Fragmentation or Inequality: Ethnic Divisions and Conflict in Sub-Saharan Africa*

Boris Gershman† — American University

Ameesh Upadhyay‡ — Knox College

February 2026

Abstract

This paper examines the relationship between local ethnic divisions and conflict in Sub-Saharan Africa. Using census subsamples and large-scale household surveys, we construct a new subnational dataset on ethnic inequality capturing group-level differences in education, asset ownership, and access to basic amenities for several hundred regions in thirty-five countries. To distinguish between deep-rooted and more recent ethnic divisions, we incorporate groups from our sample into *Ethnologue*'s linguistic tree model and generate alternative measures of both ethnic fractionalization and inequality based on ancestral languages. Our analysis, leveraging within-country variation and accounting for numerous regional characteristics, reveals a robust positive relationship between ethnic fractionalization and conflict, especially when using deeper linguistic cleavages to define distinct groups. In contrast, ethnic inequality shows no systematic association with conflict frequency or severity. These findings suggest the primacy of ethnolinguistic differences *per se*, rather than socioeconomic disparities between groups, as a driver of local conflict.

Keywords: conflict, ethnolinguistic diversity, ethnic inequality, Sub-Saharan Africa, subnational analysis

JEL Classification Numbers: D31, D74, O10, O15, Z13

*We are grateful to Timur Kuran and three anonymous referees for their valuable advice.

†Department of Economics, American University, 4400 Massachusetts Avenue NW, Washington, DC 20016-8029 (e-mail: boris.gershman@american.edu).

‡Department of Economics, Knox College, 2 East South Street, Galesburg, IL 61401 (e-mail: aupadhyay@knox.edu).

1 Introduction

Population diversity has long been viewed as an important determinant of economic growth and development. A vast empirical literature explores many dimensions of societal differences, including economic inequality, religious, cultural, and genetic diversity, with a special focus on ethnolinguistic divisions in the context of Sub-Saharan Africa.[1] Theoretically, population diversity has both costs and benefits, so its net effect on socioeconomic outcomes is unclear.[2] Given the crucial role of lasting peace in securing economic prosperity, arguably the most important social cost of population diversity is its potential to foster conflict (Arbatlı et al., 2020). This issue is particularly relevant for Africa, a diverse continent consistently displaying high levels of organized violence (Davies et al., 2024).

Despite the growing number of studies linking ethnic divisions to conflict, the overall evidence on the nature of this relationship remains mixed. To some extent, this reflects the multitude of approaches to measuring diversity. For example, while earlier studies focused on the indices of ethnolinguistic fractionalization and polarization, recent work has argued that between-group ("horizontal") ethnic inequality represents a more salient determinant of conflict (Hillesund et al., 2018). Another point of contention has been the choice of relevant group identities and the importance of accounting for cultural proximity in diversity indices (Desmet et al., 2012). Furthermore, the literature has gradually moved away from cross-country to subnational analyses, posing the challenge of constructing reliable metrics at the local level (Gershman and Rivera, 2018).

In this paper, we contribute to the ongoing debate by building a new comprehensive dataset on various dimensions of local ethnolinguistic divisions in Sub-Saharan Africa and revisiting the relationship between diversity and conflict at the subnational level. Our dataset covers 35 countries, 391 first-level administrative regions, and 853 unique ethnolinguistic groups. We use microdata from large-scale surveys to produce ethnic inequality indices capturing group-level differences in education, asset ownership, and access to basic amenities, in addition to standard measures of fractionalization and polarization. Following the literature emphasizing the importance of cultural and linguistic relatedness among groups, we distinguish between deep-rooted and contemporary ethnolinguistic cleavages across our metrics. To do so, we match all groups from survey samples to the *Ethnologue* database and aggregate our data on population composition at different tiers of the global linguistic tree, corresponding to more or less recent hypothesized ancestral languages. This

[1] See, for example, Cuesta and Wantchekon (2016) and Gershman and Rivera (2018).

[2] These trade-offs are reviewed in Alesina and La Ferrara (2005), Montalvo and Reynal-Querol (2014), Ginsburgh and Weber (2020), and Desmet et al. (2025b).

results in a hierarchy of directly comparable diversity indices, from those based on cleavages originating in the distant past to those reflecting contemporary language distinctions.

We combine our new subnational diversity database with georeferenced data on conflict events and fatalities dating back to the late 1990s, along with other regional characteristics. Our regression analysis reveals a positive relationship between ethnolinguistic fractionalization and conflict, which tends to be stronger in magnitude and statistical significance for metrics based on deep linguistic cleavages. After accounting for country fixed effects and numerous potential regional confounders, we find that, across specifications, a one-standard-deviation increase in deep-rooted fractionalization is associated with roughly a 20-35% rise in conflict events and an even larger percentage increase in fatalities. In contrast, we find that none of the ethnic inequality metrics are systematically related to conflict. Even in those cases where a significant positive association is found, it disappears after accounting for fractionalization. Similarly, there is no robust pattern linking conflict to polarization. Our results hold across model specifications and for a variety of conflict measures from alternative data sources.

Our main contribution is twofold. First, we advance the ongoing effort to leverage microdata from large-scale surveys for building standardized datasets on population diversity across countries and regions. The new subnational measures of ethnic inequality offer a broad coverage of Sub-Saharan Africa based on high-quality sources, employ a variety of underlying socioeconomic indicators, and account for ancestral relationships between ethnolinguistic groups. Our general approach to data construction is closely related to Gershman and Rivera (2018) but additionally covers ethnic inequality.[3]

We use the first-level subnational administrative region as our unit of analysis. As further discussed in section 2, this choice offers two main advantages over finer administrative subdivisions or randomly generated spatial units: the feasibility of measuring diversity with reliable survey data and the political, economic, and social relevance of these regions. Importantly, subnational units allow us to focus on within-country variation in diversity and conflict while accounting for nationwide fixed effects. Since we cover a fairly standard set of regions in their current or recent boundaries, our new data are compatible with most existing subnational-level databases and represent a ready-to-use source for future research on diversity.

[3]Fjelde and Østby (2014) is a notable earlier contribution using survey data (the Demographic and Health Surveys) to measure subnational ethnic inequality in Africa. However, the authors focus exclusively on the difference between a region's largest group and the rest of its population. In contrast, we consider all groups residing in a region and offer a more comprehensive coverage of Africa's linguistic diversity by using superior data sources and accounting for relatedness between groups.

Our survey-based approach to measuring population diversity utilizes information on actual individuals and households and differs from the increasingly popular GIS-based analyses that overlay digital maps of population counts, ethnic homelands, and measures of material wealth around the world. Although the GIS approach is simple and flexible, it is associated with multiple layers of measurement error (Gershman and Rivera, 2020). Most importantly, even the best available maps of ethnic or linguistic homelands are inaccurate since they do not account for recent migrations and are generally not designed to capture all cases of overlapping groups, especially in urban areas. This leads to substantial errors in capturing local population composition. In the context of ethnic inequality, one additionally needs fine-grained information on well-being at the ethnic-group level within regions, typically derived from night lights data (Kuhn and Weidmann, 2015; Alesina et al., 2016; Leipziger, 2023). The latter are an imperfect proxy that has been argued to be particularly poor in small and rural areas (Gibson et al., 2020), a major consideration for any analysis across Africa. Furthermore, since this approach relies on a single night lights measure, it cannot distinguish between different dimensions of ethnic inequality.[4]

Although we believe the survey-based approach yields more accurate estimates of regional diversity in Sub-Saharan Africa, it is not immune to measurement error. Privacy concerns and selection issues may affect participation and response accuracy, while sampling strategies may not adequately capture smaller or marginalized groups, limiting representativeness at the local level. Self-identified ethnicity can reflect shifting social dynamics and desirability biases, as individuals may avoid affiliation with stigmatized groups, particularly in times of conflict. In addition, response biases in questions regarding asset ownership, education, and other indicators of well-being cannot be ruled out.

Second, we contribute to the large empirical literature linking ethnolinguistic diversity and conflict.[5] Early cross-country studies (Fearon and Laitin, 2003; Collier and Hoeffler, 2004) found no robust link between fractionalization and conflict, emphasizing the role of polarization instead (Montalvo and Reynal-Querol, 2005). Later, more theoretically grounded work identified positive associations for both measures (Esteban et al., 2012), while Desmet et al. (2012) showed that the role of fractionalization becomes more pro-

[4]We illustrate the main pitfalls of the GIS approach to measuring ethnic inequality in appendix E presenting a detailed case study of Liberia.

[5]More generally, the vast literature on conflict in Africa has explored its numerous other determinants including the location of contemporary country and historical ethnic borders (Michalopoulos and Papaioannou, 2016; Depetris-Chauvin and Özak, 2024), state history (Depetris-Chauvin, 2015), weather shocks (Harari and La Ferrara, 2018; McGuirk and Nunn, 2025), and natural resources (Berman et al., 2017; Adhvaryu et al., 2021).

nounced when based on deep-rooted ethnolinguistic cleavages. Our findings align with this more recent cross-country evidence, suggesting that similar mechanisms operate at the regional level in Sub-Saharan Africa. They also complement the results on the negative link between deep-rooted diversity and local public goods provision (Gershman and Rivera, 2018).

Conceptually, the salience of fractionalization supports the view that long-standing ethnolinguistic divisions *per se* exacerbate coordination failures, mistrust, and intergroup animosity, thereby increasing the likelihood of conflict. It is also consistent with the framework of Esteban and Ray (2011a), which predicts that fractionalization should be the leading driver of conflict when competition centers on private prizes, while polarization should matter more when conflicts revolve around public prizes. In the subnational context, where access to local resources, rents, and offices is paramount, most contests are plausibly of the former type, helping to explain our empirical findings.

The null results for ethnic inequality contrast with the view that socioeconomic differences between ethnic groups, rather than diversity itself, drive conflict (Stewart, 2008). However, as discussed in the following section, while several plausible mechanisms could link ethnic inequality to conflict (Hillesund et al., 2018), the relationship remains theoretically ambiguous. Even if group disparities generate grievances, these do not automatically lead to unrest, as successful mobilization requires both organization and resources. Moreover, some research suggests that economically similar groups may be *more* prone to violence, and that the salience of ethnicity-based conflict depends on inequality within groups (Esteban and Ray, 2008; Ray and Esteban, 2017). Given this conceptual ambiguity and the presence of countervailing mechanisms, our results are not surprising.[6]

Previous empirical work on ethnic inequality has produced mixed results. Some studies find a positive association with conflict (Cederman et al., 2011; Fjelde and Østby, 2014; Houle and Bodea, 2017), while others, like ours, report no significant relationship (Østby, 2008; Huber and Mayoral, 2019).[7] These discrepancies reflect multiple differences in research design. Studies vary in their units of analysis, from grid cells and ethnic homelands to regions and countries, and the mechanisms linking inequality to conflict may not operate uniformly across these scales. Measures of ethnic inequality also differ, including

[6]An important related insight is that perceived inequality and group status do not necessarily align with actual economic gaps and can be politically manipulated (Hillesund et al., 2018), implying that objective measures of ethnic inequality may fail to capture relevant grievances.

[7]See Hillesund et al. (2018) and Hillesund and Østby (2023) for comprehensive reviews. Relatedly, Baldwin and Huber (2010) and Alesina et al. (2016) argue that country-level ethnic inequality is a stronger predictor of, respectively, public goods provision and development than fractionalization.

standard between-group indices, deviations of group means from the national average, and pairwise group comparisons. Approaches further vary in which groups and socioeconomic dimensions are included. Some research combines economic inequality with political exclusion, whereas others, like us, focus purely on socioeconomic disparities. Moreover, many studies do not explicitly compare ethnic inequality with alternative diversity measures, even though they are correlated and may capture overlapping aspects of heterogeneity. Finally, there is substantial variation in data quality and sources used to measure population composition and group-level outcomes.

While differences in empirical design complicate direct comparisons across studies, our approach offers several distinctive advantages. It builds on a newly constructed, high-quality database of diversity measures developed under consistent principles to ensure comparability across countries and indicators. Rather than selecting groups based on subjective relevance, we include all those identified in the underlying surveys and measure their cultural proximity uniformly across metrics using the structure of linguistic relationships. To capture ethnic inequality comprehensively, we rely on multiple socioeconomic indicators rather than a single dimension or a composite index. Finally, by directly comparing correlated measures of inequality, fractionalization, and polarization within the same regression framework, we identify the most salient predictors of conflict. The analysis is conducted at the level of socioeconomically and politically meaningful subnational regions, allowing us to focus on within-country variation.

Together, these features provide a coherent and internally consistent setting for reassessing how different facets of local diversity relate to conflict across Sub-Saharan Africa. Within this setting, our results indicate that the fragmentation of populations into many ethnolinguistic groups, particularly when they are more culturally distinct, is a stronger predictor of conflict than socioeconomic inequality between them. Thus, identity-based factors play a more prominent role in this relationship than group differences in socioeconomic status.

The rest of the paper is organized as follows. Section 2 provides theoretical background and motivates the empirical analysis. Section 3 introduces the new dataset. Section 4 presents our main results and robustness tests. Section 5 concludes. Further details on our dataset, definitions of all variables, and additional analyses are collected in the appendices.

2 Background and motivation

Rather than offering a comprehensive literature review, this section highlights the theoretical aspects most pertinent to our study, namely the lack of consensus on the nature of the relationship between ethnic divisions and conflict, debates around the conceptually relevant metrics of diversity, and the choice of the unit of analysis in empirical applications.

Population diversity has long been regarded as potentially both a source of creative dynamism and a driver of social tensions (Alesina and La Ferrara, 2005; Desmet et al., 2025b). Ethnolinguistic diversity has been argued to foster conflict through several channels. For example, it may erode trust and social cohesion, complicate coordination and collective action, and breed resentment when group markers align with socioeconomic differences. According to complementary "primordialist" and "instrumentalist" perspectives, tensions could arise due to either deep-rooted animosities tied to group identity or strategic activation of such divisions by elites pursuing political or economic goals (Ray and Esteban, 2017). In terms of social interactions between groups in a diverse society, "conflict theory" views them as heightening perceived threats and competition for scarce resources while "contact theory" emphasizes their positive role in reducing prejudice and building trust (Hodler et al., 2020).

2.1 Ethnic fractionalization and polarization

Much of the literature focused on how the structure of diversity – the number and relative size of groups – affects the likelihood and intensity of conflict. Early work centered on ethnolinguistic fractionalization, capturing the probability that two randomly drawn individuals in a region belong to different groups. The intuition was that highly fragmented societies, consisting of many small groups, may struggle to reconcile their distinct views, preferences, and beliefs through peaceful means of cooperation and communication, which makes them more susceptible to conflict. Furthermore, if intergroup antagonism is strong, frequent cross-group encounters may heighten friction and thus the incidence of conflict. However, just as it may prevent peaceful coordination, high fractionalization may also disperse potential alliances, raise costs for mounting a rebellion, and thus dampen conflict (Montalvo and Reynal-Querol, 2014). In addition, within a given country, a homogeneous but ethnically distinct region may have a higher incentive to secede (Desmet et al., 2025a).

In a classic contribution, Horowitz (1985) argued that conflict risk is highest not in fragmented societies, but in polarized ones where a large majority coexists with a sizable minority. In such contexts, the presence of a few large, cohesive, and mutually antagonistic groups facilitates collective mobilization through strong in-group solidarity and reduced free-riding. The polarization index and its relation to conflict have been formally investigated in the work of Esteban and Ray (1994, 1999, 2011a) and Montalvo and Reynal-Querol (2005), among others.

Unlike studies focused on specific diversity measures, the game-theoretic framework of Esteban and Ray (2011a) integrates fractionalization, polarization, and inequality within a unified model in which equilibrium conflict intensity depends on all three. In their setting, groups invest resources in a contest to seize a prize that may be public or private. Public prizes, such as political control, religious dominance, or the enforcement of collective rights, are shared equally among group members, whereas private prizes, such as access to material resources or offices, are divisible and diluted across them. The type of good at stake determines which dimension of diversity drives conflict: polarization tends to dominate when contests are over collective or identity-based goods, while fractionalization is more relevant when they concern material resources.

The Esteban-Ray approach to measuring polarization also underscores the role of intergroup distances in preferences, operationalized as linguistic proximity in their follow-up empirical work (Esteban et al., 2012). While this factor enters their model in a very specific way, other work has highlighted cultural proximity and the historical depth of ethnic divisions more broadly (Fearon, 2003; Desmet et al., 2009, 2012; Gershman and Rivera, 2018). For example, the degree of cultural proximity may mediate communication, trust, and reciprocity: shared language and history facilitate cooperation and enable sanctioning mechanisms that sustain it. In contrast, deep cleavages, corresponding to ancient linguistic splits and long histories of separation, have been argued to underpin enduring antagonism and conflict (Desmet et al., 2012).

2.2 Ethnic inequality

A parallel strand of research extends the diversity-conflict nexus by focusing on horizontal inequality defined as systematic disparities in socioeconomic outcomes across identity groups. The basic intuition is straightforward: when material advantage overlaps with ethnolinguistic divisions, coordination problems, mistrust, and grievances are amplified.

This approach dates back to the theory of relative deprivation (Gurr, 1970), which links collective violence to frustration arising when groups compare their circumstances to others and perceive systematic disadvantage. Horizontal inequalities are particularly dangerous if they couple material deprivation with strong collective identities that facilitate mobilization. Hillesund et al. (2018) summarize this logic as resting on three interlinked conditions: identity, motive, and opportunity. Identity defines the boundaries within which grievances are shared and organization becomes feasible; inequality provides the motive by generating perceptions of injustice; and opportunity shapes whether those grievances are expressed peacefully or through violence (Østby, 2008; Cederman et al., 2011).

Despite its appeal, the idea that ethnic inequality fuels conflict has been challenged. Some research suggests that violence may be most likely among groups that are economically similar rather than far apart: poorer groups often lack the resources to rebel, while rising prosperity enhances their capacity for mobilization (Ray and Esteban, 2017). Conflict can also arise when formerly dominant groups perceive the improving fortunes of their rivals as a threat to their status, making growing *equality* the source of tension. The salience of ethnic conflict may further depend on inequality within groups (Esteban and Ray, 2008, 2011b; Kuhn and Weidmann, 2015; Huber and Mayoral, 2019). Internally homogeneous groups may lack elites capable of financing and coordinating rebellion, whereas internally stratified ones can combine resources and manpower more effectively. Yet internal inequality may also blur between-group contrasts, dampening perceptions of injustice and the motivation to fight (Hillesund et al., 2018).

The conceptually relevant metric of ethnic inequality has also been debated. One standard measure is a between-group inequality index based on mean outcomes of all groups (Baldwin and Huber, 2010; Alesina et al., 2016). Another, more closely aligned with the idea of relative deprivation, compares each relevant group's average income to a reference level such as the national mean (Cederman et al., 2011). Some studies focused on inequality between two largest groups (Østby, 2008), or between a dominant group and the rest of the population (Fjelde and Østby, 2014). A further issue concerns the underlying socioeconomic indicator, which we discuss in section 3. Finally, although the degree of cultural distance between groups may be theoretically important for the reasons outlined earlier, such adjustments are absent from existing ethnic inequality measures, an omission we address explicitly in our analysis.

Taken together, the theories reviewed above suggest that ethnic fractionalization, polarization, and inequality may each capture distinct yet complementary dimensions of the relationship between diversity and conflict. Overall, existing theory remains ambiguous about which of these dimensions matters most – or even whether diversity is expected to exacerbate or dampen conflict – since the underlying mechanisms can operate in opposing directions. The depth and salience of ethnic cleavages further complicate the picture, as they are theoretically relevant across all three frameworks. Given these uncertainties, it is unsurprising that the literature has increasingly turned to empirical analysis. Yet, as discussed in the introduction, these studies have produced mixed results, in part due to differences in geographic scope, measures, data sources, and units of analysis.

2.3 Unit of analysis

Theoretical accounts linking ethnic divisions to conflict are largely agnostic about the spatial scale at which such relationships operate, and many of the underlying mechanisms can manifest at the national, regional, or more local level. Some frameworks, however, suggest scale-specific relevance. In the model of Esteban and Ray (2011a), for example, conflicts over political control, language, religion, or statewide rights resemble contests for public prizes, where polarization is expected to matter most, and pertain more naturally to the national arena. By contrast, disputes over resources or administrative control correspond to private prizes and are typically localized, implying that fractionalization should play a greater role at the subnational level.

Other mechanisms may also vary by scale. Processes such as trust erosion or failures of collective action, often emphasized in the fractionalization literature, unfold within communities and local governance structures, while horizontal inequalities linked to differential access to power are mainly national in scope. At the same time, country-level measures risk obscuring the variation in grievances and mobilization opportunities that often arise within particular groups or regions. For instance, a national index of ethnic inequality may conceal the presence of a single marginalized group capable of sustaining localized conflict, while an aggregate count of conflict events may mask their concentration in specific areas – an empirical pattern far more typical than statewide wars (Hillesund et al., 2018). Furthermore, most interactions among individuals are inherently local, and reference groups for social comparison are more likely to be nearby communities than an abstract national average.

Empirical work has evolved in response to these considerations and featured various units of analysis including countries, subnational regions, ethnicities and their homelands, ethnic dyads, and grid cells (Hillesund and Østby, 2023; Desmet et al., 2025b). Early studies examined cross-country associations between diversity and conflict, largely for reasons of data availability. More recent research has increasingly adopted subnational perspectives, sometimes focusing on specific countries. Moving below the national level not only better aligns with some of the underlying mechanisms but also allows for the inclusion of region-specific covariates and country fixed effects to account for nationwide factors such as institutions and history. Because ethnic diversity changes only slowly over time, even at the local level (Gershman and Rivera, 2018), country fixed effects are difficult to incorporate meaningfully in national panel analyses, making regional data particularly useful for isolating within-country variation.

Subnational analysis, however, presents its own challenges. Administrative divisions are endogenous and may be redrawn in part to reduce conflict or reshape local ethnic composition (Bluhm et al., 2025). In Sub-Saharan Africa, this issue is partially mitigated by the fact that many provincial and district boundaries were originally established during the colonial period (Justin and Vries, 2019). These internal subdivisions, much like state borders, were often created with little regard for local ethnic geography, even though they may have influenced subsequent spatial distribution and self-identification of ethnic groups (Müller-Crepon, 2025). Using artificial units, such as grid cells, alleviates the border endogeneity problem, but introduces other complications. First, diversity and other relevant variables must be constructed using GIS tools, whose limitations in this context were discussed earlier. Second, artificial units are arbitrary in size and shape, with no theoretical guidance on which spatial scale is meaningful and why. This reflects the modifiable areal unit problem – the sensitivity of statistical relationships to the type and configuration of spatial units.

Our choice of the first-level subnational administrative divisions in Sub-Saharan Africa is guided by both conceptual and practical considerations. It strikes a balance between the availability of reliable data and the socioeconomic and political relevance of units. They are the smallest divisions for which consistent survey-based information on population composition and socioeconomic characteristics exists across a large number of countries. Finer subdivisions would either restrict cross-country coverage or reduce representativeness. This level of aggregation also aligns well with the scale at which decisions about public goods, such as education, health, and infrastructure, are made, and where local authorities often have substantial influence. Despite the legacy of artificially imposed subnational borders, these units have since become politically salient demarcations, sometimes structuring demands and mobilization by local ethnic groups (Cunningham and Weidmann, 2010). In many African countries, regional executives or councils are directly elected, and where this is not the case, these regions nonetheless function as key arenas for fiscal distribution and administrative control. Taken together, these features make first-level subnational regions an analytically coherent and empirically tractable unit of analysis for examining the relationship between ethnic diversity and conflict.

3 Data

3.1 Ethnolinguistic groups and socioeconomic indicators

Our dataset on local population diversity covers 391 first-level subnational administrative regions across 35 countries in Sub-Saharan Africa.[8] For each country, we selected a primary source of microdata based on three main priorities. First, we preferred surveys containing more detailed listings of ethnolinguistic groups. Second, given our goal of measuring regional ethnic inequality, each chosen survey had to contain relevant socioeconomic indicators for individuals or households. Third, we aimed to maximize sample size and representativeness at the subnational level. Overall, our microdata come from three sources: census subsamples from the Integrated Public Use Microdata Series (IPUMS), Demographic and Health Surveys (DHS), and Multiple Indicator Cluster Surveys (MICS). Both the average and the median year of observation for population diversity metrics are around 2010, with some variation dictated by data quality and availability constraints. The countries in our sample, along with respective primary sources of microdata, are listed in the first two columns of appendix table B.1 (the distinction between primary and secondary sources is explained below). Note that all surveys directly report the subnational region of residence for each respondent.

We follow the methodology of Gershman and Rivera (2018) and standardize the notion of a unique ethnolinguistic group by matching the self-reported identities from original surveys to corresponding spoken language codes in the *Ethnologue* database (Lewis, 2009). This process involved two steps. First, we examined the original "raw" lists of groups in each survey and discarded entries that clearly do not represent native ethnicities. These included geographic and regional markers, nationalities, European languages (unless representing native groups like English in South Africa), various "missing data" and "other" categories, and a few further miscellaneous identifiers (e.g., "foreigner" or "from different parents").[9] Column 5 of table B.1 provides the counts of original groups for each survey – after weeding out the above categories – that likely represent ethnicities.

In the second step of the matching process, we assigned *Ethnologue* language codes to these groups. We were able to "standardize" the vast majority of survey groups – in most

[8] We use the same set of regional boundaries as Gershman and Rivera (2018). See figure 2 for a map.

[9] Notably, the 2013-14 DHS dataset for the Democratic Republic of the Congo delineates "ethnic groups" based on geographic locations making this survey unusable.

cases, based on a one-to-one match.[10] However, as can be seen from comparing the counts of standardized and original groups in columns 4 and 5, respectively, of table B.1, in a substantial number of cases, multiple survey groups were assigned the same *Ethnologue* code resulting in a shorter list of standardized ethnicities. This happened for two reasons: variations in spelling and alternate names of the same ethnic groups and the presence of multiple dialects or subgroups that are not distinctive enough linguistically to have their own *Ethnologue* code.[11] Overall, our process resulted in a total of 853 unique ethnolinguistic groups pinned down by distinct *Ethnologue* codes.

Although linking ethnic groups to standardized language classifications greatly improves cross-country comparability, it may not capture finer distinctions within linguistically homogeneous populations. In some contexts, groups sharing a language maintain distinct identities arising from historical and social divisions or based on subgroups such as clans or language dialects. While such finer cleavages may be important, there are no consistent data or objective standard for identifying and coding them across a large set of countries.

In contrast to some earlier research that focused only on "politically relevant" ethnic groups (Posner, 2004; Cederman et al., 2011), we placed no such restrictions.[12] Relatedly, in an insightful recent paper, Guarnieri (2025) shows that an increase in the cultural (linguistic) distance to the central government raises the propensity of politically relevant ethnic groups to fight over power. This result underscores that grievances may depend not just on regional characteristics, but also on the cultural affinities with ruling elites. While this is an important mechanism for understanding rebellions, we focus instead on how regional diversity and socioeconomic inequalities across all identifiable groups relate to local conflict after accounting for their linguistic proximity. The latter is made possible by our matching process that incorporates all groups in our dataset into *Ethnologue*'s language tree model, which captures the ancestral relationships between languages and allows us to

[10]The main source of unmatched groups is the Nigeria DHS (2013) survey with 318 original entries, of which we were unable to classify 54. All of them, however, are very small and altogether represent about 1.5% of the country sample. We erred on the side of caution and did not assign *Ethnologue* codes unless we were confident in the match.

[11]For example, the Bariba and Berba of Benin are both alternate names of the Baatonum language [bba]. In Cameroon, Yezoum, Mvele, Fong, and Yebekolo correspond to dialects of the Ewondo language [ewo]. Similarly, Jipal, Kwalla, and Kofyar are all dialects of the Kofyar language [kwl] of Nigeria.

[12]As discussed by Hillesund et al. (2018), a focus on politically relevant groups in the study of conflict is conceptually problematic. First, this approach excludes not-yet-mobilized groups that may nonetheless be important for assessing the risk of conflict. Second, the coding of "relevant" groups is often post-hoc, that is, based on observed conflict events, which introduces bias in favor of finding a significant pattern. Our approach avoids these potential pitfalls.

distinguish between recent and deep-rooted ethnolinguistic cleavages when constructing our diversity measures, as described in section 3.2.

We faced the following trade-off when choosing between IPUMS and DHS/MICS as microdata sources. On the one hand, IPUMS data typically provide both much larger population samples and more detailed listings of ethnolinguistic groups (if this information was collected in respective censuses). On the other hand, DHS/MICS tend to report a wider range of socioeconomic outcomes that can be used to construct a larger variety of ethnic inequality metrics. Given our priorities, we relied on IPUMS as the primary data source for 14 countries in our sample, with the remaining 21 covered by DHS or MICS. Whenever IPUMS did not report some of the socioeconomic outcomes of interest, we drew upon DHS/MICS alternatives to compute corresponding ethnic inequality indices. For example, some of the IPUMS surveys report only a subset of asset ownership indicators, in which cases we employ DHS/MICS. We refer to these auxiliary data sources as "secondary" and list them in the third column of appendix table B.1.

As column 4 of the same table makes clear, our primary and secondary data sources sometimes provide a substantially different classification of ethnolinguistic groups. To take the most extreme example, for Ghana, DHS and IPUMS record 8 and 38 standardized groups, respectively. The main reason for this type of discrepancy is that DHS surveys occasionally aggregate multiple related groups into a single ethnolinguistic cluster or simply record smaller groups as "other." The discrepancies may also reflect true sampling issues, namely, that larger-scale IPUMS are able to better capture non-populous groups relative to the less expansive DHS. Although one has to be mindful of these issues, it is not a major concern since the vast majority of our indicators are well-covered by primary data sources and we focus on those in the main text.[13]

We aimed to capture a variety of socioeconomic divisions between groups. As reviewed by Leipziger (2023), there is no universally accepted "best" measure of horizontal ethnic inequality, and previous research looked at its various economic and social dimensions. The former capture income and wealth that are often associated with the ownership and consumption of private goods, whereas the latter include education and health that typically reflect access to locally provided public goods in a developing economy setting. Exploiting the availability of multiple outcomes in our survey data, we construct different ethnic inequality metrics without making a priori assumptions regarding the most prominent sources

[13] The last row of appendix table D.2 reports the number of subnational regions covered by primary data sources for each diversity index.

of societal divisions.[14] Specifically, following the above classification, our "economic" indicators include binary measures of asset ownership (television, car, radio, flush toilet), whereas the "social" indicators are represented by educational attainment (in years). In addition, we look at household access to electricity and high-quality water sources.[15] Indicators that are measured on a categorical scale in the original surveys (such as source of water, type of latrine, or flooring material) are converted into binary variables using the "high quality" standard from Smits and Steendijk (2015). For example, we consider the binary measure of access to flush toilet (contrasted with pit latrines or other lower-quality options). Detailed definitions of all employed socioeconomic indicators are available in appendix A.

Our approach differs from the practice of aggregating socioeconomic characteristics into indices of overall household wealth using principal component analysis or related techniques (McKenzie, 2005; Smits and Steendijk, 2015). We are mainly motivated by the conceptual differences across outcomes outlined in the previous paragraph, but also the greater transparency and interpretability of resulting ethnic inequality measures. A composite wealth index combines heterogeneous domains, making it difficult to interpret what the resulting inequality metric number actually reflects. Different aspects of well-being may not move together, and computing separate metrics respects the multidimensional nature of inequality rather than collapsing it into a single number. In addition, wealth indices are typically sample-specific, which hinders cross-country or temporal comparisons, whereas outcomes such as electricity access or car ownership are more directly comparable. Index construction also involves various methodological issues, such as the selection of relevant components and an appropriate weighting scheme, that may substantially affect resulting metrics (McKenzie, 2005; Howe et al., 2008). From a data availability standpoint, not all surveys report the same outcomes or ethnic classifications, which further complicates the construction of a uniform aggregate measure of group-level well-being when multiple sources are combined. Finally, the extent of agreement between aggregate indices and conventional measures of socioeconomic position remains unclear (Howe et al., 2008; Filmer and Scott, 2012).

[14] Alcorta et al. (2018) argue that grievances triggered by educational inequality are more likely to incentivize collective action, while differences in asset ownership may have the opposite effect by limiting the ability of disadvantaged groups to rebel.

[15] We also examined but ultimately dropped the indicators of refrigerator, motorcycle, and bicycle ownership, primary and secondary school attendance, basic literacy, underweight status of children, and share of home births. These outcomes were either highly correlated with other indicators on our list or had poor availability (or both).

3.2 Measuring diversity

We focus on three prominent diversity measures: fractionalization, polarization, and between-group ethnic inequality. The index of ethnolinguistic fractionalization reflects the degree to which a region is comprised of many small groups. It is calculated as follows:

$$\text{ELF} = 1 - \sum_{i=1}^{N} s_i^2,$$

where s_i is the population share of group $i \in \{1, ..., N\}$ and N is the total number of groups. This value can be interpreted as the probability that two randomly chosen individuals from a given region belong to different ethnolinguistic groups. Fractionalization is highest when each resident of the region has a different ethnicity.

As mentioned in the introduction, some research has argued that ethnolinguistic polarization (ELP) is a more relevant predictor of conflict than ELF. Following Montalvo and Reynal-Querol (2005), we calculate the ELP index as follows:

$$\text{ELP} = 4 \sum_{i=1}^{N} s_i^2 (1 - s_i).$$

Unlike ELF, it captures the extent to which population is divided into large groups and reaches the maximum of 1 when a region is evenly split between two ethnicities.

Both ELF and ELP indices are based solely on the population shares of ethnic groups and incorporate no information on socioeconomic differences between them. The presence of such divisions is captured by ethnic inequality metrics. Following Baldwin and Huber (2010), we employ the between-group Gini index which may be calculated for any relevant socioeconomic outcome w. Suppose the mean outcomes for a pair of groups i and j are given by $\bar{w}_i$ and $\bar{w}_j$, respectively, and let $\bar{w}$ be the mean outcome for the entire region. Then, between-group inequality (BGI) index is defined as

$$\text{BGI} = \frac{1}{2\bar{w}} \sum_{i=1}^{N} \sum_{j=1}^{N} s_i s_j |\bar{w}_i - \bar{w}_j|.$$

The resulting measure may be interpreted as the (normalized) expected absolute difference in values of outcome w between two randomly chosen residents of a given region, assuming w is evenly distributed *within* each group. As explained in the previous section, we use a variety of underlying socioeconomic indicators available in our survey microdata to build corresponding BGI indices.

For most outcomes, population shares and group averages are calculated from household-level data, and ethnolinguistic groups are defined based on self-reported identities of household heads. For years of education, BGI is computed using individual-level data on adults aged 15-49 and respective ethnicity (or, in some cases, native language) identifiers. Both ELF and ELP indices are also calculated from individual-level data.[16]

The indices introduced so far treat all groups as equally distinct from an ethnolinguistic perspective. However, in reality, groups may be more or less linguistically and culturally distinct depending on the extent of shared ancestry. This may naturally affect their mutual interactions. To account for this nuance, previous studies have incorporated the information on linguistic relatedness in standard diversity measures. Here, we implement the method used by Desmet et al. (2012) and Gershman and Rivera (2018) in the national and subnational context, respectively. Specifically, we use *Ethnologue*'s linguistic tree model to aggregate basic ethnolinguistic groups to the levels of their more and less recent hypothesized ancestral language groups and compute diversity indices at each such level of aggregation $k = 1, \ldots, 13$.[17] The most disaggregated level, $k = 13$, corresponds to contemporary ethnolinguistic groups and their languages, that is, the standardized identities reported in our sample. In contrast, at the most aggregated level, $k = 1$, all ethnicities are grouped into just six major language families of the African continent: Afro-Asiatic, Niger-Congo, Nilo-Saharan, Creole, Indo-European, and Khoisan.[18] Higher aggregation levels (lower values of k) correspond to deep-rooted cleavages that emerged in the distant past, whereas standard diversity measures for $k = 13$ reflect contemporary language distinctions.

This procedure yields 13 versions of each diversity index. We denote them as ELF(k), ELP(k), and BGI(k) referring to the level of linguistic aggregation in parentheses.[19] While subnational ELF(k) and ELP(k) metrics for Sub-Saharan Africa were calculated and explored by Gershman and Rivera (2018), to the best of our knowledge, the present paper is

[16]MICS only provide the information on ethnicity and/or native language of household heads and we assume the same identity for other members of the same household.

[17]The presence of 13 aggregation tiers is dictated by *Ethnologue*'s delineation of ancestral groups for languages in our sample.

[18]The latter three families are very small and largely represent several "mixed" languages of West Africa (Creole) and indigenous South African groups (Khoisan). The only local Indo-European language is Afrikaans, and we also include small English-speaking groups native to South Africa, Botswana, and Namibia.

[19]Appendix C describes the process of constructing BGI(k) indices in detail using a specific region from our sample. Selected summary statistics for all diversity indices are reported in appendix table D.2.

the first to introduce subnational BGI(k) indices.[20] Hence, in the rest of this section, we focus on this part of our local diversity dataset. More specifically, we describe four groups of BGI indices based on educational attainment, access to electricity, ownership of television and car, which fairly represent the whole spectrum of ethnic inequality metrics in our dataset. The former two indices largely reflect inequality in access to local public goods, whereas the latter are based on the ownership of private household durables. According to a classification discussed earlier, education-based BGI represents the "social" dimension of ethnic inequality, while the other metrics focus on its "economic" dimension.

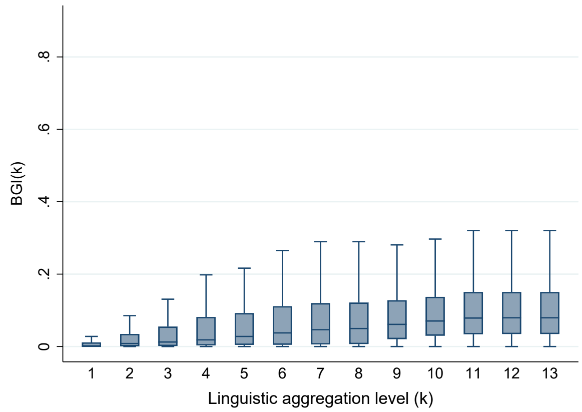
Figure 1 summarizes the distributions of BGI(k) indices for the four underlying indicators in the form of box-and-whiskers plots, sorted on the horizontal axis by the level of linguistic aggregation. Generally, BGI indices are lower for smaller k, mechanically reflecting the reduction in the number of distinct groups at upper tiers of the linguistic tree.[21] To take the most extreme example, for over a third of regions in our sample, all contemporary ethnic groups belong to a single major language family, that is, BGI(1) is trivially equal to zero. However, for larger k, we observe substantial variation in ethnic inequality indices.

Interestingly, the distributions of BGI(k) for educational attainment are more compressed compared to other indicators. On the other hand, BGI for car ownership has both the widest range and the highest median. In general, we see a higher degree of inequality in the ownership of assets that may be considered luxurious in a low-income country setting, such as car, flush toilet, and finished flooring, as opposed to more accessible television and radio. For example, average BGI(13) index for the former three categories is close to 0.3, whereas it is 0.2 for television and just 0.07 for radio (see appendix table D.2).

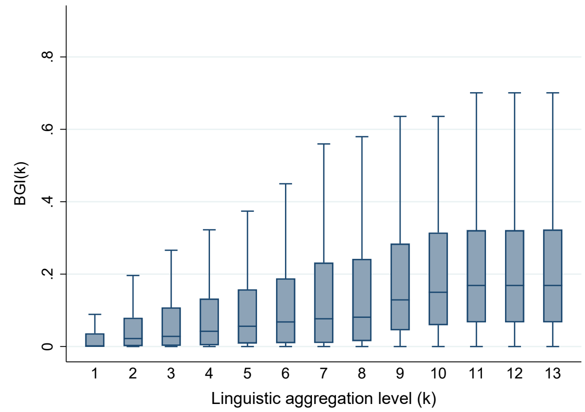
Figure 2 presents the spatial distributions of BGI(k) in electricity access for $k = 1, 5, 9, 13$. To clearly illustrate regional variation for different levels of linguistic aggregation, each panel of this figure uses graduated colors based on deciles of respective index values. It is clear from comparing the legends across panels that measured ethnic inequality is lower at higher aggregation levels, when the number of distinct groups per region is smaller. Importantly, even at a relatively high level of aggregation $k = 5$, there is a lot of variation across regions, despite the fact that numerous language groups (such as all Bantu speakers) are "merged" into one. Generally, the observed variation in BGI captures both population fragmentation into groups and their differences in electricity access. Regions with higher ethnic inequality in one outcome (e.g., electricity access) are also typically more unequal

[20]Hodler et al. (2020) develop an index of "ethnic stratification" that generalizes the Gini coefficient of inequality by weighting economic differences between pairs of individuals by their linguistic distances. This is in contrast to the BGI index that compares group means. Using the Afrobarometer data, they show that ethnic stratification measured at the town or village level is negatively related to trust.

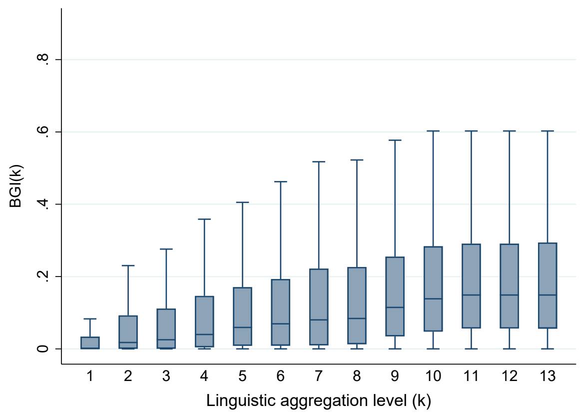
[21]Appendix C contains a formal proof for a simplified case that BGI(k) is non-decreasing in k.



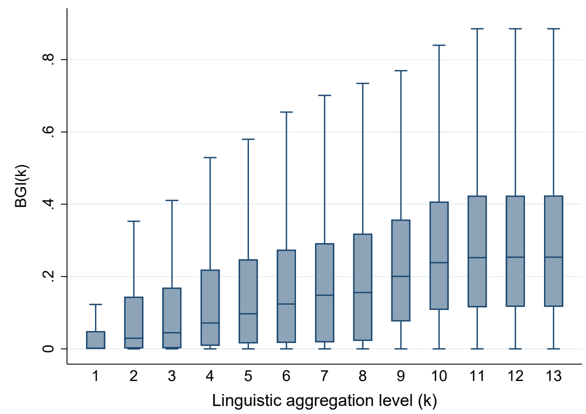
(a) Years of education



(b) Access to electricity



(c) Television ownership



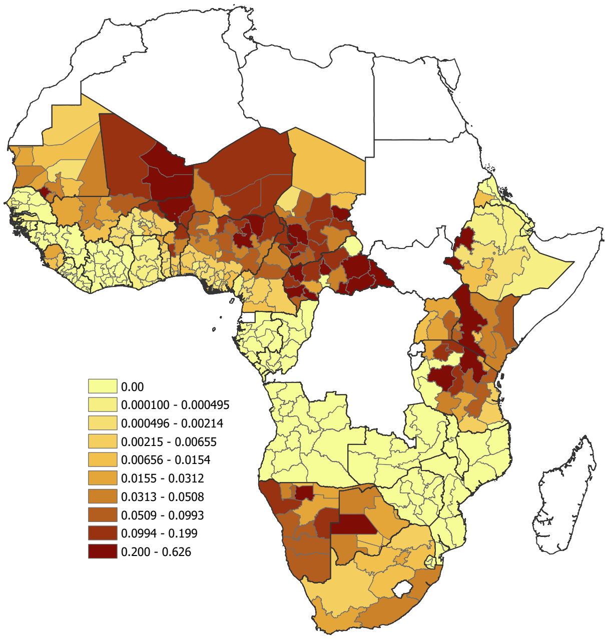
(d) Car ownership

Figure 1: Box-and-whiskers plots for selected BGI(k) indices.

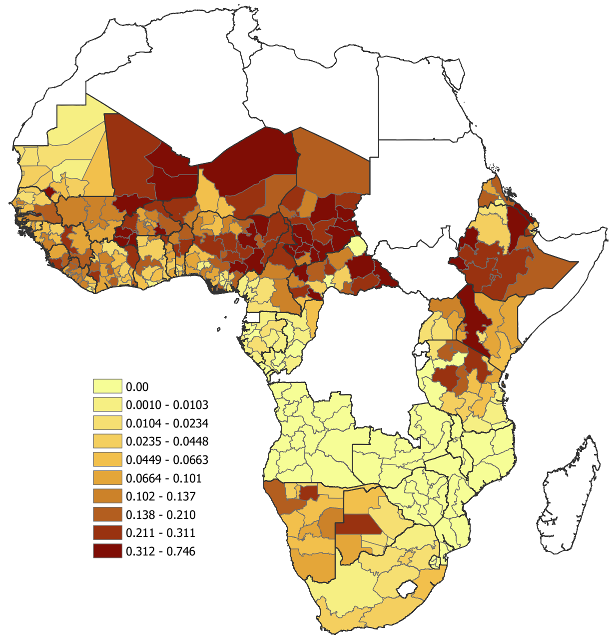
Notes. The following characteristics of the respective distributions are shown: lower quartile (bottom of the box), median (solid line segment), upper quartile (top of the box) and the lower and upper adjacent values.

in other dimensions. As shown in the appendix table D.3, pairwise correlation coefficients for BGI(k) in access to electricity and years of education are in the 0.2–0.6 range, with roughly similar interval for ethnic inequality in car ownership. The correlation is stronger for the indicator of television ownership, falling within the 0.4–0.8 range, which is intuitive, given the dependence of TV usage on electricity access.[22] Not surprisingly, we typically observe the highest pairwise correlations between BGI indices measured at the same level of linguistic aggregation since the number of distinct groups is identical at a given tier k.

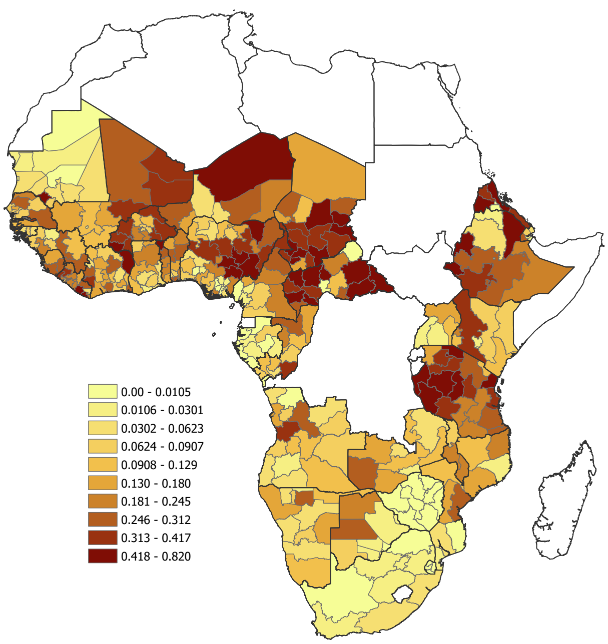
[22]The numbers are similar for access to high-quality water sources, presence of flush toilet, and radio ownership.



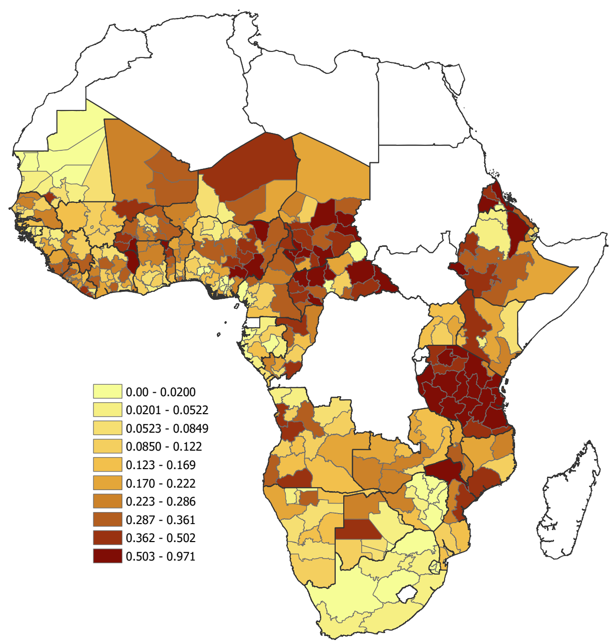
(a) BGI(1)



(b) BGI(5)



(c) BGI(9)



(d) BGI(13)

Figure 2: Spatial distribution of ethnic inequality in electricity access.

Notes. Each panel uses graduated colors based on deciles of respective index values.

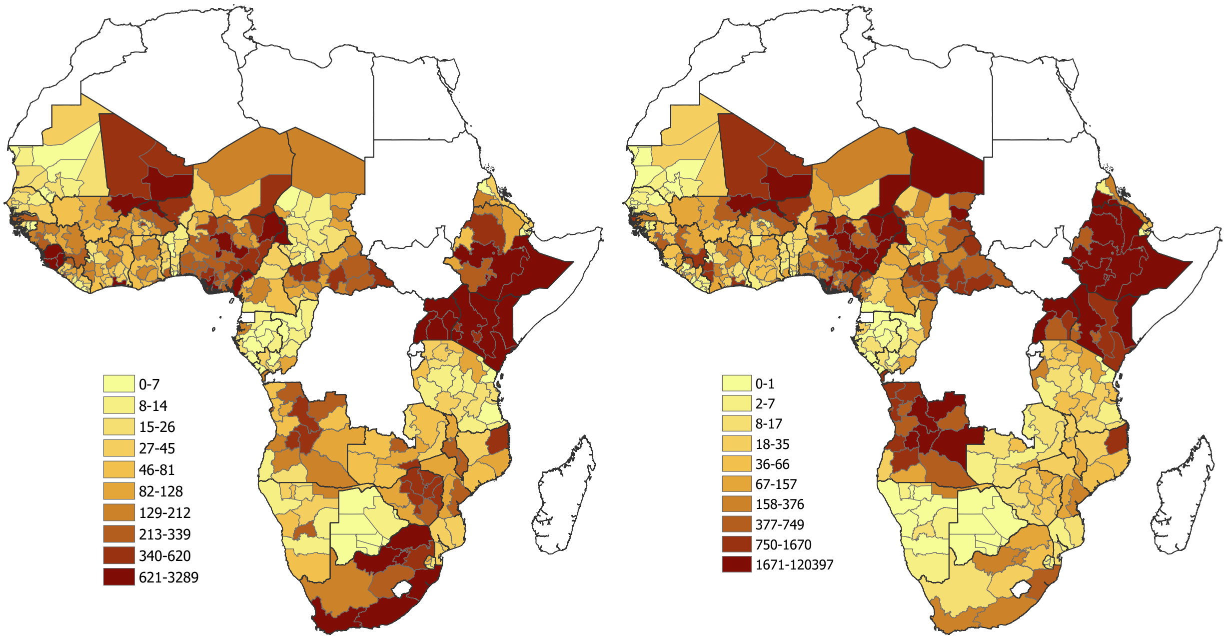
Ethnic inequality measures are moderately positively correlated with ELF at the corresponding tiers, as illustrated in the appendix figure D.3. The pairwise correlation coefficients between ELF and our baseline set of BGI indices fall within the 0.36–0.65 range. The association between BGI and ELP is generally weaker compared to ELF, as shown in figure D.4.[23]

3.3 Measuring conflict

Our primary source of georeferenced data on conflict events and fatalities is the Armed Conflict Location and Event Database (Raleigh et al., 2010), or ACLED, which has been widely used in the literature. ACLED captures a broad range of violent and non-violent actions involving political agents, including governments, rebels, militias, identity groups, political parties, external forces, rioters, protesters, and civilians. It is derived from numerous local, national, and international sources including press accounts, books, research reports, and records from humanitarian agencies. The coverage of Sub-Saharan Africa starts in January 1997 and we use the data up to June 2020. ACLED events are geographically precise to the first subnational administrative level, which makes this source well-suited for our analysis. The dataset also reports the number of fatalities per conflict event. We aggregate the counts of both events and fatalities over the entire time period for each region and use them as outcome variables in our analysis. Figure 3 maps these outcomes, reflecting the frequency and severity of conflict. There is a high within- and cross-country variation in both metrics, and they are mildly positively correlated.

In our robustness analysis, we use two additional data sources that focus on specific types of conflict, namely, the Uppsala Conflict Data Program (Sundberg and Melander, 2013), or UCDP, and the Social Conflict Analysis Database (Salehyan et al., 2012), or SCAD. In contrast to ACLED's broad coverage, UCDP only records deadly incidents associated with civil wars, and each armed conflict in the database resulted in at least 25 battle-related deaths in a calendar year. SCAD, on the other hand, excludes events that are linked to major conflicts and focuses on smaller-scale social unrest including demonstrations, riots, strikes, protests, and communal conflict. Summary statistics for all conflict variables are reported in appendix table D.1.

[23] Figures D.1 and D.2 map the spatial distribution of selected ELF(k) and ELP(k) indices.



(a) Events (b) Fatalities

Figure 3: Regional distribution of conflict events and fatalities from ACLED.

Notes. Each panel uses graduated colors based on deciles of respective counts.

4 Linking regional ethnic diversity and conflict

4.1 Empirical framework

Following Michalopoulos and Papaioannou (2016), we estimate negative binomial models of conflict incidence and severity as functions of local ethnic diversity and other characteristics.[24] Our baseline estimating equation is

$$y_{ic} = \exp(\alpha_c + \beta D_{ic} + \mathbf{X}'_{\mathbf{ic}}\boldsymbol{\gamma} + \varepsilon_{ic}),$$

where y_{ic} is one of conflict measures for region i of country c, D_{ic} is one of diversity indices from section 3.2, $\mathbf{X_{ic}}$ is a vector of regional control variables described below, α_c is the full set of country fixed effects, and ε_{ic} is the idiosyncratic error term. We estimate the model for 13 versions of each diversity index, corresponding to different levels of linguistic aggregation. This way we compare the roles of diversity based on relatively recent versus deep-rooted ethnolinguistic cleavages. The coefficient of interest is β and, given the presence of fixed effects, it is identified using within-country variation.

[24] The estimates from Poisson regressions are qualitatively similar. See appendix F.3 for a subset of our key results reproduced in this alternative framework.

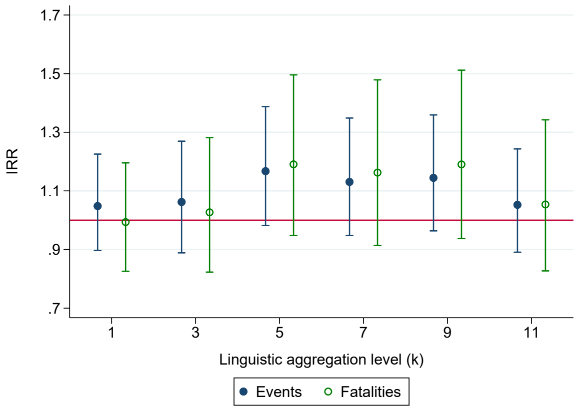
We control for a variety of relevant characteristics that may confound the main relationship of interest. Based on their importance for economic performance, conflict, and ethnolinguistic diversity (Michalopoulos, 2012), we account for both spatial average and standard deviation of agricultural suitability of land using the caloric suitability index of Galor and Özak (2016). We further include a number of geographic controls that have been shown to correlate with socioeconomic outcomes and/or population diversity, including terrain ruggedness (Nunn and Puga, 2012), malaria suitability index (Kiszewski et al., 2004; Cervellati et al., 2019), log of distance to the coast, log of surface area, and absolute latitude of region's centroid (Mitton, 2016). Given recent work on the role of genetic diversity in driving conflict (Arbatlı et al., 2020), we also add the log of distance from Addis Ababa to our list of baseline controls. To account for the potential importance of natural resource endowments (Berman et al., 2017), we include indicators for the presence of oil or gas fields and diamond mines. We also compute an indicator capturing the regional presence of ethnic homelands partitioned by state borders (Michalopoulos and Papaioannou, 2016). Additional, endogenous controls, including metrics of overall regional development and urbanization rate, are employed in robustness checks of section 4.4. Detailed definitions of all variables are provided in appendix A.

4.2 Baseline results

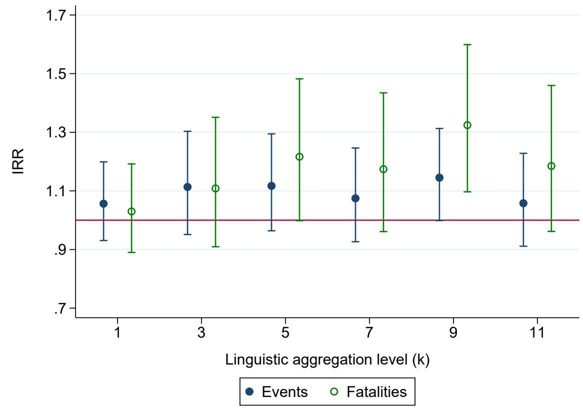
For convenience and visual clarity, we present our results in a compact graphical form similar to Gershman and Rivera (2018). Specifically, for each diversity indicator, we build a diagram showing the estimates of interest at odd-numbered tiers of linguistic aggregation, from 1 to 11 for both conflict outcomes (events and fatalities).[25] For ease of interpretation, we standardize diversity indices to have zero mean and unit standard deviation prior to estimation. We further transform the estimated β coefficients into incidence rate ratios (IRR), which are reported in the diagrams along with respective 95% confidence intervals based on robust standard errors.

Figure 4 presents estimation results for four indices of ethnic inequality introduced earlier. The vast majority of point estimates are positive, and they tend to be somewhat larger in magnitude for conflict fatalities. However, almost none of these estimates are statistically significant at the 5% level, as can be seen from the overlap of the respective confidence intervals and the reference horizontal line at IRR equal to 1 (marking the absence

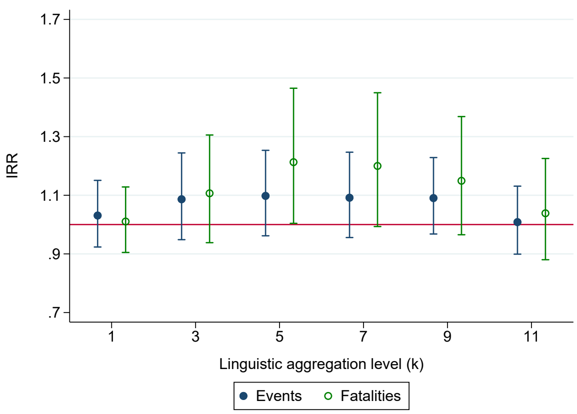
[25]We omit the results for even-numbered tiers to make the figures more transparent. The reported estimates are fully representative of the key patterns. The results for tiers 11 and 13 are virtually identical, and we omit the latter.



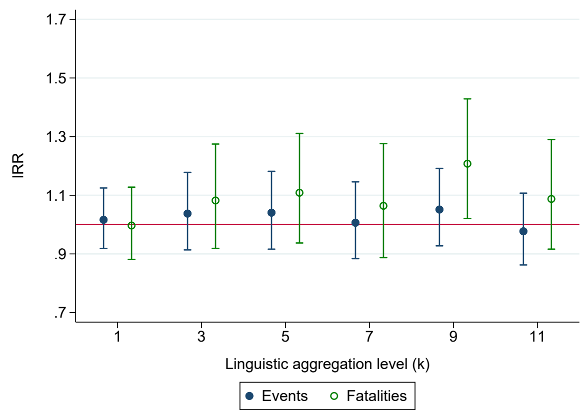
(a) Years of education



(b) Access to electricity



(c) Television ownership



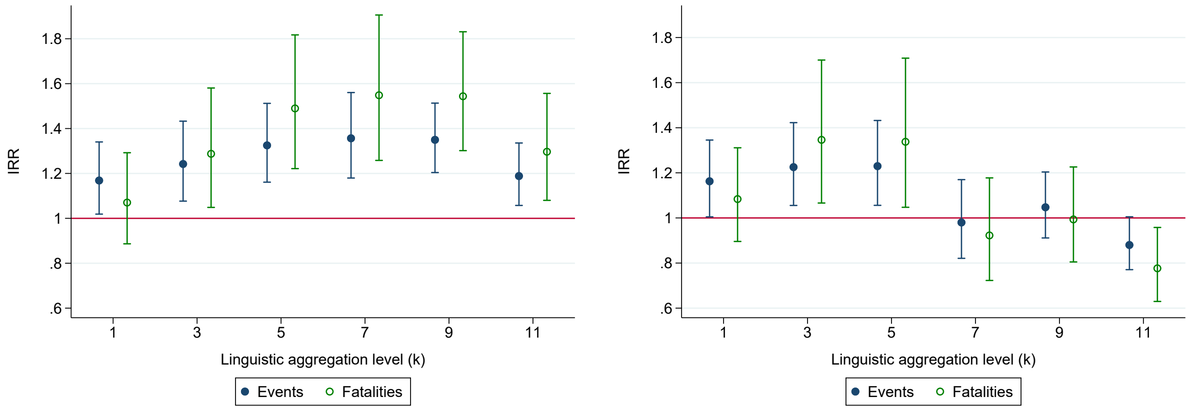
(d) Car ownership

Figure 4: Subnational ethnic inequality and conflict.

Notes. Each panel presents incidence rate ratios, along with 95% confidence intervals, based on robust standard errors. For each of the six reported levels of linguistic aggregation, a negative binomial regression is estimated, where the outcome is the sum of either conflict events or fatalities, and the right-hand-side variable of interest is a BGI(k) index capturing ethnic inequality in the dimension indicated in the figure subtitle. All regressions include country fixed effects and baseline controls described in the main text. The number of observations is 391.

of association). The results for BGI in access to electricity suggest that, other things equal, a one-standard-deviation increase in BGI(9) is associated, on average, with a 15% rise in the number of conflict events and more than a 30% increase in fatalities.

Our results for indices of ethnic inequality based on other indicators, presented in appendix F.2, are largely similar. Just as in the case of car and television ownership, there is no robust significant relationship for BGI in access to other "luxury" goods such as flush



(a) Fractionalization

(b) Polarization

Figure 5: Subnational fractionalization, polarization, and conflict.

Notes. Each panel presents incidence rate ratios, along with 95% confidence intervals, based on robust standard errors. For each of the six reported levels of linguistic aggregation, a negative binomial regression is estimated, where the outcome is the sum of either conflict events or fatalities, and the right-hand-side variable of interest is ELF(k) in panel (a) and ELP(k) in panel (b). All regressions include country fixed effects and baseline controls described in the main text. The number of observations is 391.

toilet and finished flooring, as well as high-quality water sources. Interestingly, ethnic inequality in radio ownership appears to be most strongly related to conflict, at least for some linguistic aggregation levels. However, as shown below, even those few BGI indices that are significant in these baseline regressions lose their “horse races” against ELF.

Figure 5 shows our results for ethnolinguistic fractionalization and polarization. The estimates for ELF are positive, highly statistically significant, and large, especially for $k \in [5, 9]$. At these aggregation levels, a one-standard-deviation greater ELF is associated with about 30-35% higher incidence of conflict events and 50-55% more fatalities. These results suggest that deep-rooted diversity matters more, although the estimates become smaller at tiers 1 and 3, perhaps reflecting the lack of sufficient variation in ELF at those levels.[26]

We also examined potential nonlinearities by re-estimating our baseline specification with both linear and quadratic ELF terms. The quadratic term was statistically insignificant for most values of k. In cases where it was marginally significant, the coefficient was

[26]Formal tests, conducted within the seemingly unrelated regressions framework across various specifications presented in the paper, confirm that IRRs are about 10-15% larger at tiers 5-9 relative to tier 11, a statistically significant difference. The results are weaker for tier 3, with the ratio of IRRs typically above one but statistically insignificant.

positive, suggesting that the expected number of conflict events increases with ELF at an accelerating rate. The predicted-outcomes plots are similar under the linear and quadratic specifications, and the overall evidence of a nonlinear pattern is weak.

The relationship between ELP and conflict is quite different for $k \geqslant 7$, with point estimates mostly insignificant. For higher levels of linguistic aggregation, ELP and ELF results are similar, due to a well-known tendency of these indices to be very highly correlated when the number of groups declines and ELF is low, which is what happens for smaller k (Gershman and Rivera, 2018).[27] In subsequent tests, we generally found that the estimates for ELP at larger k values are not robust and we limit the remaining exposition to the comparative analysis of ELF and BGI in their relation to conflict.[28]

Among the control variables included in our baseline specification, several turned out to be statistically significant throughout the analysis, with coefficient signs largely consistent with earlier studies of conflict in Sub-Saharan Africa. Absolute latitude and the indicator of ethnic partitioning are positively associated with conflict, although the latter is somewhat less robust to model specification in conflict events regressions. On the other hand, distances to Addis Ababa and coastline are negatively related to conflict incidence.

4.3 Fractionalization vs. inequality

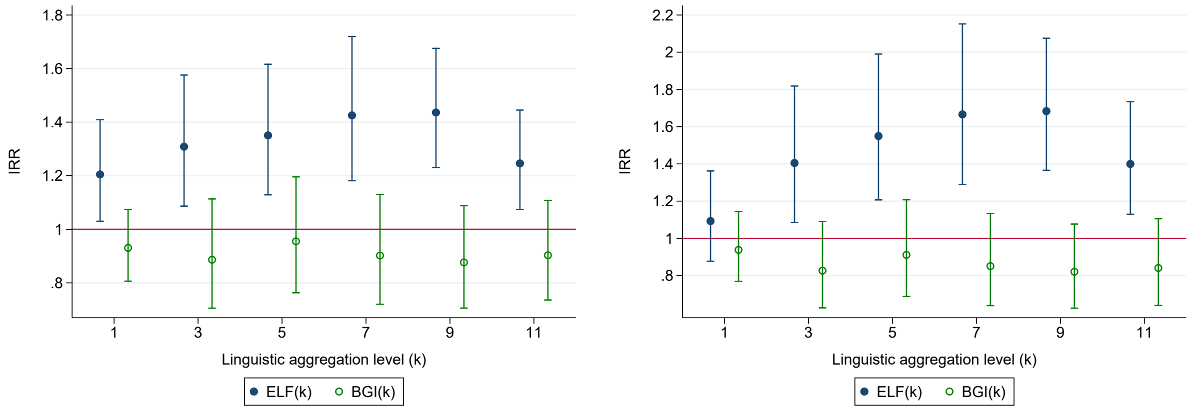
Given our baseline results, we attempt to disentangle the relationship between subnational diversity and conflict focusing on ELF(k) and BGI(k). This is especially important in light of the recent debate regarding the relative importance of these two metrics in predicting socioeconomic outcomes and the substantial positive correlation between them reported in section 3.2. Recall that, while ELF captures the ethnolinguistic fragmentation of population, BGI measures socioeconomic differences between groups.

We conduct this comparison by estimating "horse race" regressions. Specifically, we estimate the same negative binomial model as earlier, but simultaneously include both ELF and BGI indices, measured at the same level of linguistic aggregation. We present the results for BGI in years of education and electricity access, and relegate similar diagrams for other BGI indices to appendix F.4.[29]

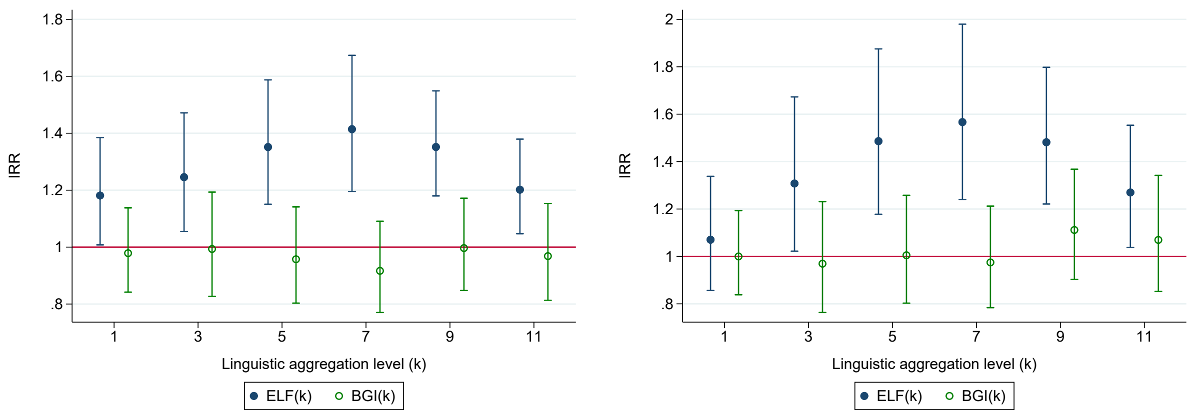
[27] The correlation between ELF(k) and ELP(k) is between 0.88 and 0.99 for $k < 7$.

[28] Appendix F.1 presents regression results for ELF and ELP indices adjusted for linguistic distances (Fearon, 2003; Desmet et al., 2009; Esteban et al., 2012; Gershman and Rivera, 2018), which is an alternative way to account for shared ancestry among different groups. Not surprisingly, these results are similar to our baseline approach.

[29] These supplementary diagrams focus on conflict events as an outcome variable. The results for fatalities are qualitatively similar.



(a) BGI in years of education and ELF as predictors of conflict events (left) and fatalities (right)



(b) BGI in electricity access and ELF as predictors of conflict events (left) and fatalities (right)

Figure 6: "Horse race" regressions for fractionalization and between-group inequality.

Notes. Each panel presents incidence rate ratios, along with 95% confidence intervals, based on robust standard errors. For each of the six reported levels of linguistic aggregation, a negative binomial regression is estimated, where the outcome is the sum of conflict events or fatalities. The right-hand side includes both ELF(k) and BGI(k) for years of education (panel a) electricity access (panel b), along with country fixed effects and baseline controls described in the main text. The number of observations is 391.

Figure 6 presents the incidence rate ratios for ELF and BGI from the "horse race" regressions. The clear "winner" here is ELF, with coefficient estimates becoming even stronger relative to the baseline in panel (a) of figure 5. In contrast, estimates for BGI are mostly negative and/or close to zero, and statistically insignificant across the board. In other words, ethnic inequality in electricity access and other outcomes (figure F.4) appears to be unrelated to conflict incidence or severity after accounting for ELF. The IRRs for ELF are generally larger for $k < 11$ implying that deeper ethnolinguistic cleavages matter

Table 1: Goodness-of-fit measures for baseline regressions

	BGI(1)	ELF(1)	Both	BGI(3)	ELF(3)	Both	BGI(5)	ELF(5)	Both
Pseudo R^2	0.090	0.091	0.091	0.090	0.092	0.092	0.090	0.093	0.093
CW R^2_{DEV}	0.617	0.622	0.622	0.619	0.625	0.625	0.619	0.631	0.631
	BGI(7)	ELF(7)	Both	BGI(9)	ELF(9)	Both	BGI(11)	ELF(11)	Both
Pseudo R^2	0.090	0.093	0.094	0.091	0.094	0.094	0.090	0.091	0.091
CW R^2_{DEV}	0.618	0.631	0.632	0.620	0.636	0.636	0.617	0.623	0.623

Notes. This table reports R^2 measures for 6 groups of negative binomial regressions. The dependent variable is the count of ACLED conflict events and each specification includes the full set of baseline controls including country fixed effects. Models in columns titled BGI(k), ELF(k), and Both include, respectively, only BGI(k) in access to electricity, only ELF(k), and both variables at the same time (a "horse race" regression). Pseudo R^2 is the standard McFadden measure, while CW R^2_{DEV} refers to the R^2 based on deviance residuals (Cameron and Windmeijer, 1996).

more for generating conflict. As mentioned earlier, weaker results for $k = 1$ could simply reflect the lack of substantial variation in diversity when aggregating ethnic groups to the level of major language families.[30]

Table 1 reports two sets of "goodness-of-fit" measures for some of our baseline regressions: regular McFadden's pseudo R^2 and the R^2 measure of Cameron and Windmeijer (1996), based on deviance residuals, that has better properties and a useful interpretation.[31] For each odd tier k between 1 and 11, R^2s are reported for regressions with only BGI(k) in electricity, only ELF(k), and both included in the model. Both sets of R^2s reflect the same findings. First, ELF is better at improving the model fit than BGI. Second, adding BGI to a specification that already contains ELF has virtually no impact on the R^2 values.

In sum, our analysis suggests that ethnolinguistic fragmentation, particularly based on deeper linguistic cleavages, has a much stronger positive association with conflict than ethnic inequality. Identity-based distinctions, rather than the presence of larger group-level differences in socioeconomic status, appear to be a better predictor of conflict.

4.4 Robustness tests

We next show the robustness of our key results. Although our baseline specification controls for a host of relevant regional characteristics and country fixed effects, there may still be

[30]As shown in appendix F.3, while Poisson regression estimates are very similar for conflict events, they are substantially larger in the case of fatalities, especially for $k = 1$.

[31]Specifically, it captures the proportionate reduction in uncertainty, measured by Kullback-Leibler divergence, due to the inclusion of regressors.

other important omitted variables. Tables 2 and 3 show that our qualitative results hold in model specifications accounting for additional factors, whether ELF and BGI enter the equation separately or jointly. Each cell in table 2 represents a different regression, while each combination of ELF(k) and BGI(k) by column in table 3 represents a different "horse race."

The first column of table 2 displays our baseline estimates from conflict events regressions, previously reported in graphical form in figures 4b and 5a. The following three columns augment these specifications by including, one by one, three metrics of average level of development and/or local public goods provision, namely, average literacy rate, electricity access, and international wealth index (IWI) of Smits and Steendijk (2015). This is done to confirm that our results are not driven by the negative relationship between deep-rooted diversity and average access to local public goods (Gershman and Rivera, 2018). The estimates show that coefficients on both ELF and BGI in electricity access are largely unaffected by the literacy and electrification variables. Controlling for IWI in column 4 has a more substantial impact on ELF estimates (without altering the qualitative results), although, due to data availability constraints, the sample size is also reduced by 15 regions.

In column 5, we add the rate of urbanization to our set of baseline controls. Urban areas tend to be regions of highest ethnic diversity and also hubs of development and local public goods provisions. Our estimates for ELF become smaller and ELF(11) loses its statistical significance, but the qualitative pattern remains intact: deep-rooted diversity is positively associated with conflict incidence. Interestingly, controlling for urbanization rate seems to mildly strengthen the BGI estimates.

Finally, the last two columns report estimates after including a regional measure of overall (rather than between-group) inequality in years of education and an index of religious fractionalization from Gershman and Rivera (2018). This has very little effect on the ELF estimates, but some BGI estimates appear stronger after including the "raw" Gini coefficient. All additional control variables, with the exception of religious fractionalization, are statistically significant when included separately, and only Gini coefficient is negatively related to conflict events. However, when all variables are accounted for simultaneously, only urbanization rate retains a statistically significant (positive) association with conflict.

Table 3 shows that our horse race results are likewise robust to alternative model specifications. The first column shows the baseline estimates corresponding to panel (b) of figure 6, while the remaining columns reflect the same sequence of sensitivity tests as in table 2. The dominance of ELF over BGI as a predictor of conflict is clear: the coefficients on ELF(3)–ELF(9) are sizable and statistically significant, whereas BGI is insignificant in every single regression.

Table 2: Robustness to alternative specifications, separately for ELF and BGI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ELF(1)	1.169**	1.221***	1.189**	1.123	1.140*	1.184**	1.167**
	(0.026)	(0.005)	(0.017)	(0.113)	(0.074)	(0.017)	(0.028)
ELF(3)	1.242***	1.292***	1.261***	1.216***	1.217***	1.256***	1.237***
	(0.003)	(0.000)	(0.002)	(0.009)	(0.008)	(0.002)	(0.003)
ELF(5)	1.325***	1.361***	1.330***	1.278***	1.254***	1.326***	1.323***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
ELF(7)	1.357***	1.377***	1.352***	1.288***	1.258***	1.354***	1.355***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)	(0.000)
ELF(9)	1.350***	1.365***	1.340***	1.296***	1.256***	1.348***	1.350***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ELF(11)	1.188***	1.191***	1.168***	1.122*	1.093	1.185***	1.183***
	(0.004)	(0.003)	(0.010)	(0.058)	(0.151)	(0.004)	(0.006)
BGI(1)	1.056	1.079	1.072	1.071	1.070	1.050	1.060
	(0.396)	(0.255)	(0.287)	(0.302)	(0.318)	(0.456)	(0.374)
BGI(3)	1.113	1.154*	1.135	1.133	1.147	1.138*	1.112
	(0.181)	(0.086)	(0.121)	(0.134)	(0.101)	(0.132)	(0.190)
BGI(5)	1.117	1.160*	1.154*	1.149*	1.143*	1.139*	1.111
	(0.141)	(0.057)	(0.064)	(0.075)	(0.087)	(0.098)	(0.160)
BGI(7)	1.075	1.130	1.119	1.078	1.110	1.110	1.068
	(0.341)	(0.124)	(0.141)	(0.333)	(0.165)	(0.183)	(0.383)
BGI(9)	1.145*	1.207**	1.205***	1.181**	1.195***	1.177**	1.141*
	(0.052)	(0.010)	(0.008)	(0.018)	(0.009)	(0.026)	(0.060)
BGI(11)	1.058	1.109	1.118	1.094	1.106	1.083	1.051
	(0.460)	(0.194)	(0.150)	(0.242)	(0.176)	(0.312)	(0.520)
Observations	391	391	391	376	391	391	391
Literacy		✓					
Electrification			✓				
IWI				✓			
Urbanization					✓		
Gini coefficient						✓	
Religious frac.							✓

Notes. a) The top panel of the table shows estimates from 42 separate negative binomial regressions of ACLED conflict events on ELF(k). The bottom panel does the same for BGI(k) in access to electricity. All regressions contain baseline controls described in the main text. IRRs are reported for each regression, along with p-values (based on robust standard errors) in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively. b) Additional controls listed at the bottom of the table are regional averages for literacy rate, electricity access, international wealth index (IWI), urbanization rate, Gini coefficient for years of education, and an index of religious fractionalization.

Table 3: Robustness to alternative specifications: ELF vs. BGI horse races

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ELF(1)	1.181**	1.227**	1.196**	1.111	1.133	1.194**	1.177**
	(0.040)	(0.012)	(0.033)	(0.205)	(0.134)	(0.030)	(0.045)
BGI(1)	0.979	0.990	0.989	1.022	1.014	0.983	0.983
	(0.779)	(0.897)	(0.883)	(0.769)	(0.856)	(0.821)	(0.828)
ELF(3)	1.246***	1.281***	1.256***	1.199**	1.193**	1.251**	1.240**
	(0.010)	(0.003)	(0.008)	(0.035)	(0.038)	(0.008)	(0.011)
BGI(3)	0.993	1.022	1.010	1.035	1.049	1.010	0.995
	(0.943)	(0.822)	(0.917)	(0.719)	(0.613)	(0.919)	(0.955)
ELF(5)	1.351***	1.366***	1.334***	1.269***	1.241***	1.344***	1.350***
	(0.000)	(0.000)	(0.000)	(0.004)	(0.008)	(0.000)	(0.000)
BGI(5)	0.957	0.991	0.994	1.017	1.025	0.970	0.958
	(0.627)	(0.919)	(0.943)	(0.856)	(0.786)	(0.740)	(0.628)
ELF(7)	1.414***	1.404***	1.377***	1.318***	1.260***	1.402***	1.413***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.009)	(0.000)	(0.000)
BGI(7)	0.917	0.959	0.961	0.951	0.997	0.931	0.917
	(0.327)	(0.650)	(0.658)	(0.576)	(0.973)	(0.428)	(0.328)
ELF(9)	1.352***	1.341***	1.311***	1.267***	1.209***	1.339***	1.352***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.008)	(0.000)	(0.000)
BGI(9)	0.997	1.046	1.054	1.058	1.092	1.017	0.997
	(0.967)	(0.601)	(0.528)	(0.500)	(0.271)	(0.840)	(0.967)
ELF(11)	1.202***	1.186**	1.153**	1.107	1.066	1.191*	1.197**
	(0.009)	(0.015)	(0.049)	(0.167)	(0.380)	(0.013)	(0.012)
BGI(11)	0.968	1.014	1.037	1.040	1.071	0.987	0.968
	(0.719)	(0.883)	(0.698)	(0.670)	(0.425)	(0.888)	(0.713)
Observations	391	391	391	376	391	391	391
Literacy		✓					
Electrification			✓				
IWI				✓			
Urbanization					✓		
Gini coefficient						✓	
Religious frac.							✓

Notes. a) The table shows estimates from 42 separate negative binomial horse race regressions of ACLED conflict events on ELF(k) and BGI(k). Each of the six panels displays results for a different level of linguistic aggregation k. All regressions contain baseline controls described in the main text. IRRs are reported for each regression along with p-values (based on robust standard errors) in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively. b) Additional controls listed at the bottom of the table are regional averages for literacy rate, electricity access, international wealth index (IWI), urbanization rate, Gini coefficient for years of education, and an index of religious fractionalization.

Our results are qualitatively unchanged after adjusting p-values for multiple hypothesis testing by controlling for the false discovery rate using the method of Benjamini et al. (2006). Specifically, appendix tables F.1 and F.2 replicate the coefficient estimates from tables 2 and 3, but report sharpened q-values of Anderson (2008) instead of regular p-values.[32] This approach offers robustness to positive dependence across tests, which is important in our setting given the substantial correlation between diversity metrics and across linguistic aggregation tiers, as well as the closely related robustness specifications. We treated all p-values from tables 2 and 3, as well as their counterparts from regressions with conflict fatalities as an outcome variable, as belonging to a single testing family. Comparison of p-values and q-values reveals only very minor differences, confirming our original conclusions.

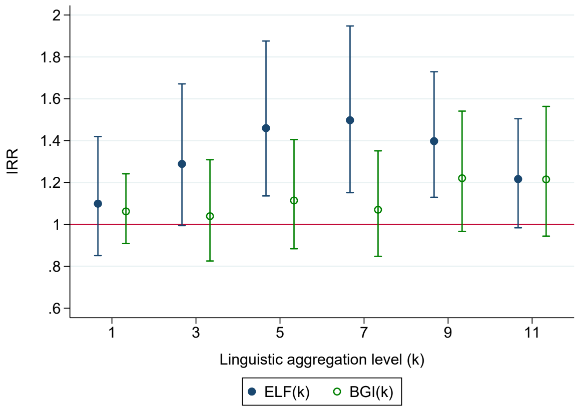
Although we cannot completely rule out the presence of some unobservable factors that may still affect our results, their robustness to a variety of potentially confounding characteristics gives us confidence. An additional empirical challenge to the observed relationship between ELF and conflict is population sorting. Indeed, it is plausible that individuals seeking security would try to leave conflict-ridden regions and avoid moving there. This would result in a *negative* association between diversity and conflict, in which case our estimates understate the true positive effect of ELF.

Furthermore, as documented in Gershman and Rivera (2018), subnational ELF in Sub-Saharan Africa is extremely persistent over time, suggesting negligible effects of population sorting. Specifically, for several countries with reliable data, they show that the correlations between ELF(k) values over several decades are well above 0.9. This persistence is stronger for lower values of k and appears just as strong for smaller administrative subdivisions.

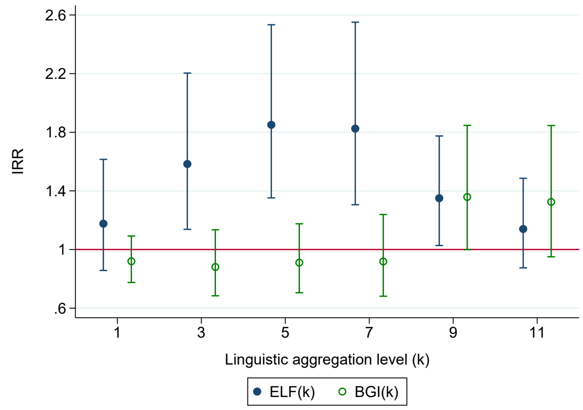
Our results are also robust to alternative measures of conflict incidence and intensity, as shown in figure 7. Panels (a) and (b) illustrate the estimates from "horse race" regressions for ELF and BGI (in electricity access) when UCDP data on deadly conflicts are used to construct the dependent variables. The overall pattern is roughly similar to that observed in panel (b) of figure 6 for ACLED data, but implies a stronger magnitude of association between deep-rooted ELF and conflict.[33] In panels (c) and (d), we report analogous results for SCAD data covering a broader scope of social unrest incidents, both violent and

[32] The false discovery rate (FDR) is the expected share of false positives among all rejected hypotheses. The q-value is the minimum FDR level at which a given hypothesis would be rejected.

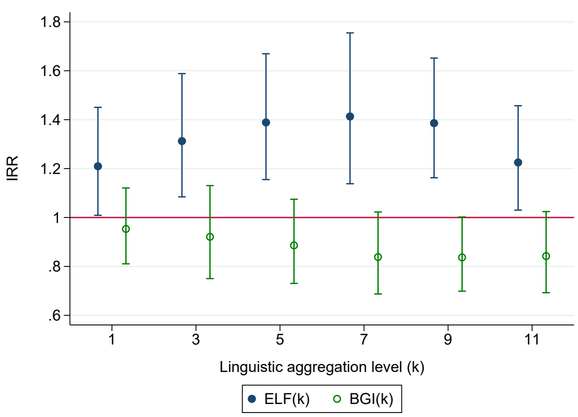
[33] In an additional robustness check, we use the Ethnic Power Relations (EPR) database (Vogt et al., 2015) to extract only those conflict events that have an ethnic dimension. In such cases, at least one side of conflict claimed to represent, recruited from, or was supported by a specific ethnic group. Our results for this subset of conflict events and fatalities are qualitatively similar to those for the unrestricted UCDP sample presented in figure 7.



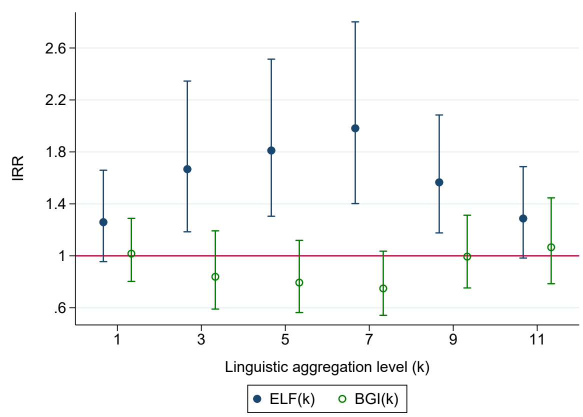
(a) UCDP events



(b) UCDP fatalities



(c) SCAD events



(d) SCAD fatalities

Figure 7: ELF vs. BGI "horse race" regressions for alternative conflict data.

Notes. Each panel presents incidence rate ratios, along with 95% confidence intervals, based on robust standard errors. For each of the six reported levels of linguistic aggregation, a negative binomial regression is estimated, where the outcome is the sum of conflict events or fatalities based on UCDP data (panels a and b) or SCAD data (panels c and d). The right-hand side includes both ELF(k) and BGI(k) for electricity access, along with country fixed effects, baseline controls, and urbanization rate. The number of observations is 391.

non-violent. Once again, the qualitative pattern is similar to our baseline, with larger IRR estimates. Thus, fragmentation (particularly when measured for deeper linguistic cleavages) but not ethnic inequality is strongly positively associated with a broad spectrum of conflict measures.[34]

[34] The regression results are qualitatively the same when using BGI in other socioeconomic outcomes.

5 Concluding remarks

This paper introduced a new dataset on subnational ethnolinguistic divisions in Sub-Saharan Africa and used it to reassess the relationship between diversity and conflict. Drawing on microdata from large-scale household surveys, we constructed comparable measures of fractionalization, polarization, and ethnic inequality across regions. To capture the depth of ethnolinguistic cleavages, we mapped all groups onto the global linguistic tree and computed each diversity measure at multiple levels of linguistic aggregation. This framework allowed for a unified assessment of how different facets and historical layers of diversity relate to conflict within countries.

Our results reveal a robust positive association between conflict and ethnolinguistic fractionalization, especially when it is defined over deep-rooted linguistic cleavages. This finding supports theoretical perspectives suggesting that coordination failures, mistrust, and enduring antagonisms among culturally distant groups trigger social tensions. It is also consistent with the Esteban and Ray (2011a) framework, which predicts that when contests revolve around private prizes, such as access to local resources, fractionalization should matter most, whereas polarization is more relevant for conflicts over public prizes, such as nationwide policies. A subnational focus fits the former case better while remaining compatible with earlier results on the empirical relevance of polarization in predicting conflict at the country level.

We find no systematic evidence that ethnic inequality, measured along multiple socioeconomic dimensions, is associated with conflict intensity or severity. This challenges claims that group-level socioeconomic disparities, rather than diversity itself, are the primary drivers of violence. The absence of a clear relationship may partly reflect the theoretical ambiguity about when and how inequality translates into mobilization and the accounts that view the inequality-conflict link as contingent on group resources, organization, and perceptions of status change (Ray and Esteban, 2017; Esteban and Ray, 2008). Our approach is also different from some earlier empirical studies that focus specifically on "politically relevant" groups and the importance of political exclusion rather than socioeconomic differences.

These conclusions apply at a particular subnational scale, first-level administrative regions, chosen based on both data availability constraints and substantive relevance. While this unit of analysis is meaningful, mechanisms linking diversity and conflict may differ at other spatial scales, and the results should not be assumed to generalize to finer or coarser levels of aggregation.

Our findings have important implications for policies related to territorial administration, redistribution, and peace. The imposition of state borders during the colonial era left a lasting imprint on African development, and the redrawing of internal administrative boundaries remains a sensitive issue today. Policymakers facing this task should weigh the potential risks of social unrest associated with greater ethnolinguistic fractionalization in the resulting regions, taking into account ancestral relationships among affected groups. Similar considerations can inform the design of refugee resettlement programs and policies addressing migrant integration. At the same time, the absence of a systematic link between conflict and ethnic inequality cautions against assuming that targeted local redistribution alone can reduce violence.

Appendices

A Description of variables

Ethnic inequality and other diversity measures

The BGI, ELF, and ELP indices are described in detail in section 3.2. The variables underlying the BGI indices were defined as follows.

Education in years. Educational attainment in years was typically recorded based on the last completed year of schooling. If this direct measure is missing, years of education were calculated from the reported level of education (e.g., primary) using country-specific schooling requirements that were in place at the time of the survey.

Access to a high-quality water source. Indicator of household access to one of the following sources: water piped into dwelling or plot, water supplied via a water tanker, bottled water. Low-quality sources include private or community wells and taps, nearby natural sources.

Flush toilet. Indicator of household access to flush toilet (as opposed to pit latrines and other lower-quality options).

Finished floor. High-quality flooring indicator, equal to one if the floor is finished (e.g., with parquet, carpet, or tiles).

Electricity access. Indicator of household access to electricity.

Asset ownership. Indicators of household ownership of radio, television, and car.

Conflict variables

ACLED events and fatalities. The regional count of conflict events and fatalities from the Armed Conflict Location and Event Database (from January 1997 to June 2020). *Source:* `https://acleddata.com/data-export-tool/` and own calculations.

UCDP events and fatalities. The regional count of conflict events and fatalities from the Uppsala Conflict Data Program's Georeferenced Event Dataset (1997–2019). *Source*: `https://ucdp.uu.se/downloads/` and own calculations.

SCAD events and fatalities. The regional count of conflict events and fatalities from the Social Conflict Analysis Database (1997–2017). *Source*: `https://www.strausscenter.org/ccaps-research-areas/social-conflict/database/` and own calculations.

Control variables

Terrain ruggedness index. Index of terrain ruggedness as constructed by Nunn and Puga (2012) at 0.5-degree resolution, averaged across cells in each region. *Source*: Gershman and Rivera (2018).

Caloric suitability index. An index capturing potential agricultural output (measured in calories) based on crops that were available for cultivation in the post-1500 CE era and available at the 5-arc-minute resolution (Galor and Özak, 2016). We calculate the mean and standard deviation of the index based on cell values within each region. *Source:* `https://ozak.github.io/Caloric-Suitability-Index` and own calculations.

Distance to the coastline. Great circle distance from a region's centroid to the closest location on the coastline. *Source:* `https://www.naturalearthdata.com` and own calculations.

Distance to Addis Ababa: Great circle distance from a region's centroid to Addis Ababa. *Source*: own calculations.

Oil and gas field indicator. A dummy variable indicating the presence of an oil or gas field in the region. *Source*: PETRODATA dataset (version 1.2) and own calculations.

Diamond occurence. A dummy variable indicating the occurence of diamonds in the region. *Source*: DIADATA dataset and own calculations.

Partitioned ethnic homeland indicator. A dummy variable equal to one, if a region intersects an ethnic homeland partitioned by state borders, as defined in Michalopoulos and Papaioannou (2016). *Source*: own calculations.

Malaria stability index. An index measuring the stability of malaria transmission based on regionally dominant vector mosquitoes, available at 0.5-degree resolution. We calculate the mean across cells in each region. *Source*: Kiszewski et al. (2004) and own calculations.

Literacy rate. Share of region's adult population (aged 15–49 years) that is literate. A person is considered literate if she can read at least part of a standard sentence or has attended secondary school. *Source*: Gershman and Rivera (2018).

Electrification. Share of region's households that have access to electricity. *Source*: Gershman and Rivera (2018).

International wealth index (IWI). Wealth index, as proposed by Smits and Steendijk (2015), averaged across households within relevant regions. *Source*: Gershman and Rivera (2018).

Urbanization rate. Share of region's households that live in urban areas. *Source*: Gershman and Rivera (2018).

Religious fractionalization. Regional index of fractionalization based on a four-way classification for religious affiliation: Christianity, Islam, traditional religion, and none. *Source*: Gershman and Rivera (2018).

Gini coefficient for years of education. Gini coefficient measuring inequality in years of education at the regional level. *Source*: own calculations.

B Microdata used to measure population diversity

Table B.1: Countries, data sources, ethnolinguistic groups, and subnational regions

Country	Primary source	Secondary source	Standardized groups (P/S)	Survey groups (P/S)	Regions
Angola	DHS (2016)		10	10	18
Benin	IPUMS (2013)		42	52	12
Botswana	IPUMS (2011)		12	11	9
Burkina Faso	IPUMS (2006)	MICS (2006)	27/18	29/19	13
Cameroon	DHS (2018)		138	173	10
Central African Rep.	MICS (2010)		9	9	17
Chad	DHS (2015)		19	19	20
Rep. of the Congo	DHS (2012)		11	11	12
Côte d'Ivoire	MICS (2006)		50	53	19
Djibouti	MICS (2006)		3	3	6
Eritrea	DHS (2002)		10	10	6
Eswatini	DHS (2007)		1	1	1
Ethiopia	IPUMS (2007)	DHS (2011)	63/48	86/59	11
Gabon	DHS (2012)		8	8	9
Gambia	DHS (2013)		8	9	8
Ghana	IPUMS (2010)	DHS (2008)	38/8	62/8	10
Guinea	DHS (2012)		6	6	8
Guinea-Bissau	MICS (2018)		8	8	9
Kenya	DHS (2014)		20	22	8
Liberia	IPUMS (2008)		17	17	15
Malawi	IPUMS (2008)		9	12	3
Mali	IPUMS (2009)	MICS (2010)	14/16	15/17	9
Mauritania	MICS (2007)		4	4	13
Mozambique	MICS (2008)		22	22	11
Namibia	DHS (2013)		8	8	13
Niger	DHS (2006)		8	8	8
Nigeria	DHS (2013)		192	318	37
Senegal	IPUMS (2013)		19	25	11
Sierra Leone	IPUMS (2004)		15	15	4
South Africa	IPUMS (2011)	DHS (2016)	11/11	11/11	9
Tanzania	DHS (1992)		98	119	21
Togo	IPUMS(2010)		31	34	5
Uganda	IPUMS (2014)		41	64	4
Zambia	IPUMS (2010)		32	61	9
Zimbabwe	DHS (2011)		2	2	10

Notes. a) DHS is Demographic and Health Surveys, MICS is Multiple Indicator Cluster Surveys, IPUMS is Integrated Public Use Microdata Series (subsamples of national censuses). b) Column 4 lists the number of unique ethnolinguistic groups (standardized based on matching to *Ethnologue* language codes) in our primary (P) and secondary (S) data sources. c) Column 5 lists the number of original (non-standardized) survey groups in the primary (P) and secondary (S) sources. d) Column 6 lists the number of subnational regions per country. See Gershman and Rivera (2018) for details on the definitions of subnational boundaries.

C Computing BGI(k)

As explained in the main text, we use *Ethnologue*'s language tree model to measure ethnic inequality at different levels of linguistic aggregation. Here, we illustrate the process using the Gash-Barka region of Eritrea as example. In our sample, there are 7 ethnic groups in this region that uniquely match to 7 languages in *Ethnologue*. These languages (in boxes) and the hypothesized ancestral relationships between them are illustrated in figure C.1, which is the relevant linguistic subtree extracted from *Ethnologue*. As can be seen, all languages spoken in the Gash-Barka region belong to two major families, Afro-Asiatic and Nilo-Saharan. This is the deepest (tier 1) cleavage in *Ethnologue*'s model beyond which it only identifies the "Proto-Human" root referring to the hypothetical common ancestor of all languages. Beyond tier 1, the tree branches out further ultimately giving rise to contemporary languages. The longer the path shared by two languages before they diverge, the more closely related they are. In our case, Tigre and Tigrigna are the closest two languages as they share 5 branches starting from tier 1, whereas any pair of languages representing two distinct major language families are the least related since they share no common branches.

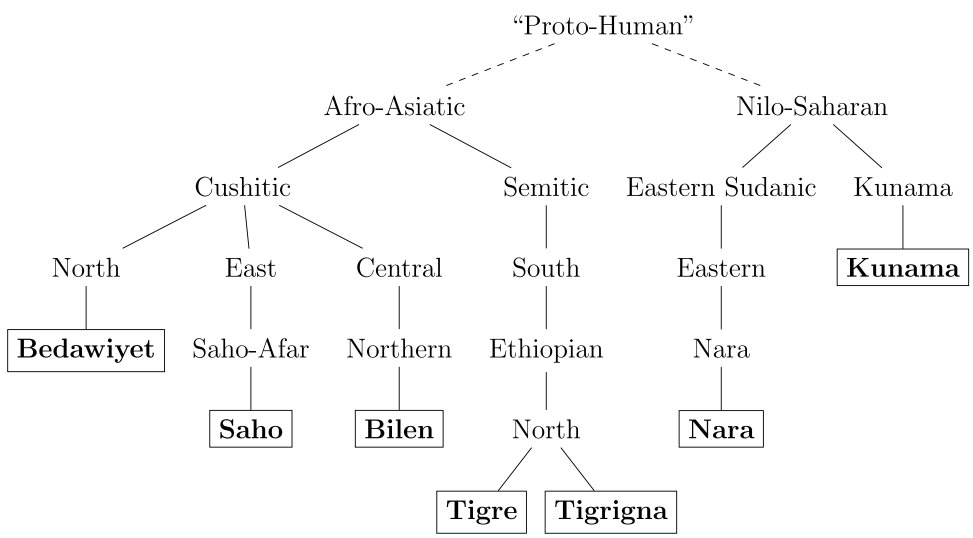


Figure C.1: Ethnolinguistic tree for languages spoken in the Gash-Barka region of Eritrea

To account for linguistic relatedness we follow the aggregation method from Desmet et al. (2012). The first step in this approach is to create "artificial" ancestral groups to equate the number of branches from the root to each extant language. The outcome of this process is shown in figure C.2, where the added ancestral groups are marked with asterisks. With each of the 6 possible tiers of the tree now properly defined, we can aggregated languages

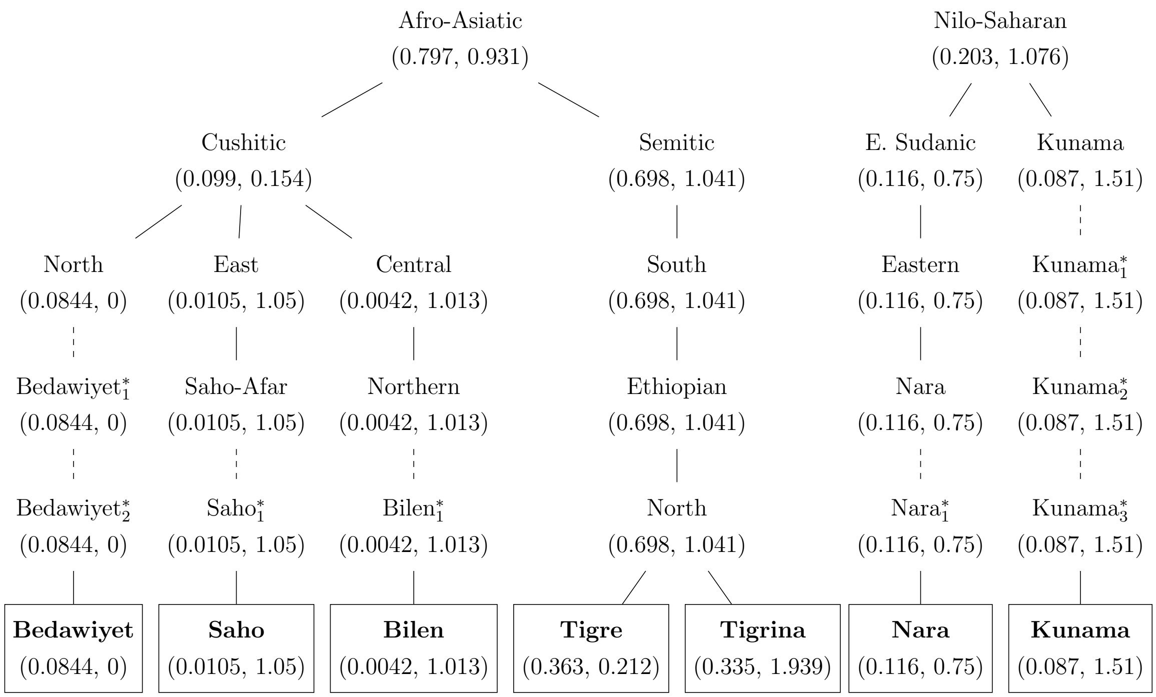


Figure C.2: Extended ethnolinguistic tree for the Gash-Barka region of Eritrea

Notes. Population shares and mean educational attainment reported in parentheses.

to the level of ancestral groups. For example, at tier 5, Tigre and Tigrigna “merge” into a single North subgroup of Ethiopian languages and at tier 2, Bedawiyet, Saho, and Bilen all “merge” into a single Cushitic subfamily. Thus, by construction, there are fewer distinct groups at higher levels of linguistic aggregation.

Figure C.2 reports other crucial elements required to measure ethnic inequality, namely, the (rounded) regional population shares and mean levels of educational attainment for each (aggregated) group. Note that both depend on the tier of the tree. For example, at tier 5, where Tigre and Tigrigna merge into North, this group becomes the largest, constituting about 70% of the region’s population, and its mean level of educational attainment is the weighted average of the respective values for the two constituent groups. We compute population shares and mean outcomes for all relevant ancestral groups, as shown in the figure. The overall regional mean educational attainment level in Gash-Barka is 0.96 years. Next, we simply use the BGI formula at each tier of the tree. For example, at tier 1, we have $\text{BGI}(1) = (2\times 0.797\times 0.203\times |0.931-1.076|)/(2\times 0.96) \approx 0.024$. Similarly, $\text{BGI}(2) \approx 0.145$, $\text{BGI}(3) = \text{BGI}(4) = \text{BGI}(5) \approx 0.148$, and $\text{BGI}(6) \approx 0.449$.

As mentioned in the main text, in our full sample, there are 13 tiers of aggregation since for some languages we count up to 12 steps from the root to the bottom of the tree. In the case of Gash-Barka, all indices for tiers 6 through 13 will be identical: additional tiers in the extended tree will simply be “copies” of tier 6.

C.1 BGI(k) is non-decreasing in k

Consider tier k of a linguistic tree. Assume that, at tier k, there are three distinct ethno-linguistic groups with population shares s_1, s_2, and s_3 and average outcomes $\bar{w}_1$, $\bar{w}_2$, and $\bar{w}_3$. Assume further that at tier $k-1$, groups 2 and 3 merge into a single ancestral group. Its population share is $s_2 + s_3$ and its average outcome is $\bar{w}_{23}$, the weighted average of $\bar{w}_2$ and $\bar{w}_3$:

$$\bar{w}_{23} = \frac{s_2}{s_2 + s_3} \cdot \bar{w}_2 + \frac{s_3}{s_2 + s_3} \cdot \bar{w}_3.$$

We claim that BGI(k) $\geqslant$ BGI($k-1$) in the same region. In our example,

$$\begin{aligned} \text{BGI}(k) &= \frac{1}{2\bar{w}} \cdot (s_1 s_2 |\bar{w}_1 - \bar{w}_2| + s_1 s_3 |\bar{w}_1 - \bar{w}_3| + s_2 s_3 |\bar{w}_2 - \bar{w}_3|), \\ \text{BGI}(k-1) &= \frac{1}{2\bar{w}} \cdot (s_1 (s_2 + s_3) |\bar{w}_1 - \bar{w}_{23}|). \end{aligned}$$

Plug in the earlier expression for w_{23} and rearrange to get

$$s_1(s_2 + s_3)|\bar{w}_1 - \bar{w}_{23}| = s_1|s_2\bar{w}_1 + s_3\bar{w}_1 - s_2\bar{w}_2 - s_3\bar{w}_3| = |s_1 s_2(\bar{w}_1 - \bar{w}_2) + s_1 s_3(\bar{w}_1 - \bar{w}_3)|$$

By triangle inequality,

$$\text{BGI}(k-1) = \frac{1}{2\bar{w}} \cdot |s_1 s_2(\bar{w}_1 - \bar{w}_2) + s_1 s_3(\bar{w}_1 - \bar{w}_3)| \leqslant \frac{1}{2\bar{w}} \cdot (s_1 s_2|\bar{w}_1 - \bar{w}_2| + s_1 s_3|\bar{w}_1 - \bar{w}_3|) \leqslant \text{BGI}(k).$$

D Summary statistics

Table D.1: Summary statistics for conflict measures

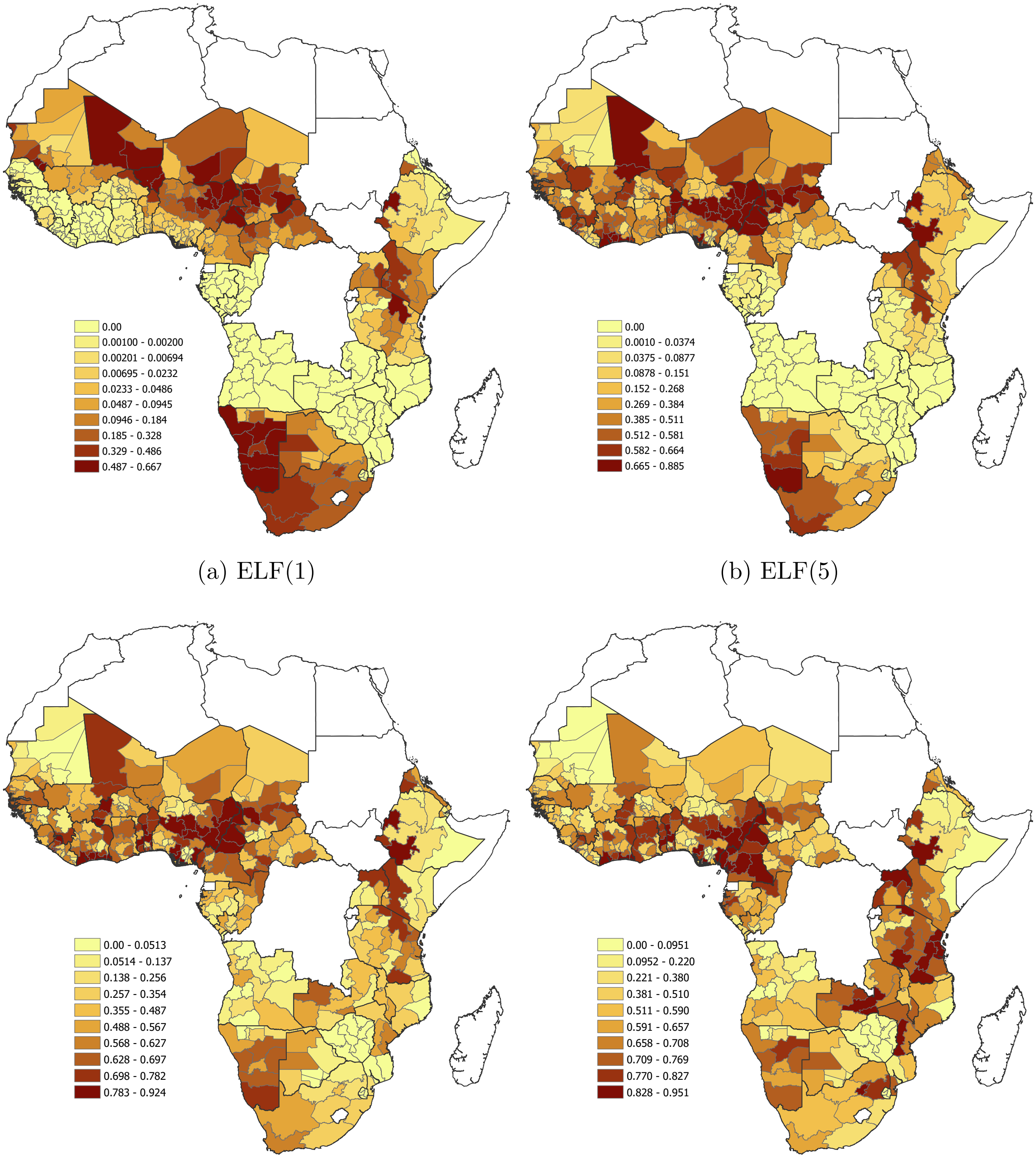
Variable	Mean	St. dev.	Min	25th pct.	Median	75th pct.	Max
ACLED events	242.5	459.8	0	20	79	262	3289
ACLED fatalities	1059.8	6786.6	0	9	57	489	120397
UCDP events	38.2	137.0	0	0	2	23	2074
UCDP fatalities	427.9	1630.9	0	0	9	242	23453
SCAD events	12.9	30.7	0	1	3	10	262
SCAD fatalities	70.0	276.0	0	0	2	26	4234

Notes. ACLED refers to the Armed Conflict and Location Event Database. UCDP stands for the Uppsala Conflict Data Program. SCAD refers to the Social Conflict Analysis Database.

Table D.2: Summary statistics for population diversity indices

Tier k	Education BGI(k)	Electricity BGI(k)	Television BGI(k)	Car BGI(k)	Water BGI(k)	Toilet BGI(k)	Floor BGI(k)	Radio BGI(k)	ELF(k)	ELP(k)
1	0.019	0.041	0.042	0.060	0.039	0.055	0.041	0.011	0.100	0.182
	(0.048)	(0.090)	(0.102)	(0.135)	(0.085)	(0.125)	(0.095)	(0.026)	(0.172)	(0.301)
2	0.037	0.073	0.078	0.110	0.070	0.089	0.080	0.021	0.180	0.318
	(0.070)	(0.119)	(0.135)	(0.179)	(0.121)	(0.161)	(0.137)	(0.037)	(0.206)	(0.342)
3	0.047	0.085	0.089	0.123	0.087	0.099	0.095	0.025	0.204	0.336
	(0.082)	(0.129)	(0.142)	(0.187)	(0.144)	(0.168)	(0.149)	(0.042)	(0.224)	(0.335)
4	0.058	0.098	0.102	0.148	0.104	0.119	0.117	0.029	0.246	0.379
	(0.088)	(0.134)	(0.147)	(0.196)	(0.155)	(0.175)	(0.163)	(0.044)	(0.242)	(0.330)
5	0.066	0.110	0.114	0.167	0.128	0.138	0.136	0.034	0.286	0.408
	(0.092)	(0.139)	(0.150)	(0.201)	(0.174)	(0.184)	(0.176)	(0.047)	(0.262)	(0.330)
6	0.076	0.128	0.129	0.190	0.148	0.159	0.157	0.040	0.319	0.415
	(0.101)	(0.152)	(0.159)	(0.213)	(0.191)	(0.199)	(0.194)	(0.052)	(0.284)	(0.323)
7	0.083	0.141	0.142	0.208	0.165	0.177	0.175	0.044	0.355	0.438
	(0.104)	(0.164)	(0.167)	(0.223)	(0.207)	(0.212)	(0.213)	(0.054)	(0.295)	(0.322)
8	0.084	0.147	0.147	0.216	0.172	0.185	0.185	0.046	0.369	0.447
	(0.104)	(0.166)	(0.170)	(0.225)	(0.211)	(0.215)	(0.216)	(0.054)	(0.296)	(0.315)
9	0.093	0.181	0.169	0.258	0.205	0.228	0.219	0.056	0.437	0.536
	(0.101)	(0.169)	(0.176)	(0.233)	(0.213)	(0.227)	(0.220)	(0.055)	(0.267)	(0.269)
10	0.101	0.202	0.184	0.285	0.231	0.254	0.243	0.064	0.493	0.561
	(0.099)	(0.176)	(0.180)	(0.234)	(0.219)	(0.231)	(0.224)	(0.060)	(0.264)	(0.248)
11	0.107	0.216	0.192	0.301	0.249	0.272	0.256	0.070	0.522	0.561
	(0.101)	(0.187)	(0.183)	(0.241)	(0.229)	(0.240)	(0.230)	(0.067)	(0.266)	(0.240)
12	0.107	0.216	0.193	0.301	0.250	0.272	0.257	0.071	0.523	0.560
	(0.101)	(0.188)	(0.183)	(0.242)	(0.229)	(0.240)	(0.231)	(0.067)	(0.267)	(0.240)
13	0.107	0.217	0.193	0.302	0.250	0.273	0.257	0.071	0.523	0.559
	(0.101)	(0.188)	(0.183)	(0.242)	(0.229)	(0.240)	(0.231)	(0.067)	(0.267)	(0.240)
N	391	391	391	391	391	391	391	391	391	391
Primary	378	391	372	361	378	391	382	372	391	391

Notes. Mean values are reported for diversity indices, with respective standard deviations in parentheses. The last row reports the number of observations covered by primary data sources.



(a) ELF(1)

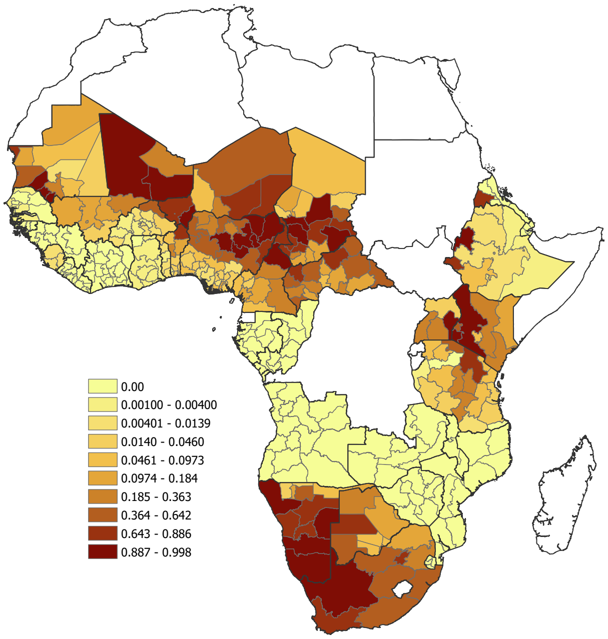
(b) ELF(5)

(c) ELF(9)

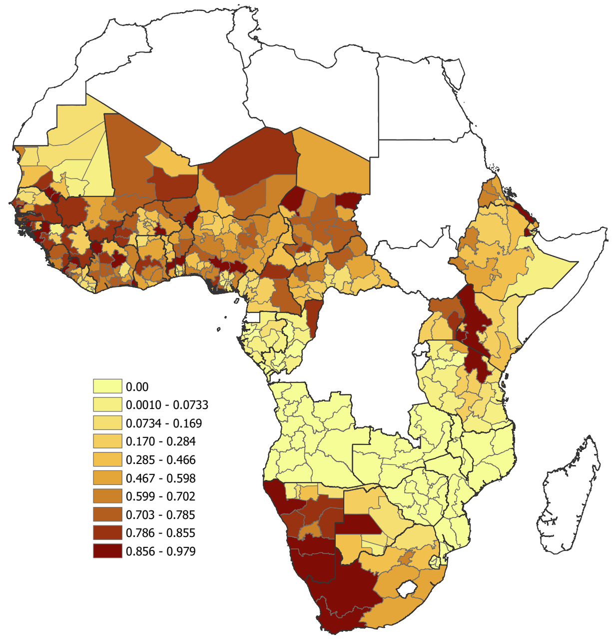
(d) ELF(13)

Figure D.1: Spatial distribution of ELF.

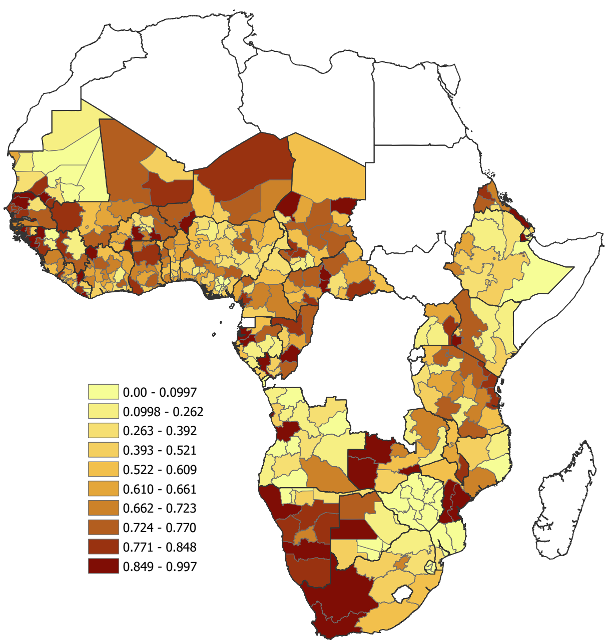
Notes. Each panel uses graduated colors based on deciles of respective index values.



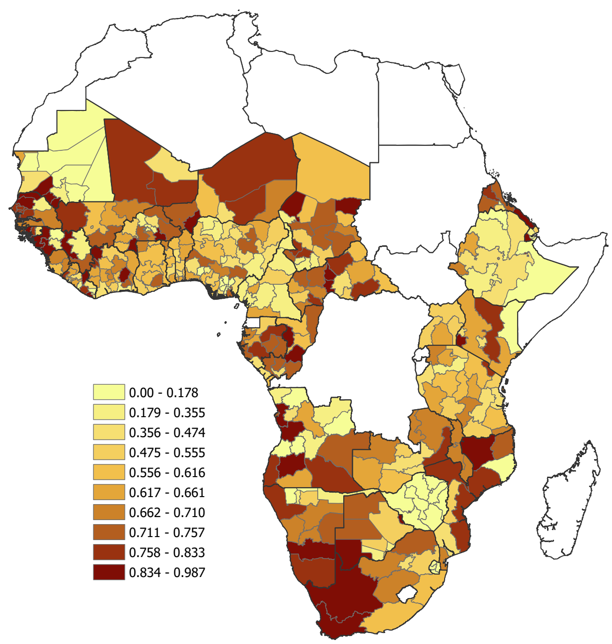
(a) ELP(1)



(b) ELP(5)



(c) ELP(9)



(d) ELP(13)

Figure D.2: Spatial distribution of ELP.

Notes. Each panel uses graduated colors based on deciles of respective index values.

Table D.3: Pairwise correlation coefficients for various BGI indices

		Education BGI				Electricity BGI				Television BGI				Car BGI			
		(1)	(5)	(9)	(13)	(1)	(5)	(9)	(13)	(1)	(5)	(9)	(13)	(1)	(5)	(9)	(13)
Education BGI	(1)	1															
	(5)	0.67	1														
	(9)	0.56	0.92	1													
	(13)	0.52	0.87	0.97	1												
Electricity BGI	(1)	0.53	0.31	0.24	0.22	1											
	(5)	0.42	0.62	0.58	0.53	0.69	1										
	(9)	0.27	0.40	0.50	0.51	0.59	0.79	1									
	(13)	0.20	0.27	0.39	0.46	0.51	0.63	0.91	1								
Television BGI	(1)	0.44	0.23	0.18	0.18	0.76	0.52	0.46	0.45	1							
	(5)	0.37	0.61	0.57	0.53	0.55	0.82	0.64	0.54	0.70	1						
	(9)	0.26	0.47	0.53	0.52	0.49	0.70	0.72	0.65	0.62	0.86	1					
	(13)	0.21	0.39	0.46	0.49	0.43	0.60	0.67	0.68	0.57	0.76	0.94	1				
Car BGI	(1)	0.30	0.13	0.09	0.10	0.61	0.40	0.36	0.35	0.71	0.49	0.43	0.39	1			
	(5)	0.29	0.53	0.51	0.48	0.41	0.64	0.50	0.40	0.47	0.76	0.65	0.56	0.63	1		
	(9)	0.16	0.37	0.45	0.48	0.33	0.50	0.64	0.61	0.34	0.56	0.61	0.56	0.51	0.80	1	
	(13)	0.12	0.27	0.37	0.45	0.27	0.38	0.59	0.64	0.29	0.45	0.53	0.57	0.44	0.67	0.93	1

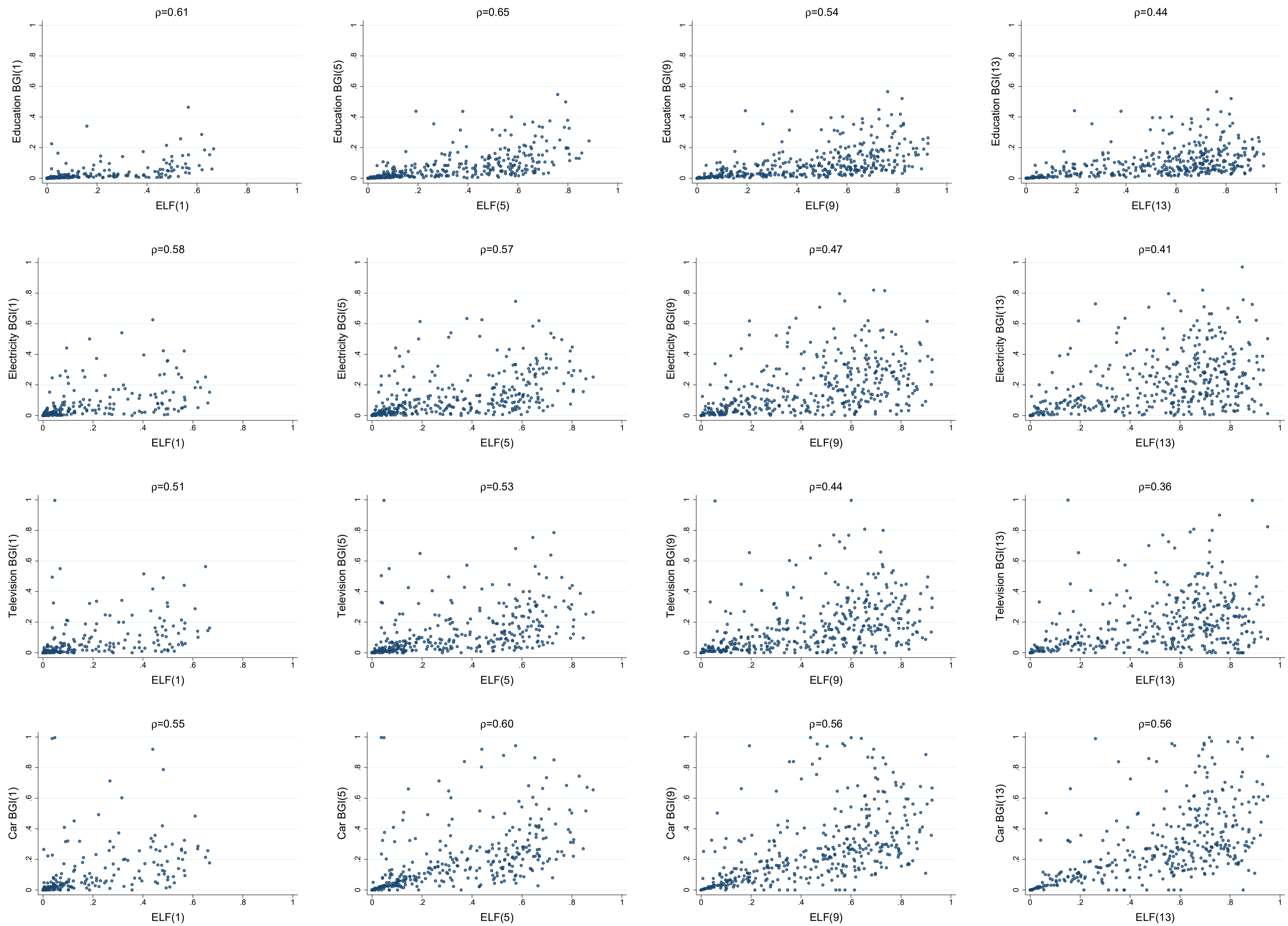


Figure D.3: Scatterplots and pairwise correlation coefficients (ρ) for selected ELF(k) and BGI(k) indices.

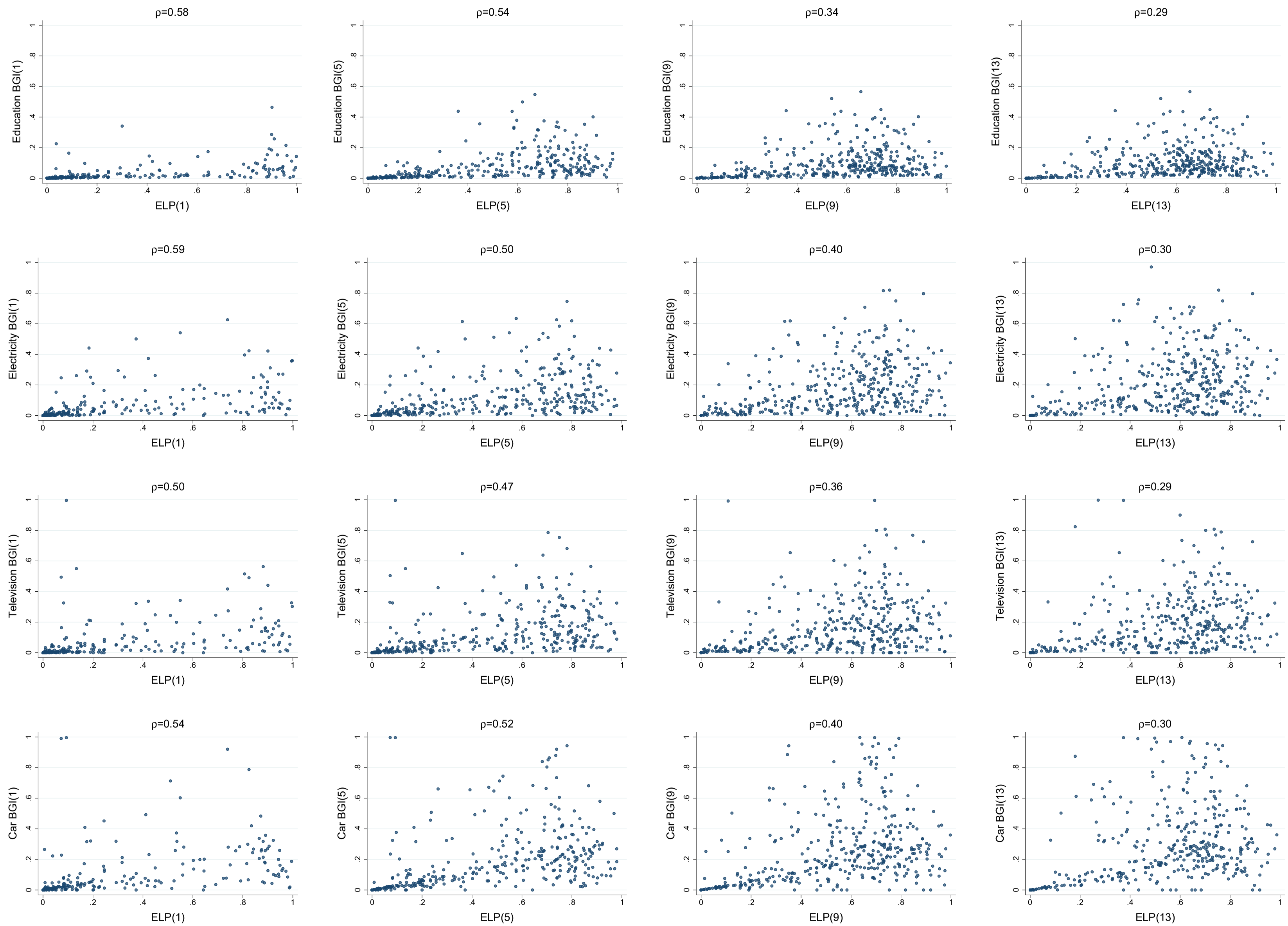


Figure D.4: Scatterplots and pairwise correlation coefficients (ρ) for selected ELP(k) and BGI(k) indices.

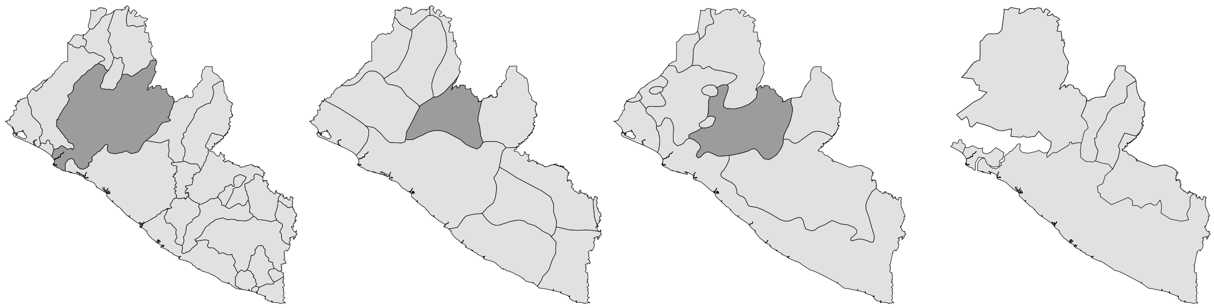
E GIS approach to measuring subnational BGI

A popular simpler alternative to measuring ethnolinguistic diversity via household surveys is to rely on geographic information systems. In this section, we compare these two approaches to highlight some of the key limitations of GIS methods in estimating local ethnic inequality. Given many variations of both input data and inequality metrics used in the literature, a comprehensive comparison is outside the scope of this mini-investigation. Instead, we present a case study of regional BGI in Liberia to illustrate several main issues.

The GIS approach requires four layers of spatial data: 1) polygons representing the units of analysis (subnational regions in our case), 2) polygons representing ethnolinguistic groups, 3) a high-resolution population raster with estimated counts of people residing in each relevant grid cell, 4) a high-resolution raster with estimated levels of economic activity in each relevant grid cell. The first component is the least controversial and we simply use the same set of regional boundaries as in our main analysis.

There are several possible options for the second component. The World Language Mapping System (WLMS) is a GIS companion to *Ethnologue* that aims to spatially represent the "homelands" of all known non-extinct languages around the globe (we use version 16 published in 2011). It is very detailed and links polygons to *Ethnologue* language codes, which we also use to standardize ethnolinguistic groups recorded in survey data. Most polygons in WLMS are non-overlapping, but there are certain areas where languages homelands intersect. While this reflects the qualitative situation on the ground, users ultimately have to make rather arbitrary decisions about how to divide the total population size or measured economic activity in such overlapping areas between coexisting groups, with unclear effect on the quality of resulting metrics. We note the importance of this issue in adding noise to GIS-based measures but set it aside for our case study since in the case of Liberia, WLMS polygons are all non-overlapping (after dropping the English language).

Figure E.1(a) shows the WLMS polygons for Liberia. Overall, these represent 26 unique *Ethnologue* codes, compared to 17 in the 2008 IPUMS dataset that we use in our survey-based analysis. Of the 9 "extra" groups in WLMS, 4 are very small, with the estimated total number of speakers between 7 and 20 thousand in the entire country. Furthermore, *Ethnologue* disaggregates the Grebo macrolanguage into 5 groups of dialects with unique codes, whereas the Liberian census only lists a single corresponding ethnicity, Grebo. Similarly, *Ethnologue* has distinct codes for Eastern and Western Krahn languages, whereas the sole corresponding ethnicity from the census is Krahn. Despite these minor differences, the WLMS and IPUMS group classifications are similar, especially once aggregated at deeper tiers of the language tree.



(a) WLMS (b) Murdock (c) GREG (d) GeoEPR

Figure E.1: Ethnolinguistic homelands of Liberia.

Notes. The figure shows maps of ethnolinguistic groups in Liberia according to 4 sources: World Language Mapping System (WLMS), Murdock's 1959 map of precolonial Africa, Geo-Referencing of Ethnic Groups (GREG) database, and Geo-Referencing Ethnic Power Relations (GeoEPR) database. The homeland of Kpelle is highlighted in the first three panels.

Other commonly used maps of ethnolinguistic groups are Murdock's 1959 atlas of Africa (Murdock, 1959), GREG (Weidmann et al., 2010), and GeoEPR (Vogt et al., 2015), all shown in figure E.1 for Liberia. The Murdock map aimed to represent African tribal homelands before the colonial era and covers 14 non-overlapping groups. GREG (Geo-Referencing of Ethnic Groups) is the digitized version of *Atlas Narodov Mira*, a global map of ethnic homelands compiled by Soviet ethnographers in 1960s. It is substantially less detailed compared to either WLMS or the Murdock map for Africa and only covers 9 groups in Liberia, including a large polygon labeled as "Bantu-speaking Pygmy tribes" (most likely an error since no such groups have been documented in the country). Finally, GeoEPR (Geo-Referencing Ethnic Power Relations) is a database popular in political science that focuses only on "politically relevant" groups. GeoEPR contains 5 polygons in Liberia, including a small area assigned to "Americo-Liberians," plus an "Indigenous Peoples" polygon covering most of the country.

Clearly, available digital maps of ethnolinguistic groups vary substantially and any particular choice requires justification. Notably, even groups present across databases are mapped differently. The homeland of Kpelle, the largest group in the Liberia, highlighted in figure E.1 for three sources (it is absent from GeoEPR) is an example of a relatively *good* alignment across sources. Overall, WLMS provides the most detailed and up-to-date picture that is conceptually aligned with the purpose of our study. Hence, we proceed with WLMS for the most "apples-to-apples" comparison, with a caveat that, like other sources, it was not designed to capture recent migrations and ethnic diversity of urban areas.

(a) LandScan 2008 (b) DMSP 2008 (c) VIIRS 2012

Figure E.2: Population and night lights data for Liberia.

Notes. Panel (a) shows the LandScan 2008 data on population counts. Zero values are grayed out, while the rest vary from 1 to 56,883. Panels (b) and (c) show the DMSP 2008 and VIIRS 2012 nightlights data, converted to binary scale: gray and blue cells correspond to no lights and some lights, respectively. Polygons in each panel represent 15 counties of Liberia.

There are several sources of high-resolution population data that rely on somewhat different methodologies to produce their estimates. We use the LandScan rasters, produced by the Oak Ridge National Laboratory and available annually, including for our reference year 2008. As figure E.2(a) shows, Liberia is mostly sparsely populated, with substantial population concentrated in two coastal counties, Montserrado (including the Monrovia capital area hosting about a third of the country's total population) and Margibi just to the east.

Regional and ethnolinguistic polygons, along with the LandScan raster, allow to easily estimate local population shares of all ethnic groups, a necessary input for calculating BGI. The intersection of regional and WLMS language polygons yields 73 "partitions." Meanwhile, a cross-tabulation by region and ethnicity in IPUMS gives 249 combinations (despite IPUMS having 9 fewer groups in total). Thus, even with the most detailed ethnolinguistic map as input, the GIS approach substantially underestimates the presence of Liberia's groups across its regions. In some cases, this presence is modest in terms of regional population shares, and the GIS approximation works reasonably well. But in other cases, it does not. For example, in the urbanized Margibi county, according to WLMS/LandScan, there are only two groups, Bassa and Kpelle, with regional shares of 14 and 86%, respectively. In contrast, IPUMS documents 17 groups in Margibi, with Bassa and Kpelle constituting 20 and 47%. The measurement error gets much worse for the Montserrado county. According to WLMS/LandScan, virtually the entire county (98.5%) are the Kpelle, whereas

IPUMS reveals its extreme diversity, with ELF(13) = 0.9 and the population share of Kpelle at only 19%.[35] Clearly, the GIS approach fails to capture the diversity of urban areas. Since ethnolinguistic groups "merge" into ancestral families at deeper tiers of the language tree, the scope for GIS measurement error becomes smaller and the discrepancy with survey-based indices is less severe (Gershman and Rivera, 2020). Still, for example, in Montserrado, IPUMS-based ELF(5) remains high at 0.7.

The final required component in the BGI calculation is a high-resolution "economic activity" raster. Recent literature has predominantly relied on the night lights data as a proxy for economic activity.[36] The idea is that luminosity at night reflects total economic activity in a grid cell, and a proxy for group-level average income or material well-being may be computed by dividing the total amount of light detected within a group's regional partition by its population (Alesina et al., 2016). For this exercise, we use the improved version of the 2008 annual composite DMSP-OLS night lights data, adjusted for the issues of blooming and top coding (Chiovelli et al., 2025).[37] As an alternative, we use better-quality higher-resolution VIIRS data, which are only available from 2012 onward. Panels (b) and (c) of figure E.2 show both rasters, transformed for the purpose of illustration to a binary "no vs. some lights" scale. In both rasters, about 99.5% of all cells are zeros, even though the more sensitive VIIRS sensors are able to detect more bright spots across the country than DMSP. This is reflective of the generally low performance of the night lights data in identifying economic activity in rural areas (Gibson et al., 2020), which is important for any local-level analysis in Sub-Saharan Africa.

According to DMSP data, there are no lights at all in 62 out of 73 region-language partitions and in 7 entire counties out of 15 (trivially resulting in zero measured BGI for those regions). When using VIIRS data, the number of partitions and full counties without any light falls to 49 and 2, respectively. Both sources pick up plenty of light around Monrovia, but since virtually the entire county and its population fall within a single homeland, BGI is estimated near 0.

[35] On the Murdock map, Montserrado is almost fully within the Bassa homeland. In GREG, it is split between the mislabeled "Bantu-speaking pygmy tribes" and Gola, while in GeoEPR, the only politically relevant group present in the county are the "Americo-Liberians."

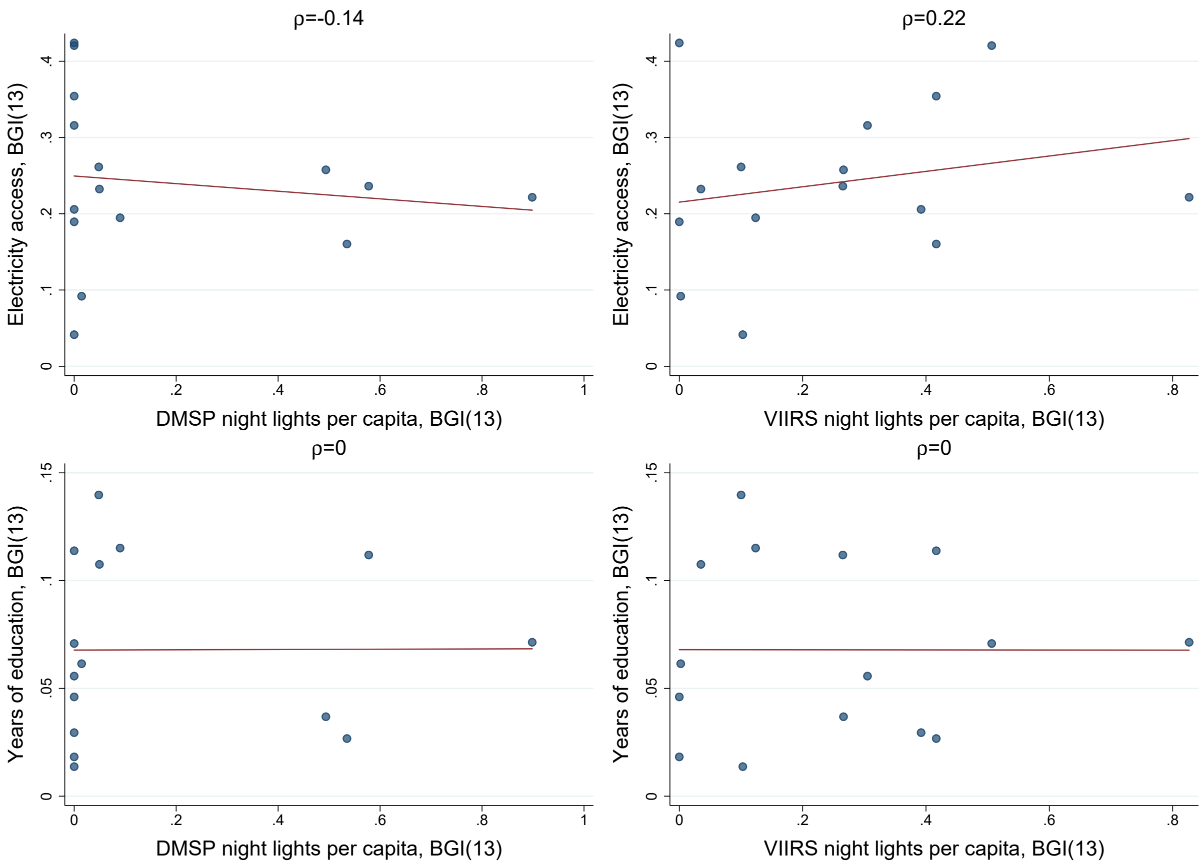
[36] Earlier studies made use of the G-Econ database, but its coarse resolution and reliance on official statistics at the regional level makes it unsuitable for subnational-level analysis, especially in poorer nations (Cederman et al., 2011, 2015).

[37] The "raw" DMSP data were used for the purpose of estimating country-level ethnic inequality by Alesina et al. (2016), Cederman et al. (2015), and Kuhn and Weidmann (2015), among others.

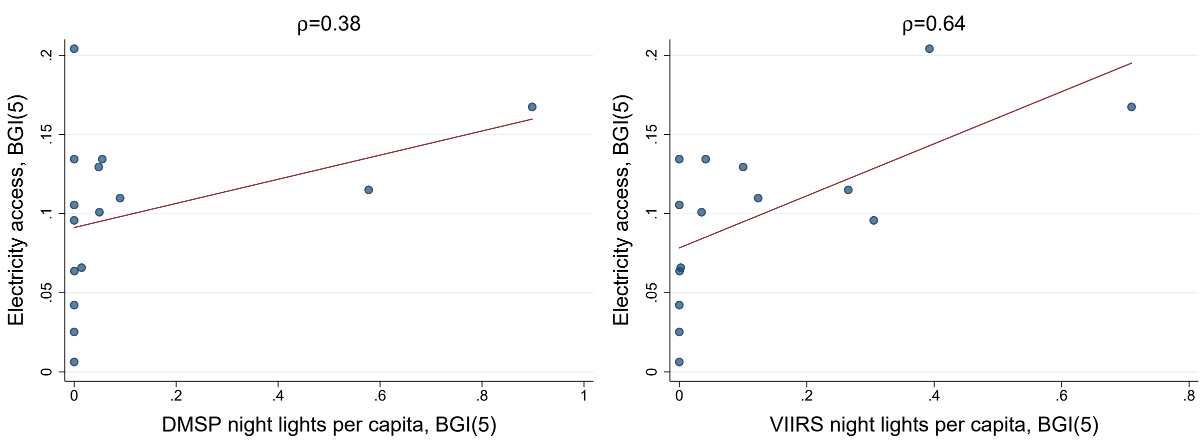
Overall, in rural regions of Liberia, where WLMS and LandScan do a better job at capturing regional ethnolinguistic composition, night lights data do a poor job in detecting economic activity. And in urbanized regions, where lights are more easily detected from space, GIS tools cannot generally give an accurate ethnolinguistic composition. Both issues have a major impact on the estimated BGI.

Figure E.3(a) compares BGI(13) indices based on night lights and IPUMS data on electricity access and educational attainment. As expected, DMSP measures for many counties are bunched around 0 and align very poorly with survey-based indices. VIIRS measures are a clear improvement and yield a modest positive correlation of 0.22 with survey-based BGI in electricity access (and zero correlation with BGI in education). Both lights-based indices are moderately positively correlated with BGI in television ownership (coefficients of 0.33 and 0.42, respectively) but that is not the case for car ownership (-0.28 and 0.12). The GIS and survey-based approaches are somewhat more aligned at deep tiers of the language tree. For example, at tier 5, pairwise correlations between BGI in electricity access and the lights-based measures increase to 0.38 and 0.64 for DMSP and VIIRS, respectively. As may be seen from panel (b) of figure E.3, much of this improvement is driven by a combination of mostly very low BGI values (due to the merging of groups into ancestral families) and a few higher-value outliers.

In sum, even the highest-quality combination of GIS inputs (WLMS, LandScan, and VIIRS) yields BGI indices that are at best moderately aligned with survey-based measures. The main challenges are not specific to Liberia and reflect the difficulty of accurately capturing both ethnolinguistic composition and fine differences in material well-being using GIS.



(a) BGI(13): electricity access, years of education, and night lights per capita



(b) BGI(5): electricity access and night lights per capita

Figure E.3: BGI across Liberian counties: surveys vs. GIS.

Notes. This figure compares BGI indicators based on the IPUMS 2008 dataset and the GIS approach using night lights data. Each point is one of 15 counties in Liberia. Best fitting line is shown for each scatterplot and ρ denotes bivariate correlation coefficient.

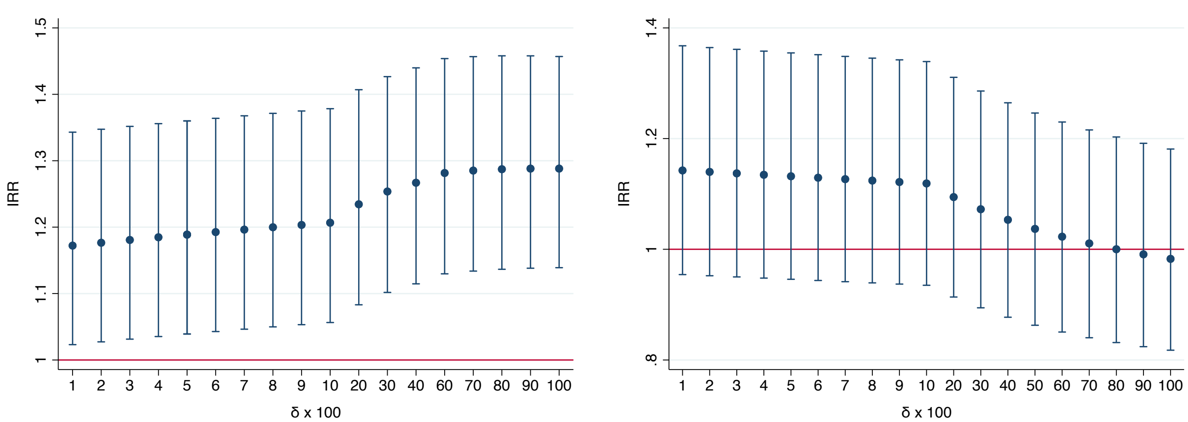
F Additional regression results

F.1 ELF and ELP indices adjusted for linguistic distance

An alternative way to account for shared ancestry of ethnolinguistic groups in the ELF and ELP indices is to add linguistic distance "weights" in their calculation. As proposed by Fearon (2003),

$$\text{ELF}_\delta = \sum_{i=1}^{N}\sum_{j=1}^{N} s_i s_j \tau_{ij}, \quad \tau_{ij} = 1 - (l/m)^\delta,$$

where l is the number of tree branches shared by languages i and j, m is the maximum possible number of such branches (equal to 13 for our Sub-Saharan African sample), and δ is a parameter that determines how fast the distance declines as the number of common branches goes up. A similar adjustment can be applied to the polarization index. While Fearon (2003) used $\delta = 0.5$, other studies preferred $\delta = 0.05$ (Desmet et al., 2009; Esteban et al., 2012). Since there is no well-grounded reason to fix δ at any given value, we follow Gershman and Rivera (2018) and compute ELF_δ and ELP_δ indices for 19 different values of δ, from 0.01 to 0.1 with step 0.01, and from 0.1 to 1 with step 0.1. Figure F.1 presents the results from baseline regressions for these alternative indices.

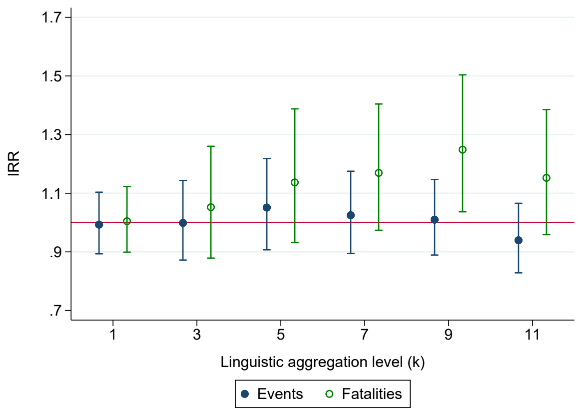


(a) ELF_δ (b) ELP_δ

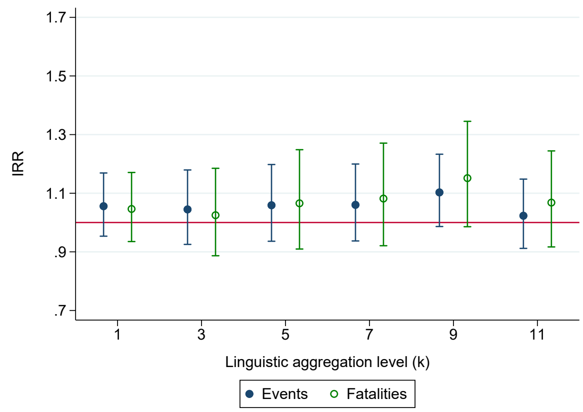
Figure F.1: Baseline regression results for alternative ELF and ELP measures.

Notes. Each panel presents incidence rate ratios, along with 95% confidence intervals, based on robust standard errors. For each reported value of δ, a negative binomial regression is estimated, where the outcome is the sum of ACLED conflict events and the right-hand-side variable of interest is either ELF_δ (panel a) or ELP_δ (panel b). All regressions include country fixed effects and baseline controls described in the main text. The number of observations is 391.

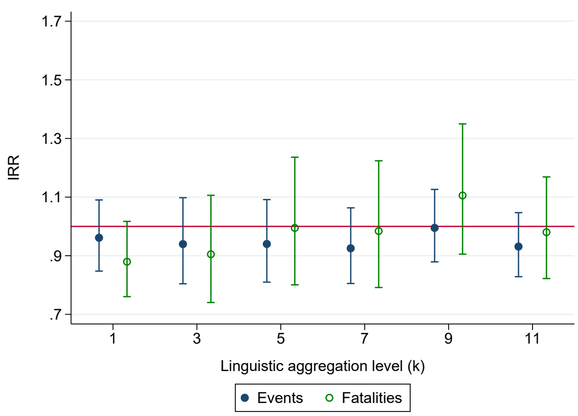
F.2 Further ethnic inequality indices and conflict



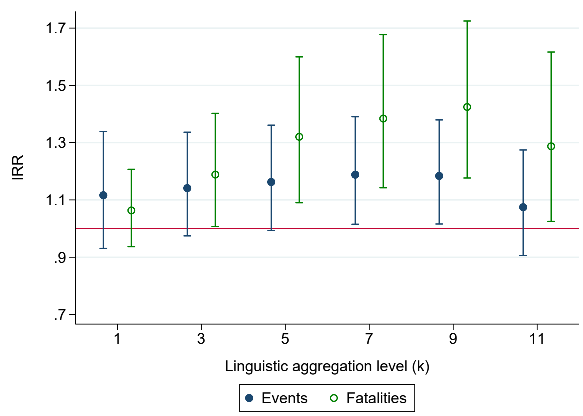
(a) Access to a high-quality water source



(b) Flush toilet ownership



(c) Finished floor ownership

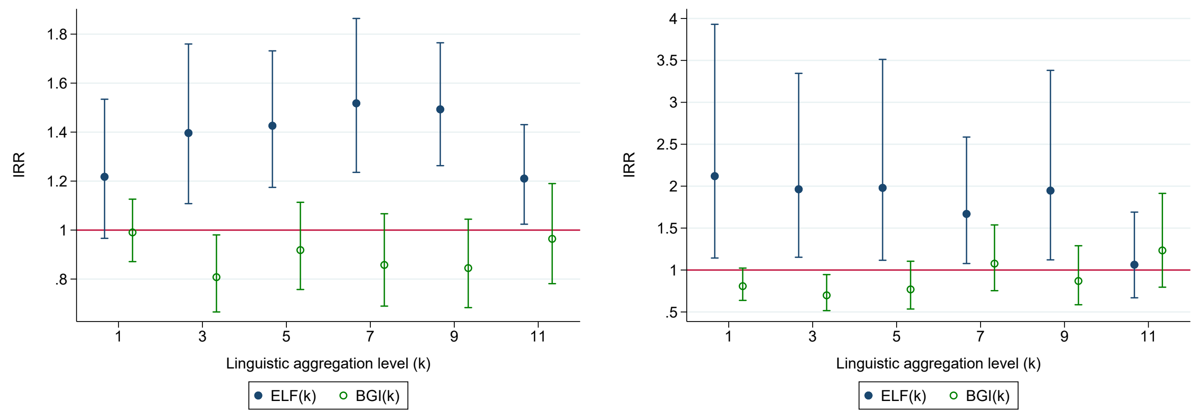


(d) Radio ownership

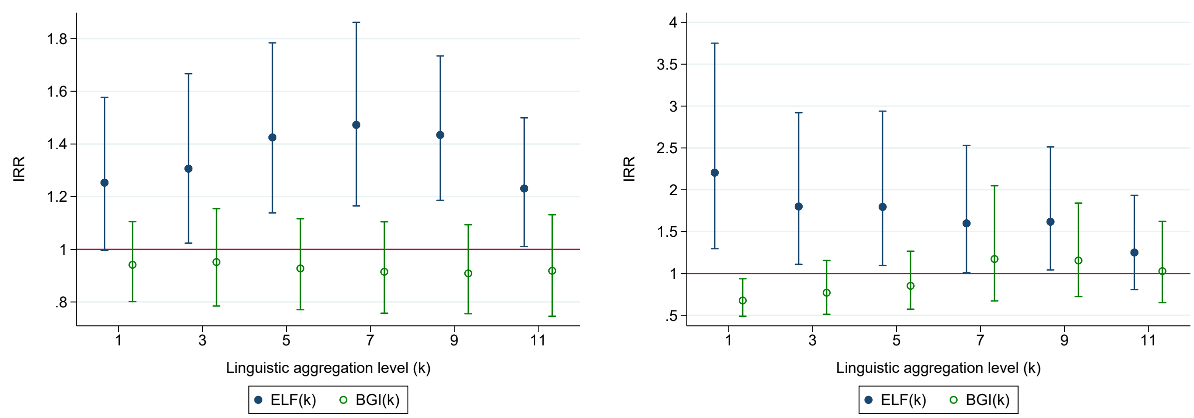
Figure F.2: Subnational ethnic inequality and conflict: additional BGI(k) measures.

Notes. Each panel presents incidence rate ratios, along with 95% confidence intervals, based on robust standard errors. For each of the six reported levels of linguistic aggregation, a negative binomial regression is estimated, where the outcome is the sum of either conflict events or fatalities, and the right-hand-side variable of interest is a BGI(k) index capturing ethnic inequality in the dimension indicated in the figure subtitle. All regressions include country fixed effects and baseline controls described in the main text. The number of observations is 391.

F.3 ELF versus BGI: Poisson regressions



(a) BGI in years of education and ELF as predictors of conflict events (left) and fatalities (right)

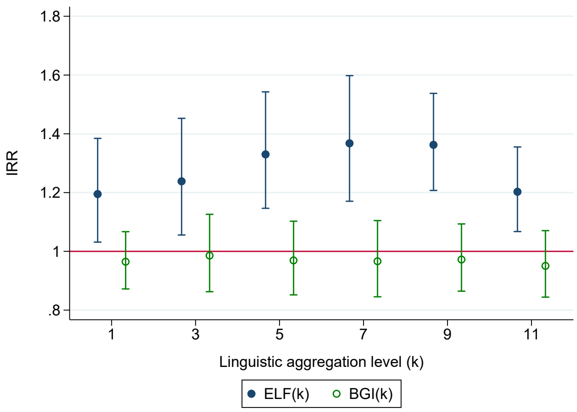


(b) BGI in electricity access and ELF as predictors of conflict events (left) and fatalities (right)

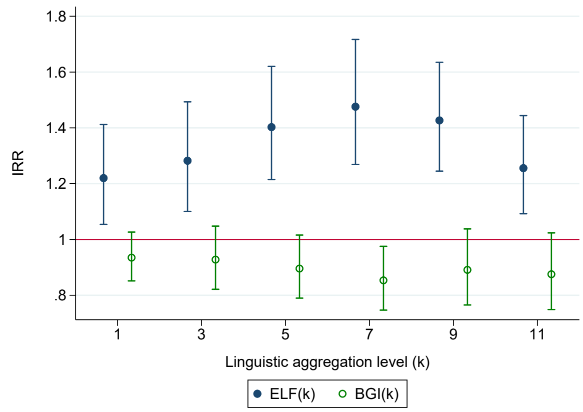
Figure F.3: Poisson "horse race" regressions for ELF and BGI.

Notes. Each panel presents incidence rate ratios, along with 95% confidence intervals, based on robust standard errors. For each of the six reported levels of linguistic aggregation, a Poisson regression is estimated, where the outcome is the sum of conflict events or fatalities. The right-hand side includes both ELF(k) and BGI(k) for years of education (panel a) electricity access (panel b), along with country fixed effects and baseline controls described in the main text. The number of observations is 391.

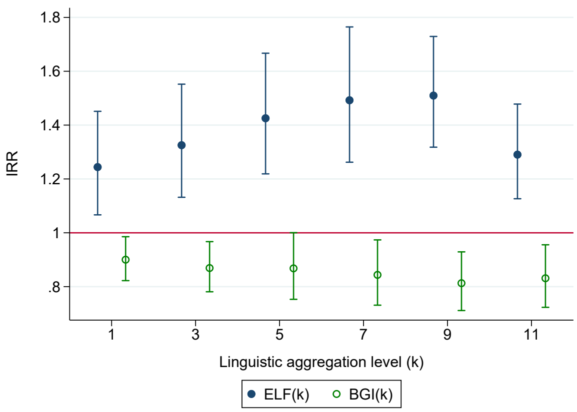
F.4 ELF versus BGI: Further "horse race" regressions



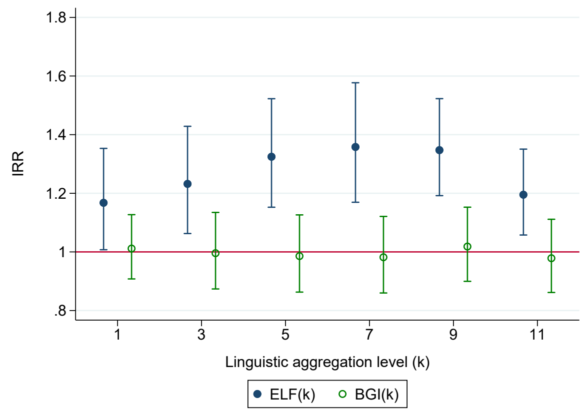
(a) Television ownership



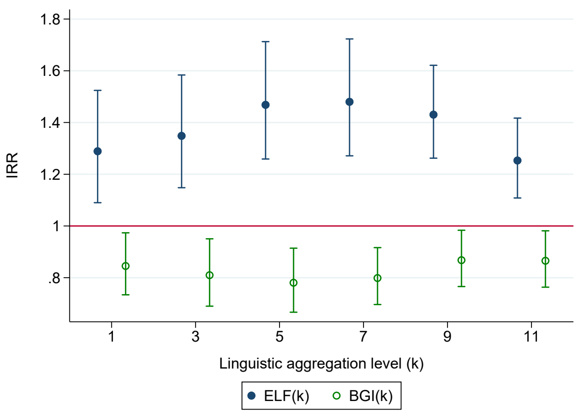
(b) Car ownership



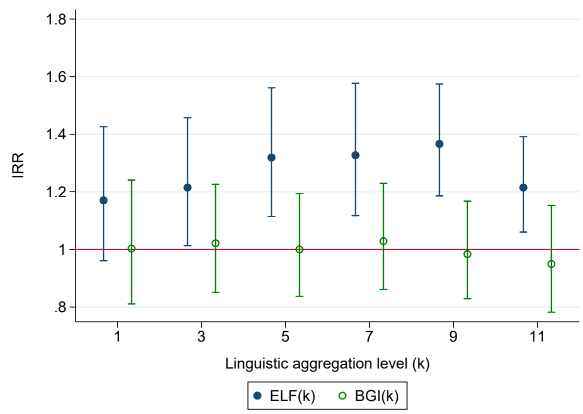
(c) Access to a high-quality water source



(d) Flush toilet ownership



(e) Finished floor ownership



(f) Radio ownership

Figure F.4: "Horse race" regressions for additional BGI(k) measures.

Notes. See figure 6 notes. The outcome variable is the count of conflict events.

F.5 Controlling for multiple hypothesis testing

Table F.1: Robustness to alternative specifications, separately for ELF and BGI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ELF(1)	1.169^{**}	1.221^{**}	1.189^{**}	1.123	1.140^{*}	1.184^{**}	1.167^{**}
	(0.042)	(0.013)	(0.030)	(0.119)	(0.084)	(0.030)	(0.044)
ELF(3)	1.242^{***}	1.292^{***}	1.261^{***}	1.216^{**}	1.217^{**}	1.256^{***}	1.237^{***}
	(0.009)	(0.001)	(0.007)	(0.019)	(0.018)	(0.007)	(0.009)
ELF(5)	1.325^{***}	1.361^{***}	1.330^{***}	1.278^{***}	1.254^{***}	1.326^{***}	1.323^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)	(0.001)	(0.001)
ELF(7)	1.357^{***}	1.377^{***}	1.352^{***}	1.288^{***}	1.258^{***}	1.354^{***}	1.355^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.007)	(0.001)	(0.001)
ELF(9)	1.350^{***}	1.365^{***}	1.340^{***}	1.296^{***}	1.256^{***}	1.348^{***}	1.350^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
ELF(11)	1.188^{***}	1.191^{***}	1.168^{**}	1.122^{*}	1.093	1.185^{**}	1.183^{**}
	(0.010)	(0.009)	(0.021)	(0.071)	(0.142)	(0.011)	(0.015)
BGI(1)	1.056	1.079	1.072	1.071	1.070	1.050	1.060
	(0.302)	(0.226)	(0.254)	(0.259)	(0.268)	(0.277)	(0.296)
BGI(3)	1.113	1.154^{*}	1.135	1.133	1.147	1.138^{*}	1.112
	(0.167)	(0.095)	(0.123)	(0.130)	(0.111)	(0.126)	(0.171)
BGI(5)	1.117	1.160^{*}	1.154^{*}	1.149^{*}	1.143^{*}	1.139	1.111
	(0.134)	(0.071)	(0.077)	(0.085)	(0.096)	(0.110)	(0.150)
BGI(7)	1.075	1.130	1.119	1.078	1.110	1.110	1.068
	(0.277)	(0.125)	(0.134)	(0.274)	(0.153)	(0.208)	(0.300)
BGI(9)	1.145^{*}	1.207^{**}	1.205^{**}	1.181^{**}	1.195^{**}	1.177^{**}	1.141^{*}
	(0.067)	(0.021)	(0.018)	(0.032)	(0.019)	(0.041)	(0.073)
BGI(11)	1.058	1.109	1.118	1.094	1.106	1.083	1.051
	(0.343)	(0.175)	(0.142)	(0.217)	(0.163)	(0.268)	(0.380)
Observations	391	391	391	376	391	391	391
Literacy		✓					
Electrification			✓				
IWI				✓			
Urbanization					✓		
Gini coefficient						✓	
Religious frac.							✓

Notes. a) The top panel of the table shows estimates from 42 separate negative binomial regressions of ACLED conflict events on ELF(k). The bottom panel does the same for BGI(k) in access to electricity. All regressions contain baseline controls described in the main text. IRRs are reported for each regression, along with sharpened q-values in parentheses (Anderson, 2008), as described in the main text. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively. b) Additional controls listed at the bottom of the table are regional averages for literacy rate, electricity access, international wealth index (IWI), urbanization rate, Gini coefficient for years of education, and an index of religious fractionalization.

Table F.2: Robustness to alternative specifications: ELF vs. BGI horse races

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ELF(1)	1.181*	1.227**	1.196**	1.111	1.133	1.194**	1.177*
	(0.054)	(0.024)	(0.048)	(0.185)	(0.130)	(0.045)	(0.061)
BGI(1)	0.979	0.990	0.989	1.022	1.014	0.983	0.983
	(0.511)	(0.559)	(0.550)	(0.508)	(0.545)	(0.525)	(0.528)
ELF(3)	1.246**	1.281***	1.256***	1.199**	1.193*	1.251**	1.240**
	(0.021)	(0.010)	(0.019)	(0.049)	(0.053)	(0.019)	(0.023)
BGI(3)	0.993	1.022	1.010	1.035	1.049	1.010	0.995
	(0.580)	(0.525)	(0.566)	(0.478)	(0.427)	(0.566)	(0.580)
ELF(5)	1.351***	1.366***	1.334***	1.269***	1.241**	1.344***	1.350***
	(0.001)	(0.001)	(0.002)	(0.010)	(0.019)	(0.002)	(0.002)
BGI(5)	0.957	0.991	0.994	1.017	1.025	0.970	0.958
	(0.434)	(0.566)	(0.580)	(0.545)	(0.511)	(0.491)	(0.434)
ELF(7)	1.414***	1.404***	1.377***	1.318***	1.260**	1.402***	1.413***
	(0.001)	(0.001)	(0.001)	(0.005)	(0.019)	(0.001)	(0.001)
BGI(7)	0.917	0.959	0.961	0.951	0.997	0.931	0.917
	(0.270)	(0.447)	(0.449)	(0.408)	(0.580)	(0.323)	(0.270)
ELF(9)	1.352***	1.341***	1.311***	1.267***	1.209**	1.339***	1.352***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.0179)	(0.001)	(0.001)
BGI(9)	0.997	1.046	1.054	1.058	1.092	1.017	0.997
	(0.580)	(0.424)	(0.385)	(0.370)	(0.237)	(0.537)	(0.580)
ELF(11)	1.202**	1.186**	1.153*	1.107	1.066	1.191**	1.197**
	(0.019)	(0.028)	(0.065)	(0.155)	(0.298)	(0.026)	(0.023)
BGI(11)	0.968	1.014	1.037	1.040	1.071	0.987	0.968
	(0.478)	(0.550)	(0.472)	(0.457)	(0.322)	(0.553)	(0.478)
Observations	391	391	391	376	391	391	391
Literacy		✓					
Electrification			✓				
IWI				✓			
Urbanization					✓		
Gini coefficient						✓	
Religious frac.							✓

Notes. a) The table shows estimates from 42 separate negative binomial horse race regressions of ACLED conflict events on ELF(k) and BGI(k). Each of the six panels displays results for a different level of linguistic aggregation k. All regressions contain baseline controls described in the main text. IRRs are reported for each regression, along with sharpened q-values in parentheses (Anderson, 2008), as described in the main text. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively. b) Additional controls listed at the bottom of the table are regional averages for literacy rate, electricity access, international wealth index (IWI), urbanization rate, Gini coefficient for years of education, and an index of religious fractionalization.

References

Adhvaryu, Achyuta, James Fenske, Gaurav Khanna, and Anant Nyshadham, "Resources, Conflict, and Economic Development in Africa," *Journal of Development Economics*, 2021, *149*, 102598.

Alcorta, Ludovico, Jeroen Smits, and Haley J. Swedlund, "Inequality and Ethnic Conflict in Sub-Saharan Africa," *Social Forces*, 2018, *97* (2), 769–792.

Alesina, Alberto and Eliana La Ferrara, "Ethnic Diversity and Economic Performance," *Journal of Economic Literature*, 2005, *43* (3), 762–800.

_ , Stelios Michalopoulos, and Elias Papaioannou, "Ethnic Inequality," *Journal of Political Economy*, 2016, *124* (2), 428–488.

Anderson, Michael L., "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 2008, *103* (484), 1481–1495.

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

Baldwin, Kate and John D. Huber, "Economic Versus Cultural Differences: Forms of Ethnic Diversity and Public Goods Provision," *American Political Science Review*, 2010, pp. 644–662.

Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli, "Adaptive Linear Step-Up Procedures That Control the False Discovery Rate," *Biometrika*, September 2006, *93* (3), 491–507.

Berman, Nicolas, Mathieu Couttenier, Dominic Rohner, and Mathias Thoenig, "This Mine is Mine! How Minerals Fuel Conflicts in Africa," *American Economic Review*, 2017, *107* (6), 1564–1610.

Bluhm, Richard, Roland Hodler, and Paul Schaudt, "Political Border Designs and the Salience of Ethnicity in Politics," March 2025. Working Paper, Department of Economics, University of St. Gallen.

Cameron, A. Colin and Frank A. G. Windmeijer, "R-Squared Measures for Count Data Regression Models with Applications to Health-Care Utilization," *Journal of Business & Economic Statistics*, 1996, *14* (2), 209–220.

Cederman, Lars-Erik, Nils B. Weidmann, and Kristian Skrede Gleditsch, "Horizontal Inequalities and Ethnonationalist Civil War: A Global Comparison," *American Political Science Review*, 2011, pp. 478–495.

_ , _ , **and Nils-Christian Bormann**, "Triangulating horizontal inequality: Toward improved conflict analysis," *Journal of Peace Research*, 2015, *52* (6), 806–821.

Cervellati, Matteo, Giorgio Chiovelli, and Elena Esposito, "Bite and Divide: Malaria and Ethnolinguistic Diversity," 2019. CEPR Discussion Paper No. DP13437.

Chiovelli, Giorgio, Stelios Michalopoulos, Elias Papaioannou, and Tanner Regan, "Illuminating the Global South," *Economic Journal*, 2025, *forthcoming.*

Collier, Paul and Anke Hoeffler, "Greed and Grievance in Civil War," *Oxford Economic Papers*, 2004, *56* (4), 563–595.

Cunningham, Kathleen Gallagher and Nils B. Weidmann, "Shared Space: Ethnic Groups, State Accommodation, and Localized Conflict1," *International Studies Quarterly*, 2010, *54* (4), 1035–1054.

Davies, Shawn, Garoun Engström, Therése Pettersson, and Magnus Öberg, "Organized Violence 1989–2023, and the Prevalence of Organized Crime Groups," *Journal of Peace Research*, 2024, *61* (4), 673–693.

de la Cuesta, Brandon and Leonard Wantchekon, "Is Language Destiny? The Origins and Consequences of Ethnolinguistic Diversity of Sub-Saharan Africa," in Victor Ginsburgh and Shlomo Weber, eds., *The Palgrave Handbook of Economics and Language*, Palgrave Macmillan UK, 2016, chapter 18, pp. 513–537.

Depetris-Chauvin, Emilio, "State History and Contemporary Conflict: Evidence from Sub-Saharan Africa," October 2015. SSRN Working Paper.

_ **and Ömer Özak**, "(De facto) Historical Ethnic Borders and Contemporary Conflict in Africa," May 2024. SSRN Working Paper.

Desmet, Klaus, Ignacio Ortuño-Ortín, and Ömer Özak, "Is Secessionism Mostly About Income or Identity? A Global Analysis of 3,153 Subnational Regions," *Economic Journal*, May 2025, *135*, 1261–1299.

_ , _ , **and Romain Wacziarg**, "The Political Economy of Linguistic Cleavages," *Journal of Development Economics*, 2012, *97* (2), 322–338.

— , — , and — , "On the Measurement of Social Heterogeneity," June 2025. NBER Working Paper No. 33896.

— , **Shlomo Weber, and Ignacio Ortuño-Ortín**, "Linguistic Diversity and Redistribution," *Journal of the European Economic Association*, December 2009, *7* (6), 1291–1318.

Esteban, Joan and Debraj Ray, "On the Measurement of Polarization," *Econometrica*, July 1994, *62* (4), 819–851.

— **and** — , "Conflict and Distribution," *Journal of Economic Theory*, 1999, *87* (2), 379–415.

— **and** — , "On the Salience of Ethnic Conflict," *American Economic Review*, December 2008, *98* (5), 2185–2202.

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

— **and** — , "A model of ethnic conflict," *Journal of the European Economic Association*, 2011, *9* (3), 496–521.

— , **Laura Mayoral, and Debraj Ray**, "Ethnicity and Conflict: An Empirical Study," *American Economic Review*, June 2012, *102* (4), 1310–42.

Fearon, James D., "Ethnic and Cultural Diversity by Country," *Journal of Economic Growth*, June 2003, *8* (2), 195–222.

— **and David D. Laitin**, "Ethnicity, Insurgency, and Civil War," *American Political Science Review*, 2003, pp. 75–90.

Filmer, Deon and Kinnon Scott, "Assessing Asset Indices," *Demography*, 2012, *49* (1), 359–392.

Fjelde, Hanne and Gudrun Østby, "Socioeconomic Inequality and Communal Conflict: A Disaggregated Analysis of Sub-Saharan Africa, 1990–2008," *International Interactions*, 2014, *40* (5), 737–762.

Galor, Oded and Ömer Özak, "The Agricultural Origins of Time Preference," *American Economic Review*, 2016, *106* (10), 3064–3103.

Gershman, Boris and Diego Rivera, "Subnational Diversity in Sub-Saharan Africa: Insights From a New Dataset," *Journal of Development Economics*, 2018, *133*, 231–263.

_ and _ , "Measuring Regional Ethnolinguistic Diversity in Sub-Saharan Africa: Surveys vs. GIS," *The World Bank Economic Review*, 2020, *34* (S1), S40–S45.

Gibson, John, Susan Olivia, and Geua Boe-Gibson, "Night Lights In Economics: Sources and Uses," *Journal of Economic Surveys*, 2020, *34* (5), 955–980.

Ginsburgh, Victor and Shlomo Weber, "The Economics of Language," *Journal of Economic Literature*, June 2020, *58* (2), 348–404.

Guarnieri, Eleonora, "Cultural Distance and Ethnic Civil Conflict," *American Economic Review*, April 2025, *115* (4), 1338–68.

Gurr, Ted Robert, *Why Men Rebel*, Princeton, NJ: Princeton University Press, 1970.

Harari, Mariaflavia and Eliana La Ferrara, "Conflict, Climate, and Cells: A Disaggregated Analysis," *Review of Economics and Statistics*, 10 2018, *100* (4), 594–608.

Hillesund, Solveig and Gudrun Østby, "Horizontal Inequalities, Political Violence, and Nonviolent Conflict Mobilization: A Review of the Literature," *Journal of Economic Surveys*, 2023, *37* (5), 1589–1635.

_ , Karim Bahgat, Gray Barrett, Kendra Dupuy, Scott Gates, Håvard Mokleiv Nygård, Siri Aas Rustad, Håvard Strand, Henrik Urdal, and Gudrun Østby, "Horizontal Inequality and Armed Conflict: A Comprehensive Literature Review," *Canadian Journal of Development Studies*, 2018, *39* (4), 463–480.

Hodler, Roland, Sorawoot Srisuma, Alberto Vesperoni, and Noémie Zurlinden, "Measuring Ethnic Stratification and Its Effect on Trust in Africa," *Journal of Development Economics*, 2020, *146*, 102475.

Horowitz, Donald L., *Ethnic Groups in Conflict*, Berkeley, CA: University of California Press, 1985.

Houle, Christian and Cristina Bodea, "Ethnic Inequality and Coups in Sub-Saharan Africa," *Journal of Peace Research*, 2017, *54* (3), 382–396.

Howe, Laura D., James R. Hargreaves, and Sharon R.A. Huttly, "Issues in the Construction of Wealth Indices for the Measurement of Socio-Economic Position in Low-Income Countries," *Emerging Themes in Epidemiology*, 2008, *5* (3), 1–14.

Huber, John D. and Laura Mayoral, "Group Inequality and the Severity of Civil Conflict," *Journal of Economic Growth*, 2019, *24* (1), 1–41.

Justin, Peter Hakim and Lotje De Vries, "Governing Unclear Lines: Local Boundaries as a (Re)source of Conflict in South Sudan," *Journal of Borderlands Studies*, 2019, *34* (1), 31–46.

Kiszewski, Anthony, Andrew Mellinger, Andrew Spielman, Pia Malaney, Sonia Ehrlich Sachs, and Jeffrey Sachs, "A Global Index Representing the Stability of Malaria Transmission," *American Journal of Tropical Medicine and Hygiene*, 2004, *70* (5), 486–498.

Kuhn, Patrick M. and Nils B. Weidmann, "Unequal We Fight: Between- and Within-Group Inequality and Ethnic Civil War," *Political Science Research and Methods*, 2015, *3* (3), 543–568.

Leipziger, Lasse Egendal, "Measuring Ethnic Inequality: An Assessment of Extant Cross-National Indices," *British Journal of Political Science*, 2023, *53* (2), 652–673.

Lewis, M. Paul, ed., *Ethnologue: Languages of the World*, 16 ed., Dallas: SIL International, 2009.

McGuirk, Eoin F. and Nathan Nunn, "Transhumant Pastoralism, Climate Change, and Conflict in Africa," *Review of Economic Studies*, January 2025, *92* (1), 404–441.

McKenzie, David J., "Measuring Inequality With Asset Indicators," *Journal of Population Economics*, 2005, *18* (2), 229–260.

Michalopoulos, Stelios, "The Origins of Ethnolinguistic Diversity," *American Economic Review*, 2012, *102* (4), 1508–39.

_ **and Elias Papaioannou**, "The Long-Run Effects of the Scramble for Africa," *American Economic Review*, 2016, *106* (7), 1802–48.

Mitton, Todd, "The Wealth of Subnations: Geography, Institutions, and Within-Country Development," *Journal of Development Economics*, 2016, *118*, 88–111.

Montalvo, José G. and Marta Reynal-Querol, "Ethnic Polarization, Potential Conflict, and Civil Wars," *American Economic Review*, 2005, *95* (3), 796–816.

_ **and** _ , "Cultural Diversity, Conflict, and Economic Development," in Victor A. Ginsburgh and David Throsby, eds., *Handbook of the Economics of Art and Culture*, Vol. 2, Elsevier, 2014, chapter 18, pp. 485–506.

Müller-Crepon, Carl, "Building Tribes: How Administrative Units Shaped Ethnic Groups in Africa," *American Journal of Political Science*, 2025, *69* (2), 406–422.

Murdock, George P., *Africa: Its Peoples and Their Cultural History*, New York: McGraw-Hill Book Company, 1959.

Nunn, Nathan and Diego Puga, "Ruggedness: The Blessing of Bad Geography in Africa," *Review of Economics and Statistics*, February 2012, *94* (1), 20–36.

Østby, Gudrun, "Polarization, Horizontal Inequalities and Violent Civil Conflict," *Journal of Peace Research*, 2008, *45* (2), 143–162.

Posner, Daniel N., "Measuring Ethnic Fractionalization in Africa," *American Journal of Political Science*, 2004, *48* (4), 849–863.

Raleigh, Clionadh, Andrew Linke, Håvard Hegre, and Joakim Karlsen, "Introducing ACLED: An Armed Conflict Location and Event Dataset," *Journal of Peace Research*, 2010, *47* (5), 651–660.

Ray, Debraj and Joan Esteban, "Conflict and Development," *Annual Review of Economics*, 2017, *9*, 263–293.

Salehyan, Idean, Cullen S. Hendrix, Jesse Hamner, Christina Case, Christopher Linebarger, Emily Stull, and Jennifer Williams, "Social Conflict in Africa: A New Database," *International Interactions*, 2012, *38* (4), 503–511.

Smits, Jeroen and Roel Steendijk, "The International Wealth Index (IWI)," *Social Indicators Research*, 2015, *122* (1), 65–85.

Stewart, Frances, ed., *Horizontal Inequalities and Conflict: Understanding Group Violence in Multiethnic Societies*, Houndmills, UK: Palgrave Macmillan, 2008.

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.

Weidmann, Nils B., Jan Ketil Rød, and Lars-Erik Cederman, "Representing Ethnic Groups in Space: A New Dataset," *Journal of Peace Research*, July 2010, *47* (4), 491–499.