2017, *37* (12), 4302–4315.

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

**Galor, Oded and Ömer Özak**, "The agricultural origins of time preference," *American Economic Review*, 2016, *106* (10), 3064–3103.

**_ , _ , and Assaf Sarid**, "Geographical roots of the coevolution of cultural and linguistic traits," *Available at SSRN 3284239*, 2018.

**Gomes, Joseph Flavian**, "The health costs of ethnic distance: evidence from sub-Saharan Africa," *Journal of Economic Growth*, 2020, *25* (2), 195–226.

**Grollemund, Rebecca, Simon Branford, Koen Bostoen, Andrew Meade, Chris Venditti, and Mark Pagel**, "Bantu expansion shows that habitat alters the route and pace of human dispersals," *Proceedings of the National Academy of Sciences*, 2015, *112* (43), 13296–13301.

**Guariso, Andrea and Thorsten Rogall**, "Rainfall inequality, political power, and ethnic conflict in Africa," 2017.

**Guarnieri, Eleonora and Ana Tur-Prats**, "Cultural distance and conflict-related sexual violence," *The Quarterly Journal of Economics*, 2023, *138* (3), 1817–1861.

**Harari, Mariaflavia and Eliana La Ferrara**, "Conflict, climate, and cells: a disaggregated analysis," *Review of Economics and Statistics*, 2018, *100* (4), 594–608.

**Harbom, Lotta, Erik Melander, and Peter Wallensteen**, "Dyadic dimensions of armed conflict, 1946—2007," *Journal of peace research*, 2008, *45* (5), 697–710.

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

**Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack Porter**, "Valid $t$-Ratio Inference for IV," *American Economic Review*, 2022, *112* (10), 3260–3290.

**Lewis, M Paul, Gary F Simons, and Charles D Fennig**, "Ethnologue: Languages of the World," *Dallas, Texas: SIL International. Online version: http://www.ethnologue.com*, 2020.

**Luca, Giacomo De, Roland Hodler, Paul A Raschky, and Michele Valsecchi**, "Ethnic favoritism: An axiom of politics?," *Journal of Development Economics*, 2018, *132*, 115–129.

**Michalopoulos, Stelios**, "The origins of ethnolinguistic diversity," *American Economic Review*, 2012, *102* (4), 1508–39.

_ **and Elias Papaioannou**, "National institutions and subnational development in Africa," *The Quarterly journal of economics*, 2014, *129* (1), 151–213.

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

**Montalvo, José G and Marta Reynal-Querol**, "Ethnic polarization, potential conflict, and civil wars," *American economic review*, 2005, *95* (3), 796–816.

**Montalvo, Jose G and Marta Reynal-Querol**, "Ethnic diversity and growth: Revisiting the evidence," *Review of Economics and Statistics*, 2017, pp. 1–43.

**Morelli, Massimo and Dominic Rohner**, "Resource concentration and civil wars," *Journal of Development Economics*, 2015, *117*, 32–47.

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

**Müller-Crepon, Carl, Yannick Pengl, and Nils-Christian Bormann**, "Linking Ethnic Data from Africa," 2020.

**Nickell, Stephen**, "Biases in dynamic models with fixed effects," *Econometrica: Journal of the econometric society*, 1981, pp. 1417–1426.

**Nunn, Nathan and Leonard Wantchekon**, "The slave trade and the origins of mistrust in Africa," *American Economic Review*, 2011, *101* (7), 3221–52.

**Obradovich, Nick, Ömer Özak, Ignacio Martín, Ignacio Ortuño-Ortín, Edmond Awad, Manuel Cebrián, Rubén Cuevas, Klaus Desmet, Iyad Rahwan, and Ángel Cuevas**, "Expanding the measurement of culture with a sample of two billion humans," *Journal of the Royal Society Interface*, 2022, *19* (190), 20220085.

**Robbeets, Martine and Alexander Savelyev**, *Language Dispersal Beyond Farming*, John Benjamins Publishing Company, 2017.

**Smith, Michael Garfield**, *The plural society in the British West Indies*, Univ of California Press, 1965.

**Spolaore, Enrico and Romain Wacziarg**, "Ancestry, language and culture," in "The Palgrave handbook of economics and language," Springer, 2016, pp. 174–211.

**_ and _**, "War and relatedness," *Review of Economics and Statistics*, 2016, *98* (5), 925–939.

**_ and _**, "The political economy of heterogeneity and conflict," Technical Report, National Bureau of Economic Research 2017.

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

**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.

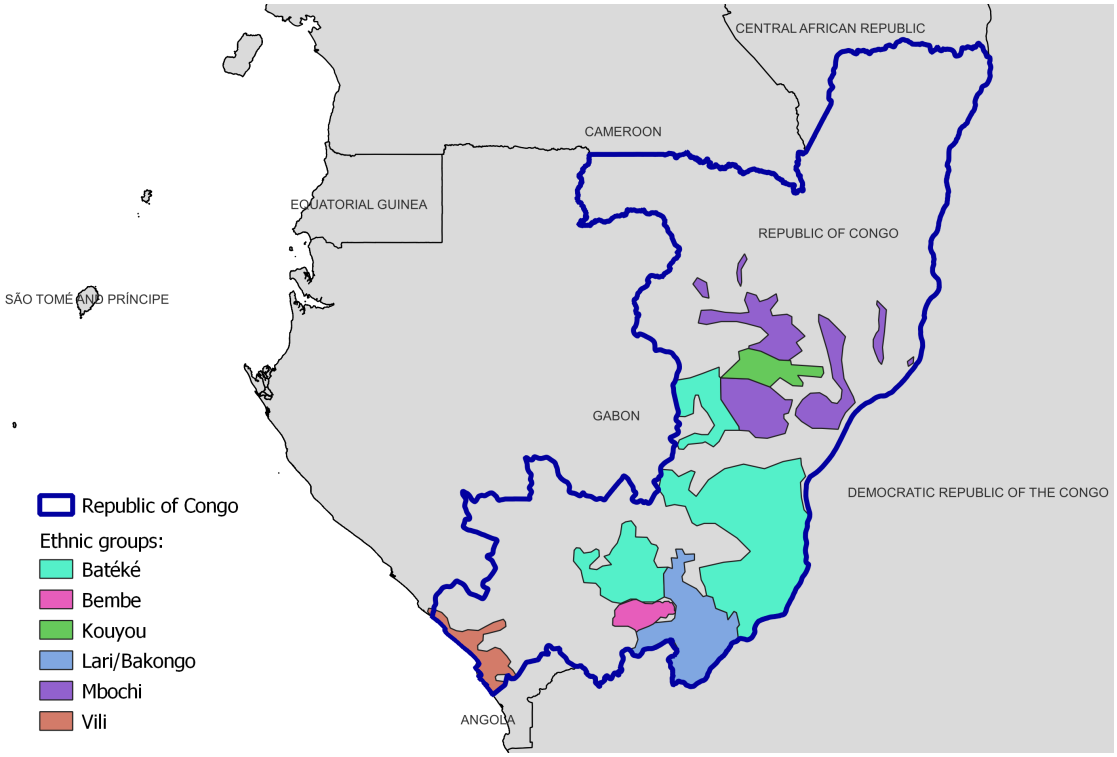
**Wucherpfennig, Julian, Nils B. Weidmann, Luc Girardin, Lars-Erik Cederman, and Andreas Wimmer**, "Politically Relevant Ethnic Groups across Space and Time: Introducing the GeoEPR Dataset," *Conflict Management and Peace Science*, 2011, *28* (5), 423–437.

**_ , Nils W Metternich, Lars-Erik Cederman, and Kristian Skrede Gleditsch**, "Ethnicity, the State, and the Duration of Civil War," *World Politics*, 2012, *64* (1), 79–115.

# Appendix (Intended for Online Publication)

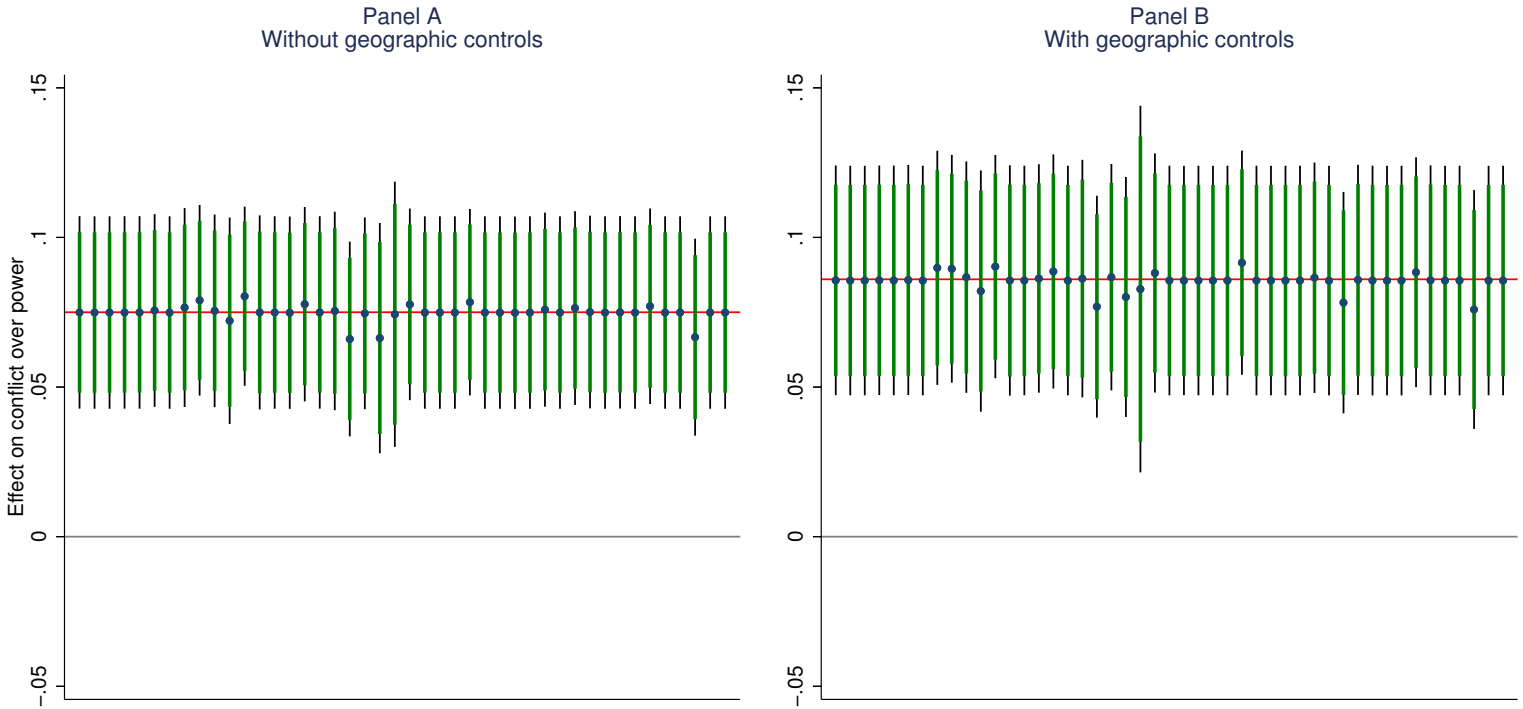
## Appendix A: Additional Figures and Tables

FIGURE A-1: *Ethnic Groups and Ethnic Settlements in Congo*



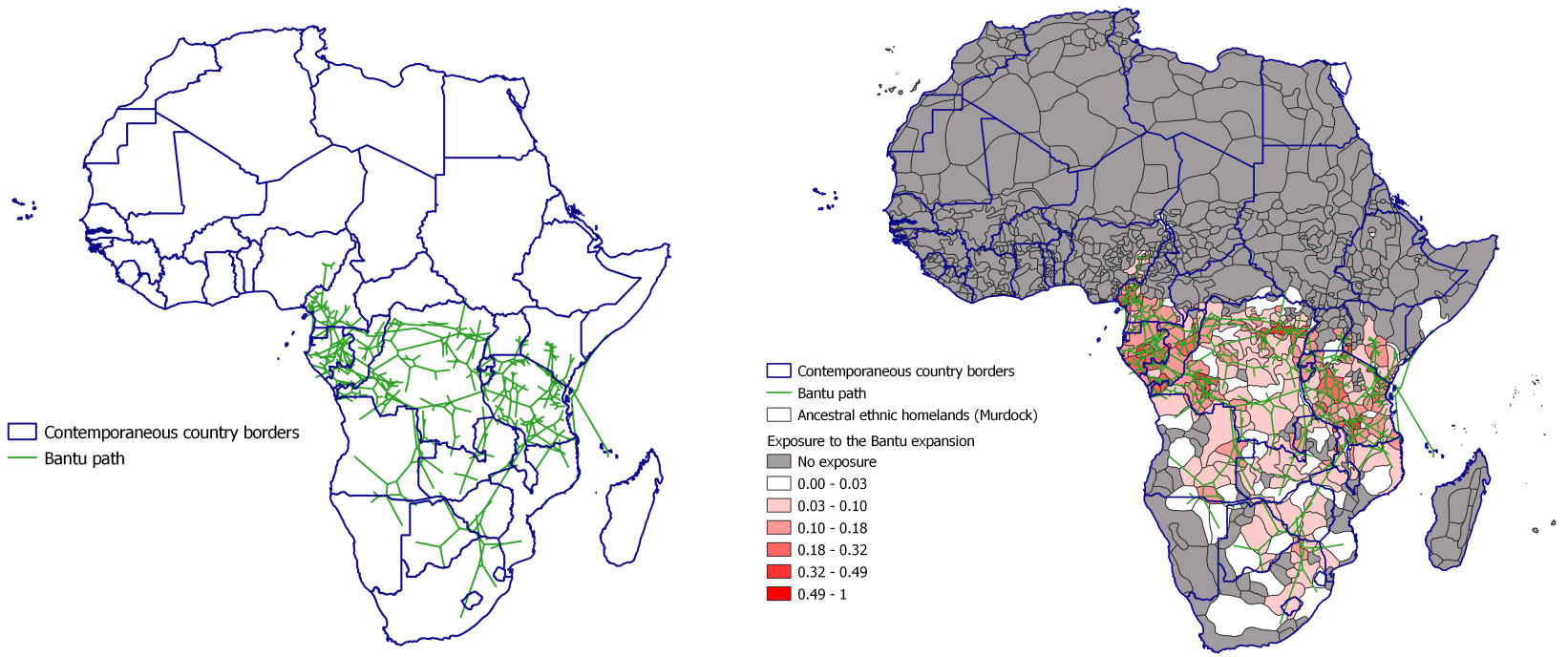

*Notes*: The figure illustrates the six politically relevant ethnic groups (Lari/Bakongo, Kouyou, Mbochi (proper). Bateke, Bemba, Vili) in Congo-Brazzaville and their settlements. Source: Ethnic Power Relations (EPR) Core (Wucherpfennig et al., 2012) and GEO-EPR (Wucherpfennig et al., 2011).

FIGURE A-2: *Checking for Outliers: Dropping one Country at a Time*




*Notes*: The figures show the stability of the estimates when dropping one country at a time from the sample. The red horizontal line indicates the baseline coefficient; green and black vertical lines indicate 90- and 95-percent confidence intervals, respectively. All specifications include country, year, country-year, and ethnicity fixed effects as well as ethnicity-specific time trends. Panel B includes geographic controls.

FIGURE A-3: *The Path of the Bantu Expansion and the Murdock Map*




*Notes*: The figure on the left shows the historical migration route of the Bantu tribe reconstructed by Grollemund et al. (2015) and contemporaneous country borders. See Appendix B for additional details on the route reconstruction. The figure on the right overlays historical migration route with the map of pre-colonial ethnic settlements constructed by Murdock. Different colors denote the intensity of exposure of each ethnic group to the Bantu expansion, computed as the the length of the path crossing an ethnic homeland divided by the size of the ethnic settlement, all normalized to range between 0 and 1.

TABLE A-1: *Data Extract and Illustration of the Identifying Variation*

| country | year | potential rebel | Government: ethnic group 1 [*status*] | ethnic group 2 [*status*] | ethnic group 3 [*status*] | ethnic group 4 [*status*] | ethnic group 5 [*status*] | linguistic distance | conflict over power |
|---|---|---|---|---|---|---|---|---|---|
| Congo | 1995 | Lari/Bakongo | Bembe [*Senior partner*] | Bateke [*Junior partner*] | Kouyou [*Junior partner*] | Vili [*Junior partner*] | Lari/Bakongo [*Junior partner*] | 0.136 | 0 |
| Congo | 1996 | Lari/Bakongo | Bembe [*Senior partner*] | Bateke [*Junior partner*] | Kouyou [*Junior partner*] | Vili [*Junior partner*] | Lari/Bakongo [*Junior partner*] | 0.136 | 0 |
| Congo | 1997 | Lari/Bakongo | Bembe [*Senior partner*] | Bateke [*Junior partner*] | Kouyou [*Junior partner*] | Vili [*Junior partner*] | Lari/Bakongo [*Junior partner*] | 0.136 | 0 |
| Congo | 1998 | Lari/Bakongo | Mbochi (proper) [*Senior partner*] | Bateke [*Junior partner*] | Kouyou [*Junior partner*] | | | 0.202 | 1 |
| Congo | 1999 | Lari/Bakongo | Mbochi (proper) [*Senior partner*] | Bateke [*Junior partner*] | Kouyou [*Junior partner*] | | | 0.202 | 1 |
| Congo | 2000 | Lari/Bakongo | Mbochi (proper) [*Senior partner*] | Bateke [*Junior partner*] | Kouyou [*Junior partner*] | | | 0.202 | 0 |

*Notes*: The table illustrates a data extract for the potential rebel Lari/Bakongo in Congo between 1995 and 2000. The identifying variation is within-ethnicity changes in cultural distance resulting from government changes. In this case, the change in the government coalition occurred in 1998 and Lari/Bakongo experienced an increase in cultural distance and engaged in conflict against the government.

TABLE A-2: *Cultural Distance and Ethnic Civil Conflict: Robustness Tests*

| | Dependent variable: Ethnic conflict over power | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Linguistic distance | 0.043** | 0.037** | 0.086*** | 0.086*** | 0.086** | 0.083*** | 0.091*** | 0.084*** | 0.058* | 0.043** |
| | (0.021) | (0.017) | (0.018) | (0.021) | (0.040) | (0.018) | (0.020) | (0.019) | (0.031) | (0.019) |
| Mean of dep. var. | 0.042 | 0.042 | 0.048 | 0.048 | 0.048 | 0.048 | 0.072 | 0.051 | 0.042 | 0.048 |
| Robustness test | *Groups split across countries* | | *Cluster by country* | *Cluster by country-year* | *Cluster by ethnicity* | *Climatic controls* | *At least 1 conflict* | *At least 1 gov. change* | *Balanced panel* | $\lambda = 0.05$ |
| Distance type | w | uw | w | w | w | w | w | w | w | w |
| Country-year FE | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| Ethn.-country FE | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| Ethn.-year FE | yes | yes | | | | | | | | |
| Ethn. year trend | | | yes | yes | yes | yes | yes | yes | yes | yes |
| Geographic controls | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| Climatic controls | | | | | | yes | | | | |
| Observations | 7,434 | 7,434 | 9,990 | 9,990 | 9,990 | 9,990 | 6,667 | 8,370 | 8,866 | 9,990 |
| Adjusted R-squared | 0.836 | 0.836 | 0.500 | 0.503 | 0.504 | 0.506 | 0.504 | 0.486 | 0.539 | 0.504 |

*Notes*: For a description of all explanatory variables, refer to notes in Table 1. The sample includes 44 African countries, 57 years (1961-2017), and 265 ethnic groups. Two-way clustered standard errors by year and country are reported in parenthesis, unless otherwise indicated in columns 3, 4, and 5. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

TABLE A-5: *Dissatisfaction with Government Performance*

| | Dependent variable: Thinks the government is performing badly | | |
|---|---|---|---|
| | Average across items | | |
| | (1) | (2) | (3) |
| Linguistic distance | 0.079*** | 0.079*** | 0.059*** |
| | (0.020) | (0.020) | (0.019) |
| Mean of dep. var. | 0.571 | 0.571 | 0.571 |
| Country-year FE | yes | yes | yes |
| Ethnicity-country FE | yes | yes | yes |
| Ethnicity-country year trends | yes | yes | yes |
| Individual controls | | yes | yes |
| Geographic controls | | | yes |
| Observations | 165,092 | 165,092 | 165,092 |
| Adjusted R-squared | 0.176 | 0.177 | 0.177 |

*Notes*: The unit of observation is an individual belonging to one of 904 distinct ethnic groups in 28 African countries interviewed between 1999 and 2018 in 7 rounds of the Afrobarometer survey. The dependent variable is a binary variable capturing the extent to which the person thinks the government is performing very badly or fairly badly in handling the following matters: managing the economy, creating jobs, keeping prices stable, narrowing income gap, reducing crime, improving basic health services, addressing educational needs, improving water and sanitation services, ensuring enough to eat, fighting corruption, reducing conflict, combating malaria, combating HIV, maintaining roads and bridges, providing reliable electric supply, managing rivers, promoting equal rights/opportunities for women, addressing needs of youth, protecting rights and promoting opportunities for disabled. The dependent variable takes the mean across all items. *Linguistic distance* captures the weighted linguistic distance between a respondent and the ethnic groups at the government. *Linguistic distance* is standardized. For a description of geographic controls, refer to the notes in Table 1. Individual controls include age, age squared, a dummy for female and a dummy for residence in a rural area. Standard errors clustered at the ethnicity level are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

TABLE A-3: *Cultural Distance and Ethnic Civil Conflict: Lagged Conflict and Anticipatory Effects*

| | Ethnic conflict over power | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Linguistic distance$_{t-1}$ (lag) | | | | 0.051** | |
| | | | | (0.023) | |
| Linguistic distance$_t$ | 0.086*** | 0.046*** | 0.036** | | 0.074** |
| | (0.019) | (0.013) | (0.017) | | (0.030) |
| Linguistic distance$_{t+1}$ (lead) | | | | | 0.011 |
| | | | | | (0.032) |
| Conflict over power$_{t-1}$ | | 0.572*** | 0.649*** | | |
| | | (0.082) | (0.123) | | |
| Mean of dep. var. | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 |
| Country-year fixed effects | yes | yes | yes | yes | yes |
| Ethnicity fixed effects | yes | yes | yes | yes | yes |
| Ethnicity year trend | yes | yes | yes | yes | yes |
| Geographic controls | yes | yes | yes | yes | yes |
| Observations | 9,990 | 9,676 | 9,394 | 9,676 | 9,676 |
| Adjusted R-squared | 0.507 | 0.676 | 0.319 | 0.515 | 0.515 |

*Notes*: The unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government for gaining power and 0 otherwise. *Linguistic distance$_t$* captures weighted linguistic distance between a potential rebel and the ethnic groups at the government. Conflict$_{t-1}$ is a binary variable that is equal to 1 if an ethnic group was involved in a conflict over power in the previous year. Since one caveat with the specification in column 2 is that the inclusion of a lagged dependent variable in a fixed effects model might generate Nickell bias (Nickell, 1981), in column 3 I instrument the first lag of the dependent variable (Conflict$_{t-1}$) with the second lag (Conflict$_{t-2}$). Linguistic distance$_{t-1}$ and Linguistic distance$_{t+1}$ denote a lagged and lead measure of linguistic distance, respectively. For a description of all explanatory variables, refer to notes in Table 1. The sample includes 44 African countries, 57 years (1961-2017), and 265 ethnic groups. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

TABLE A-4: *IV: Inference with Lee et al. (2022) tF Procedure*

| | Conflict over government power | |
|---|---|---|
| | (1) | (2) |
| Linguistic distance | 0.200*** | 0.149*** |
| | (0.037) | (0.029) |
| tF 0.05 se | [0.058] | [0.036] |
| Mean of dep. var | 0.050 | 0.050 |
| Distance type | w | uw |
| Country-year FE | yes | yes |
| Ethnicity FE & trends | yes | yes |
| Controls | yes | yes |
| Observations | 3,775 | 3,775 |

*Notes*: The unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting over power. *Linguistic distance* captures linguistic distance between each potential rebel and the ethnic groups at the government. For a description of geographic controls, refer to the notes in Table 1. The sample includes 17 African countries in the Bantu region, 57 years (1961-2017), and 95 distinct ethnic groups. Two-way clustered standard errors by year and country, adjusted for the low number of clusters using the number of countries, are reported in parenthesis and standard errors further re-adjusted using Lee et al. (2022) *tF* procedure are reported in square brackets. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

TABLE A-6: *Discontent with Government Performance Index: Principal Component Analysis Loadings*

| *How well or badly would you say the current government is handling the following matters:* | Loadings | | | | | | |
|---|---|---|---|---|---|---|---|
| | Round 1 1999-2001 | Round 2 2002-2003 | Round 3 2005-2006 | Round 4 2008-2009 | Round 5 2011-2013 | Round 6 2014-2015 | Round 7 2016-2017 |
| Managing the economy | | 0.297 | 0.327 | 0.260 | 0.273 | 0.291 | 0.223 |
| Improving living standards of the poor | | | | 0.278 | 0.277 | 0.302 | 0.234 |
| Creating jobs | 0.581 | 0.281 | 0.310 | 0.266 | 0.260 | 0.294 | 0.235 |
| Keeping prices stable | 0.587 | 0.282 | 0.301 | 0.231 | 0.225 | 0.276 | 0.217 |
| Narrowing the income gap | 0.563 | 0.278 | 0.306 | 0.260 | 0.254 | 0.284 | 0.227 |
| Reducing crime | | 0.288 | 0.292 | 0.242 | 0.244 | 0.257 | 0.221 |
| Improving basic health services | | 0.291 | 0.310 | 0.261 | 0.266 | 0.278 | 0.245 |
| Addressing educational needs | | 0.290 | 0.303 | 0.243 | 0.255 | 0.271 | 0.239 |
| Improving water and sanitation services | | 0.221 | 0.276 | 0.249 | 0.245 | 0.270 | 0.223 |
| Ensuring enough to eat | | 0.268 | 0.317 | 0.268 | 0.262 | 0.283 | 0.231 |
| Fighting corruption | | 0.295 | 0.312 | 0.254 | 0.261 | 0.279 | 0.226 |
| Reducing conflict | | 0.290 | | | 0.251 | | |
| Combating HIV | | 0.253 | 0.249 | 0.198 | 0.208 | | |
| Combating malaria | | 0.264 | | | | | |
| Maintaining roads and bridges | | | | 0.247 | 0.243 | 0.254 | 0.204 |
| Providing reliable electric supply | | | | 0.251 | 0.237 | 0.260 | 0.205 |
| Protecting rivers and forests | | | | 0.254 | | | |
| Promoting equal rights and opportunities for women | | | | 0.228 | 0.231 | | 0.227 |
| Preventing election violence | | | | | | | 0.230 |
| Preventing or resolving violent community conflict | | | | | | | 0.242 |
| Countering violence from armed extremists | | | | | | | 0.231 |
| Addressing needs of youth | | | | | | | 0.246 |
| Protecting rights, promoting opportunities for disabled | | | | | | | 0.238 |
| Common variance explained by first principal component | 62% | 37% | 39% | 37% | 36% | 44% | 38% |
| N (respondents) | 12,666 | 15,281 | 19,661 | 19,462 | 36,904 | 46,134 | 13,361 |

*Notes*: The table illustrates the loadings of each variable from the Afrobarometer survey used for the principal component analysis. Each variable is binary and is equal to 1 if the respondent thinks the government is performing badly or fairly badly in handling each policy goal. Sample sizes denote the set of respondent answering each question for which information on ethnicity is available.

TABLE A-7: *Preferences over Policy Index: Principal Component Analysis Loadings*

| *What are the most important problems facing this country that government should address?* | Loadings | | | | | | |
|---|---|---|---|---|---|---|---|
| | Round 1 1999-2001 | Round 2 2002-2003 | Round 3 2005-2006 | Round 4 2008-2009 | Round 5 2011-2013 | Round 6 2014-2015 | Round 7 2016-2017 |
| Health | 0.138 | 0.101 | 0.239 | 0.350 | 0.244 | 0.367 | 0.326 |
| Infrastructure | 0.195 | 0.208 | 0.240 | 0.149 | 0.344 | 0.186 | 0.229 |
| Agriculture/Food | 0.445 | -0.132 | 0.137 | -0.026 | -0.222 | -0.127 | -0.042 |
| Governance | -0.104 | 0.621 | 0.427 | 0.387 | 0.186 | 0.337 | 0.243 |
| The economy | -0.695 | -0.722 | -0.771 | -0.737 | -0.667 | -0.663 | -0.711 |
| Services | 0.413 | 0.121 | 0.243 | 0.387 | 0.535 | 0.501 | 0.522 |
| Other | 0.128 | -0.060 | -0.106 | 0.079 | -0.068 | -0.061 | -0.005 |
| No pressing problems | 0.253 | 0.062 | 0.140 | 0.070 | 0.051 | 0.088 | 0.047 |
| Common variance explained by first principal component | 21% | 21% | 20% | 21% | 21% | 23% | 21% |
| N (ethnic groups) | 144 | 232 | 292 | 373 | 585 | 525 | 557 |

*Notes*: The table illustrates the loadings of each variable from the Afrobarometer survey used for the principal component analysis used to construct an index capturing preferences over policy at the ethnic group level. Each variable captures the fraction of respondents within each ethnic group stating that a given domain should be the one that the government should address. The options that respondents can provide are more fine-grained than the above-listed domains. For example, under the "health" domain, respondents can list health, AIDS, or sickness/disease. I follow the grouping in the Afrobarometer questionnaires (e.g., see https://www.afrobarometer.org/wp-content/uploads/2022/02/alg_r6_questionnaire.pdf) to aggregate up the various options into the fewer domains.

## Appendix B: Additional Analyses

### B-1 Estimation Using Partitioned Ethnicities

This section describes in detail the procedure adopted to obtain the results in columns 1 and 2 of Table A-2. My main estimation strategy does not allow to account for unobserved time-varying ethnic-specific shocks. If these occurred simultaneously to a change in government and were correlated to an ethnic group's decision to rebel, they could confound my results. To address this, I re-run the analysis on a subset of ethnic groups that were split by country borders during the Scramble for Africa. This strategy, similar to the one employed by Michalopoulos and Papaioannou (2014) and analogous to the one adopted by Dickens (2018), exploits the fact that the same ethnicity is simultaneously exposed to different governments in different countries. As the ethnic identity of governments changes in some countries but not others, the quasi-random allocation of borders provides an exogenous source of within-group variation in cultural distance to the government.[46] As cultural distance of partitioned groups varies over time and between countries, this variation allows me to control for a full set of ethnicity-year fixed effects.

While this estimation strategy constitutes a more tightly-controlled empirical exercise, it has the disadvantage of restricting the sample to a peculiar set of ethnicities, i.e., those that were split into multiple countries. As Michalopoulos and Papaioannou (2016) show, partitioned groups are considerably more likely to experience political violence, ethnic wars, and government-led discrimination when compared to non partitioned ethnicities. For this reason, one should interpret the results presented in this section as a robustness exercise in support of the baseline findings.

I merge contemporary ethnicities to the Murdock Atlas, which contains information on ethnic settlements prior to European contact and therefore provides the most reliable source to identify groups that were partitioned during the Scramble for Africa.[47] The map in Figure B-1 displays the partitioned ethnicities that I successfully merged with the EPR ethnic groups. This restricted sample includes 84 distinct Murdock ethnic groups partitioned across 31 African countries.[48]

I estimate the following empirical model:

$$\text{Conflict}_{rct} = \eta_{c,r} + \lambda_{c,t} + \zeta_{r,t} + \beta \text{LD}_{rct} + \Gamma \text{G}_{rct} + \epsilon_{rct} \tag{10}$$

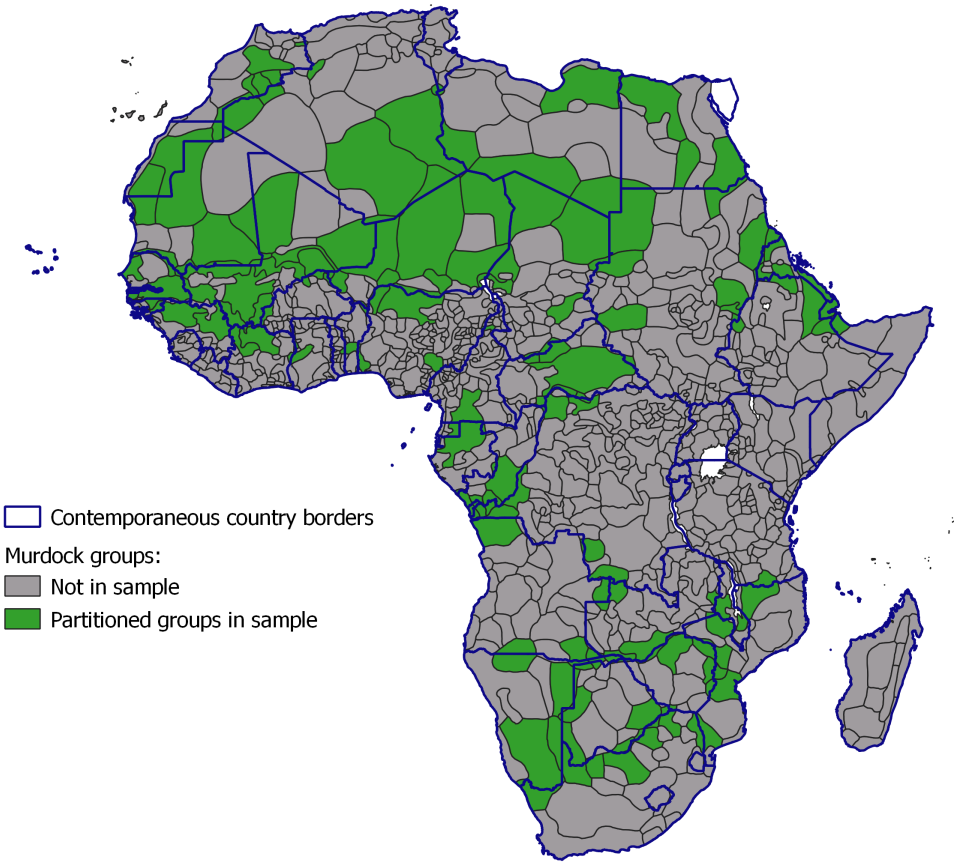
The dependent variable is a measure of conflict for ethnic group $r$ in country $c$ and year $t$. The main independent variable, $\text{LD}_{rct}$ is a measure of linguistic distance to the

[46] For additional evidence and details on the randomness in the allocation of colonial boundaries, see the discussion in Michalopoulos and Papaioannou (2016) and in Dickens (2018).

[47] This is also the data source employed by Michalopoulos and Papaioannou (2016), who were the first exploiting partitioned groups for identification.

[48] Note that my merging procedure is successful for 41 percent of the ethnicities that Michalopoulos and Papaioannou (2016) classify as partitioned.

FIGURE B-1: *Murdock Ethnic Groups Split by Country Borders*




*Notes*: The figure displays the ethnic groups in the Murdock Ethnographic Atlas that were split by country borders and that I include in the difference-in-differences analysis. I do not include groups that I cannot successfully merge to at least two groups in the EPR dataset belonging to different countries.

government in period $t$. As the main specification, I add a full set of country-year fixed effects ($\lambda_{c,t}$), as well as a set of ethnicity-country fixed effects ($\eta_{c,r}$). Moreover, since the same potential rebel is present in multiple countries, this specification controls for a full set of ethnicity-year fixed effects ($\zeta_{r,t}$), and thus accounts for any time-variant ethnic-specific shocks. $G_{rct}$ indicates the same set of geographic controls outlined in section 4.

### B-2 Francois et al.'s (2015) Data on the Ethnic Composition of the Government

As a robustness test, I use Francois et al.'s (2015) data on the ethnicity of African cabinet members to construct an alternative measure of linguistic distance. For 15 out of the 44 countries included in my estimation sample, Francois et al. (2015) provide information on the ethnic identity of more than 90% of national ministers from the 1960s (i.e., independence) up to 2004. I manually match each cabinet minister's ethnicity to one or more Ethnologue languages (up to 8). Conveniently, Francois et al. (2015) provide information on the number of top and lower government positions occupied by each ethnic group in a given country and year. This information allows me to generate a linguistic distance measure similar to the one I employ in my main specification and assign a higher weight to ethnicities occupying top government positions.[49]

I then re-run my baseline specification using this alternative measure. The sample ($N = 2,763$) includes 15 countries (Benin, Cameroon, Cote d'Ivoire, Democratic Republic of Congo, Gabon, Ghana, Guinea, Liberia, Nigeria, Republic of Congo, Sierra Leone, Tanzania, Togo, Kenya, and Uganda) and 91 potential rebels (out of the 265 in my main sample). Reassuringly, my baseline linguistic distance measure and this alternative one are highly correlated (correlation coefficient: 0.75).

Table B-1 shows that results hold for this subsample of countries and years when employing my baseline linguistic distance measure (columns 1-3). Reassuringly, the linguistic distance measure based on Francois et al.'s (2015) data (columns 4-6) leads to similar results. Coefficients are larger in columns 4 and 5; albeit not statistically significant, the coefficient on column 6 is still positive and large in magnitude. Despite the restricted coverage, these findings suggest that my main results are robust to employing a linguistic distance measure relying on an alternative data source on the ethnic composition of the executive.

### B-3 Instrument Construction and Validity

To construct the Bantu instrument, I start by constructing a measure of exposure to the Bantu expansion for each ethnic group's homeland. I merge contemporary ethnicities to their ancestral homelands in the Murdock map, the most ancient record of ethnic settlements in Africa. I then overlay ethnic groups' ancestral homelands on the historical

[49] In a similar vein as for my EPR-based measure, when constructing the weights I assume that holding a top-government position is equivalent to holding two lower-government positions.

TABLE B-1: *Robustness using Francois et al.'s (2015) Data on Ethnic Composition of the Government*

| | Dependent variable: Ethnic conflict over power | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Linguistic distance$^{\text{Baseline}}$ | 0.047** | 0.031*** | 0.050** | | | |
| | (0.019) | (0.005) | (0.018) | | | |
| Linguistic distance$^{\text{Francois et al.}}$ | | | | 0.080** | 0.091** | 0.038 |
| | | | | (0.035) | (0.038) | (0.044) |
| Mean of dep. var. | 0.037 | 0.037 | 0.037 | 0.037 | 0.037 | 0.037 |
| Country-year fixed effects | yes | yes | yes | yes | yes | yes |
| Ethnicity fixed effects | yes | yes | yes | yes | yes | yes |
| Ethnicity year trend | | | yes | | | yes |
| Geographic controls | | yes | yes | | yes | yes |
| Observations | 2,763 | 2,763 | 2,763 | 2,763 | 2,763 | 2,763 |
| Adjusted R-squared | 0.348 | 0.352 | 0.438 | 0.346 | 0.348 | 0.429 |

*Notes*: the unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government and 0 otherwise. *Linguistic distance*$^{Baseline}$ captures linguistic distance between each potential rebel and the ethnic groups at the government based on the EPR dataset. *Linguistic distance*$^{Francois\ et\ al.}$ captures linguistic distance between each potential rebel and the ethnic groups at the government based on the Francois et al.'s (2015) data. The linguistic distance measures are standardized. Geographic controls include the logged geodesic distance between each ethnic group and the government, absolute distance in elevation, absolute distance in ruggedness, and absolute distance in the caloric suitability index (CSI). The sample includes 91 distinct ethnic groups in 15 African countries over 44 years (1961-2004). Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

migration route of the Bantu reconstructed by Grollemund et al. (2015).[50] The left panel of Figure A-3 illustrates the route of the Bantu migration and contemporaneous country borders. For each ethnic group in the Murdock map, I construct a Bantu Index capturing the intensity of exposure to the Bantu migration route. I calculate the length of the path crossing each ethnic homeland, divide it by the size of the settlement, and then normalize this index to range between 0 and 1:

$$\text{Bantu Index}_{EA} = \frac{\text{Path Length}}{\text{Homeland Area}} \tag{11}$$

Figure A-3 (right) shows the distribution of the Bantu index across Africa. There is considerable within-country, cross-ethnicity variation in ethnic groups' exposure to the Bantu expansion. Since each EPR group is associated to multiple groups in the

[50]See Appendix B for details on how Grollemund et al. (2015) reconstructed the route.

Ethnographic Atlas, I proceed as follows to obtain an index at the EPR-group level:

$$\text{Bantu Index}_{EPR} = \sum_{i=1}^{N} \alpha_i \text{Bantu Index}_{EA_i} \tag{12}$$

where $\alpha_i$ is a weight reflecting the accuracy of the match between each Murdock group and the EPR group.[51]

Next, I construct the instrument, i.e., the absolute distance in the Bantu Index between a potential rebel and the government. To this end, I follow the same procedure I adopted for the linguistic distance measure. First, I compute the Bantu distance between a potential rebel (r) and each ethnic group forming the government ($g_i$):

$$\text{BD}_{rg_i} = |\text{Bantu Index}_r - \text{Bantu Index}_{g_i}| \tag{13}$$

Second, I compute a weighted Bantu distance between each potential rebel and the government:

$$BD^W = \sum_{i=1}^{N} p_{g_i} \times BD_{rg_i} \tag{14}$$

where $N$ denotes the total number of ethnicities forming the government and $p_{g_i}$ is a weight reflecting the position of power of group $g_i$ in the coalition. Alternatively, I construct the unweighted version:

$$BD^{UW} = \sum_{i=1}^{N} \frac{BD_{rg_i}}{N} \tag{15}$$

*B-3.1 Instrument Validity*

The validity of the IV estimates rests on the assumption that the differential exposure to the Bantu expansion affects conflict today only through its impact on cultural distance, conditional on the controls included in the regression. The first concern is that geographical differences between ancestral homelands might have determined the path of the Bantu migration and, at the same time, might also affect the likelihood of conflict between groups today. First, Grollemund et al. (2015) emphasize how the route of the expansion was mostly determined by emerging savannah corridors, which appeared in the rainforest due to a climatic shock and then disappeared once the climatic crisis was over. Despite this suggestive evidence of exogeneity in the way the path evolved, my preferred specification addresses this concern by controlling for a set of geographic

[51] The accuracy is given by the percentage of nodes in the EPR-group linguistic tree covered by the Murdock-group linguistic tree. $\sum_{i=1}^{N} \alpha_i = 1$. If the match is equally accurate between each Murdock group, then Bantu $\text{Index}_{EPR} = \sum_{i=1}^{N} \frac{\text{Bantu Index}_{EA_i}}{N}$. See Section C-2 for details on the LEDA package, the merging procedure, and the accuracy of the match.

controls.

Another concern could be that the Bantu, in gradually expanding, intentionally and systematically avoided pre-existing tribes with certain characteristics that make them more (or less) prone to conflict today. Suppose the Bantu avoided particularly bellicose tribes, and favored territories inhabited by more peaceful populations. Then my instrument would be systematically picking up differences in bellicosity. If these differences have a direct effect on conflict today, the exclusion restriction is violated.

I examine whether this example is at play in my setting by delving deeper in the reduced-form analysis. If the Bantu systematically avoided certain territories along unobservable characteristics correlated with contemporary conflict, then only being more (or less) affected to the Bantu expansion should have a reduced-form effect on conflict, but not both. I test this by splitting the absolute distance in the Bantu index into two: being *more* exposed to the Bantu expansion than the ethnicities at the government, and being *less* exposed.

I run an alternative reduced-form estimation using these two measures, instead of the absolute difference as in my main specification. As shown in Table B-2, I do not find evidence of a different reduced-form effect of being more versus less exposed to the Bantu expansion. Both coefficients hold the same sign, and the F statistics indicates that these coefficients are not different from each other. This suggests that what matters for conflict between ethnic groups is their *absolute* distance in the Bantu index—as conjectured in section 6.1—and not the fact of being more (or less) affected by the Bantu migration compared to the other group.

TABLE B-2: *Reduced Form: Splitting the Bantu Distance into Two*

| | Ethnic conflict over power | |
|---|---|---|
| | (1) | (2) |
| Bantu more | 0.055 | 0.049 |
| | (0.038) | (0.044) |
| Bantu less | 0.056 | 0.054 |
| | (0.055) | (0.054) |
| F statistics (Bantu more - Bantu less=0) | 0.00 | 0.00 |
| Prob > F | 0.989 | 0.955 |
| Country-year fixed effects | yes | yes |
| Ethnicity fixed effects | yes | yes |
| Ethnicity year trends | yes | yes |
| Geographic controls | | yes |
| Observations | 3,775 | 3,775 |
| Adjusted R-squared | 0.546 | 0.546 |

*Notes*: the unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government and 0 otherwise. *Bantu more* and *Bantu less* is a continuous variable that denotes whether the rebel was more or less exposed to the Bantu expansion than the government, respectively. Two-way clustered standard errors by year and country are reported in parenthesis, and are adjusted for the low number of clusters using the number of countries. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

## Appendix C: Data Description, Sources, and Summary Statistics

### C-1 Data Sources

**Ethnic groups:** Data on ethnic groups comes from the Ethnic Power Relations (EPR) dataset family (Wucherpfennig et al., 2012; Vogt et al., 2015). EPR lists each politically relevant ethnic group for each country in each year, and the respective access (or lack thereof) to executive government power. EPR defines an ethnic group as an identity group that defines itself or is defined by others along linguistic, religious or racial characteristics.
Source: https://icr.ethz.ch/data/epr/core/

**Linguistic distance:** I retrieve data on languages (up to 3) spoken by each ethnic group from the Ethnic Dimensions Dataset. I compute a measure of linguistic distance (called cladistic distance) using information on linguistic trees coming from the Ethnologue database (Lewis et al., 2020). I describe the methodology for the construction of the linguistic distance measure in section 3.
Sources: https://icr.ethz.ch/data/epr/ed/ and https://www.ethnologue.com/browse/names

**Conflict:** Data on ethnic civil conflicts comes from the UCDP/PRIO Armed Conflict Dataset (version 20.1) (Harbom et al., 2008; Allansson et al., 2017), with information on the identity of the combatants involved in a conflict (rebel groups and governments). Information on the ethnic identity of rebel groups comes from the ACD2EPR dataset (Vogt et al., 2015).
Sources: https://ucdp.uu.se/downloads/ and https://icr.ethz.ch/data/epr/acd2epr/

**Geodesic distance:** I compute geodesic distances between each group's centroid, which I calculate based on ethnic settlements provided by the GeoEPR dataset (Wucherpfennig et al., 2011). Throughout the analysis, I use log geodesic distance.
Source: https://icr.ethz.ch/data/epr/geoepr/

**Absolute difference in elevation:** I combine GeoEPR with data on elevation at the grid level (2.5 arc-minute resolution) provided by the Worldclim Global Climate Database (Fick and Hijmans, 2017). I measure average elevation of each ethnic group, and then calculate the absolute distance in elevation between each ethnic group and the ethnicities forming the government.
Source: https://www.worldclim.org/data/worldclim21.html

**Absolute difference in ruggedness:** I produce a measure of ruggedness for each ethnic group using the standard deviation of the Worldclim elevation data. I calculate the absolute distance in ruggedness between each ethnic group and the ethnicities forming the government.
Source: https://www.worldclim.org/data/worldclim21.html

**Absolute difference in the Caloric Suitability Index:** Data on the potential agricultural output come from the Caloric Suitability Index (CSI) provided by Galor and Özak (2016). This measure reflects the potential caloric yield of a grid cell based on the Global Agro-Ecological Zones (GAEZ) project of the Food and Agriculture Organization (FAO). The Index results from a combination of climatic and grographic variables unaffected by human activity. I use the pre-1500 average CSI measure that includes cells with zero productivity. I calculate the average CSI for each ethnic group, and calculate the absolute distance in CSI between each ethnic group and the ethnicities forming the government.
Source: https://ozak.github.io/Caloric-Suitability-Index/

**Absolute difference in temperature:** Temperature data comes from the WorldClim Global Climate Database. I compute a time-varying measure of mean temperature for each group, and calculate the absolute distance in mean temperature between each ethnic group and the ethnicities forming the government.
Source: https://www.worldclim.org/data/worldclim21.html

**Absolute difference in rainfall:** Rainfall data comes from the WorldClim Global Climate Database. I compute a time-varying measure of average rainfall for each group, and calculate the absolute distance in mean rainfall between each ethnic group and the ethnicities forming the government.
Source: https://www.worldclim.org/data/worldclim21.html

**Ancestral homelands of contemporary ethnicities:** I retrieve the geolocation of ethnic groups' ancestral homelands from the Murdock Map, digitized by Nunn and Wantchekon (2011). I merge EPR ethnicities to the corresponding Murdock groups' counterparts using the LEDA R-package (see details in Section C-2.)
Source: https://scholar.harvard.edu/nunn/pages/data-0

**Rainfall in the growing season:** I construct a measure at the ethnic group and year level that captures the amount of rain received during the main crop's growing season, similar to the one proposed by Guariso and Rogall (2017). I retrieve information about the starting and ending month of the growing season for the group's settlement's main crop as well as the average amount of rain experienced during those months from the GrowUp platform, which in turn exploits data from the PRIO-GRID dataset. The final variables I exploit in the analysis are the following: the absolute distance in rainfall between each potential rebel and the groups forming the government whenever the potential rebel receives more rain than the government; and the absolute distance in rainfall between each potential rebel and the groups forming the government whenever the potential rebel receives less rain than the government.
Source: https://growup.ethz.ch/rfe

**Presence of oil fields:** Data on the total number of oil field in each ethnic group's

settlement area comes from the GrowUp platform, which in turn exploits data from the PRIO-GRID dataset. This variable is similar to the one employed by Morelli and Rohner (2015). The final variables I exploit in the analysis are the following: the absolute distance in the number of oil fields between each potential rebel and the groups forming the government whenever the potential rebel holds more oil fields than the government; and the absolute distance in the number of oil fields between each potential rebel and the groups forming the government whenever the potential rebel holds more oil fields than the government.
Source: https://growup.ethz.ch/rfe

**The Bantu expansion route.** Data on the Bantu expansion route comes from work by Grollemund et al. (2015). They reconstructed the route by collecting the indigenous geographic location of 424 Bantu languages, including now-extinct languages. Using Bayesian techniques on a sample of 100 lexical items, they construct a pyhlogenetic tree connecting all these languages. Based on this, they reconstruct the probable ancestral geographical locations of each of the internal nodes of the phylogenetic tree through a model calibrated using archaeological evidence (e.g., node 1 is dated back to 4000-5000 before present in the Grassfields region of Cameroon, on the basis of an archaeological site called Shum Laka, the principal site associated with Bantu homelands). Using other archeological sites, they calibrate other branching points of the tree. using a Brownian motion model, they infer the ancestral latitude and longitude for each internal node of the tree, which they then connect using straight lines. Additional methodological details can be found in Grollemund et al. (2015) and in the article's Supplementary Appendix.

**Opinions on government performance:** Data on individuals' opinion on the performance of the current government comes from seven rounds of the Afrobarometer survey (Afrobarometer Data, 1999-2017), conducted between 1999 and 2017 in 27 African countries. See Table A-6 for additional details on the questions asked in each round. Refer to Section 7 for details on how I construct a measure of an individual's degree of disagreement with policies implemented by the current government. I manually merge the self-reported ethnolinguistic affiliation of Afrobarometer respondents to the corresponding Ethnologue language.)
Source: http://afrobarometer.org/data

**Age:** Age of respondent at the time of survey.

**Gender:** An indicator variable equal to one if a respondent is female.

**Rural:** An indicator variable for rural locations.

**Opinions on the role of government:** Data on individuals' opinion on the general role of government comes from seven rounds of the Afrobarometer survey (Afrobarometer Data, 1999-2017), conducted between 1999 and 2017 in 27 African countries. Refer to Section 7 and Table A-7 for details on how I construct a measure of an ethnic group's

view of the role of government. I manually merge the self-reported ethnolinguistic affiliation of Afrobarometer respondents to the corresponding Ethnologue language.
Source: http://afrobarometer.org/data

**National budget spent on public (vs particularistic) goods:** I construct a binary variable at the country and year level (*Public goods*) based on information included in the V-Dem dataset, which collects a variety of indicators coded by country experts (see Coppedge, Michael and Gerring, John and Knutsen, Carl Henrik and Lindberg, Staffan I. and others (2023) for details). The variable I use to construct the *Public goods* indicator comes from the following question asked to country experts: "Considering the profile of social and infrastructural spending in the national budget, how "particularistic" or "public goods" are most expenditures?". The V-Dem dataset considers particularistic spending as narrowly targeted on a specific corporation, sector, social group, region, party, or set of constituents. Such spending may be referred to as "pork", "clientelistic", or "private goods." Instead, V-Dem considers public-goods spending as "intended to benefit all communities within a society, even though it may be means-tested so as to target poor, needy, or otherwise underprivileged constituents." V-Dem suggests experts to consider the entire budget of social and infrastructural spending when answering this question. The variable *Public goods* takes value 1 if "most social and infrastructure expenditures are public-goods but a significant portion (e.g., 1/4 or 1/3) is particularistic" or "all social and infrastructure expenditures are public-goods in character. Only a small portion is particularistic.". The variable *Public goods* takes value 0 if "most all of the social and infrastructure expenditures are particularistic" or "most social and infrastructure expenditures are particularistic, but a significant portion (e.g. 1/4 or 1/3) is public-goods" or "social and infrastructure expenditures are evenly divided between particularistic and public-goods programs."
Source: https://v-dem.net/data/the-v-dem-dataset/country-year-v-dem-fullothers-v13/

**Distribution of public services across social groups:** I construct a binary variable at the country and year level (*Equal distribution*) based on information included in the V-Dem dataset, which collects a variety of indicators coded by country experts (see Coppedge, Michael and Gerring, John and Knutsen, Carl Henrik and Lindberg, Staffan I. and others (2023) for details). The variable I use to construct the *Equal distribution* indicator comes from the following question asked to country experts: "Are basic public services, such as order and security, primary education, clean water, and healthcare, distributed equally across social groups?". V-Dem considers a social group as "differentiated within a country by caste, ethnicity, language, race, region, religion, migration status, or some combination thereof." V-Dem emphasizes that "social group identity is contextually defined and is likely to vary across countries and through time. Nonetheless, at any given point in time there are social groups within a society that are understood—by

those residing within that society—to be different, in ways that may be politically relevant." Given the relevance of ethnicity in African politics (see Francois et al. (2015) and literature therein cited), this measure is arguably appropriate for my setting. The variable *Public goods* takes value 1 if public services are distributed somewhat equally, relatively equally, or equally and zero otherwise.
Source: https://v-dem.net/data/the-v-dem-dataset/country-year-v-dem-fullothers-v13/

### C-2 Merging Ethnic Groups Across Datasets

To merge EPR ethnic groups to their ancestral homelands in the Murdock Map, I use a recent R-package created by Müller-Crepon et al. (2020) called Linking Ethnic Data for Africa (LEDA). Since ethnic categories vary considerably across datasets, Müller-Crepon et al. (2020) created an algorithm that systematically links ethnic groups across 11 dataset. Using a dictionary-based linking procedure, they match more than 8,100 ethnicities via the list of known language families, languages, and dialects from the 16th edition of the Ethnologue database.

Since links are formed through the linguistic tree, the LEDA package allows the researcher to choose the level of precision for each match. One option is to use the so-called *set overlap* rule, which generates a link between any two groups that share at least one language node at a specified level of the language tree. The higher the level specified, the higher is the completeness, but the lower the precision of the match. Throughout the analysis, I opt for precision of matches, and choose the *dialect* level (i.e., the lowest level) of the linguistic tree of ethnic group in dataset A for generating matches to the ethnic group(s) in dataset B.

Even after choosing the matching procedure that favors precision, some matches remain more precise than others. Conveniently, for each match, the package provides an indication of the degree of precision by indicating the extent to which the linguistic tree of an ethnic group in dataset A overlaps with the linguistic tree of an ethnic group in dataset B. Whenever LEDA matches an ethnic group in dataset A (e.g., the EPR dataset) to multiple groups in dataset B (e.g., the Murdock Map), I employ this accuracy measure as a weight when calculating statistics for group in dataset A based on multiple groups in dataset B (e.g., when computing a measure of Bantu exposure for each EPR group based on the Bantu Index of multiple Murdock groups). Groups in dataset B that constitute more accurate matches will get more weight than groups constituting weaker links.[52]

For additional information on the LEDA package, see:
http://www.carlmueller-crepon.org/project/ethnic_matching/

[52]This occurs often in my setting. Because the definition of ethnic group in the EPR dataset tends to be broader than the one in the Murdock Map, an EPR group is usually linked to multiple groups in the Murdock Map.

### C-3 Summary Statistics

TABLE C-1: *Number of Government Changes by Country*

| Country | Number of changes | Country | Number of changes |
|---|---|---|---|
| Angola | 0 | Madagascar | 2 |
| Benin | 8 | Malawi | 1 |
| Botswana | 0 | Mali | 1 |
| Burundi | 0 | Mauritania | 0 |
| Cameroon | 2 | Mauritius | 8 |
| Central African Republic | 5 | Morocco | 0 |
| Chad | 7 | Mozambique | 1 |
| Comoros | 2 | Namibia | 1 |
| Congo | 9 | Niger | 10 |
| Congo, DRC | 8 | Nigeria | 9 |
| Cote d'Ivoire | 5 | Senegal | 2 |
| Djibouti | 2 | Sierra Leone | 6 |
| Egypt | 0 | South Africa | 1 |
| Equatorial Guinea | 0 | South Sudan | 0 |
| Eritrea | 1 | Sudan | 2 |
| Ethiopia | 2 | Tanzania | 1 |
| Gabon | 2 | The Gambia | 0 |
| Ghana | 7 | Togo | 5 |
| Guinea | 5 | Uganda | 4 |
| Guinea-Bissau | 6 | Zambia | 1 |
| Kenya | 6 | Zimbabwe | 4 |
| Liberia | 2 | | |
| Libya | 0 | **Total** | **138** |

*Notes*: The table reports the number of changes in the ethnic identity of governments experienced by each country in the sample over the period 1961-2017.

TABLE C-2: *Summary Statistics*

| | Mean (1) | Std. Dev. (2) | Min (3) | Max (4) | Obs. (5) |
|---|---|---|---|---|---|
| *Conflict Variables* | | | | | |
| Conflict over government power | 0.048 | 0.213 | 0 | 1 | 9,990 |
| Conflict over territory | 0.024 | 0.153 | 0 | 1 | 9,990 |
| *Linguistic Distance* | | | | | |
| Linguistic Distance$^{W}$ | 0.441 | 0.310 | 0 | 1 | 9,990 |
| Linguistic Distance$^{UW}$ | 0.442 | 0.306 | 0 | 1 | 9,990 |
| *Geographic Controls* | | | | | |
| Log geodesic distance$^{W}$ | 5.524 | 1.271 | 0 | 7.284 | 9,990 |
| Log geodesic distance$^{UW}$ | 5.519 | 1.284 | 0 | 7.274 | 9,990 |
| Absolute distance ruggedness$^{W}$ | 132.9 | 153.3 | 0 | 1,223 | 9,990 |
| Absolute distance ruggedness$^{UW}$ | 135.2 | 154.9 | 0 | 1,238 | 9,990 |
| Absolute distance elevation$^{W}$ | 216.2 | 239.2 | 0 | 1,385 | 9,990 |
| Absolute distance elevation$^{UW}$ | 216.8 | 237.9 | 0 | 1,329 | 9,990 |
| Absolute distance CSI$^{W}$ | 581.1 | 1149 | 0 | 9,209 | 9,990 |
| Absolute distance CSI$^{UW}$ | 582.4 | 1128 | 0 | 9,206 | 9,990 |

*Notes*: The sample includes 265 distinct ethnic groups in 44 African countries over a period of 57 years (1961-2017). The superscript $W$ indicates average distances between potential rebels and each ethnic group in the government coalition weighted by the role of ethnic groups in power (where senior partners receive double the weight of each junior partner); the superscript $UW$ indicates unweighted average distances between potential rebels and each ethnic group in the government coalition.

TABLE C-3: *Summary Statistics: Bantu Region*

| | Mean (1) | Std. Dev. (2) | Min (3) | Max (4) | Obs. (5) |
|---|---|---|---|---|---|
| *Conflict Variables* | | | | | |
| Conflict over power | 0.050 | 0.218 | 0 | 1 | 3,775 |
| Conflict over territory | 0.008 | 0.087 | 0 | 1 | 3,775 |
| *Linguistic Distance* | | | | | |
| Linguistic Distance$^{W}$ | 0.308 | 0.266 | 0 | 1 | 3,775 |
| Linguistic Distance$^{UW}$ | 0.313 | 0.268 | 0 | 1 | 3,775 |
| *Bantu Distance* | | | | | |
| Bantu Distance$^{W}$ | 0.049 | 0.049 | 0 | 0.435 | 3,775 |
| Bantu Distance$^{UW}$ | 0.055 | 0.055 | 0 | 0.596 | 3,775 |
| *Geographic Controls* | | | | | |
| Log geodesic distance$^{W}$ | 5.577 | 1.347 | 0 | 7.284 | 3,775 |
| Log geodesic distance$^{UW}$ | 5.580 | 1.346 | 0 | 7.274 | 3,775 |
| Absolute distance ruggedness$^{W}$ | 154.0 | 120.6 | 0 | 676.1 | 3,775 |
| Absolute distance ruggedness$^{UW}$ | 158.8 | 128.2 | 0 | 688.7 | 3,775 |
| Absolute distance elevation$^{W}$ | 260.2 | 238.5 | 0 | 1,385 | 3,775 |
| Absolute distance elevation$^{UW}$ | 265.5 | 238.0 | 0 | 1,327 | 3,775 |
| Absolute distance CSI$^{W}$ | 342.4 | 601.5 | 0 | 3,646 | 3,775 |
| Absolute distance CSI$^{UW}$ | 377.1 | 663.7 | 0 | 3,440 | 3,775 |

*Notes*: The sample includes 86 distinct ethnic groups in 17 African countries corresponding to regions covered by Bantu languages according to the Guthrie classification (Angola, Botswana, Burundi, Cameroon, Congo, Congo DRC, Equatorial Guinea, Gabon, Kenya, Malawi, Mozambique, Rwanda, South Africa, Tanzania, Uganda, Zambia, Zimbabwe) over a period of 57 years (1961-2017). The superscript $W$ indicates average distances between potential rebels and each ethnic group in the government coalition weighted by the role of ethnic groups in power (where senior partners receive double the weight of each junior partner); the superscript $UW$ indicates unweighted average distances between potential rebels and each ethnic group in the government coalition.