Subnational Diversity in Sub-Saharan Africa: Insights from a New Dataset*

Boris Gershman†
American University

Diego Rivera
American University

January 2018

Abstract

This paper presents a new dataset on subnational ethnolinguistic and religious diversity in Sub-Saharan Africa covering 36 countries and almost 400 first-level administrative units. We use population censuses and large-scale household surveys to compile detailed data on the ethnolinguistic composition of each region and match all reported ethnicities to *Ethnologue*, a comprehensive catalog of world languages. This matching allows us to standardize the notion of an ethnolinguistic group and account for relatedness between language pairs, a correlate of shared history and culture, when calculating diversity indices. Exploiting within-country variation provided by our new dataset, we find that local public goods provision, as reflected in metrics of education, health, and electricity access, is negatively related to ethnolinguistic diversity, but only if the underlying basic languages are first aggregated into larger families or if linguistic distances between groups are taken into consideration. In other words, only deep-rooted diversity, based on cleavages formed in the distant past, is strongly inversely associated with a range of regional development indicators. Furthermore, we show that subnational diversity has been remarkably persistent over the past two-three decades implying that population sorting in the short to medium run is unlikely to bias our main findings.

Keywords: African development, ethnolinguistic diversity, public goods provision, religious diversity, subnational analysis

JEL Classification Numbers: H41, O10, O15, Z12, Z13

---

*We are grateful to Nathan Nunn and two anonymous referees for their valuable advice. We also thank seminar and conference participants at the 2016 ASREC conference at Chapman University, 2016 "Development Economics and Policy" conference at Heidelberg University, 2016 "Ethnicity and Diversity" conference at Universidad Carlos III de Madrid, 2016 ABCDE conference on "Data and Development Economics" in Washington DC, 2017 CSAE conference at the University of Oxford, 2017 workshop on "Culture, Economic Development, and Diversity" in Groningen, American University, and Florida International University. A special thanks goes to Bridgette Wellington for support with the DHS and to Jeroen Smits for sharing the data on the international wealth index. Quamrul Ashraf, Lars-Erik Cederman, Klaus Desmet, Avital Livny, Omar McDoom, Ömer Özak, and Rutger Schilpzand contributed useful comments.

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

1 Introduction

Ever since the seminal contribution of Easterly and Levine (1997), ethnic diversity has been one of the most thoroughly explored deep determinants of economic development in general and Africa's "growth tragedy" in particular.[1] Despite the growing number of rigorous empirical studies, the overall evidence remains mixed and the debate continues, with special attention given to the issues of data quality and the choice of appropriate diversity and development metrics.

This paper presents a new high-quality subnational-level dataset on ethnolinguistic diversity covering 36 countries and almost 400 first-level administrative units in Sub-Saharan Africa. We first use the available population censuses and large-scale household surveys to extract detailed information on regional ethnolinguistic composition in each country. We next standardize the notion of an ethnolinguistic group by matching reported ethnicities to *Ethnologue*, a comprehensive catalog of world languages. Beyond providing a benchmark for defining unique groups, this matching also incorporates our dataset into *Ethnologue*'s family tree model which captures the historical structure of relationships between languages. Finally, based on the distribution of 750 ethnolinguistic groups across regions in our sample, we produce a variety of diversity metrics, namely fractionalization and polarization indices adjusted for linguistic similarity or calculated at different levels of linguistic aggregation. Therefore, we explore both recent and deep cleavages in the ethnolinguistic structure of each region's population.

Having compiled this new dataset, we use it to examine the association between regional diversity and various development indicators, with a particular focus on local public goods provision as reflected by access to schooling, health facilities, and electricity. Our analysis shows that diversity indices based on fully disaggregated lists of ethnolinguistic groups, as they are provided in the original surveys, are not significantly related to subnational development in the vast majority of specifications. However, once linguistic relatedness is taken into consideration, a striking robust pattern emerges. Diversity indices that are calculated for groups aggregated into larger ethnolinguistic families or that are directly adjusted for linguistic similarities between groups turn out to be significantly negatively related to local public goods provision. In other words, only deep-rooted diversity, driven by cleavages formed in the distant past, is strongly connected to a range of contemporary development outcomes.

[1] See Cuesta and Wantchekon (2016) for a recent overview of research on ethnolinguistic diversity in economics and political science focusing on Sub-Saharan Africa.

In specifications that account for a host of geographic characteristics, urbanization rate, and country fixed effects, our regression estimates imply that a one-standard-deviation increase in deep-rooted diversity, as measured by either fractionalization or polarization index, is associated with a deterioration in educational and health outcomes, such as literacy rate and prevalence of child malnutrition, in the range of 0.1–0.2 standard deviations. When household access to electricity is used as an outcome variable, the relevant standardized point estimates are more modest, not exceeding 0.09 in absolute value. These findings are robust to excluding regions with less reliable data on ethnolinguistic composition, highly urbanized areas, and administrative units containing capital cities. Standard stress-tests imply that, in order to completely explain away our findings, selection on unobservables would have to be of a larger magnitude than selection on observable characteristics and actually bias our coefficients of interest in the opposite direction.

Our results for broader indicators of regional development are mixed. Nighttime luminosity, a metric highly correlated with electricity access, is negatively associated with the whole range of diversity indices, and the magnitude of respective standardized coefficient estimates is in the range between 0.075 and 0.15. However, the results for income per capita and household wealth are largely insignificant, highlighting the importance of differentiating between various types of development indicators in the studies of diversity. The negative relationship to deep-rooted diversity only emerges in the analyses of outcomes capturing local public goods provision.

In order to investigate whether population sorting is likely to bias our estimates, we explore the dynamics of subnational diversity. Specifically, for five countries in our sample, we calculate and compare regional ELF indices at different points in time separated by two-three decades. The correlation between these pairs of indices is close to 0.97 on average, that is, subnational diversity is remarkably persistent. Furthermore, the tiny observed changes in diversity are completely unrelated to contemporary economic activity, consistent with the absence of significant population sorting across regions in the short to medium run.

Finally, in addition to ethnolinguistic diversity, the main subject of this paper, we also briefly explore subnational religious divisions. We construct religious diversity indices for the regions in our sample and show that, first, they are not systematically related to any development indicators and, second, their inclusion in our main specifications does not alter any reported findings on ethnolinguistic diversity.

This study contributes to the large literature on diversity and economic performance. Our first contribution is the new subnational-level dataset that we argue is superior to existing alternatives. While there are several standard national-level datasets on diversity

that are employed in cross-country analyses (Alesina et al., 2003; Fearon, 2003; Desmet et al., 2012), there have been only a few attempts to systematically examine the ethnolinguistic composition of subnational regions, notably by Alesina and Zhuravskaya (2011) and Gerring et al. (2015). As we make clear below, our database improves upon these efforts in several major ways. First, it covers a much larger sample of countries and first-level administrative regions in Sub-Saharan Africa. Second, we employ more recent and/or higher quality data sources, including national censuses that account for more than 50% of our sample. Third, unlike earlier studies, we thoroughly examine all groups listed in each original survey and match them to the corresponding *Ethnologue* language codes thereby standardizing the notion of an ethnolinguistic group. Fourth and most importantly, in addition to standard fractionalization and polarization measures, we construct two sets of diversity indices accounting for linguistic relatedness between groups. To the best of our knowledge, this is the first study providing such indices at the subnational level, a crucial step forward which, as it turns out, makes all the difference for the empirical significance of regional diversity.[2]

Our second contribution is the new analysis of the relationship between ethnolinguistic diversity and development outcomes. Conceptually, the nature of this relationship is not a priori clear since there are multiple channels through which diversity may affect socioeconomic performance, both positively and negatively.[3] On the one hand, high ethnic diversity may be associated with conflicting preferences and beliefs breeding mistrust, social antagonism, and lack of cooperation, which result in diminished public goods provision. On the other hand, diversity may bring together a variety of complementary skills boosting productivity. Whether the net impact of diversity is positive or negative is ultimately an empirical question, the answer to which may depend on the regional context, the chosen unit of analysis, diversity index, and the type of socioeconomic outcome. Complicating matters, diversity may itself be responsive to local environment and shaped in part by migration of people searching for better economic opportunities or fleeing conflict.

Early cross-country empirical studies mainly found a negative association between ethnic diversity and a variety of performance indicators including income per capita and eco-

[2]In addition, our methodology is in many ways preferable to the approach based on combining digital maps of ethnolinguistic groups with disaggregated population data, which is prone to measurement error due to inaccurate "homeland" boundaries, ad hoc aggregation of groups, noisy imputed regional population shares, and inability to capture high diversity in urban areas (Gershman and Rivera, 2018).

[3]Miguel and Gugerty (2005), Alesina and La Ferrara (2005), Habyarimana et al. (2007), Esteban and Ray (2011), Ashraf and Galor (2013), among many others, discuss various mechanisms plausibly linking diversity to social and economic outcomes.

nomic growth, quality of governance and institutions, public goods provision, human and social capital.[4] In addition, some authors emphasized the importance of interaction effects between diversity, political institutions, and income. For instance, Collier (2000) shows that ethnic diversity is only negatively related to economic growth in non-democracies. This result is corroborated by the analysis in Alesina and La Ferrara (2005) who find a positive interaction effect between diversity and income per capita in standard growth regressions. Their interpretation is that the beneficial role of diversity is more likely to manifest itself in countries that are richer and have better institutions. More recently, Ashraf and Galor (2013) found a hump-shaped relationship between genetic diversity, a fundamental determinant of ethnic diversity, and contemporary income per capita, a pattern consistent with the presence of both positive and adverse effects of diversity on productivity.

An important aspect of the debate on measurement that emerged in the cross-country literature is the importance of accounting for group similarities when calculating diversity indices. Fearon (2003) offered the first country-level dataset in which fractionalization measures were adjusted for linguistic distances between groups. Desmet et al. (2009) showed that this adjustment matters in applications: in their analysis, only the indices accounting for linguistic distances are negatively related to redistribution. Desmet et al. (2012) suggested an alternative approach to capture relatedness between linguistic groups by first aggregating them into larger families and then measuring diversity for these deeper divisions. They further showed that the choice of aggregation level makes a difference for the empirical relationship between diversity and development outcomes across countries. Our paper directly contributes to this line or research by constructing both varieties of indices accounting for the structure of ethnolinguistic cleavages at the subnational level for Sub-Saharan Africa and showing that such adjustments are indeed crucial for the accurate analysis of the association between local diversity and public goods provision.

Given the well-known drawbacks of cross-country studies lumping together heterogeneous nations from around the world without being able to control for all relevant country-specific characteristics, a wave of research focused instead on within-country variation in diversity across regions, districts, or even smaller units of analysis. Miguel and Gugerty (2005) establish a negative relationship between ethnic diversity and local public goods

[4]See Easterly and Levine (1997), La Porta et al. (1999), Collier (2000), Alesina et al. (2003), Alesina and La Ferrara (2005), Bjørnskov (2007), among others. An extensive literature in political science and economics focuses on the relationship between diversity and conflict, see Fearon and Laitin (2003), Montalvo and Reynal-Querol (2005), Esteban et al. (2012), and references therein.

provision, namely school funding and water well maintenance, in rural western Kenya.[5] Glennerster et al. (2013) find no significant association between ethnic diversity and public goods provision across chiefdoms in Sierra Leone. Yet another recent study examines a similar question exploiting the variation across districts in Zambia and finds a positive relationship between ethnic heterogeneity and certain welfare outcomes related to publicly provided goods and services (Gisselquist et al., 2016). Gerring et al. (2015) compile a microlevel dataset on developing countries from around the world and show that, while there is a negative association between ethnic diversity and socioeconomic outcomes at the country level, it disappears or even reverses the sign at the regional and district levels.[6]

As can be seen from this sample of recent studies, evidence on the relationship between diversity and development indicators at the subnational level is largely inconclusive. Notably, none of the cited papers make any attempt to account for similarities between various groups when measuring ethnic diversity. In contrast, our analysis shows that, unlike commonly used metrics, it is precisely the indices adjusted for linguistic relatedness that are systematically negatively related to a range of development outcomes in a broad sample of African regions. This finding is consistent with the notion that the extent of dissimilarity between groups matters for the ultimate impact of diversity on cooperation, collective action, and the provision of local public goods.

More generally, our paper contributes to the growing literature on subnational development exploiting within-country variation to establish robust determinants of economic outcomes. For instance, Gennaioli et al. (2013) construct a large dataset on regions from 110 countries and show that differences in regional human capital account for a large share of variation in subnational income per capita. Hodler and Raschky (2014) find that regions in a broad sample of countries are better developed, as measured by higher nighttime luminosity, if the current political leader was born there, in line with the idea of regional favoritism. Mitton (2016) explores the role of geography and local institutions in explain-

[5]Miguel (2004) shows that, in contrast, there is no such relationship in a nearby district in Tanzania suggesting that nation-building reforms in that country mitigated the adverse effects of ethnic diversity on public goods provision.

[6]Robinson (2018) uses Afrobarometer surveys for 16 countries and census data for Malawi to show that regional ethnic diversity is associated with lower "ethnocentric" trust. Beyond Africa, Alesina et al. (2015) show that deforestation is positively related to the degree of ethnic fractionalization across Indonesian districts. Algan et al. (2016) exploit the rules of public housing allocation in France to establish that ethnic diversity measured at the apartment block level induces "social anomie" leading to increased vandalism and lower levels of building maintenance. Beugelsdijk et al. (2018) establish that greater diversity of cultural values is negatively associated with economic performance and local public goods provision in a sample of European regions.

ing regional development. We complement this literature by revisiting the connection between ethnolinguistic diversity and socioeconomic development at the subnational level. Although our analysis focuses on Sub-Saharan Africa, the same approach can be used to expand the coverage to other regions of the world in future research.

The rest of the paper is organized as follows. The following section provides a detailed description of our new dataset. Section 3 reports the main findings on the relationship between subnational diversity and development indicators. Section 4 examines the persistence of regional diversity. Section 5 conducts a robustness analysis and section 6 concludes. Appendices contain additional results, information on data sources, descriptive statistics, definitions of all variables, and examples clarifying the process of constructing adjusted diversity measures.

2 The new dataset

2.1 Basic principles

In all empirical work on diversity, the quality of data on ethnolinguistic composition is naturally the first-order consideration. We followed several basic principles to fix the list of ethnolinguistic groups for each country and region in our sample. First, other things equal, we chose the data source with the most detailed list of groups. That is, we are agnostic about which groups are more important in terms of their political influence, historical presence in the region's territory, or any other criteria, and we treat them all equally.[7] Having picked the source with the longest list of groups we then matched them to their respective spoken languages as documented in *Ethnologue*. This matching serves as a standardization device, where the existence of a distinct spoken language effectively defines group identity. In the vast majority of cases, ethnic groups in Sub-Saharan Africa have their own corresponding languages, often with a similar name. In some cases, however, multiple ethnic groups speak exactly the same language or closely related dialects that do not have their own *Ethnologue* classification codes. Such groups are assigned the same language code and thus merge into a single category.[8]

[7] In contrast, Posner (2004) offers an ELF index based only on "politically relevant" ethnic groups in African countries. While this approach makes sense when analyzing policy-mediated effects of diversity at the national level, it is unnecessarily restrictive when broader transmission channels are considered.

[8] For example, the Senegalese census of 2002 distinguishes between the Toucouleur and the Fulani people, two closely related ethnic groups. The Fulani speak Pulaar, while the Toucouleur speak a dialect of that language which does not have its own *Ethnologue* code. Hence, we classify both groups as Pulaar.

As explained in more detail further below, we take into account the relatedness between ethnolinguistic groups when measuring diversity. The *Ethnologue* matching plays a crucial role in this process since it automatically integrates our groups into a linguistic family tree model, which enables the calculation of proximity between any pair of languages. It also allows to trace the historical roots of contemporary languages and aggregate them into families representing their more or less distant common ancestors.

The group standardization process also helped us to identify the cases in which the original data sources lumped multiple ethnic groups together implicitly assuming that they are identical or at least very similar to each other. Whenever we were unable to identify the unique language representing such cluster in a particular region, we marked the whole category as "other," acknowledging our inability to pin down the relevant groups. This process made us completely discard the available data on regional ethnolinguistic composition of the Democratic Republic of the Congo (DRC). For example, the 2013–2014 DHS survey aggregates ethnic groups in the DRC in a few categories based on their geographic location, an arbitrary classification considering the enormous ethnic diversity in this country.[9]

Finally, selecting the unit of analysis is also important in any study of diversity. Our choice of the first-level subnational administrative regions as basic units is prompted by two main factors: data availability and the political and economic relevance of these divisions. First, local diversity should ideally be measured using regionally representative data. Given the scarcity of detailed data on subnational ethnolinguistic composition in Sub-Saharan Africa, the crudest first-level administrative division enables the widest possible country coverage. Employing lower-level subnational units would harm either the property of representativeness or country coverage (or both). Similarly, given the ultimate goal of exploring the relationship between diversity and development indicators, the latter need to be both available and representative for the same units of analysis, and our choice again appears to be the most appropriate.

Second, we are well aware of the "modifiable areal unit problem" which refers to the potential sensitivity of results to the definition of spatial units of analysis and thus motivates the selection of such units based on the underlying theoretical considerations rather than simply data availability. Fortunately, in our case, the two criteria are well-aligned. Most of our development indicators reflect the extent and quality of the local provision of

[9]The 2010 MICS survey offers a more natural five-category aggregation of ethnic groups (Bantu, Sudanese, Nilotic, Hamitic, and Pygmies), but such classification still does not allow to perform the *Ethnologue* matching since each of these categories contains a variety of distinct languages.

public goods such as schools, hospitals, and access to electricity. Since relevant policies are often implemented or at least affected by authorities at the first level of administrative division, the latter is a natural and well-defined choice for outlining within-country boundaries. Having said that, we readily admit that regional boundaries are endogenous and sometimes designed with a certain distribution of ethnic groups in mind. Furthermore, we do not claim that our results would necessarily hold under alternative choices of the basic unit of analysis. Table A.1 in the appendix provides the number of regions for each of the 36 countries in our sample and describes the employed sets of administrative boundaries.[10]

2.2 Data sources

In search of the data on ethnolinguistic composition of subnational regions in Sub-Saharan Africa we reviewed close to 200 surveys and reports containing such information. When available, our preferred source in almost all cases was either a national census or its subsample offered by the Integrated Public Use Microdata Series (IPUMS) project, since censuses cover most of the country's population and, as we found out, typically provide the most detailed lists of ethnolinguistic groups relative to other types of surveys.[11] Overall, we were able to find census or IPUMS data on the ethnic composition of 202 regions in 22 countries, which constitutes more than 50% of our full sample.

In the absence of census data, we used large-scale household surveys, mostly various waves of DHS (Demographic and Health Surveys) and MICS (Multiple Indicator Cluster Surveys). Our choice of the best available survey was based on multiple criteria including the number of listed ethnic groups, sample size, proper subnational coverage, and the regional population shares of unidentified "other" ethnicities. We also preferred more recent surveys to the older ones, other things equal. Overall, data for 13 countries are based on either DHS or MICS.[12] For the single remaining country, Zimbabwe, the above sources did not provide enough detail and we instead relied on one of the World Health Organization (WHO) surveys. Table A.2 in the appendix summarizes the information on data sources and shows survey years for each country in the sample.

[10]The number of subnational units per country varies from 3 in Malawi to 37 in Nigeria.

[11]For instance, the 2010 IPUMS dataset for Ghana contains 38 unique ethnolinguistic groups, whereas the most comprehensive DHS survey from 2014 lists only 8 of them. In cases when both census and large-scale household surveys were available and used similar classification of ethnolinguistic groups, we found that the resulting regional diversity measures were generally very close.

[12]In cases when both male and female surveys were available, we combined them together and calculated population shares of listed ethnic groups based on the comprehensive samples of respondents.

After carefully considering all of the available data sources, we were unable to cover several countries in Sub-Saharan Africa. In addition to the DRC case mentioned above, no reliable data on subnational ethnic composition could be found for Sudan and South Sudan, Somalia, Lesotho, Rwanda, Burundi, Madagascar, and small island nations.[13] In total, our database covers 36 countries, 398 regions, and 750 unique ethnolinguistic groups.

2.3 Measuring diversity

Consider a region hosting N ethnolinguistic groups and let s_i be the share of group i in the region's population, so that $\sum_{i=1}^{N} s_i = 1$. Given this population structure, the most commonly used diversity measure is the index of ethnolinguistic fractionalization, defined as $\text{ELF} = 1 - \sum_{i=1}^{N} s_i^2$, which captures the probability that two randomly chosen residents in the region belong to distinct ethnolinguistic groups.

By construction, the standard ELF index treats all groups as equally distinct. This is problematic if one believes, for instance, that the degree of cooperation between groups depends on how different they are in terms of history, language, or culture. Literature going back to the seminal paper by Greenberg (1956) suggested to adjust the standard ELF index by incorporating information on group similarities. The generalized version of the ELF index may be calculated as $\sum_{i=1}^{N} \sum_{j=1}^{N} s_i s_j \tau_{ij}$, where τ_{ij} is the "distance" between groups i and j. Such index may be interpreted as the expected distance between two randomly selected residents of the region. A natural question is then how to measure the distance for each pair of groups.

A popular solution to this problem in recent studies has been to calculate linguistic distance between groups (Ginsburgh and Weber, 2016).[14] The so-called cladistic approach, popularized by Laitin (2000) and Fearon (2003), is the most readily applicable in our setting and allows us to calculate linguistic distances for *any* pair of *Ethnologue* languages in a

[13] Somalia and Lesotho are generally thought to be ethnically homogeneous countries, unless finer subdivisions such as clans are taken into account. No data on ethnic composition are available for Rwanda after the 1994 events. Madagascar is diverse and, unlike most of continental Sub-Saharan Africa, is populated by groups speaking Austronesian languages. As for the five omitted North African countries, none of the available surveys have satisfactory data on their regional ethnolinguistic composition. The population of these countries largely represents a mix of Arab and Berber ethnicities, with a few small minorities, and Arabic is predominantly used in communication.

[14] Spolaore and Wacziarg (2016a) explore the measures of genetic, linguistic, religious, and cultural distances between countries. They find that all of these measures are positively correlated and argue that genetic (ancestral) distance represents a summary statistic for various cultural traits transmitted intergenerationally, including language.

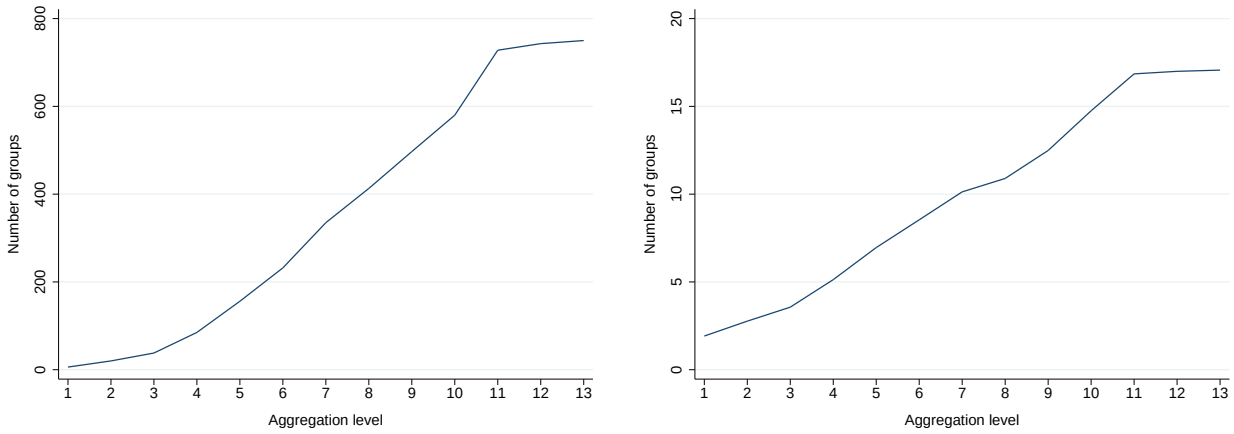
straightforward manner. Specifically, this is done by comparing the relative positions of each two ethnolinguistic groups on the linguistic family tree, a standard model of language origination adopted by *Ethnologue*. The formula proposed by Fearon (2003) assumes that $\tau_{ij} = 1 - (l/m)^{\delta}$, where l is the number of 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. Since it is not a priori clear how to choose δ, we try 19 different values in our calculations, from 0.01 to 0.1 with step 0.01, and from 0.1 to 1 with step 0.1.[15] The higher the value of δ, the more important linguistic distance is, so that in the limit, as $\delta \to \infty$, all different groups are treated as completely distinct and the adjusted measure converges to the basic unweighted ELF index. Overall, our approach yields 19 "distance-adjusted" indices, denoted as ELF_{δ}, where the subscript corresponds to the value of δ used to calculate linguistic distances.[16]

Desmet et al. (2012) suggested an alternative way to take the relatedness between languages into account when calculating diversity indices, also based on the linguistic family tree model. The idea is to aggregate ethnolinguistic groups to the level of different tiers of the linguistic tree, thereby exploring major cleavages originating at different points in history. Higher levels of aggregation correspond to deeper linguistic divisions, and the very first tier of the tree represent major language families. In the context of our data for Sub-Saharan Africa, 750 unique language codes belong to just six major families, according to the *Ethnologue* classification: Afro-Asiatic (121), Niger-Congo (557), Nilo-Saharan (61), Khoisan (4), Creole (4), and Indo-European (3).[17] Each family then branches out and reaches the depth of at most 12 subdivisions. Hence, we calculate thirteen indices and use the notation ELF(k) for an index calculated at the k-th aggregation level, where ELF(13)

[15]This range includes $\delta = 0.5$ used by Fearon (2003) and $\delta = 0.05$ preferred by Desmet et al. (2009) and Esteban et al. (2012).

[16]Appendix B considers an example from our dataset to clarify the procedure of constructing a distance-adjusted ELF index. An obvious drawback of the cladistic approach is that the language family tree model treats all nodes and branches in the same way without providing any details about the timing or importance of language splits. The lexicostatistical approach directly compares the vocabularies of language pairs and yields more continuous measures of linguistic distances. Reassuringly, for a limited sample of countries whose populations speak Indo-European languages, the two approaches produce highly correlated metrics (Spolaore and Wacziarg, 2016a).

[17]The only "local" Indo-European language in our dataset is Afrikaans spoken in South Africa, Namibia, Botswana, and Zimbabwe. The other two are English, with small native groups in South Africa, Botswana, and Namibia, and German, with a small group in Namibia only. Note that foreigners are excluded from our lists of ethnolinguistic groups.



(a) Total number of groups in the sample

(b) Average number of groups per region

Figure 1: Number of ethnolinguistic groups by aggregation level.

corresponds to the standard ELF index. In order to produce these metrics, we construct extended regional linguistic trees in which the paths from each language to the root are of the same length, as described in appendix B.

Panel (a) of Figure 1 shows how the number of distinct groups in our dataset changes depending on the level of aggregation, from just 6 fundamental language families at the first level to 750 languages in the fully disaggregated case.[18] Note that the number of groups increases dramatically up to level 11 and stays roughly the same at the next two levels, since very few languages actually reach the full depth of 13 branches from the root. Panel (b) shows the average number of unique ethnolinguistic groups per region depending on the level of aggregation. The pattern is qualitatively similar to the one shown in panel (a), with a steady increase in the number of groups from about 2 per region at level 1 to roughly 17 at levels 11–13.

In the context of Sub-Saharan Africa, the Niger-Congo language family, comprising 557 out of 750 groups in our dataset, occupies a special place. Its major branch are the Bantu languages spoken by ethnic groups populating a vast stretch of the African continent. Figure 2 illustrates the structure of the Niger-Congo family and the place of Bantoid languages in our dataset by showing the largest subdivisions of the family at each level of aggregation. For example, the largest subgroup of the Niger-Congo family is the

[18]Table A.1 in the appendix shows the number of unique ethnolinguistic groups for each country in our sample which varies from 3 in Djibouti and Swaziland to 192 in Nigeria, with an average of 25 groups per country. To put our total number of 750 groups in perspective, the widely used country-level datasets on ethnolinguistic diversity, offered by Alesina et al. (2003) and Fearon (2003), contain 1055 and 822 groups, respectively, for the *entire world*.

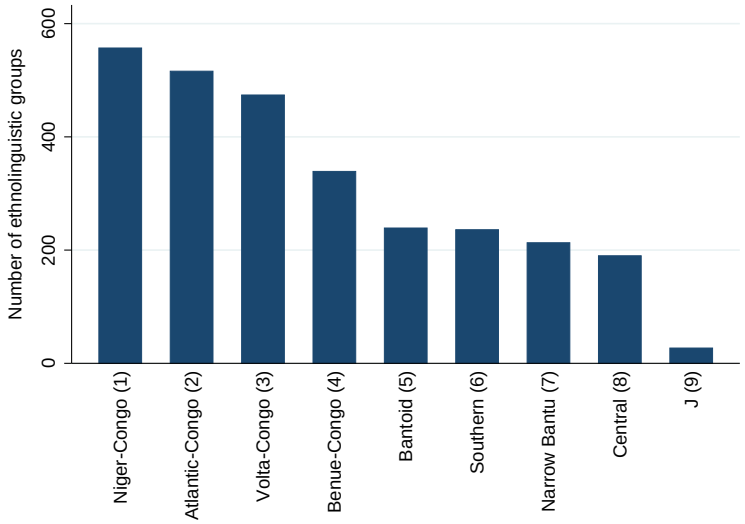


Figure 2: Number of groups in major subdivisions of the Niger-Congo family.

Atlantic-Congo subfamily comprising 516 languages. The Atlantic-Congo subfamily splits into further groups at the third level of the linguistic tree, among which the Volta-Congo subfamily is the largest (474 languages). The Bantoid languages at the fifth aggregation level contain 239 groups, whereas the narrowly defined Bantu family at the seventh level covers 213 groups. In other words, roughly between a quarter and a third of all languages in our dataset fall into the Bantu family depending on how narrowly it is defined. As is clear from Figure 2, a sharp drop occurs after the eighth aggregation level, when the Central subfamily of the Narrow Bantu (190 languages) splits into 14 distinct subgroups of which the largest one (J) contains just 27 languages. Thus, there is a drastic "unification" of languages between level 9 and levels 5–8. As a result, whenever diversity is measured at aggregation level 8 or higher, all the numerous Bantu languages merge together and are treated as a single group.

As argued in a series of both theoretical and empirical studies, an alternative measure of diversity may be better suited for capturing antagonism between groups, or their propensity to engage in conflict, namely the index of ethnolinguistic polarization.[19] Given the population structure set up at the beginning of this section, the basic index of ethnolinguistic polarization proposed by Reynal-Querol (2002) in her analysis of conflict is given by $\text{ELP} = 4\sum_{i=1}^{N} s_i^2(1 - s_i)$, where, as before, s_i is the regional population share of group i. This index captures how far the societal structure is from the perfectly polarized

[19]See Esteban and Ray (1994; 2011), Reynal-Querol (2002), Montalvo and Reynal-Querol (2005), and Esteban et al. (2012).

population consisting of two equal-sized groups. The idea, supported by formal theoretical models of conflict and going back to Horowitz (1985), is that the existence of a sizable ethnic minority alongside the dominant group substantially increases the likelihood of ethnic conflict.

Since the standard polarization index does not take into account the relatedness between groups, we also calculate adjusted ELP indices, as we did in the case of ELF. The distance-adjusted ELP index equals $4\sum_{i=1}^{N}\sum_{j=1}^{N} s_i s_j^2 \tau_{ij}$, where τ_{ij} is computed in the same way as above, and is in fact a version of the polarization measure developed by Esteban and Ray (1994). Altogether, for each region we calculate nineteen distance-adjusted ELP_δ indices and thirteen $\text{ELP}(k)$ indices for different levels of linguistic aggregation and use the same notation conventions as earlier.

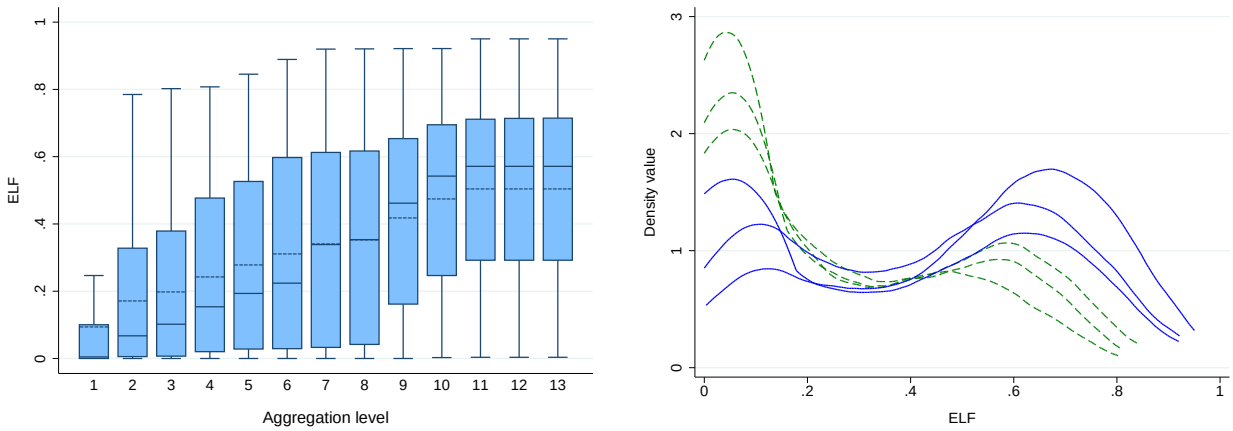
2.4 Exploring the new dataset

Figure 3 provides summary statistics for the $\text{ELF}(k)$ indices. The box-and-whiskers plots in panel (a) demonstrate that, as the level of linguistic aggregation increases (that is, as k decreases), the average and median values of the $\text{ELF}(k)$ index (represented by dashed and solid horizontal segments, respectively) go down, since the effective number of groups falls. For example, the mean is about 0.5 at level 13, 0.3 at level 5, and just 0.1 for $k = 1$. The range of values that $\text{ELF}(k)$ indices take is quite wide in the majority of cases and almost covers the whole $(0, 1)$ interval at levels 11–13.[20]

Beyond summary statistics, panel (b) of Figure 3 presents kernel density estimates of the distributions of $\text{ELF}(k)$ for selected aggregation levels. For the most aggregated cases ($k = 3$, 4, and 5, shown as dashed green curves), there is a clear concentration of values in the lower tail of the domain. However, as the level of aggregation decreases ($k = 7$, 9, and 13, shown as solid blue curves), a clear bimodal distribution emerges, with many regions falling both in the low- and high-diversity ends of the domain.

Figure 4 visualizes the regional distribution of $\text{ELF}(k)$ indices for four different levels of aggregation. The standard ELF index corresponds to ELF(13), which is displayed in panel (d). What is immediately clear from this map is that there is a lot of variation in diversity within countries. For instance, highly diverse Ethiopia masks the very unequal distribution of ethnolinguistic heterogeneity across its regions. Some areas, such as SNNPR (Southern Nations, Nationalities, and Peoples' Region) in the southwest and the chartered cities of Addis Ababa and Dire Dawa, are highly diverse with ELF(13) indices equal to 0.9, 0.7, and

[20] As mentioned earlier, very few languages actually reach the depth of 12 or 13 branches from the root, which leaves regional population composition stable for $k > 10$.



(a) Box-and-whiskers plots

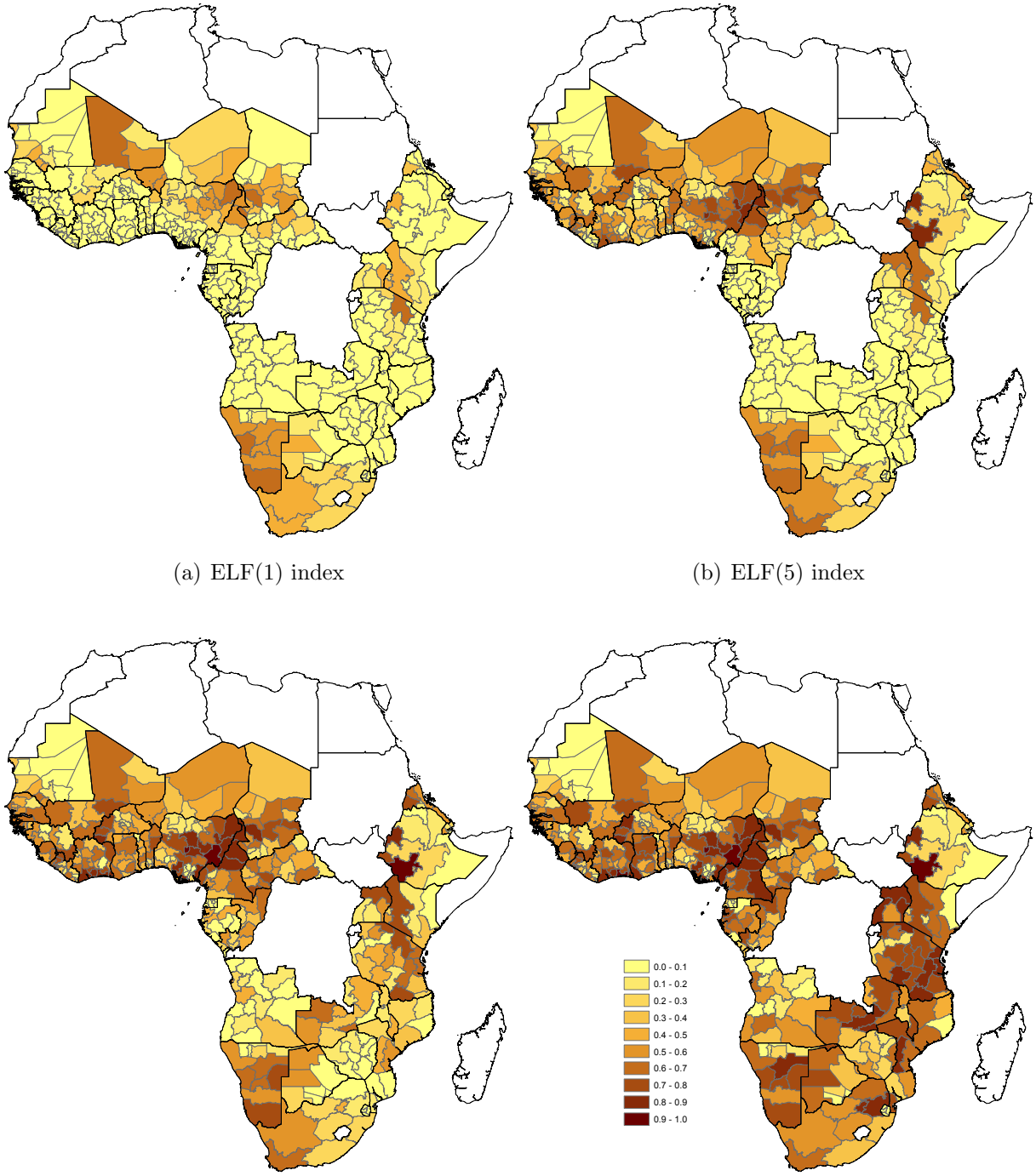
(b) Kernel density estimates

Figure 3: Descriptive statistics for ELF(k) indices.

Notes. The box-and-whiskers plots in panel (a) contain the following information: interquartile range (IQR), where the bottom (top) of the box corresponds to the lower (upper) quartile of the distribution, mean value (dashed segment), median value (solid segment), and the adjacent values representing the most extreme values within the range of 1.5×IQR from lower and upper quartiles. The kernel density plots in panel (b) correspond to $k = 3$, 4, 5, 7, 9, and 13, sorted from top to bottom by the density value at 0.

0.68, respectively, while other regions, such as Afar in the northeast and Somali in the east, are quite uniform, with ELF(13) indices at just 0.19 and 0.03, respectively. Thus, within a single country, ELF(13) varies from 0.03 to 0.9. In contrast, countries like Cameroon and Zimbabwe are more uniformly diverse across subnational regions.

Comparison of different panels in Figure 4 shows that varying the level of linguistic aggregation indeed matters. As k decreases, the map of subnational diversity becomes more pale. The highest contrast is naturally observed between ELF(13) and ELF(1) indices, the two extremes: while the map in panel (d) is quite dark on average, there are only a few darker regions in panel (a). The latter represent the areas in which ethnic groups speaking languages from some of the six fundamental families coexist. In fact, in panel (a), it is possible to discern the frontiers of these major language families in Sub-Saharan Africa passing through countries like Nigeria, Chad, Mali, Kenya, and Namibia. Furthermore, comparison of the top two panels to the bottom ones reveals that, at a sufficiently high level of linguistic aggregation, some countries, including Angola, Gabon, Mozambique, Republic of the Congo, Zambia, and Zimbabwe, become almost uniformly pale reflecting the common origin of the Bantu-speaking peoples populating these nations. Overall, high diversity dissipates as deeper ethnolinguistic cleavages are considered.



(a) ELF(1) index

(b) ELF(5) index

(c) ELF(9) index

(d) ELF(13) index

Figure 4: Regional distribution of ELF(k) indices.

Notes. Here and in the figures to follow below, the same color-coding scheme, as reflected by the legend seen in panel (d), applies to all maps. This enables direct comparison of regional diversity measures based on alternative indices.

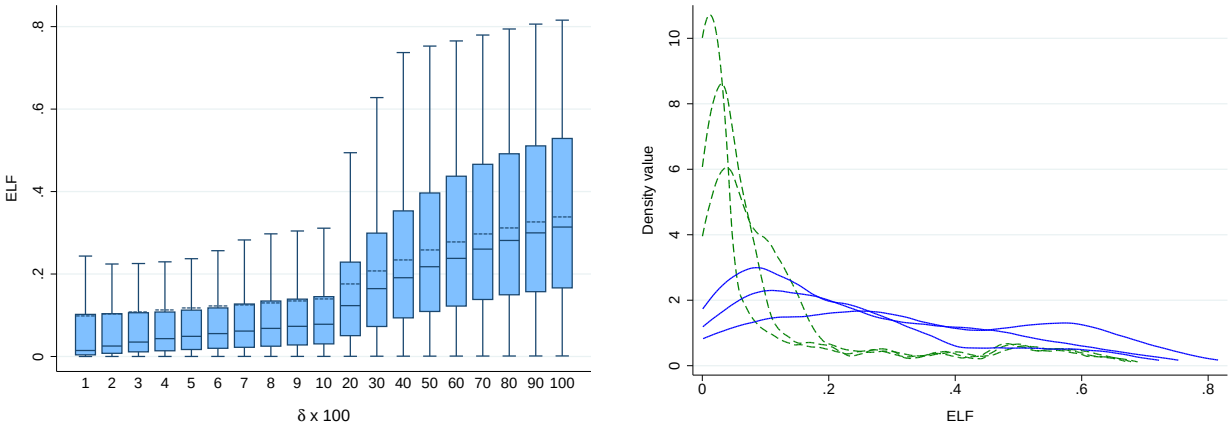
Figure 5 shows the descriptive statistics for ELF_δ indices. The evolution of the box-and-whiskers plots in panel (a) is rather monotonic in δ, with a visible jump at the point when the step increases from 0.01 to 0.1 for $\delta \in [0.1, 1]$. Clearly, median and mean fractionalization increases in δ, since higher value of that parameter implies heavier discounting of linguistic closeness and stronger emphasis on distinctiveness of groups. To take the two values of δ used in the literature, the mean of $\text{ELF}_{0.05}$ is 0.12, while the mean of $\text{ELF}_{0.5}$ is more than twice as large. Yet the correlation between the two indices (0.8) is quite high. Panel (b) of Figure 5 shows the kernel density plots for distance-adjusted ELF indices for selected values of δ. As one would expect, for low values of δ (dashed green curves), the distribution is collapsing toward the left end of the domain reflecting the increasing closeness between groups. In contrast, for values of delta above 0.1 (solid blue curves), the distribution looks much more uniform covering a broad range of fractionalization levels.

Figure 6 shows the spatial distribution of ELF_δ indices for $\delta = 0.05$ and $\delta = 0.5$. Relative to the striking contrast between panels in Figure 4, these two maps are rather similar. Yet, naturally, the map is more pale for $\delta = 0.05$, while for $\delta = 0.5$, diversity hotspots are more visible. Similar to the maps for $\text{ELF}(k)$ indices measured at high aggregation levels, dark spots in panel (a) capture those regions where diversity is more deeply rooted, that is, where local ethnic groups belong to completely distinct language families. In general, the relationship between ELF_δ and $\text{ELF}(k)$ indices is not straightforward, as illustrated in Figure D.1 in the appendix.

Analogous descriptive figures and maps for polarization indices are provided in appendix C. Much of the same intuition having to do with the effects of aggregation and accounting for linguistic relatedness applies to ELP indices. However, the concepts of fractionalization and polarization capture different dimensions of diversity and clearly, the regional distributions of corresponding indices do not look the same. Overall, the relationship between $\text{ELF}(k)$ and $\text{ELP}(k)$ indices measured at the same aggregation level represents a well-known inverted-U pattern, as can be seen in the top row of Figure D.2.[21] The association between ELF_δ and ELP_δ indices for the same values of δ, illustrated in the bottom row of Figure D.2, is not as clear-cut, and the inverted U is much less pronounced.

Overall, we argue that our new dataset is the best attempt so far to systematically describe subnational ethnolinguistic diversity in Sub-Saharan Africa. The following section briefly compares our contribution to alternative approaches.

[21] Since the range of values for ELF indices contracts at higher aggregation levels, the downward section of the inverted U disappears for smaller k and the correlation between $\text{ELF}(k)$ and $\text{ELP}(k)$ increases dramatically. Table D.1 in the appendix shows pairwise correlations for selected ELF and ELP indices.

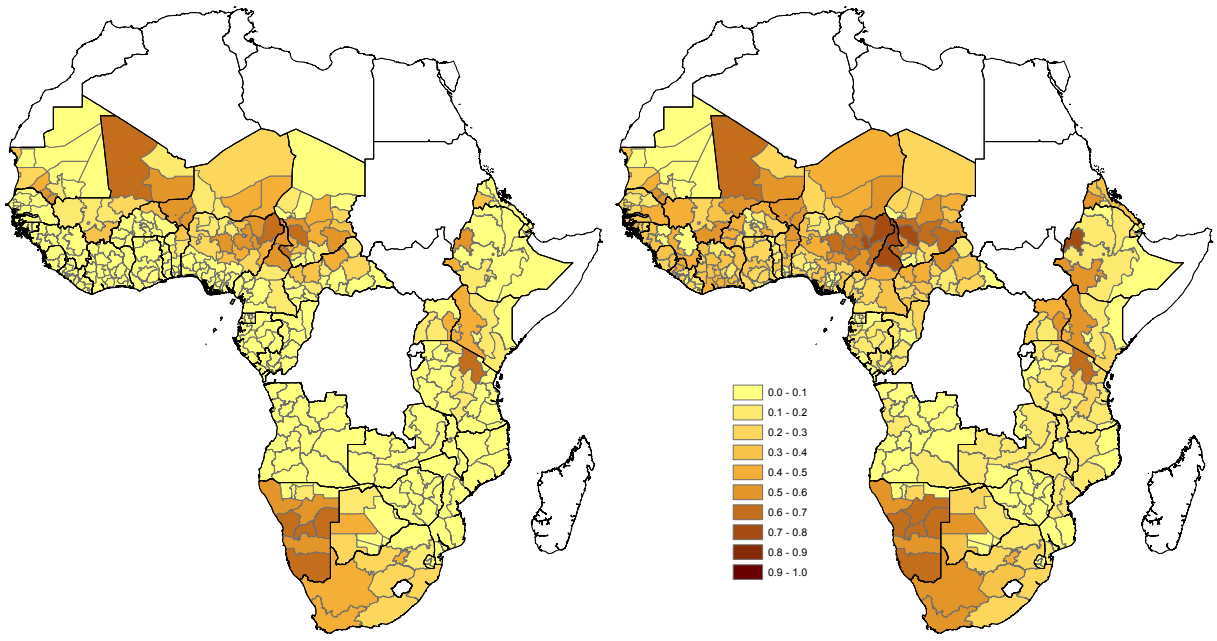


(a) Box-and-whiskers plots

(b) Kernel density estimates

Figure 5: Descriptive statistics for ELF_δ indices.

Notes. The box-and-whiskers plots in panel (a) are constructed in the same way as the ones in Figure 3. The kernel density plots in panel (b) correspond to $\delta = 0.01$, 0.05, 0.1, 0.3, 0.5, and 1 (sorted from top to bottom by the density value at 0).



(a) $\text{ELF}_{0.05}$ index

(b) $\text{ELF}_{0.5}$ index

Figure 6: Regional distribution of ELF_δ indices.

2.5 Comparison to other approaches

An important earlier effort to build a large-scale subnational-level dataset on ethnolinguistic composition is the study by Alesina and Zhuravskaya (2011) henceforth referred to as AZ.[22] Although their dataset has a worldwide coverage, it only includes 23 out of 36 countries from our sample, some of them with non-standard or outdated subnational divisions, and mostly relies on the DHS surveys for Sub-Saharan Africa. A fair share of those sources are substantially less detailed relative to those that we used. For example, according to the AZ dataset, there are only six ethnic groups in Côte d'Ivoire: Akan, Kru, Northern Mande, Southern Mande, Voltaic, and "foreign workers." Setting aside the issue of counting the latter as a valid ethnic group, the first five items on the list represent ethnolinguistic *families*. By comparison, our source provides information on 50 well-defined groups for this country.[23]

In several cases, the original lists of ethnic groups were shortened by AZ leading to severe distortions in the implied population composition of subnational units. For instance, while the 2005 DHS survey for Ethiopia contains over 60 ethnicities, only 9 of them remain in the AZ dataset.[24] Among the curtailed groups are the Afar people of Ethiopia, who, according to the 2007 census, constituted 90% of population in the namesake Afar region. Similarly, over 3/4 of the population in the Gambela region are coded as "other" by AZ just because the Nuer and the Anuak peoples representing the majority of this region's population were dropped from the list of ethnicities.

A recent paper by Gerring et al. (2015) is another attempt to construct a subnational-level dataset on ethnic diversity for a broad sample of countries. It also covers 23 out of 36 countries in our sample and fully relies on the DHS data (female sample only) to back out regional ethnic composition. This dataset appears to have the same drawbacks as the AZ case. Specifically, the number of ethnic groups per country is systematically lower not just relative to our new dataset, but even compared to the original DHS surveys.[25] Furthermore, the authors did not distinguish between groups that represent well-defined

[22]Note that the authors do not perform any type of subnational analysis, but rather use their dataset to construct *national-level* measures of segregation. Nevertheless, the raw data from AZ are readily available and can be used to construct subnational-level diversity indices. Overall, the correlation between the ELF index based on AZ data and our ELF(13) index in the common sample of 140 regions is around 0.75.

[23]Similarly, AZ identify 10 groups in Burkina Faso, 14 in Kenya, 7 in Senegal (counting "non-Senegalese"), and 3 in Zimbabwe. The respective numbers in our dataset are 27, 29, 18, and 19.

[24]In the case of Tanzania, the original list of about 100 groups in the 1992 DHS survey was cut to 29.

[25]For example, the authors identify 6 groups in Côte d'Ivoire, 14 in Ethiopia, 7 in Ghana, 11 in Nigeria, and 20 in Uganda. The respective numbers in our dataset are 50, 64, 38, 192, and 39.

ethnicities, language families, or geographically defined categories. For instance, in the case of Chad, the paper claims to have identified 14 groups. Upon closer examination, it turns out that two of these groups are "other Chadian ethnos" and "alien," while several other categories represent geographic divisions of Chad such as Tandjile, Lac Iro, and Mayo-Kebbi. Similarly, the DHS data on "ethnic groups" for the DRC used by the authors represent geographic regions, as discussed in section 2.1 above.

Our methodology substantially improves upon these previous attempts to measure subnational diversity in Sub-Saharan Africa based on population surveys. We cover a larger sample of countries and use sources containing much more detailed lists of ethnolinguistic groups. Our standardization process via *Ethnologue* matching guarantees that all included ethnolinguistic groups represent well-defined comparable entities rather than language clusters or geographic areas. Finally, we construct subnational diversity indices that take into account linguistic relatedness between groups, a crucial refinement which, as shown in section 3 below, is key to understanding the relationship between subnational diversity and a range of development indicators.

An alternative approach to constructing local diversity metrics used in recent empirical studies relies on geographic information systems (GIS).[26] Indeed, it is straightforward to combine one of the available digital maps of ethnolinguistic groups with high-resolution population data and a set of regional boundaries to back out the implied ethnolinguistic composition of each region and compute the diversity measures of interest. In Gershman and Rivera (2018), we follow this approach to calculate a series of ELF(k) indices and compare them to our survey-based analogues. We find that the GIS-based indices derived from the World Language Mapping System (a digital map of *Ethnologue* languages) perform the best in matching survey-based benchmarks. The correlation between ELF(13) indices stands at 0.55 and exceeds 0.75 for $k < 8$. The correspondence improves in the sample excluding urbanized regions, which is expected, since the available ethnolinguistic maps are unable to capture high diversity of cities and urban areas. Other sources of bias inherent in the GIS approach include out-of-date and likely inaccurate "traditional" boundaries of groups and noisy high-resolution population data.[27] Despite its deficiencies, the GIS

[26]See, for example, Kuhn and Weidmann (2015), Alesina et al. (2016), and Desmet et al. (2016). The latter paper is especially notable for using an elaborate algorithm to compute local population shares of various linguistic groups.

[27]In Gershman and Rivera (2018), we show that the use of GIS-based diversity indices leads to attenuation bias in an exercise from section 3.2. Furthermore, in "horse-race" regressions, the relevant GIS-based indices reduce in absolute value and lose statistical significance, while their survey-based counterparts remain intact, indicating that the latter represent less noisy measures of regional diversity.

approach can be quite useful, particularly in cases when no reliable survey data are available or when non-standard regions are used as basic units of analysis.

As an additional point of reference, we have used our sources of data on ethnolinguistic composition to construct *country-level* ELF(k) indices and compared them to those from commonly used datasets. The correlations between the resulting ELF(13) index and fractionalization indices from Alesina et al. (2003) and Desmet et al. (2012) are 0.93 and 0.9, respectively, in the common sample of 36 African countries. Our ELF(k) indices are tightly related to their counterparts in Desmet et al. (2012), with correlation coefficients around 0.9 for most values of k and reaching 0.97 for $k = 8$. Thus, at the country level, our data sources yield diversity indices broadly similar to measures available from earlier studies.

2.6 Religious diversity

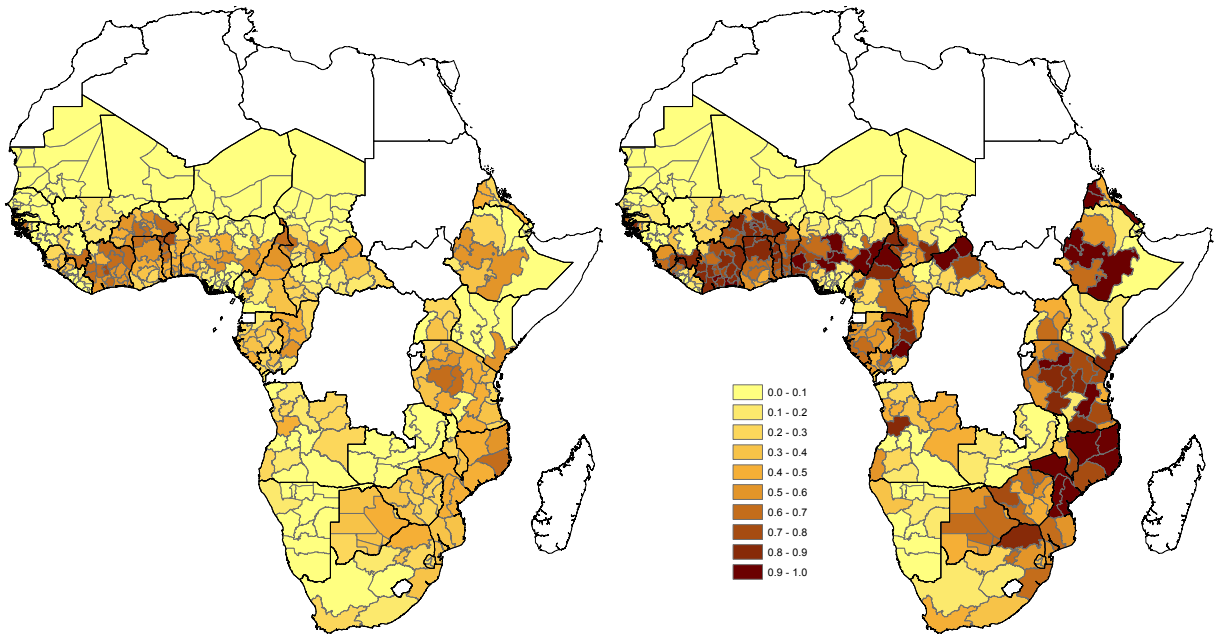
Although the main focus of this paper is on ethnolinguistic diversity, naturally, ethnicity and language are not the only societal cleavages potentially important for aggregate social and economic outcomes. To complement our main analysis, we followed the basic principles from section 2.1 to construct a new subnational-level dataset on religious diversity for our baseline sample of countries.

In this exercise, we adopt a simple four-way classification for religion: Christianity, Islam, "traditional" religion, and none.[28] This approach allows a uniform treatment of all countries in our sample, since some of our sources provide only such coarse classification, but has a clear downside of underestimating religious diversity when important divisions exist within major religions in the form of different denominations. We were able to find high-quality data on subnational religious composition for almost all countries in our baseline sample and produced simple religious fractionalization (RF) and religious polarization (RP) indices.[29]

Just like its counterparts for ethnolinguistic diversity, Figure 7 makes it clear that religious diversity is not evenly distributed withing countries. In the northern part of Sub-Saharan Africa, countries like Chad, Mali, Niger, and Senegal are almost uniformly

[28] Apart from the respondents explicitly classified as following animist or traditional religion, this category also includes such local religions as Vodoun in Benin and Badimo in Botswana. Our suspicion is that the "none" group, which is surprisingly sizable in many countries, sometimes may also contain the practitioners of traditional religions.

[29] Table A.2 lists the data sources for each country. In the cases of Djibouti and Mauritania, where the data are technically unavailable, but Islam is the official state religion by law, we assume that the entire population is Muslim. Excluding these two countries from the analysis does not qualitatively change the results. No data could be found for Equatorial Guinea.



(a) RF index (b) RP index

Figure 7: Regional distribution of religious diversity indices.

Muslim. Just to the south of the Sahara desert, there is a clear belt where Islam coexists with Christianity, and many regions in countries such as Burkina Faso, Ethiopia, Ghana, and Nigeria are highly polarized. Substantial presence of traditional religion in countries like Benin, Burkina Faso, and Guinea-Bissau adds to religious diversity in this region. In east and southeast Africa, Tanzania and Mozambique are examples of countries containing multiple regions with a substantial Christian/Muslim split. Most countries in central and south Africa are predominantly Christian, and diversity hotspots in those areas mostly reflect a substantial presence of non-religious respondents.[30]

Visual comparison of maps for ethnolinguistic and religious diversity reveals no clear pattern in their relationship. Indeed, for instance, the correlation between RF and ELF(k) indices ranges from -0.17 for ELF(1) to 0.26 for ELF(13). In section 3.6, we compare the relative roles of ethnolinguistic and religious divisions in explaining the variation in subnational economic performance.

Although fractionalization and polarization indices have dominated the empirical literature on ethnolinguistic and religious diversity, numerous other measures have been offered.

[30]To see whether the inclusion of the "none" group makes a big difference, we recalculated our indices by dropping those respondents. The correlation coefficients between pairs of RF and RP indices based on alternative classifications are equal to 0.89 and 0.85, respectively.

In Gershman and Rivera (2017), we construct subnational indices capturing the overlap between ethnic and religious cleavages and examine their relationship to conflict. In the same vein, Desmet et al. (2017) construct country-level measures of the overlap between ethnicity and culture. Another line of work instead focuses on ethnic inequality and its implications for public goods provision, development, and conflict (Baldwin and Huber, 2010; Kuhn and Weidmann, 2015; Alesina et al., 2016). We leave the analysis of these and other dimensions of subnational ethnic and religious diversity for future research.

3 Subnational diversity and development

3.1 Data and empirical strategy

In this section, we use our new dataset to revisit the connection between diversity and regional development indicators, with a focus on those reflective of local public goods provision. Specifically, we compile a dataset on regional educational outcomes, health indicators, and access to electricity, all of which are consistently measured at the subnational level using large-scale household surveys.[31] We also examine broader measures of economic development, namely nighttime luminosity, income per capita, and household wealth.

In order to account for possible confounding factors, we supplement our dataset with an array of relevant regional characteristics which primarily includes exogenous geographic factors linked to ethnolinguistic diversity and/or economic development in earlier studies. Throughout our analysis, we control for the absolute latitude of the region's centroid, surface area, ocean access (landlocked indicator), terrain ruggedness, capital city dummy, distance from capital city, as well as mean and standard deviation of land suitability for agriculture.

Mitton (2016) shows that ruggedness, ocean access, and absolute latitude all have significant explanatory power for within-country differences in per-capita income. Furthermore, in the African context, ruggedness and ocean access are important factors having a lasting influence on development through their connection to the intensity of historical slave trade (Nunn and Wantchekon, 2011; Nunn and Puga, 2012). Region's area and its remoteness from capital city are likely associated with both socioeconomic outcomes and ethnolinguistic diversity. In Sub-Saharan Africa, distance to capital city also captures the strength and penetration of formal national-level institutions in the regions (Michalopoulos and Pa-

[31] In most cases, these were the waves of DHS and MICS conducted around 2010, see Table A.2 in the appendix for a complete list of sources used for each country.

paioannou, 2014). Land suitability for agriculture is a fundamental determinant of early development, whereas its standard deviation is included to account for the relationship between variability in geographical endowments and ethnic diversity (Michalopoulos, 2012). Finally, since the provision of public goods varies enormously between rural and urban areas, and the latter are also likely to be more diverse, we additionally control for regional urbanization rates in all of our regressions.[32]

Our baseline specification is a simple linear model at the regional level:

$$y_i = \alpha_c + \beta D_i + X_i'\Gamma + \varepsilon_i,$$

where y_i is one of the development outcomes, D_i is one of the diversity indices, X_i' is a vector of control variables, α_c is the full set of country fixed effects, and ε_i is an idiosyncratic component. Note that, unlike cross-country studies, our setup accounts for nation-specific characteristics and thus exploits within-country variation in diversity to estimate the relationships of interest.[33]

3.2 Education

We start with our findings on the relationship between educational outcomes and diversity. The first two indicators are literacy rate and the share of population with secondary education or above.[34] The second set of outcomes contains net school attendance ratios measuring participation in primary or secondary schooling among children of appropriate age.[35] Since the results are very similar across these four indicators, we only report those for literacy rate and net secondary school attendance.

We run 64 separate regressions for each educational outcome: 26 for ELF(k) and ELP(k) indices (13 each) and 38 for ELF$_\delta$ and ELP$_\delta$ indices (19 each). All regressions include the full set of controls described above and country fixed effects. Given the abundance of

[32]Using regional population density instead of urbanization rate does not qualitatively affect our findings. In section 5, we show the robustness of our results to the exclusion of urban regions.

[33]Motivated by the findings of Ashraf and Galor (2013) we also experimented with a quadratic specification in diversity, but found no systematic evidence of significant nonlinearities.

[34]A person is considered literate if she can read at least part of a standard sentence or has attended secondary school. Both measures are constructed for population between 15 and 49 years old. Due to the design of DHS surveys, in some cases the data were only available for females. Since the results are very similar for gender-specific and combined samples, we use these two indicators based on the female samples to maximize regional coverage.

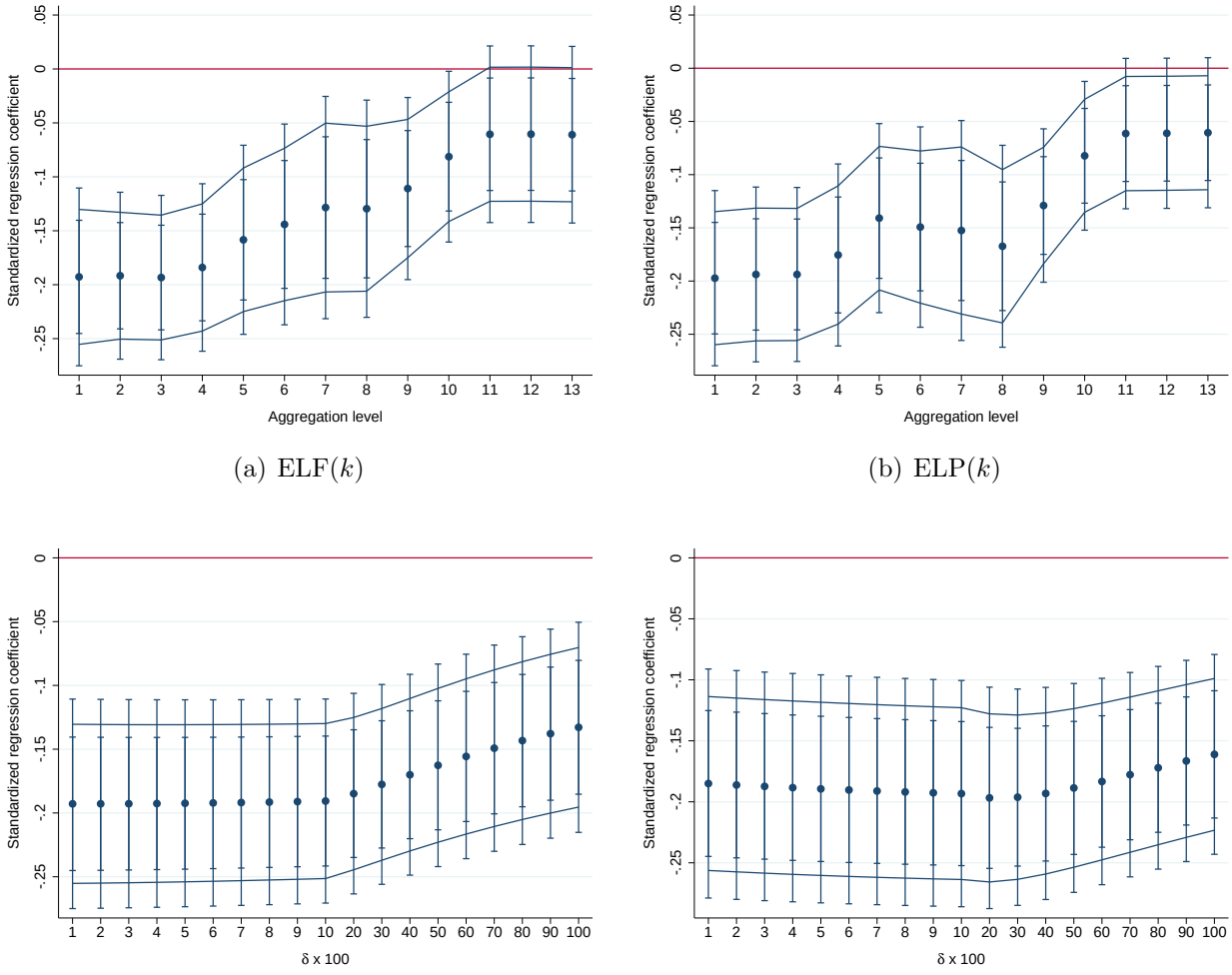
[35]In contrast, gross attendance ratio measures participation in schooling among individuals of any age between 5 and 24. The results for net and gross attendance ratios are virtually identical.

diversity indices, we present most of our estimates in an intuitive graphical form instead of using regression tables. First, for each specification, we calculate the standardized estimate of β, along with the corresponding 90%, 95%, and 99% confidence intervals based on robust standard errors. We next display these estimates as functions of either the level of linguistic aggregation, $k = 1, \ldots, 13$, or the parameter $\delta \in [0.01, 1]$, and we connect the 95% confidence intervals with linear segments for visual clarity.[36] Thus, these diagrams summarize the results of thirteen and nineteen regressions, respectively. Standardization of coefficient estimates makes it easy to compare the relative magnitude, or "economic significance," of various diversity measures. Specifically, each point estimate reflects the average change (in standard deviations) in the left-hand-side outcome variable associated with an increase in the corresponding diversity index by one standard deviation, other things equal.[37]

Figure 8 summarizes the relationship between regional literacy rate and ethnolinguistic diversity in our sample and is reflective of the patterns observed for other educational outcomes. Panel (a) shows that the estimated coefficients for ELF(k) indices are negative and strongly statistically significant all the way up to aggregation levels 11–13. Furthermore, the magnitude of the estimates tends to decrease as the classification of ethnolinguistic groups gets finer. It is highest for the crudest aggregation levels, from 1 to 3, with the standardized coefficient estimates around -0.19, and gradually decreases to -0.11 for $k = 9$, -0.08 for $k = 10$, and -0.06 at levels 11–13. Panel (b) shows that a similar pattern holds for the ELP(k) indices, all of which are strongly negatively related to literacy rate, especially for aggregation levels 1–10, with coefficient estimates roughly similar to those in the ELF case. Overall, the relationship between diversity and literacy is *weakest* when ethnic groups are not aggregated at any considerable level, a common choice in the literature. In contrast, deeper linguistic cleavages seem to represent more important divisions.

[36] As shown in appendix E, standard errors clustered at the country level are typically larger than the robust ones, but this adjustment does not qualitatively alter the conclusions reported below. As an alternative, we also constructed the standard errors adjusted for spatial correlation using the method of Conley (1999) for two different "distance windows," 200 and 400 kilometers (these cutoffs roughly correspond to an average of 5 and 10 neighbors, respectively, for each region, as defined by centroid coordinates). In specifications with literacy rate as the outcome variable and various ELF indices as diversity measures, these turn out to be substantially smaller than the clustered standard errors and somewhat larger than the robust ones.

[37] Annobon, a tiny island in the Gulf of Guinea and one of the provinces of Equatorial Guinea, drops out from our sample, since it is not covered by the raster files used to produce some of the baseline geographic controls. Other than that, the sample size (indicated in the caption of each diagram) depends solely on the availability of data on outcome variables.



(a) ELF(k)

(b) ELP(k)

(c) ELF$_\delta$

(d) ELP$_\delta$

Figure 8: Ethnolinguistic diversity and literacy rate ($n = 397$).

Notes. Panels (a) and (b) graphically represent the outcomes of thirteen regressions each. In panel (a), regional literacy rate is regressed on ELF(k) indices, one at a time, the full set of controls, and country fixed effects. For each aggregation level $k = 1, \ldots, 13$, the standardized point estimate of the coefficient on ELF(k) index is displayed, along with respective 90%, 95% (connected by linear segments), and 99% confidence intervals based on robust standard errors. Panel (b) is constructed in the same way for ELP(k) indices. Panels (c) and (d) graphically represent the outcomes of nineteen regressions each. The estimating equations have the same specification as for panels (a) and (b), but with ELF$_\delta$ and ELP$_\delta$ indices on the right-hand side. Standardized coefficient estimates and their respective confidence bands are shown for different values of $\delta \times 100$, the key parameter used in adjustment for linguistic similarity between pairs of languages. Sample size n is indicated in the caption. The figures for all other outcomes in this section are constructed in the same way.

Panels (c) and (d) of Figure 8 show the results for ELF_δ and ELP_δ, the indices of diversity adjusted for linguistic distances. In both cases, diversity is highly statistically significant and negatively associated with regional literacy rate, regardless of the specific value of δ. In the case of ELF, the standardized coefficient estimate is monotonically decreasing in δ in absolute value, from around -0.19 for $\delta = 0.01$ to roughly -0.13 for $\delta = 1$, with an obvious kink at $\delta = 0.1$, where the step increases from 0.01 to 0.1. For ELP, there is in fact a barely visible non-monotonicity in the relationship between the standardized coefficient estimate and δ: its magnitude is slightly increasing for $\delta < 0.3$ and decreasing for larger values of δ. Overall, the results for both ELF_δ and ELP_δ are stronger for smaller values of δ, that is, if linguistic similarity is emphasized or, in other words, the distinction between languages is discounted more heavily.[38]

Figure 9 shows the scatterplots of residuals corresponding to selected regressions from Figure 8. The first four plots illustrate the case of $\text{ELF}(k)$ for $k = 2, 5, 9, 13$. It is clear from this sequence of diagrams that the negative relationship between regional literacy rate and diversity dissipates as the level of linguistic aggregation decreases. The last two plots show the estimated pattern for ELF_δ indices when $\delta = 0.05$ and $\delta = 0.5$. As discussed earlier, in both cases the coefficient of interest is statistically significant and negative, but the slope is somewhat steeper for lower δ. It is also clear from the scatterplots in Figure 9 that the estimated relationships are not driven by any outliers.

The results for net secondary school attendance ratio reported in Figure 10 are quite similar to those for the literacy rate. Again, the association with diversity is sharpest when ELF and ELP indices take into account linguistic relatedness, especially for small values of k and δ. The standardized coefficient estimates on diversity in the strongest specifications are roughly in the range between -0.18 and -0.12.

3.3 Health

In addition to educational outcomes, we have constructed two regional health indicators. The first one is the share of births happening at home rather than at specialized facilities like clinics and hospitals. This measure largely reflects the accessibility of such health facilities and is an important measure of household well-being since home births in developing countries pose substantial risks to both the mother and the newborn (Montagu et al., 2011). Our second health indicator is the share of moderately underweight children under the age of five, a metric of child malnutrition. A child falls into this category if his or

[38] Sharper results for lower values of δ are consistent with the choice of $\delta = 0.05$ by Desmet et al. (2009) and Esteban et al. (2012) in their country-level analyses over $\delta = 0.5$ originally used by Fearon (2003).

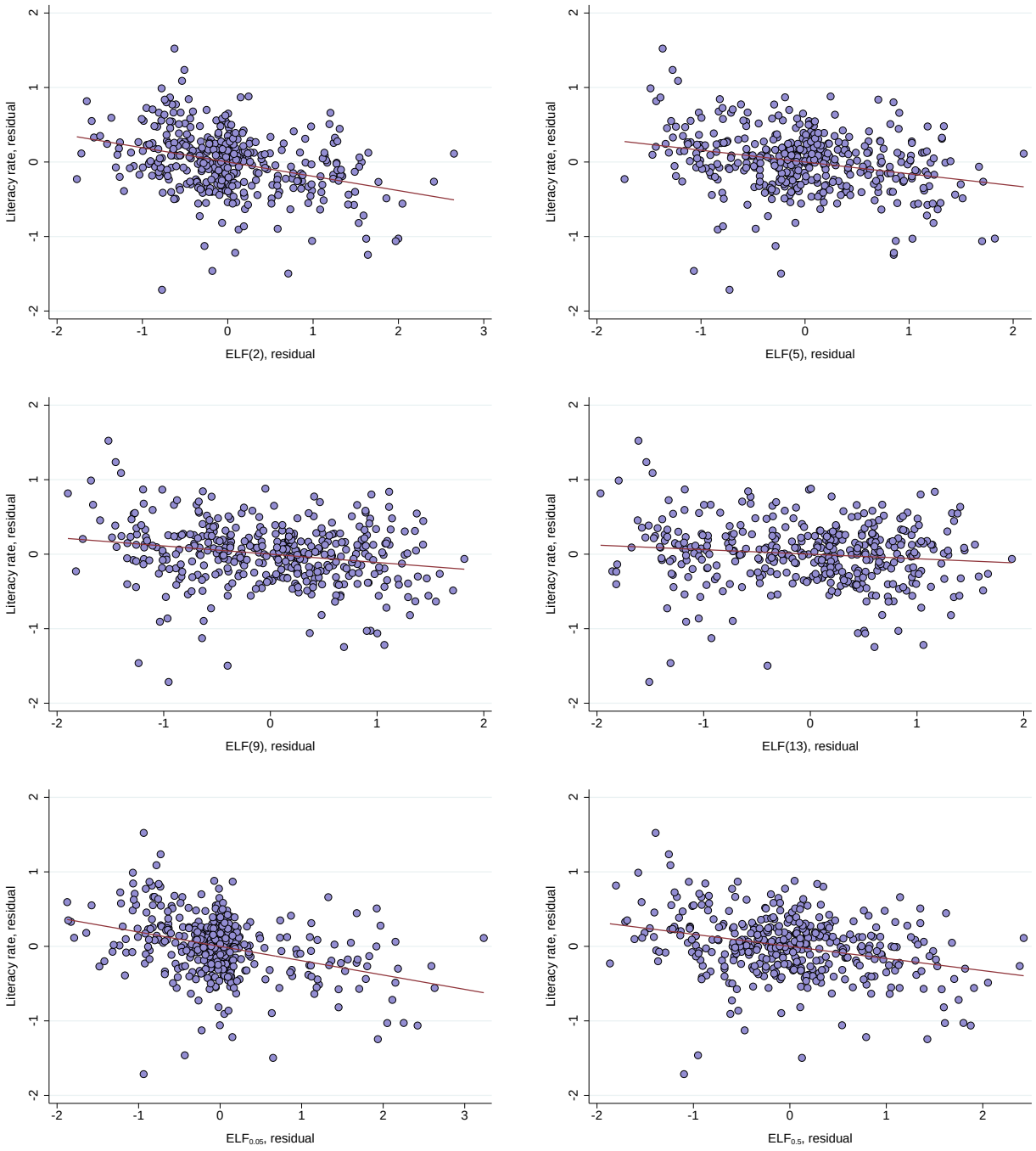
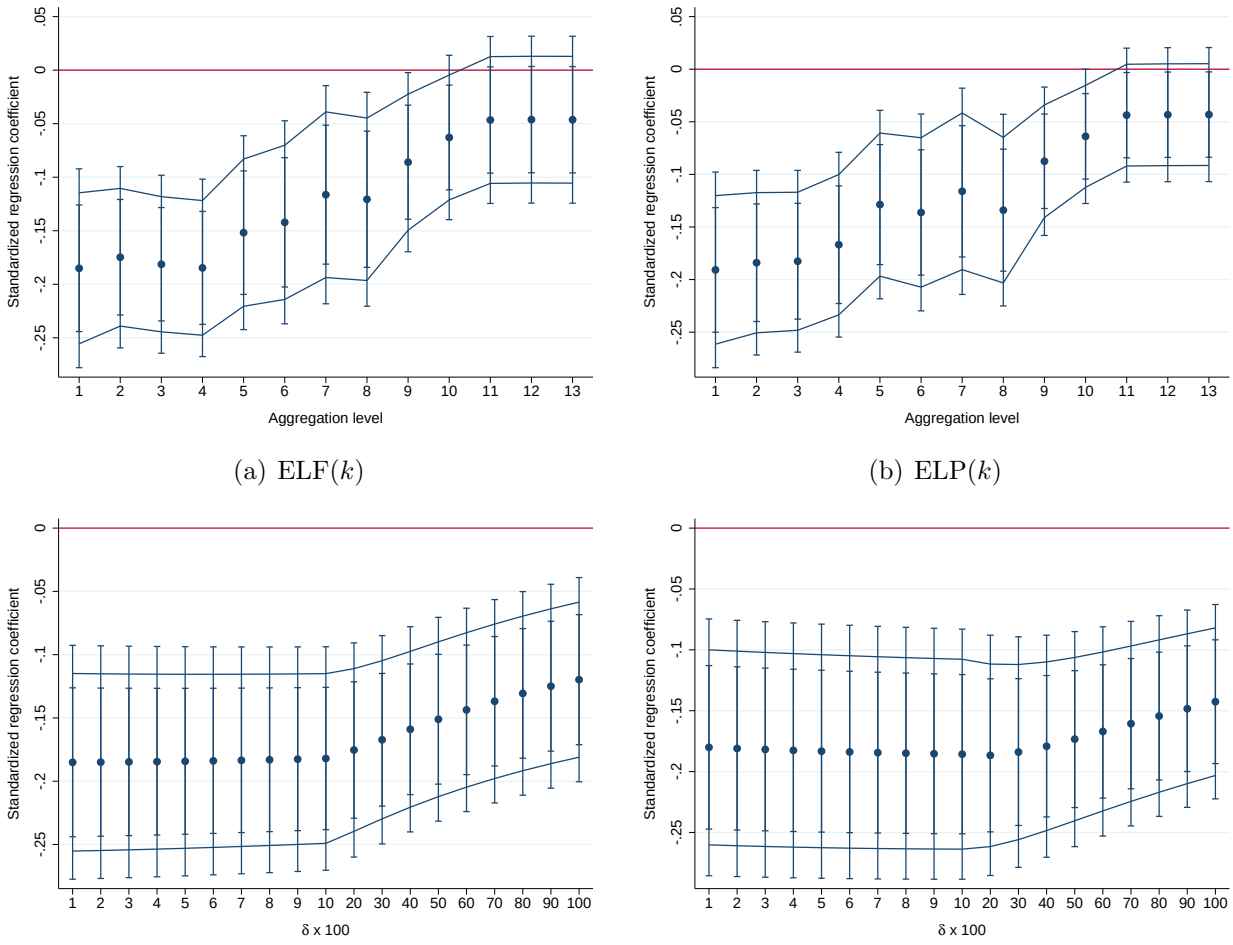


Figure 9: Ethnolinguistic diversity and literacy rate: scatterplots of residuals.

Notes. Each diagram represents a scatterplot of residuals from regressing regional literacy rate and respective diversity measures on the baseline set of control variables specified in the text and the full set of country fixed effects. Both variables are standardized to be consistent with Figure 8 specifications.



(a) ELF(k)

(b) ELP(k)

(c) ELF$_\delta$

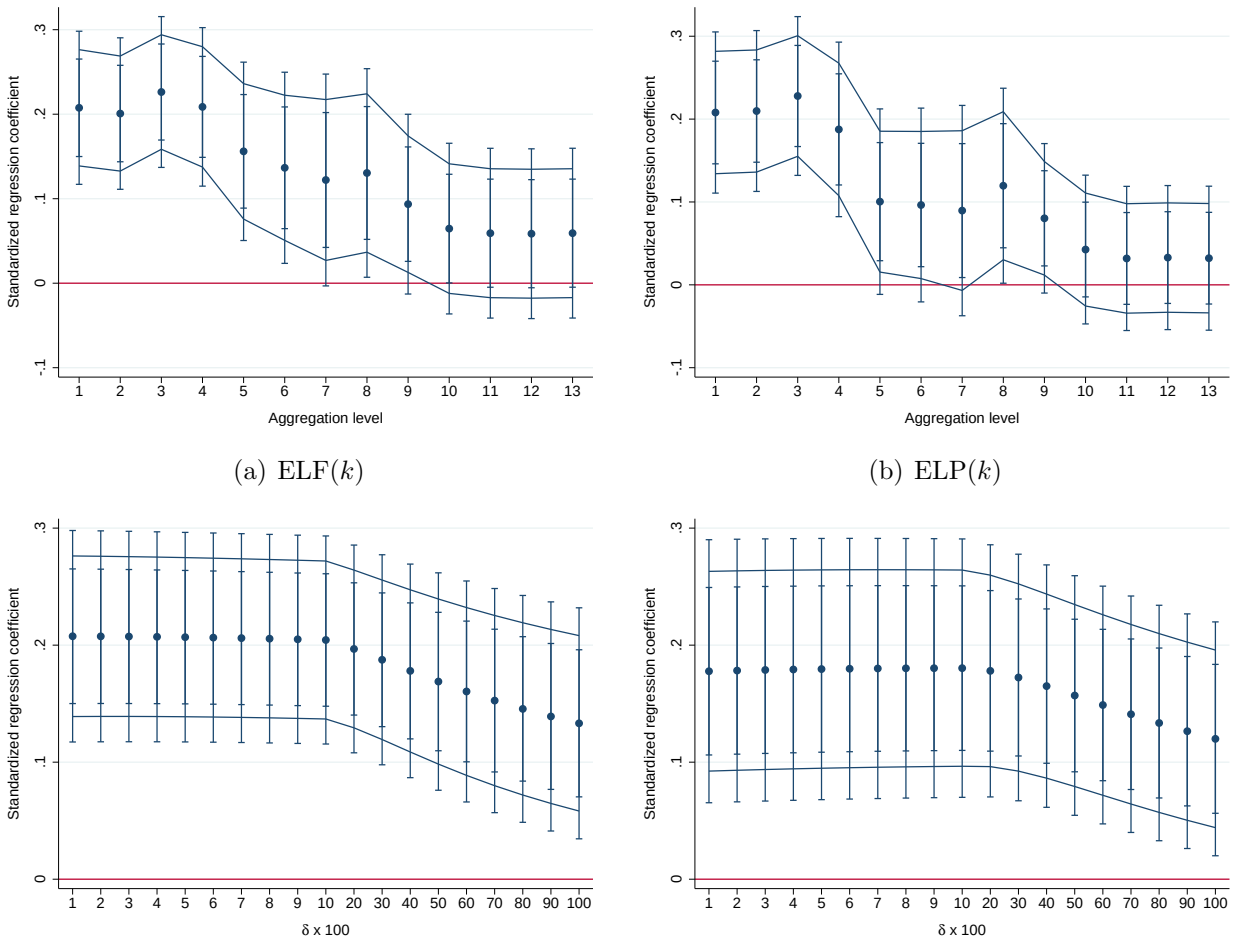
(d) ELP$_\delta$

Figure 10: Ethnolinguistic diversity and net secondary school attendance ratio ($n = 388$).

her weight is two standard deviations below the median value in the reference population, which is reflective of both acute and chronic malnutrition.[39]

We use the same graphical approach as above to display our estimation results. Figure 11 summarizes the outcomes of 64 regressions uncovering the relationship between regional ethnolinguistic diversity and the share of home births. Overall, the qualitative patterns are very similar to the earlier findings on education. Other things equal, higher diversity

[39]In contrast, stunting captures the long-term effect of inadequate nutrition and chronic illness, but is not sensitive to short-term changes in dietary intake. Wasting represents malnutrition in the period immediately preceding the survey and may also result from a recent episode of illness causing weight loss. The results for regional prevalence of stunting and wasting are qualitatively similar to those reported below and are omitted.



(a) ELF(k) (b) ELP(k)

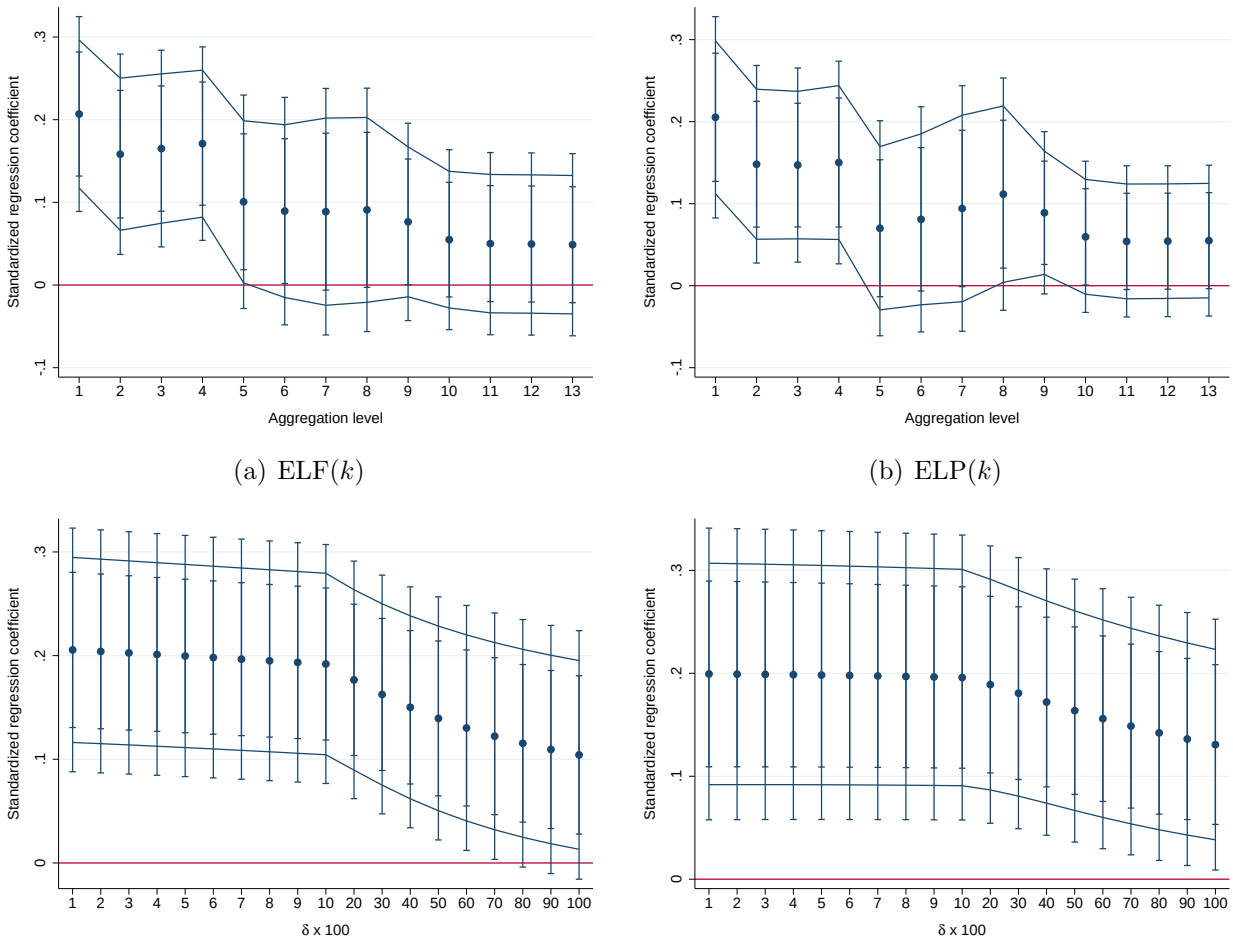
(c) ELF$_\delta$ (d) ELP$_\delta$

Figure 11: Ethnolinguistic diversity and home births ($n = 382$).

is associated with a larger share of home births and this relationship is strongest, both in terms of magnitude and statistical significance, at cruder levels of linguistic aggregation.[40] In the case of ELF(k) indices, as shown in panel (a), the standardized coefficient estimates hover around 0.2 for levels 1–4, fall somewhere between 0.11 and 0.15 for levels 5–8, drop below 0.1 at $k = 9$, and lose statistical significance thereafter. Distance-adjusted diversity indices are all highly statistically significant and the magnitude of respective estimates is larger for small values of δ, as can be seen in panels (c) and (d).

Figure 12 shows the estimation results for child malnutrition. As seen in panels (a) and (b), only the indices calculated at top four aggregation levels are strongly positively related

[40]The coefficient estimates on diversity are positive in these health regressions, since, by construction, lower values of the respective outcome variables are reflective of better access to health facilities and nutrition. Thus, Figures 11 and 12 look like flipped versions of the diagrams for educational outcomes.



(a) ELF(k)

(b) ELP(k)

(c) ELF$_\delta$

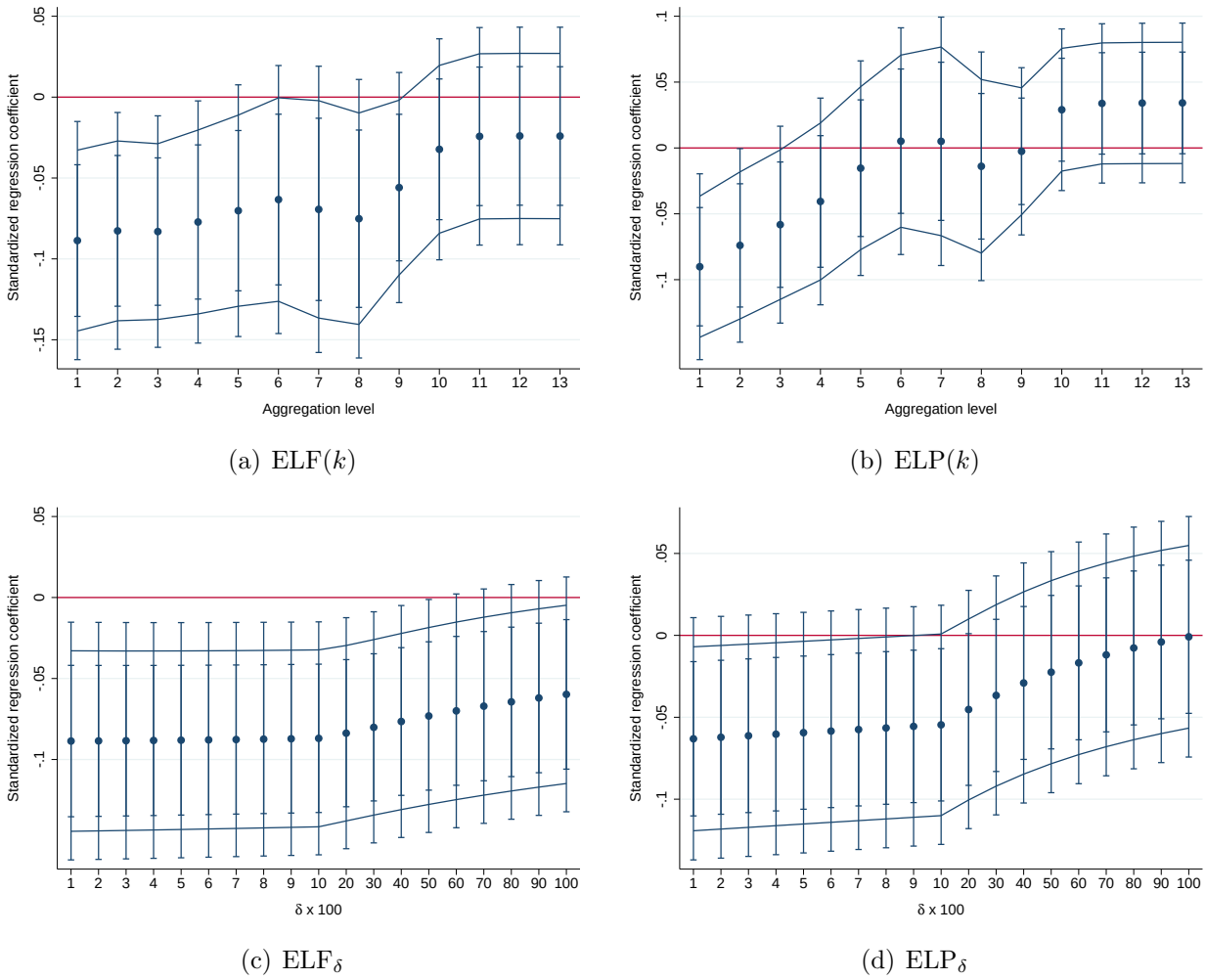
(d) ELP$_\delta$

Figure 12: Ethnolinguistic diversity and child malnutrition ($n = 397$).

to the regional share of underweight children. As for the distance-adjusted indices, as before, the estimates are stronger for smaller values of δ. In all regressions with statistically significant results, the standardized coefficients of interest are roughly in the range between 0.1 and 0.2, depending on the type of diversity index.

3.4 Electricity and lights

Our final indicator capturing local public goods provision is the share of region's households that have access to electricity. Electrification is an important element of basic infrastructure and naturally affects the living standard of local population. In addition, we also construct a regional measure of average nighttime luminosity. Henderson et al. (2012) have shown



(a) ELF(k)

(b) ELP(k)

(c) ELF$_\delta$

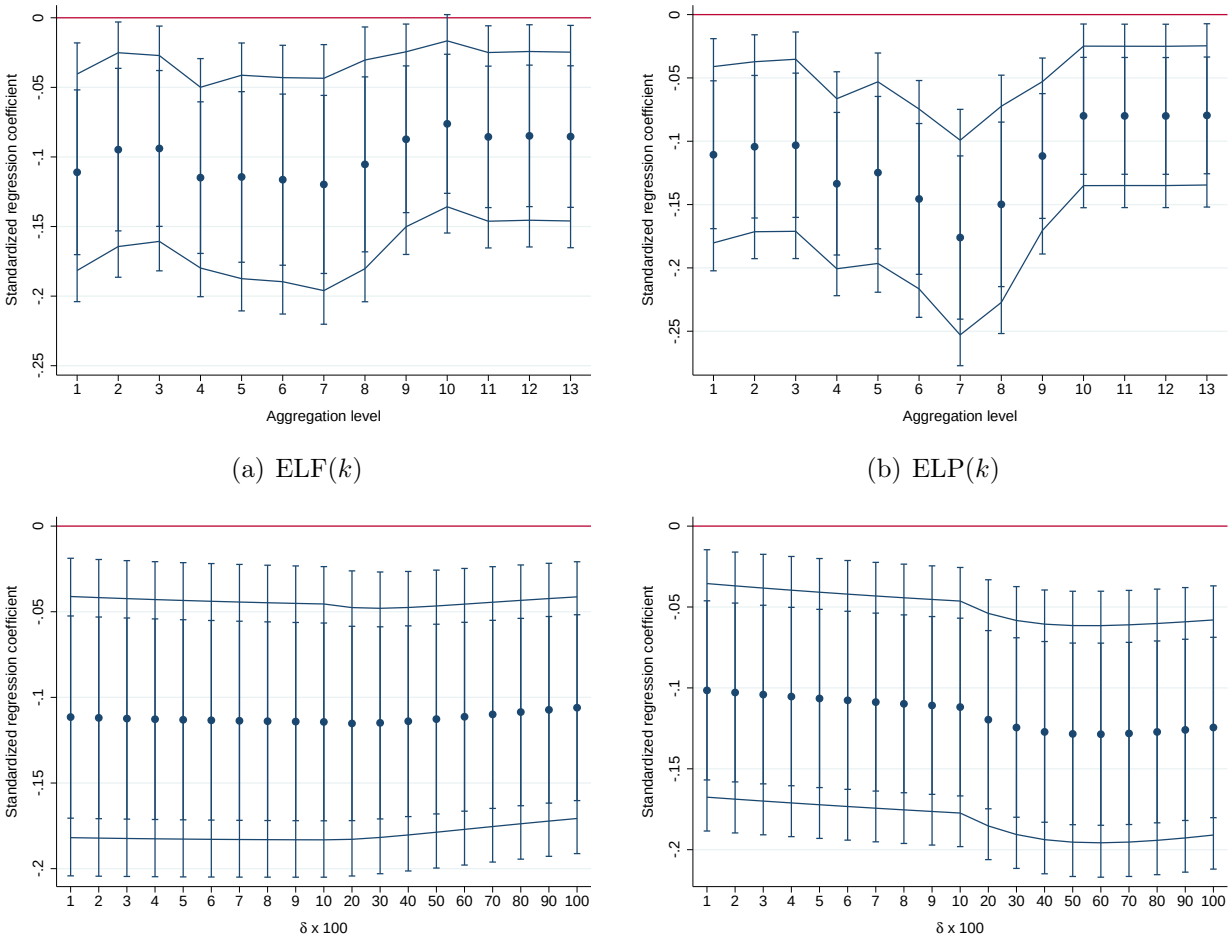
(d) ELP$_\delta$

Figure 13: Ethnolinguistic diversity and access to electricity ($n = 391$).

this measure to be correlated with traditional metrics of economic activity such as GDP per capita, and recent research has employed it when standard indicators are not available, as is the case for subnational regions in most developing countries.[41] Following this literature, we calculate mean luminosity for each region in our sample and then take the natural logarithm of 0.01 plus the lights index averaged across years 2010 and 2011. Not surprisingly, our measures of luminosity and access to electricity are tightly connected, with the correlation coefficient of 0.72.

We next run the same type of regressions as above and show the estimates in Figures 13 and 14. The results for electricity access are in line with those reported for educational and health outcomes. Both ELF(k) and ELP(k) indices, but especially the former, are

[41] See, for example, Michalopoulos and Papaioannou (2013; 2014) and Hodler and Raschky (2014).



(a) ELF(k)

(b) ELP(k)

(c) ELF$_\delta$

(d) ELP$_\delta$

Figure 14: Ethnolinguistic diversity and nighttime lights ($n = 397$).

negatively related to electricity access at high aggregation levels. Similarly, the estimates for ELF$_\delta$ indices are larger in magnitude for low values of δ, while the corresponding ELP$_\delta$ indices are only statistically significant (at the 5% level) for $\delta < 0.1$. The relevant point estimates are generally lower in magnitude compared to the cases of educational and health indicators, roughly falling in the range between -0.09 and -0.05.

Interestingly, as shown in panel (a) of Figure 14, nighttime luminosity is significantly negatively related to ELF(k) indices for any value of k. The pattern is similar for polarization indices, with the magnitude of the coefficient estimates being highest in the mid-range of aggregation levels. Distance-adjusted diversity indices are also all negatively related to nighttime lights. For ELF$_\delta$, we observe a similar pattern as before, although the coefficient estimates remain very similar at different values of δ. In the case of ELP$_\delta$, the magnitude

of estimated coefficients is in fact slightly larger at higher values of δ, and this relationship is non-monotonic overall. The standardized point estimates in most cases of interest are around -0.1 reaching as low as -0.17 for ELP(7).

As mentioned earlier, despite its high correlation with electricity access, nighttime luminosity is considered to be a good proxy for the overall level of economic development more broadly. In the following section, we complement our analysis with direct measures of regional income per capita and household wealth.

3.5 Income and wealth

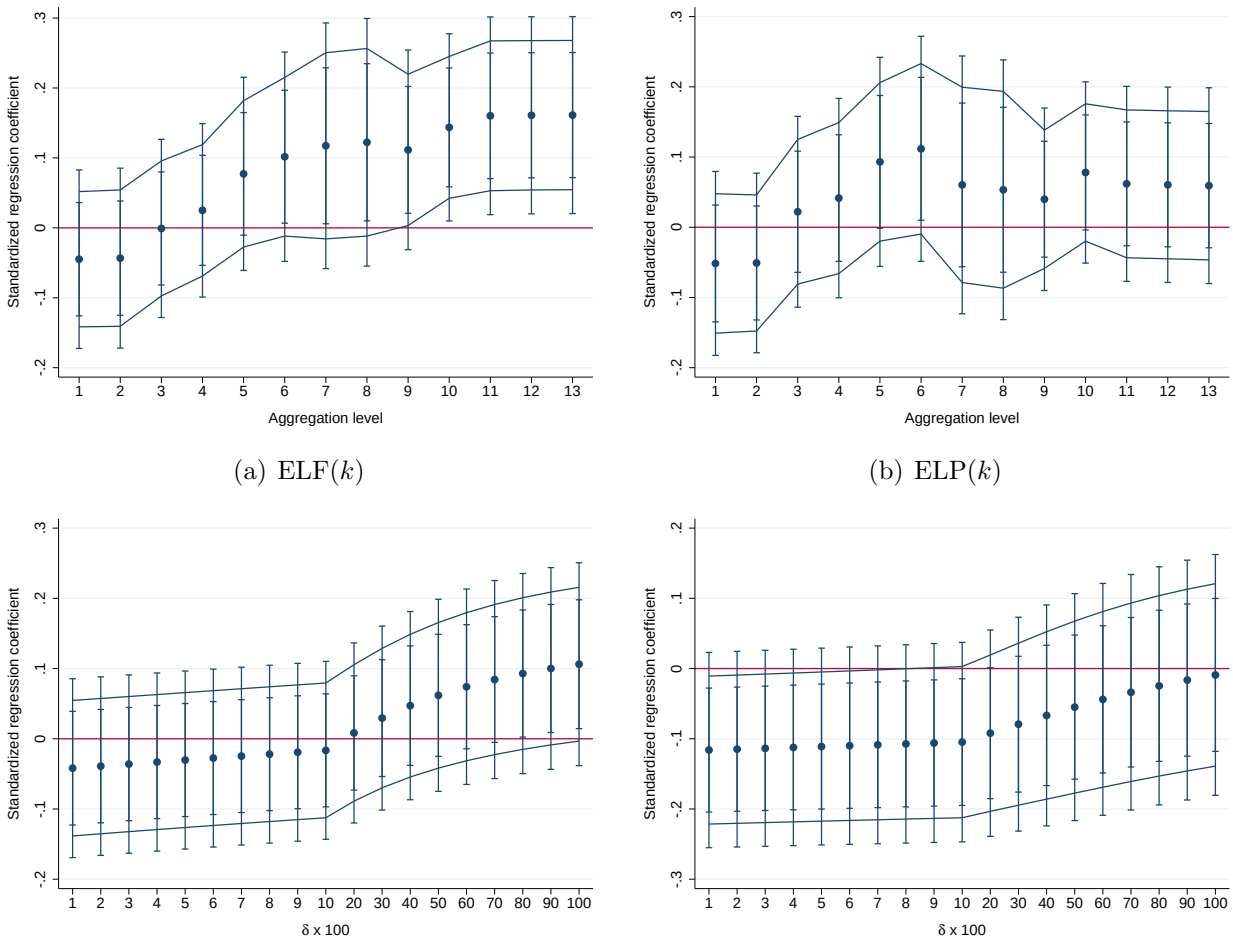
We first compile a dataset on gross regional product (GRP) per capita for countries in our sample by supplementing the figures from Gennaioli et al. (2013) with those from Mitton (2016) to cover Gambia, Mozambique, and Zimbabwe.[42] Both sources provide estimates of GRP per capita in current purchasing power parity (PPP) dollars for the year 2005. Unfortunately, the pooled dataset only covers 198 regions in 18 countries, that is, about half of our sample. Still, we estimate our main specifications for this restricted sample, with natural logarithm of GRP per capita used as an outcome variable.

As shown in Figure 15, the results are quite different from the patterns observed earlier. Relative to the baseline estimates for literacy rate in Figure 8, it appears that the sets of coefficients in four panels are "shifted up" vertically. As a result, many point estimates become *positive* and insignificant at the 5% level, with an intriguing exception of ELF(k) measured at the highest levels of disaggregation.[43] Interestingly, when we re-estimate our models for public goods outcomes in this restricted subsample of 198 regions, the qualitative results reported above remain intact (and are in fact stronger in terms of economic significance), suggesting that the different findings for income per capita are not necessarily driven by a selected subset of regions.

Given the well-known issues with the quality of official GDP statistics for Sub-Saharan Africa (Young, 2012; Jerven, 2013), their limited availability, and additional problems with subnational measures such as the lack of proper regional PPP conversion rates, we also construct an alternative survey-based metric of local development capturing average household wealth. Specifically, we employ the international wealth index (IWI), an asset-based index of material well-being measured at the household level (Smits and Steendijk, 2015). As explained by the authors of the index, it is conceptually similar to the wealth

[42]Data for Nigerian states come from the Canback Global Income Distribution Database (C-GIDD).

[43]As shown in Table 2 of section 5, the latter finding is substantially weakened in the "rural" sample, with ELF(13) barely significant at the 10% level.



(a) ELF(k)

(b) ELP(k)

(c) ELF$_\delta$

(d) ELP$_\delta$

Figure 15: Ethnolinguistic diversity and log GRP per capita ($n = 198$).

scores typically included in the DHS and MICS surveys, but has an important feature of comparability both across countries and over time, as it is based on a common set of assets. These include seven consumer durables (TV, refrigerator, phone, bicycle, car, cheap utensils, and expensive utensils), three housing characteristics (number of sleeping rooms, floor material, toilet facility), access to water supply and electricity.[44] Given how basic some of these assets are, IWI is particularly well-suited for capturing wealth in low-income countries and thus fits our analysis really well.

[44]The ownership of consumer durables and access to electricity are binary variables, while other components are measured on a three-category quality scale. IWI is derived using principal component analysis and is scaled to vary between 0 and 100, where a household with an index of 100 enjoys the full set of assets, all of the highest quality.

We connect household-level IWI data to the corresponding surveys closest to our benchmark year 2010 and construct regional indices of average wealth.[45] Next, we estimate our baseline specifications yet again, now with regional IWI as an outcome variable. As seen in Figure 16, the coefficient estimates displayed in the four panels are small in magnitude and statistically insignificant.[46]

Overall, the measures of income per capita and wealth explored in this section yield largely insignificant results. This finding is a warning against hasty generalizations of our findings for indicators reflecting primarily local public goods provision to broader metrics of regional development. The negative association with deep-rooted ethnolinguistic diversity only transpires in the context of outcomes whose production requires explicit collective effort and participation of local authorities.

3.6 Ethnolinguistic and religious diversity

In this section, we enhance our analysis by briefly exploring the role of religious diversity within the framework used so far. Specifically, we run a series of "horse-race" regressions which add the indices of religious fractionalization or polarization introduced in section 2.6 to our baseline model specifications.

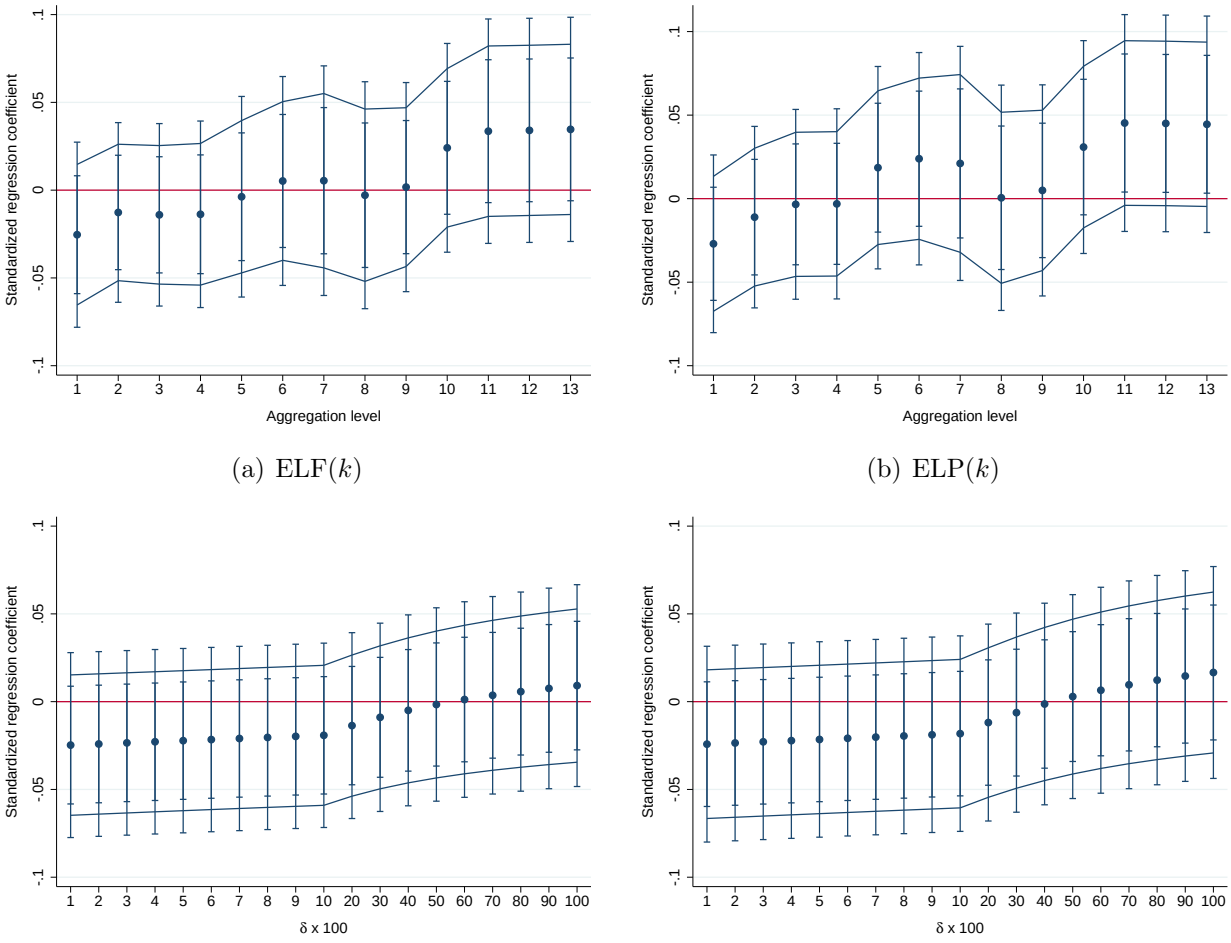
Figure 17 shows the results of this exercise for regional literacy rate. As can be seen from panels (a) through (d) of this figure, religious diversity appears to be unrelated to regional literacy and does not significantly affect any of the estimates for ethnolinguistic diversity. This pattern largely holds for all other outcomes explored above, except for electricity access and the share of home births, both of which are weakly negatively related to religious fractionalization in a few selected "horse races."

Overall, subnational religious diversity seems largely unrelated to our development indicators, nor does its addition to the model change any of our earlier findings.[47]

[45]The surveys used for each country are listed in Table A.2 in the appendix. Household-level IWI files were obtained from Jeroen Smits and online at `https://globaldatalab.org/iwi`. No data are currently available for Botswana and Eritrea. The correlation coefficient for regional IWI and log GRP per capita is equal to 0.79 and, remarkably, coincides with the correlation between *country-level* IWI and log GNI per capita as reported by Smits and Steendijk (2015) for a sample of 87 countries.

[46]Interestingly, as shown in section 5, the results for IWI become somewhat closer to those reported for public goods outcomes in the "rural" sample. Since IWI supplements the data on the ownership of private goods with some elements reflective of local public goods provision (access to electricity and water), the negative association between the latter and deep-rooted diversity is presumably driving this result.

[47]A more comprehensive analysis of this important societal cleavage, including its interaction with ethnic divisions, is beyond the scope of this paper, but is undertaken in Gershman and Rivera (2017).



(a) ELF(k)

(b) ELP(k)

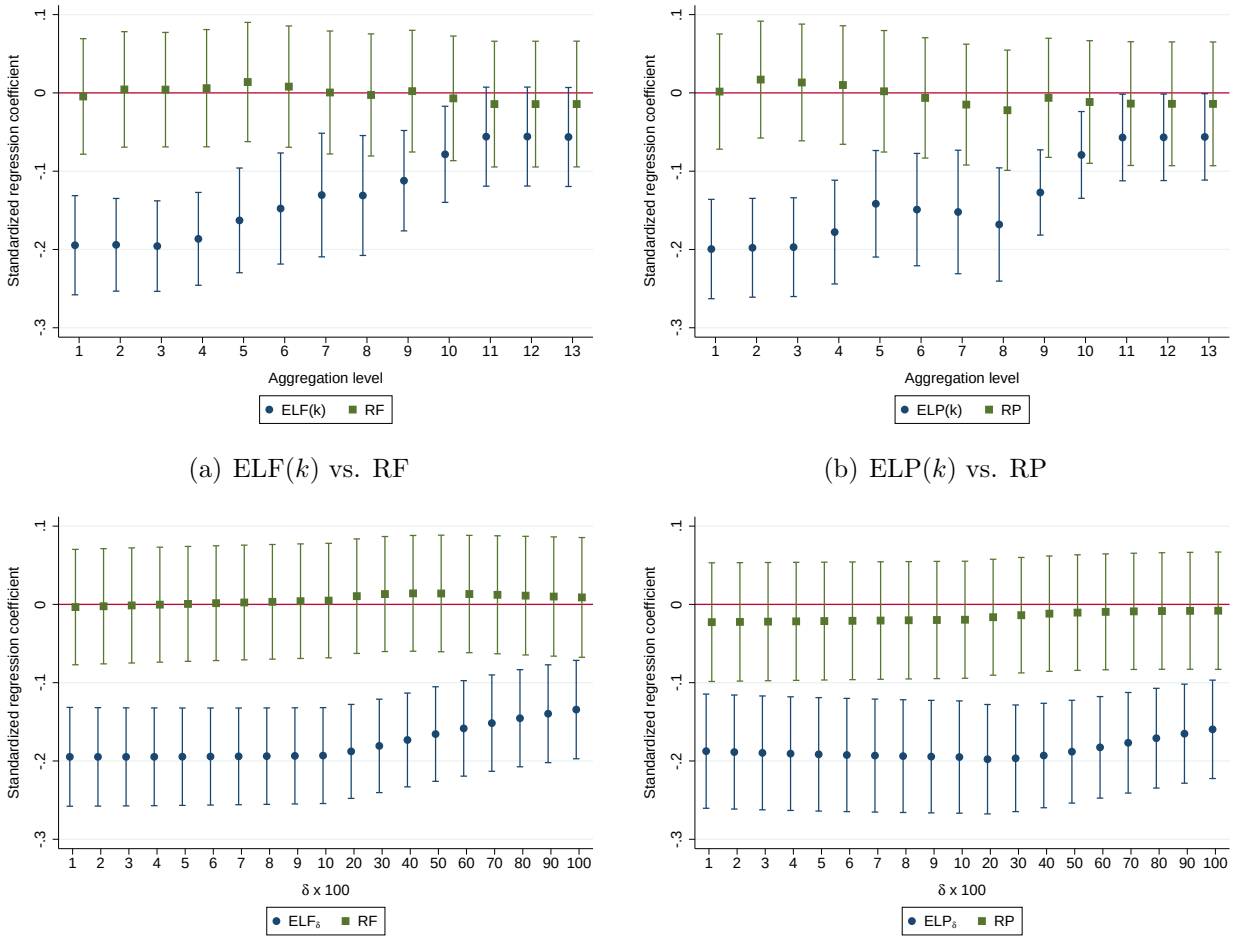
(c) ELF$_\delta$

(d) ELP$_\delta$

Figure 16: Ethnolinguistic diversity and household wealth ($n = 382$).

3.7 Main results in perspective

The empirical analysis presented in this section shows that indices of ethnolinguistic diversity adjusted for relatedness between languages are strongly negatively associated with a range of development indicators reflecting local public goods provision. Educational and health outcomes, as well as access to electricity, are to a large extent determined by the presence of schools, hospitals, and power lines, some of the basic elements of socioeconomic infrastructure. In this regard, our findings contribute most directly to the sizable empirical literature on the relationship between diversity and public goods provision going back to Alesina et al. (1999) and Miguel and Gugerty (2005).



(a) ELF(k) vs. RF

(b) ELP(k) vs. RP

(c) ELF$_\delta$ vs. RF

(d) ELP$_\delta$ vs. RP

Figure 17: Ethnolinguistic vs. religious diversity: "horse-race" regressions for literacy rate

Notes. This figure reports the estimates from models equivalent to those in Figure 8, but with indices of religious fractionalization or polarization added to the list of regressors. The point estimates are accompanied by 95% confidence intervals based on robust standard errors.

Several mechanisms have been offered to rationalize the negative association between the two. Among other things, differences between ethnic groups may harm collective action and undermine public goods provision due to conflicting preferences, prejudice against other groups, difficulties to communicate and trust each other in the absence of shared language and culture, as well as inability to impose sanctions on non-co-ethnics for failing to cooperate.[48] In contrast, homogeneous communities may be better at generating public goods since common language, social networks, history of interactions, and shared cultural heritage make it easier for people to cooperate and punish free-riders. While our empirical setting does not pin down specific mechanisms underlying the established nega-

tive relationship, the main finding on the importance of accounting for group similarity is consistent with all of the above channels. Indeed, the extent of dissimilarity or closeness between groups should strengthen or alleviate the impact of diversity on public goods provision. Larger differences, as captured by linguistic distances (which are correlated with cultural distance in a broader sense), make mutual understanding and cooperation, as well as participation in social networks of non-co-ethnics, especially problematic and hinder the provision of public goods.

Our null results for measures of income per capita and household wealth underscore the importance of differentiating between the types of development indicators in studies of ethnolinguistic diversity. The absence of significant association with these two broad metrics of economic progress may reflect, in particular, the net result of positive and negative effects of diversity, as discussed in the introduction. In contrast, the negative impact of diversity dominates when we focus narrowly on outcomes capturing the provision of local public goods which relies heavily on collective effort and cooperation.

It is also interesting to compare our main findings to those reported by Desmet et al. (2009) and Desmet et al. (2012) for a global cross-section of countries. First, like the former study, we find that distance-adjusted diversity measures are more relevant than basic ELF and ELP indices, although we look at development indicators, rather than the level of redistribution. Second, we, too, observe that the level of linguistic aggregation matters. However, in contrast to our results, Desmet et al. (2012) report that *finer* distinctions between languages appear to be more important in the case of public goods provision and economic growth at the country level, whereas deeper divisions matter more in the analysis of civil conflict and redistribution.[49]

In the absence of exogenous variation in ethnolinguistic diversity, we cannot claim to have identified the causal effects of diversity. The following two sections aim to alleviate the usual concerns that reverse causality, omitted variable bias, and measurement error may qualitatively alter our main results.

[48]Similar mechanisms may apply not just at the community level, but also at the level of local administrations in charge of budget allocation and other policy decisions. Furthermore, ethnic favoritism could lead to suboptimal distribution of resources in a heterogeneous region. See Habyarimana et al. (2007) and Gisselquist et al. (2016) for a discussion of alternative theories.

[49]In a recent paper, Spolaore and Wacziarg (2016b) find that countries whose populations are more closely related in terms of genes, language, or religion are also more likely to engage in interstate military conflict. The authors emphasize that this result is specific to international conflict and argue that the opposite relationship is likely to hold for the case of within-country disputes over common goods, policies, or government control.

4 Persistence of subnational diversity

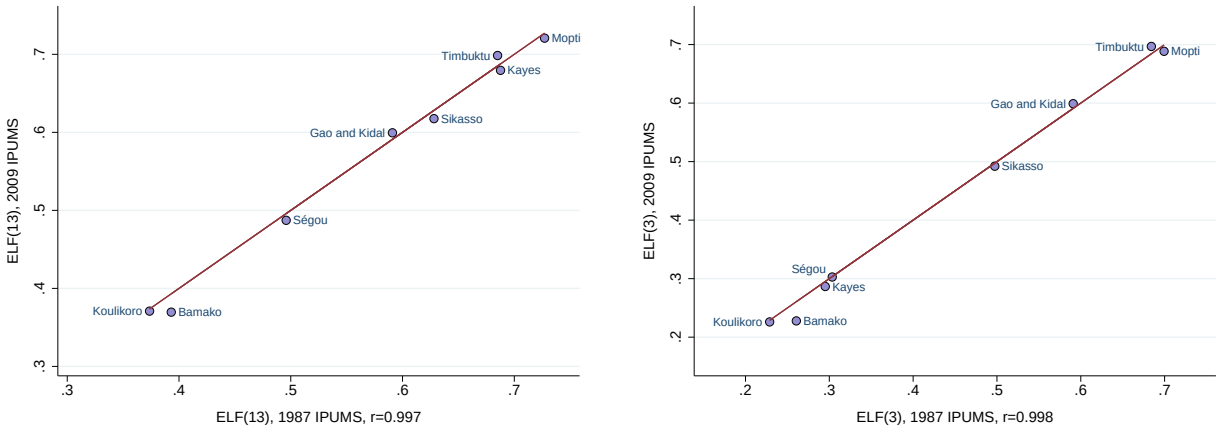
It is a priori sensible to argue that a quest for better public goods provision may lead to population sorting, which would boost diversity in regions offering broader economic opportunities, particularly urban areas. This argument seems especially relevant in the context of a subnational analysis, since changing residence within a country is presumably less costly than moving abroad. If such sorting takes place, our estimates understate the true negative effect of diversity.

We address this issue in the following ways. First, we control for regional urbanization rates, distance to the capital, and capital city indicator in all of our regressions. In section 5 below, we also perform robustness checks showing that our main results hold and become even more economically significant in subsamples that completely exclude urban or capital regions. Second, in this section, we explicitly examine the persistence of subnational diversity. To this end, we construct regional ELF indices for the cases in which both recent and older high-quality data on ethnolinguistic composition are available for identically defined subnational regions. We were able to find such data for five countries in Sub-Saharan Africa: Mali, Zambia, Liberia, Kenya, and Gabon. In each case, we work with consistent administrative boundaries and two surveys separated by a long time period: 22 years for Mali (1987–2009), 20 years for Zambia (1990–2010), 34 years for Liberia (1974–2008), 25 years for Kenya (1989–2014), and 19 years for Gabon (1993–2012).

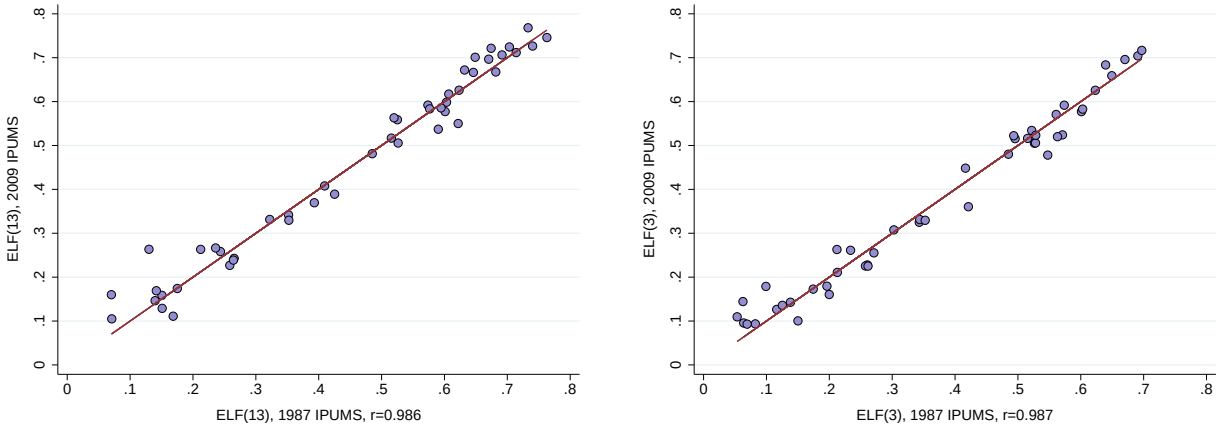
The results of our analysis are graphically presented in Figures 18–22 comparing regional ELF(k) indices over time for these five countries.[50] In all cases, the observed degree of persistence is remarkable. The case of Mali is particularly interesting since data for consistent boundaries are available for both first- and second-level administrative units. As panel (a) of Figure 18 demonstrates, both ELF(13) and ELF(3) indices remain virtually unchanged between 1987 and 2009, with the correlation coefficient (r) across 8 regions exceeding 0.99. What is even more astonishing, diversity turns out to be almost as persistent across 47 second-level subnational units of Mali within the same time frame, as seen in panel (b). Figure 19 shows that a similar pattern holds for Zambia, where the correlation between regional ELF(10) and ELF(13) indices in 1990 and 2010 exceeds 0.98 and 0.99, respectively.[51]

[50] Note that in these figures, as well as those in appendix F, the solid line represents the 45-degree line.

[51] Gisselquist et al. (2016) examine ethnolinguistic diversity in Zambia at the finer, district level. They document that internal migration in Zambia is fairly low, with the vast majority of movers relocating within districts. Furthermore, they find that the change in diversity between 2000 and 2010 is very weakly correlated with health and educational outcomes, lending little support to the idea of residential sorting.



(a) First-level subnational regions



(b) Second-level subnational regions

Figure 18: Persistence of regional diversity in Mali, 1987–2009.

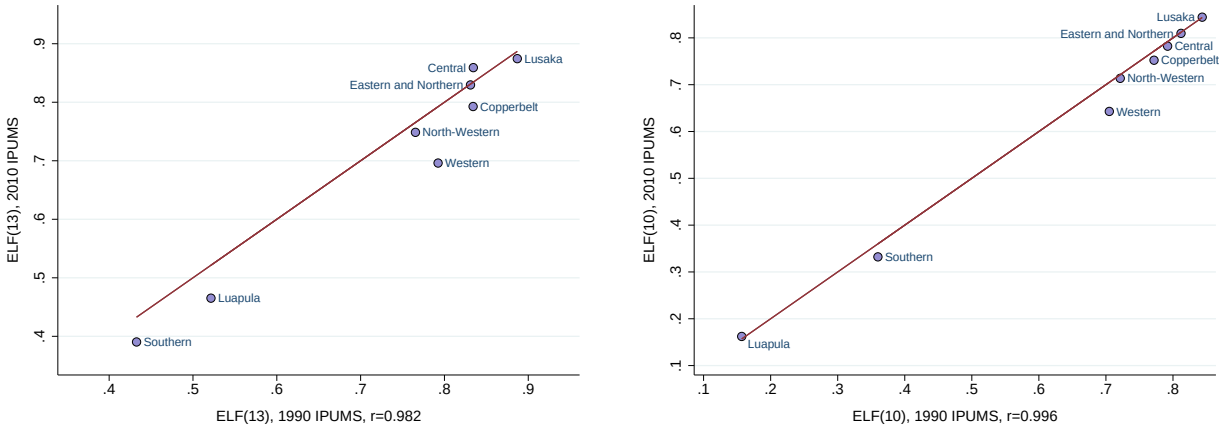


Figure 19: Persistence of regional diversity in Zambia, 1990–2010.

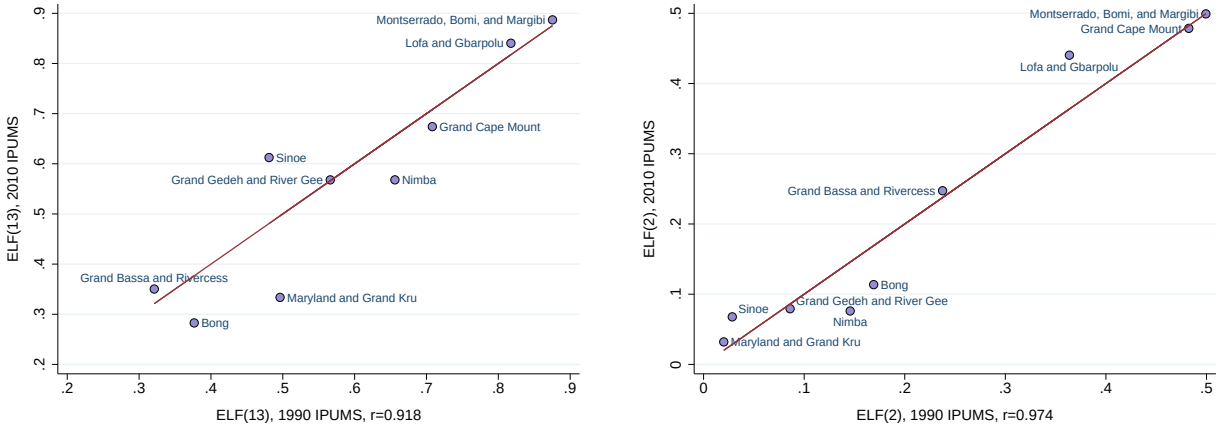


Figure 20: Persistence of regional diversity in Liberia, 1974–2008.

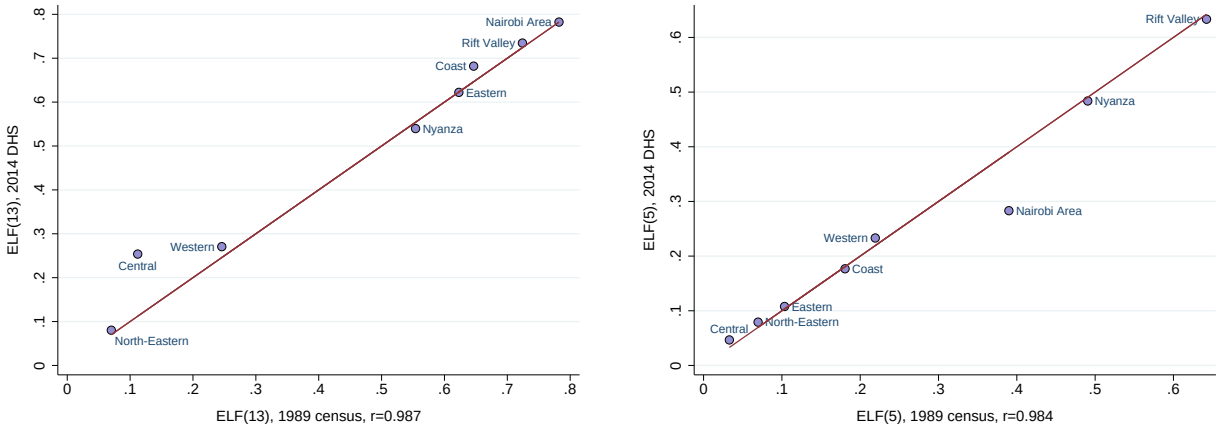


Figure 21: Persistence of regional diversity in Kenya, 1989–2014.

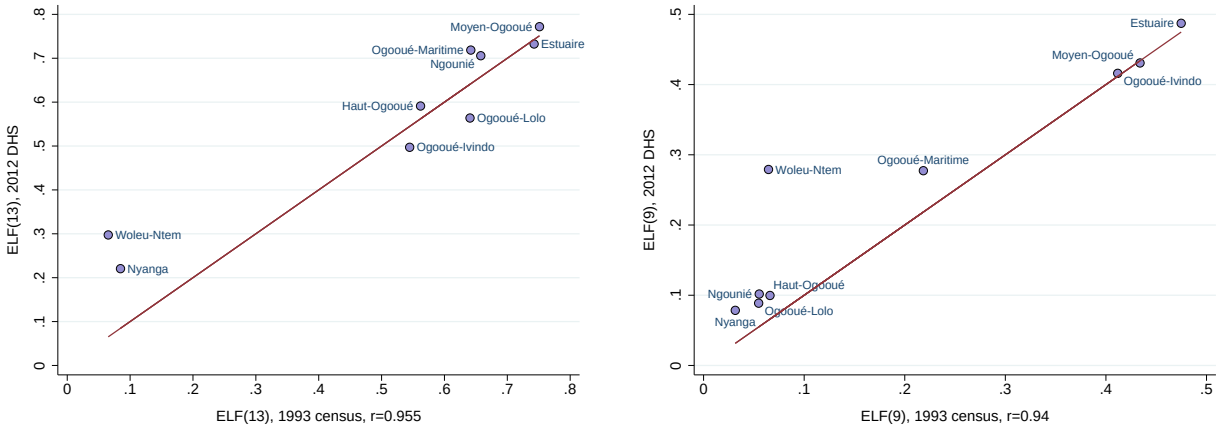


Figure 22: Persistence of regional diversity in Gabon, 1993–2012.

In Liberia, a country ravaged by civil wars from 1989 to 2003, the correlation coefficient between regional ELF(13) indices in 1974 and 2008 is equal to 0.92 and reaches 0.97 for ELF(2), as seen in Figure 20. These are remarkable numbers considering the direct impact of civil war on the displacement of population.[52] Figures 21 and 22 show that subnational diversity has also been very persistent in Kenya and Gabon.[53] Similar patterns are observed for ELF_δ indices, as shown in appendix F for the cases of Mali and Kenya. Overall, across all couples of ELF(k) and ELF_δ indices calculated for the five countries at the first subnational level, the average value of the pairwise correlation coefficient is close to 0.97.

Finally, in order to see whether the tiny observed changes in diversity are systematically associated with local economic performance, we correlate them with the regional measure of nighttime luminosity, defined earlier in section 3.4.[54] Figure 23 illustrates the relationship between changes in selected ELF(k) indices and nighttime lights for the pooled sample of five countries and 42 regions. None of the cases yield a statistically significant association that would be consistent with residential sorting.

Overall, the direct evidence presented above implies that subnational diversity is both remarkably persistent over time and generally unresponsive to regional economic performance. Of course, given the scarcity of available data, the scope of this exercise is rather limited, both in terms of the sample size and the time frame. Still, it serves to partly mollify the concern that our estimates are biased by population sorting in the short to medium run, as captured by a period of two-three decades.

5 Robustness analysis

This section further explores the sensitivity of our results by conducting several robustness checks. Our first "stress-test" has to do with the quality of original data on regional ethnolinguistic composition. As explained earlier, we carefully treat the cases in which the best available data do not permit to credibly establish the ethnic identity of survey respondents. In general, the regional shares of such unidentified residents in our dataset

[52] Glennerster et al. (2013) document high persistence of diversity at the chiefdom level in Sierra Leone between 1963 and 2004. Apparently, despite forced migration during the civil war in this country, the vast majority of movers returned to their home chiefdoms after the end of conflict.

[53] Interestingly, this high correlation holds despite the fact that the initial-year ELF indices are based on census reports while the final-year indices are based on DHS surveys.

[54] This indicator is particularly attractive since we are working with regional boundaries made consistent over relevant 20-30-year periods. The IPUMS project provides the corresponding digital maps, which allows us to recalculate nighttime luminosity using properly adjusted boundaries.

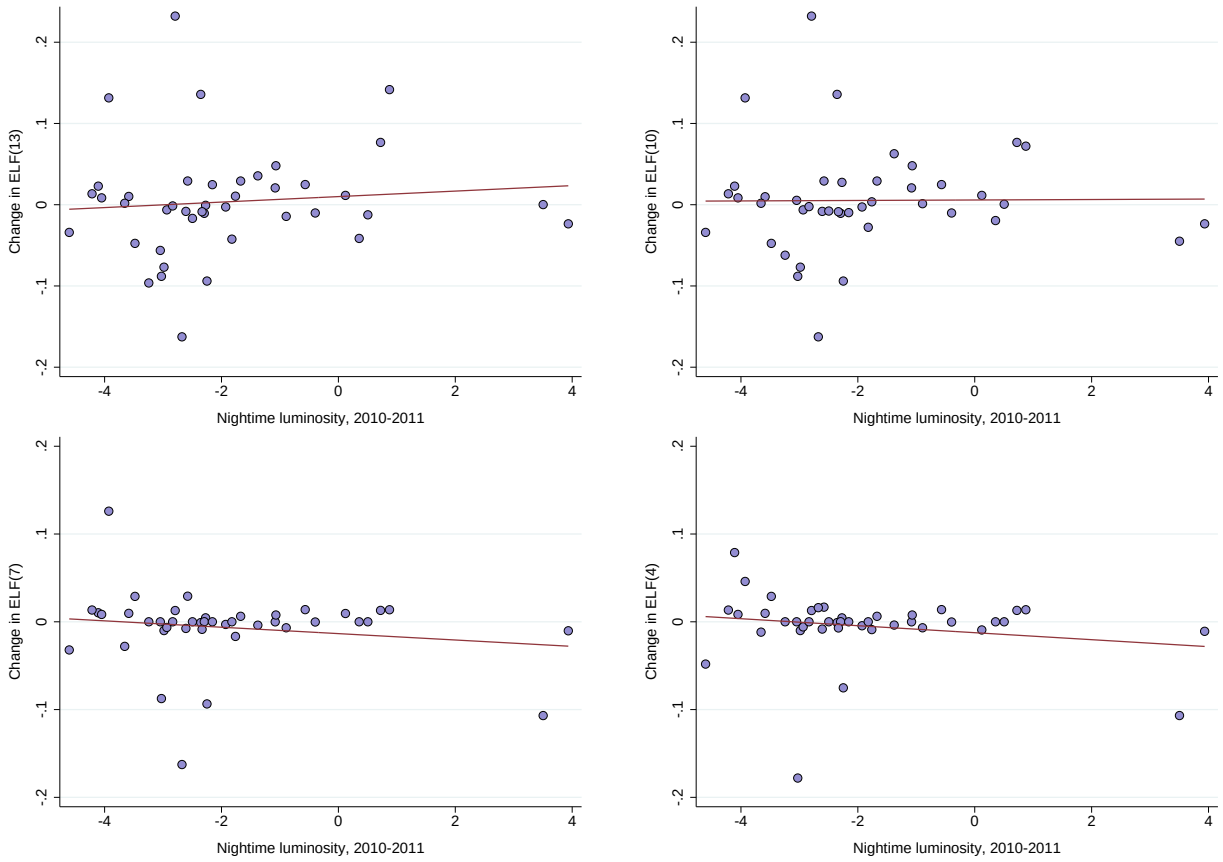


Figure 23: Change in ELF(k) and nighttime luminosity.

are very low, with a mean of 1.5%. However, to make sure this measurement error does not distort our results, we rerun a subset of specifications from section 3 after excluding the regions where "others" constitute more than 5% of the population.

Table 1 compares the estimates for the full and "trimmed" samples. Each standardized coefficient estimate and the corresponding robust standard error in this table represent a different regression, in which one of the three dependent variables indicated in the top row is regressed on the diversity index specified in the first column, the full set of regional controls, and country fixed effects. The estimates in the high-quality subsample are very similar to the baseline and often larger in magnitude. Thus, our results are essentially unaffected by the exclusion of regions with less accurate information on ethnolinguistic composition.

The second set of robustness checks considers the sensitivity of our main results to the exclusion of regions with urbanization rates of 50% or higher and those containing capital cities. As discussed earlier, elements of basic infrastructure are more readily available in

Table 1: Robustness to measurement error in diversity indices

	Literacy rate		Home births		Nighttime lights	
Share of "others"	Any	<5%	Any	<5%	Any	<5%
	(1)	(2)	(3)	(4)	(5)	(6)
ELF(3)	-0.193^{***}	-0.200^{***}	0.226^{***}	0.238^{***}	-0.094^{***}	-0.095^{***}
	(0.029)	(0.031)	(0.034)	(0.036)	(0.034)	(0.036)
ELF(8)	-0.130^{***}	-0.145^{***}	0.130^{***}	0.121^{**}	-0.105^{***}	-0.101^{**}
	(0.039)	(0.041)	(0.048)	(0.051)	(0.038)	(0.040)
ELF(13)	-0.061^{*}	-0.078^{**}	0.059	0.052	-0.085^{***}	-0.088^{***}
	(0.032)	(0.034)	(0.039)	(0.042)	(0.031)	(0.033)
$ELF_{0.05}$	-0.192^{***}	-0.206^{***}	0.207^{***}	0.214^{***}	-0.113^{***}	-0.113^{***}
	(0.031)	(0.032)	(0.035)	(0.037)	(0.035)	(0.040)
$ELF_{0.5}$	-0.163^{***}	-0.176^{***}	0.169^{***}	0.169^{***}	-0.113^{***}	-0.112^{***}
	(0.031)	(0.032)	(0.036)	(0.038)	(0.034)	(0.036)
ELP(3)	-0.194^{***}	-0.205^{***}	0.228^{***}	0.246^{***}	-0.103^{***}	-0.106^{***}
	(0.032)	(0.033)	(0.037)	(0.038)	(0.035)	(0.037)
ELP(8)	-0.167^{***}	-0.178^{***}	0.120^{***}	0.113^{**}	-0.150^{***}	-0.138^{***}
	(0.037)	(0.039)	(0.045)	(0.047)	(0.039)	(0.041)
ELP(13)	-0.061^{**}	-0.062^{**}	0.032	0.024	-0.080^{***}	-0.075^{**}
	(0.027)	(0.029)	(0.034)	(0.036)	(0.028)	(0.029)
$ELP_{0.05}$	-0.189^{***}	-0.212^{***}	0.180^{***}	0.203^{***}	-0.107^{***}	-0.103^{***}
	(0.036)	(0.038)	(0.043)	(0.044)	(0.033)	(0.038)
$ELP_{0.5}$	-0.189^{***}	-0.198^{***}	0.157^{***}	0.166^{***}	-0.128^{***}	-0.122^{***}
	(0.033)	(0.035)	(0.040)	(0.041)	(0.034)	(0.037)
Observations	397	363	382	348	397	363

Notes. a) Each cell shows the standardized coefficient estimate for the corresponding diversity index (and the respective robust standard error) in a regression of the outcome variable indicated in the top row on that diversity index, the full set of regional controls, and country fixed effects. Thus, this table shows estimates from 60 separate regressions. b) ***, **, and * denote statistical significance at the 1, 5, and 10% level, respectively.

urban areas which at the same time often display higher levels of diversity. The first six columns of Table 2 indicate that, when literacy rate, home births, and nighttime lights are used as outcome variables, our estimates on diversity typically become even stronger in the sample of less urbanized areas. In contrast, the already weak results for log GRP per capita become even weaker in the rural sample, so that ELF(13) only remains statistically significant at the 10% level, still with a positive coefficient estimate. In the case of IWI,

Table 2: Robustness to the exclusion of urban and capital regions

	Literacy rate		Home births		Nighttime lights		Log GRP per capita		IWI	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
ELF(3)	-0.212^{***}	-0.198^{***}	0.256^{***}	0.234^{***}	-0.143^{***}	-0.130^{***}	-0.088	-0.031	-0.060^{*}	-0.012
	(0.036)	(0.030)	(0.045)	(0.037)	(0.046)	(0.040)	(0.073)	(0.049)	(0.033)	(0.023)
Adjusted R^2	0.803	0.834	0.685	0.710	0.698	0.702	0.709	0.790	0.823	0.867
ELF(5)	-0.170^{***}	-0.160^{***}	0.182^{***}	0.156^{***}	-0.157^{***}	-0.143^{***}	0.026	0.060	-0.037	0.002
	(0.042)	(0.035)	(0.052)	(0.043)	(0.050)	(0.044)	(0.081)	(0.056)	(0.036)	(0.025)
Adjusted R^2	0.792	0.825	0.664	0.690	0.700	0.704	0.705	0.791	0.821	0.867
ELF(8)	-0.144^{***}	-0.135^{***}	0.129^{**}	0.130^{**}	-0.178^{***}	-0.148^{***}	0.075	0.104	-0.026	0.008
	(0.048)	(0.040)	(0.062)	(0.051)	(0.050)	(0.045)	(0.100)	(0.074)	(0.041)	(0.028)
Adjusted R^2	0.786	0.819	0.653	0.684	0.702	0.703	0.707	0.794	0.821	0.867
ELF(13)	-0.076^{*}	-0.067^{**}	0.055	0.061	-0.183^{***}	-0.119^{***}	0.132^{*}	0.145^{**}	-0.004	0.034
	(0.040)	(0.032)	(0.051)	(0.041)	(0.042)	(0.038)	(0.077)	(0.058)	(0.036)	(0.028)
Adjusted R^2	0.779	0.813	0.647	0.678	0.711	0.702	0.716	0.804	0.820	0.868
$ELF_{0.05}$	-0.196^{***}	-0.190^{***}	0.219^{***}	0.211^{***}	-0.147^{***}	-0.142^{***}	-0.097	-0.059	-0.063^{*}	-0.017
	(0.040)	(0.032)	(0.046)	(0.036)	(0.045)	(0.040)	(0.072)	(0.048)	(0.033)	(0.023)
Adjusted R^2	0.801	0.834	0.677	0.706	0.700	0.705	0.710	0.792	0.823	0.868
$ELF_{0.5}$	-0.176^{***}	-0.168^{***}	0.185^{***}	0.175^{***}	-0.177^{***}	-0.151^{***}	-0.002	0.035	-0.044	0.002
	(0.039)	(0.031)	(0.048)	(0.038)	(0.044)	(0.039)	(0.078)	(0.053)	(0.035)	(0.024)
Adjusted R^2	0.797	0.829	0.668	0.697	0.707	0.707	0.704	0.790	0.822	0.867
Observations	302	361	293	348	302	361	154	180	293	348
Rural only	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Capitals excluded	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Notes. a) This table shows estimates from 60 separate regressions, see the notes to Table 1. b) "Rural only" sample excludes regions with urbanization rate of 50% or higher. c) "Capitals excluded" sample drops the regions containing national capitals. d) Full set of controls and country fixed effects are included in each model. e) ***, **, and * denote statistical significance at the 1, 5, and 10% level, respectively.

the indices of deep-rooted diversity, ELF(3) and $\text{ELF}_{0.05}$, gain some statistical significance relative to the baseline sample, thus becoming more in line with specifications for our indicators of local public goods provision.

Finally, we explore the potential for omitted variable bias. There are two features in our analysis that directly alleviate this concern. First, we explicitly control for a variety of relevant confounding characteristics. Second, subnational-level analysis permits the inclusion of country fixed effects which account for nationwide factors. Nevertheless, it is still possible that there are certain important, possibly unobservable omitted regional factors that bias our estimates. A common approach to evaluating the magnitude of this problem is to check how the coefficients of interest and the overall explained variance in the outcome variable change with the inclusion of controls, relative to more parsimonious specifications.

Following common practice, we perform two robustness tests developed, respectively, by Altonji et al. (2005) and Oster (2018), for a subset of baseline regressions from section 3. To conduct a version of the former test, we follow the procedure from Nunn and Wantchekon (2011) to calculate the AET ratios, equal to $\beta_F/(\beta_R - \beta_F)$, where β_F is the coefficient estimate on diversity in the "full" baseline specification and β_R is the one from the "restricted" model including only country fixed effects. We next follow the insight of Oster (2018) who argues that, in addition to movements in the estimated coefficients, it is important to take into account the changes in R^2, and suggests an enhanced version of the test. Specifically, we calculate Oster's delta values which are interpreted as the degree of selection on unobservables relative to observables that would be necessary to completely explain away the baseline result.[55]

As shown in Table 3, both the AET ratios and Oster's delta values are typically negative and substantial in magnitude, suggesting that selection on unobservables would have to be both much more important and actually operate in the opposite direction relative to observables in order to entirely explain away our results.[56] In other words, if anything, our regional controls strengthen, rather than attenuate the estimates of interest, and, if those controls are "representative" of a broader range of relevant factors, omitted variable bias is unlikely to overturn our findings.

[55] We adopt the most conservative value of $R_{\max} = 1$ in this calculation, see Oster (2018) for details. As in the case of the AET ratios, country fixed effects are treated as nuisance parameters.

[56] As expected, the ELF indices measured at higher levels of linguistic disaggregation are more sensitive, since the magnitude of estimated coefficients in those cases is lowest.

Table 3: Robustness to bias from unobservable factors

	Literacy rate		School attendance		Home births		Nighttime lights	
	AET	Oster	AET	Oster	AET	Oster	AET	Oster
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ELF(1)	−30.49	197.27	−14.84	−48.64	−14.87	−28.73	−6.04	−17.98
ELF(3)	−34.84	34.39	−13.31	−85.40	−8.92	−13.59	−4.86	−14.77
ELF(5)	−3.54	−3.53	−2.50	−3.30	−2.26	−1.75	−1.35	−3.31
ELF(7)	−1.56	−1.36	−1.09	−1.27	−1.10	−0.78	−0.77	−1.82
$ELF_{0.05}$	−14.69	−30.36	−8.51	−16.66	−8.65	−10.15	−3.65	−9.89
$ELF_{0.5}$	−3.26	−3.35	−2.18	−2.88	−2.32	−1.87	−1.15	−2.80

Notes. Columns labeled "AET" report the Altonji et al. (2005) ratios, as described in the main text. Columns labeled "Oster" report delta values from Oster (2018), under the assumption that $R_{\max} = 1$.

6 Concluding remarks

This paper introduces a new dataset on subnational ethnolinguistic and religious diversity in Sub-Saharan Africa. Our dataset relies on high-quality data sources such as population censuses and large-scale household surveys and standardizes ethnolinguistic groups based on their spoken languages. Most importantly, we construct a set of diversity measures accounting for linguistic relatedness between groups, previously unavailable at the subnational level.

We show that whenever standard indices of ethnolinguistic fractionalization or polarization are adjusted for linguistic distances, a robust negative relationship emerges between diversity and a variety of educational and health outcomes, as well as access to electricity and nighttime luminosity. Similar relationship holds for the indices of deep-rooted diversity based on cleavages formed in the distant past.

Our findings stress the importance of accounting for group similarities when measuring subnational diversity and are consistent with the view that the degree of distinctiveness between groups indeed matters for aggregate outcomes requiring collective action such as local public goods provision. Interestingly, our results do not carry over to regional measures of income per capita and household wealth, underscoring the need to differentiate between the types of development indicators in studies of diversity.

Appendices

A Countries, regions, and ethnolinguistic groups

Table A.1: Countries, regions, ethnolinguistic groups, and administrative boundaries

Country	Regions/Groups	Boundaries	Country	Regions/Groups	Boundaries
Angola	18/10	Current	Kenya	8/29	02/2013
Benin	12/8	Current	Liberia	15/17	Current
Botswana	9/15	Current	Malawi	3/8	Current
Burkina Faso	13/27	Current	Mali	9/16	Current
Cameroon	10/43	Current	Mauritania	13/4	10/2014
Central African Republic	17/9	Current	Mozambique	11/22	Current
Chad	20/19	08/2012	Namibia	13/14	07/2013
Republic of the Congo	12/11	Current	Niger	8/8	Current
Côte d'Ivoire	19/50	08/2011	Nigeria	37/192	Current
Djibouti	6/3	Current	Senegal	11/18	08/2008
Equatorial Guinea	7/5	07/2015	Sierra Leone	4/15	Current
Eritrea	6/10	Current	South Africa	9/11	Current
Ethiopia	11/64	Current	Swaziland	4/3	Current
Gabon	9/8	Current	Tanzania	21/98	04/2002
Gambia	8/9	Current	Togo	5/5	Current
Ghana	10/38	Current	Uganda	4/39	Current
Guinea	8/6	Current	Zambia	9/33	10/2011
Guinea-Bissau	9/13	Current	Zimbabwe	10/19	Current

Notes. a) Current corresponds to 12/2015. b) The official first-level administrative divisions of Botswana include 9 rural and 7 urban districts (2 cities, 4 towns, and 1 township). We integrate the urban locations into corresponding rural districts. c) As of 02/2008, the first-level administrative division of Chad included 22 regions. Our dataset merges together the northern regions of Borkou, Ennedi, and Tibesti in one (as was officially the case before 02/2008). In 09/2012, Ennedi region was split into Ennedi Est and Ennedi Ouest. d) In 09/2011, Côte d'Ivoire was reorganized from nineteen regions into fourteen districts. e) In 08/2015, Djibloho province split from Wele-Nzas in Equatorial Guinea. f) In 03/2013, Kenya was reorganized from eight provinces to 47 counties. g) In 11/2014, Nouakchott region of Mauritania was split into Nouakchott Nord, Nouakchott Ouest, and Nouakchott Sud. h) In 09/2008, Kaffrine region of Senegal split from Kaolack, Kédougou region split from Tambacounda, and Sédhiou region split from Kolda. i) In 08/2013, the Kavango region of Namibia was split into Kavango East and Kavango West. j) In 05/2002, Manyara region of Tanzania split from Arusha. Four new regions were created in 03/2012. In addition, our dataset merges the five regions of Zanzibar archipelago into one. k) In 11/2011, Muchinga province was formed from five districts of the Northern province and one district of the Eastern province in Zambia. l) Information on the changes of administrative boundaries is taken from `statoids.com`. m) The count of ethnolinguistic groups is the number of unique three-letter *Ethnologue* codes per country in our sample.

Table A.2: Sources of primary survey data on ethnicity, religion, and outcomes

Country	Ethnicity	Religion	PG outcomes	IWI
Angola	DHS (2016)	DHS (2016)	DHS (2016)	DHS (2011)
Benin	Census (2013)	Census (2013)	DHS (2012)	DHS (2012)
Botswana	Census (2001)	IPUMS (2011)	Multiple	None
Burkina Faso	IPUMS (2006)	Census (2006)	DHS (2010)	DHS (2010)
Cameroon	DHS (2004)	Census (2005)	DHS (2011)	DHS (2011)
Central African Republic	Census (2003)	MICS (2010)	MICS (2010)	MICS (2010)
Chad	DHS (2015)	Census (2009)	MICS (2010)	MICS (2010)
Republic of the Congo	DHS (2012)	Census (2007)	DHS (2012)	DHS (2012)
Côte d'Ivoire	MICS (2006)	DHS (2012)	DHS (2012)	DHS (2012)
Djibouti	MICS (2006)	None	MICS (2006)	MICS (2006)
Equatorial Guinea	Census (1994)	None	MICS (2000)	MICS (2000)
Eritrea	DHS (2002)	DHS (2002)	Multiple	None
Ethiopia	Census (2007)	Census (2007)	DHS (2011)	DHS (2011)
Gabon	Census (1993)	DHS (2012)	DHS (2012)	DHS (2012)
Gambia	Census (2003)	Census (2003)	DHS (2013)	DHS (2013)
Ghana	IPUMS (2010)	Census (2010)	MICS (2011)	MICS (2011)
Guinea	DHS (2012)	DHS (2012)	DHS (2012)	DHS (2012)
Guinea-Bissau	Census (2009)	Census (2009)	MICS (2014)	MICS (2014)
Kenya	Census (1989)	DHS (2014)	DHS (2009)	DHS (2009)
Liberia	IPUMS (2008)	IPUMS (2008)	DHS (2013)	DHS (2013)
Malawi	Census (2008)	Census (2008)	DHS (2010)	DHS (2010)
Mali	Census (2009)	IPUMS (2009)	MICS (2010)	DHS (2006)
Mauritania	MICS (2007)	None	MICS (2007)	MICS (2007)
Mozambique	MICS (2008)	Census (2007)	DHS (2011)	DHS (2011)
Namibia	Census (2001)	DHS (2013)	DHS (2013)	DHS (2013)
Niger	Census (2001)	Census (2012)	DHS (2012)	DHS (2012)
Nigeria	DHS (2013)	DHS (2013)	MICS (2011)	DHS (2013)
Senegal	IPUMS (2002)	IPUMS (2002)	DHS (2011)	DHS (2011)
Sierra Leone	IPUMS (2004)	IPUMS (2004)	MICS (2010)	MICS (2010)
South Africa	Census (2011)	IPUMS (2001)	Multiple	GHS (2014)
Swaziland	Census (1976)	DHS (2007)	MICS (2010)	MICS (2010)
Tanzania	DHS (1992)	DHS (2005)	DHS (2010)	DHS (2010)
Togo	DHS (2014)	DHS (2014)	MICS (2010)	DHS (2014)
Uganda	IPUMS (2002)	IPUMS (2002)	DHS (2011)	DHS (2011)
Zambia	IPUMS (2010)	IPUMS (2010)	DHS (2007)	DHS (2007)
Zimbabwe	WHO (2003)	DHS (2011)	DHS (2011)	DHS (2011)

Notes. The "PG outcomes" (public goods outcomes) column refers to educational and health indicators, as well as access to electricity. DHS is Demographic and Health Surveys; MICS is Multiple Indicator Cluster Surveys; IPUMS is Integrated Public Use Microdata Series (subsamples of national censuses); WHO is World Health Organization (World Health Survey); GHS is General Household Survey (South Africa). For Botswana, data on PG outcomes come from IPUMS (2011), Census (2011) reports, National Survey on Literacy (2003), and MICS (2000); for Eritrea, data on PG outcomes come from DHS (2002) and Population and Health Survey (2010); for South Africa, data on PG outcomes come from IPUMS (2011), Census (2011) reports, GHS (2010), and DHS (2003).

B Diversity measures based on linguistic trees

This appendix explains the construction of adjusted diversity indices using the Gash-Barka region of Eritrea as an example. The 2002 DHS survey identifies 7 distinct ethnic groups in this region: Bilen (0.004), Hedareb (0.084), Kunama (0.087), Nara (0.116), Saho (0.011), Tigre (0.363), and Tigrinya (0.335), where regional population shares are indicated in parentheses. These ethnic groups are uniquely matched to their spoken languages indexed by three-letter *Ethnologue* codes: Bilen (byn), Bedawiyet (bej), Kunama (kun), Nara (nrb), Saho (ssy), Tigre (tig), Tigrigna (tir). This matching and the *Ethnologue* database allow us to build a linguistic tree for the Gash-Barka region, as shown in Figure B.1.

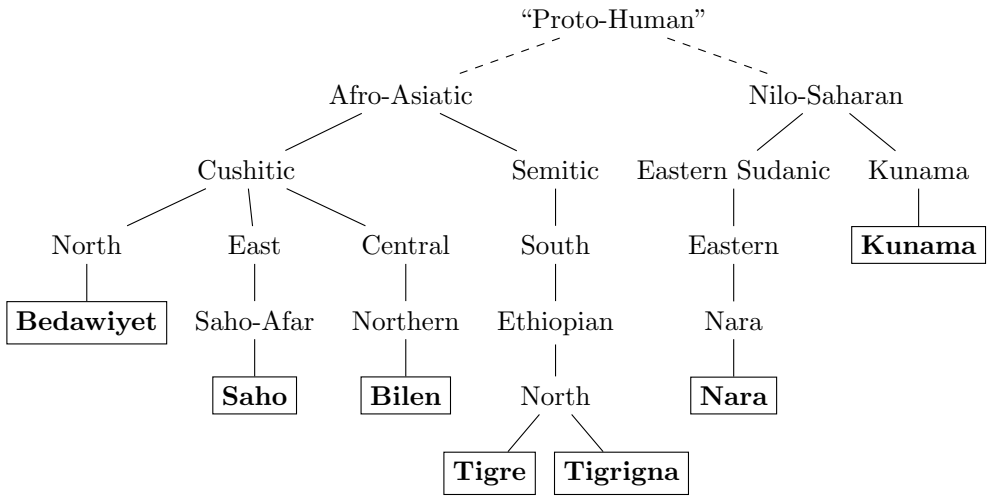


Figure B.1: Ethnolinguistic tree for the Gash-Barka region, Eritrea.

In order to construct distance-adjusted diversity indices we first compute the linguistic distances $\tau_{ij} = 1-(l/m)^{\delta}$ for each pair of languages i and j, where l is the number of shared branches and $m = 13$ is the maximum possible number of shared branches. If we take $\delta = 0.05$, the distance between Bedawiyet and Saho is equal to $1-(2/13)^{0.05} \approx 0.09$, while the distance between Tigre and Tigrigna is $1-(5/13)^{0.05} \approx 0.047$. In these calculations, the common language family (Afro-Asiatic) is counted as a shared branch coming from the hypothetical common ancestor of all languages known as "Proto-Human." For another pair of languages, Kunama and Saho, there are no shared branches, and the distance between the two is set equal to 1, as in the standard ELF index. In this fashion, the full distance matrix for all seven languages is constructed and then used to calculate the distance-adjusted ELF index. For $\delta = 0.05$, such index is equal to 0.354, as opposed to the standard ELF index of 0.728. For $\delta = 0.5$, the adjusted index is equal to 0.532.

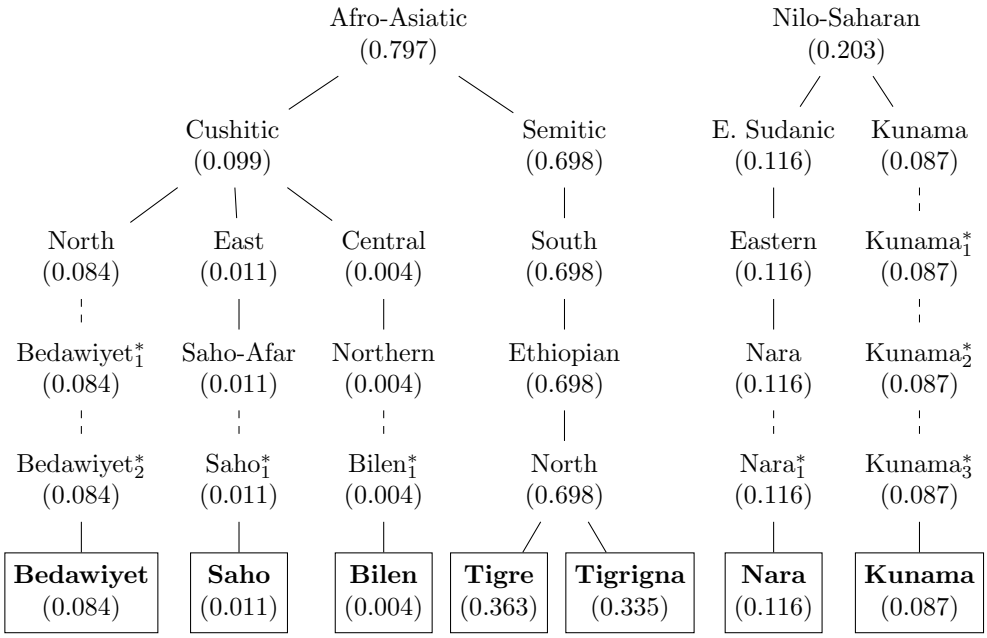
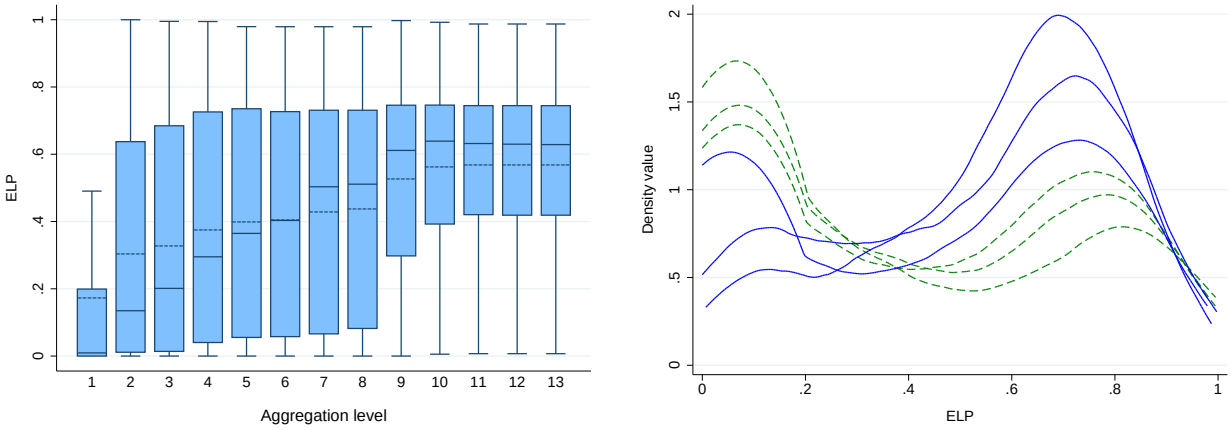


Figure B.2: Extended ethnolinguistic tree for the Gash-Barka region, Eritrea.

This linguistic tree for the Gash-Barka region can also be used to construct the ELF(k) indices from Desmet et al. (2012). Each horizontal tier of the tree represents a possible aggregation level k, where the coarsest level 1 corresponds to major language families. The problem is that different languages have the paths of varying depths starting from the "Proto-Human" root. For instance, Bedawiyet is located at level 4, while Tigre is located at level 6. To overcome this issue, we construct an extended linguistic tree such that each basic language is separated from the root by an equal number of branches. We do this by adding "fictitious" intermediate languages, as shown in Figure B.2. The assumption is that all languages went through intermediate stages before reaching their current shape. Once the tree is reconstructed, we can calculate ELF indices at 6 different levels of aggregation after adding up the population shares accordingly. For instance, moving from level 6 to level 5, Tigre and Tigrigna merge into their parent "North" subdivision which is now a group covering almost 70% of the region's population. As a result, the ELF index decreases from 0.728 at level 6 to 0.485 at level 5. Upon reaching level 1, we observe that 79.7% of the region's population speak Afro-Asiatic languages, while 20.3% speak Nilo-Saharan languages, yielding the ELF(1) index of 0.324. We calculate the diversity indices at 13 levels of aggregation since the deepest path for languages in our sample has 13 branches. For the Gash-Barka region, ELF indices at levels below 6 are all equal to 0.728.

C Descriptive statistics for polarization indices

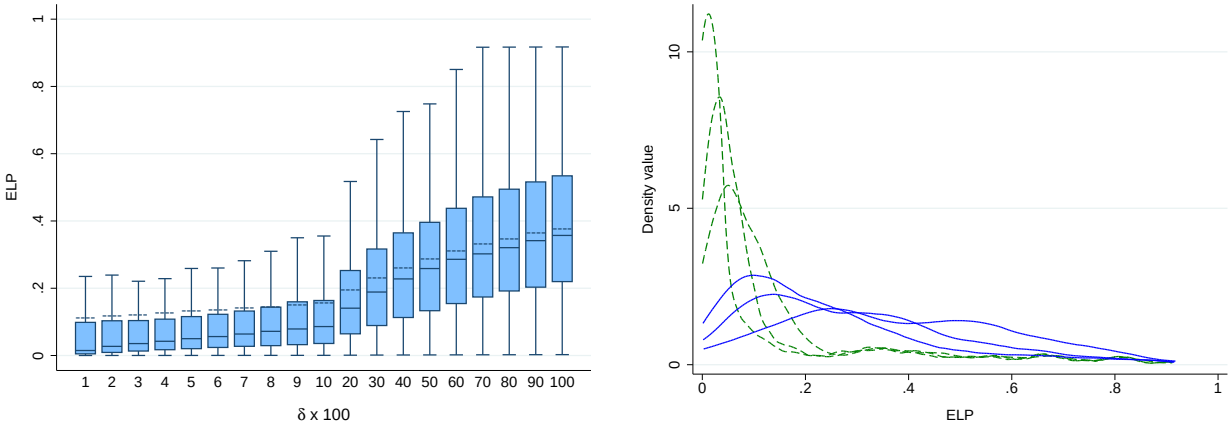


(a) Box-and-whiskers plots

(b) Kernel density estimates

Figure C.1: Descriptive statistics for ELP(k) indices.

Notes. The box-and-whiskers plots in panel (a) contain the following information: interquartile range (IQR), where the bottom (top) of the box corresponds to the lower (upper) quartile of the distribution, mean value (dashed segment), median value (solid segment), and the adjacent values representing the most extreme values within the range of 1.5×IQR from lower and upper quartiles. The kernel density plots in panel (b) correspond to $k = 3$, 4, 5, 7, 9, and 13, sorted from top to bottom by the density value at 0.

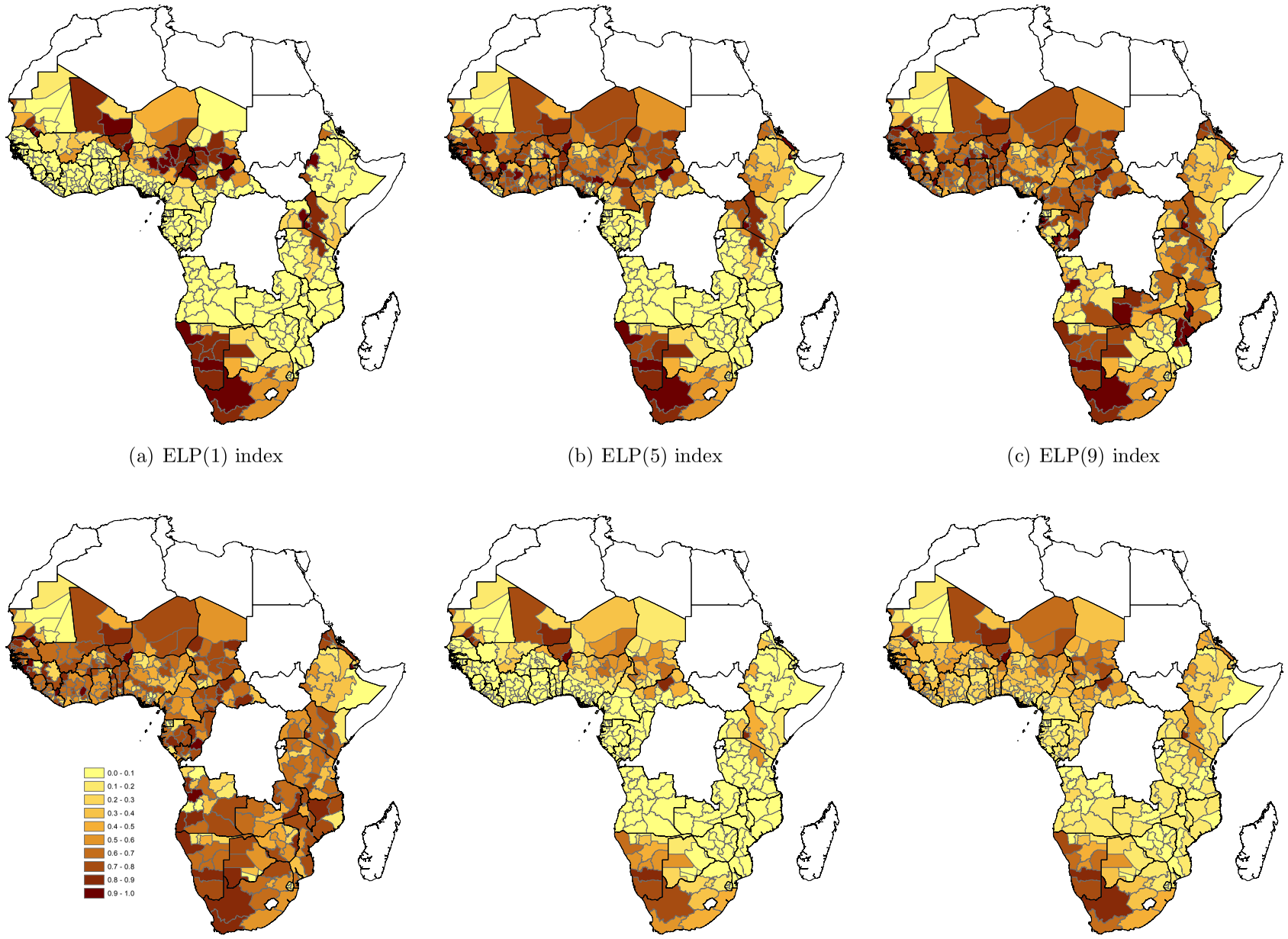


(a) Box-and-whiskers plots

(b) Kernel density estimates

Figure C.2: Descriptive statistics for ELP$_\delta$ indices.

Notes. The box-and-whiskers plot in panel (a) is constructed in the same way as the one in Figure 3. The kernel density plots in panel (b) correspond to $\delta = 0.01$, 0.05, 0.1, 0.3, 0.5, and 1 (sorted from top to bottom by the density value at 0).



(a) ELP(1) index (b) ELP(5) index (c) ELP(9) index

(d) ELP(13) index (e) $ELP_{0.05}$ index (f) $ELP_{0.5}$ index

Figure C.3: Regional distribution of ELP indices.

D Relationship between selected diversity indices

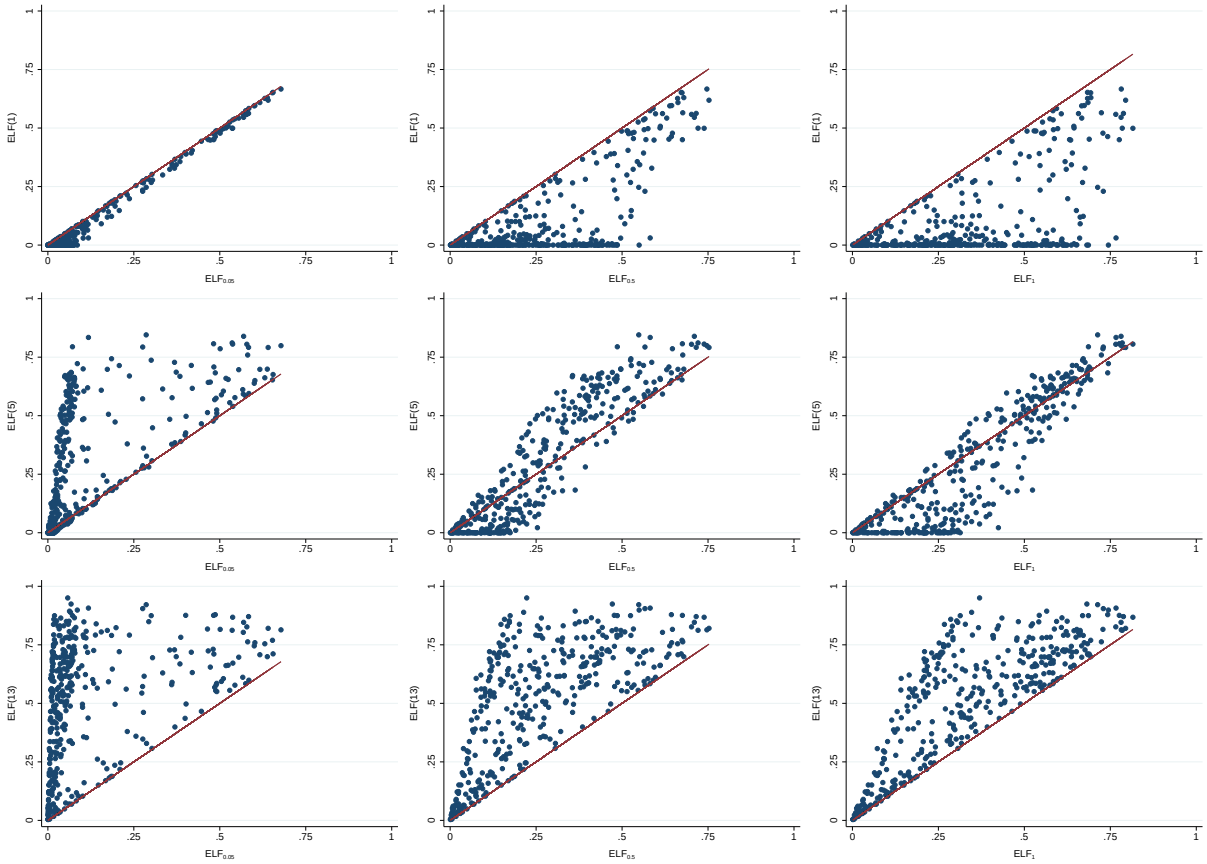


Figure D.1: Relationship between various ELF indices.

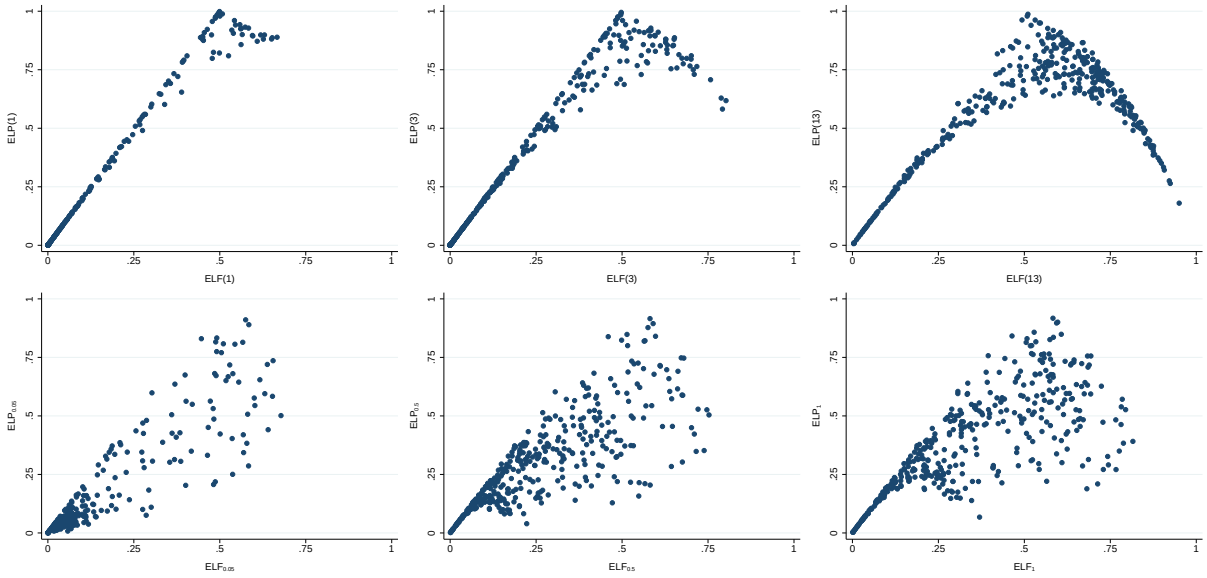
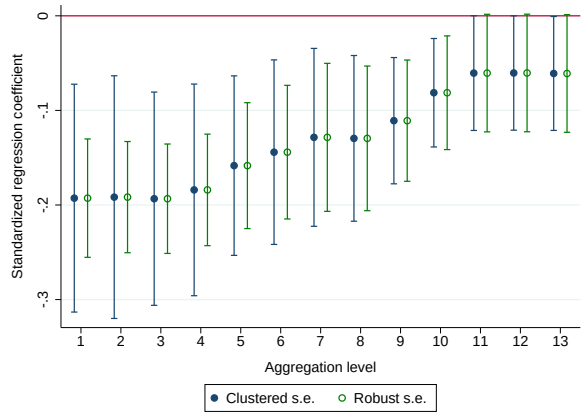


Figure D.2: Relationship between various ELF and ELP indices.

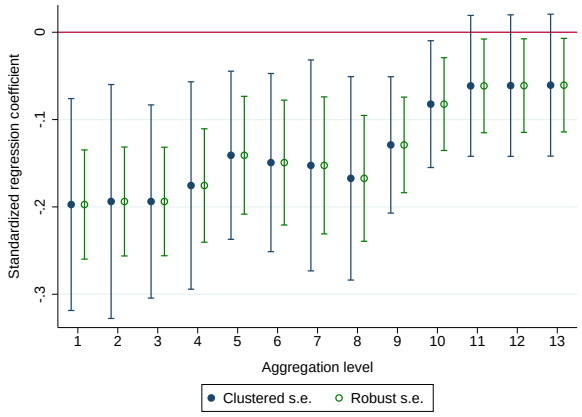
Table D.1: Pairwise correlation coefficients for various diversity indices

	ELF(1)	ELF(3)	ELF(5)	ELF(7)	ELF(9)	ELF(13)	ELP(1)	ELP(3)	ELP(5)	ELP(7)	ELP(9)	ELP(13)	$ELF_{0.05}$	$ELF_{0.5}$	$ELP_{0.05}$	$ELP_{0.5}$
ELF(1)	1.00															
ELF(3)	0.72	1.00														
ELF(5)	0.54	0.85	1.00													
ELF(7)	0.41	0.74	0.91	1.00												
ELF(9)	0.34	0.62	0.78	0.85	1.00											
ELF(13)	0.28	0.44	0.55	0.60	0.83	1.00										
ELP(1)	0.99	0.71	0.53	0.41	0.34	0.27	1.00									
ELP(3)	0.61	0.95	0.83	0.74	0.61	0.42	0.61	1.00								
ELP(5)	0.44	0.75	0.92	0.86	0.70	0.47	0.45	0.81	1.00							
ELP(7)	0.38	0.63	0.75	0.86	0.69	0.44	0.38	0.69	0.87	1.00						
ELP(9)	0.25	0.41	0.48	0.56	0.77	0.65	0.25	0.45	0.58	0.70	1.00					
ELP(13)	0.18	0.30	0.33	0.39	0.49	0.63	0.18	0.33	0.42	0.53	0.73	1.00				
$ELF_{0.05}$	0.99	0.78	0.63	0.51	0.44	0.35	0.98	0.68	0.53	0.46	0.31	0.23	1.00			
$ELF_{0.5}$	0.72	0.91	0.92	0.88	0.84	0.70	0.71	0.86	0.83	0.74	0.59	0.45	0.80	1.00		
$ELP_{0.05}$	0.92	0.68	0.50	0.38	0.28	0.16	0.92	0.63	0.51	0.46	0.32	0.26	0.91	0.65	1.00	
$ELP_{0.5}$	0.68	0.81	0.75	0.69	0.57	0.37	0.68	0.82	0.82	0.81	0.65	0.59	0.73	0.81	0.80	1.00

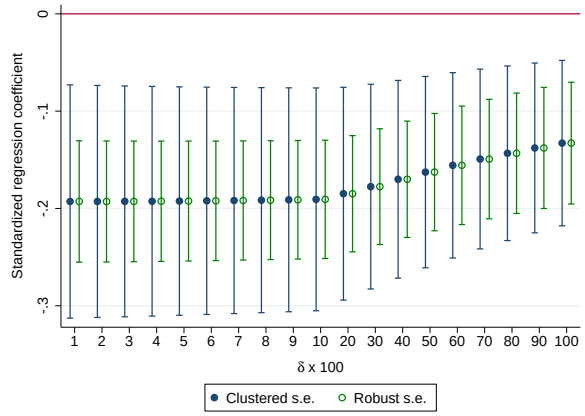
E Results with clustered standard errors



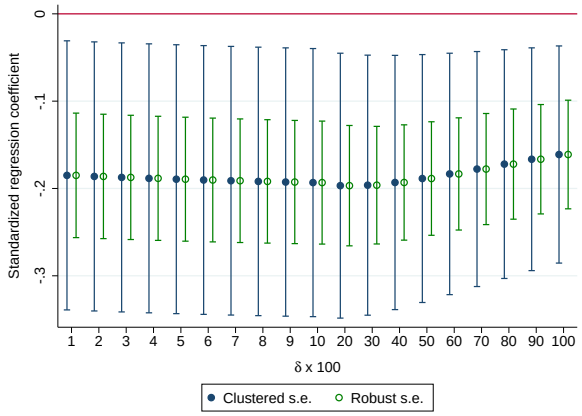
(a) Literacy rate and ELF(k)



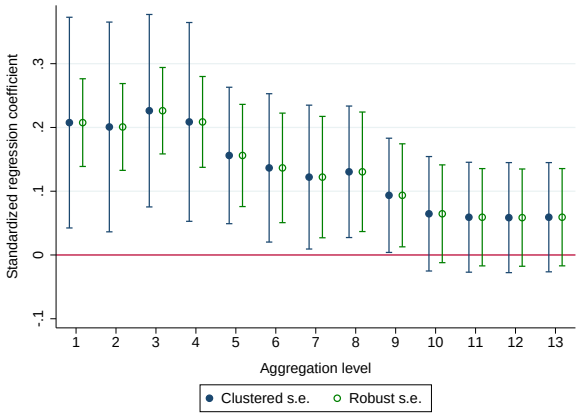
(b) Literacy rate and ELP(k)



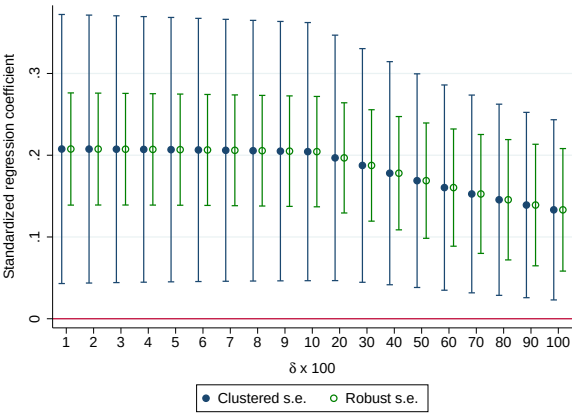
(c) Literacy rate and ELF$_\delta$



(d) Literacy rate and ELP$_\delta$



(e) Home births and ELF(k)

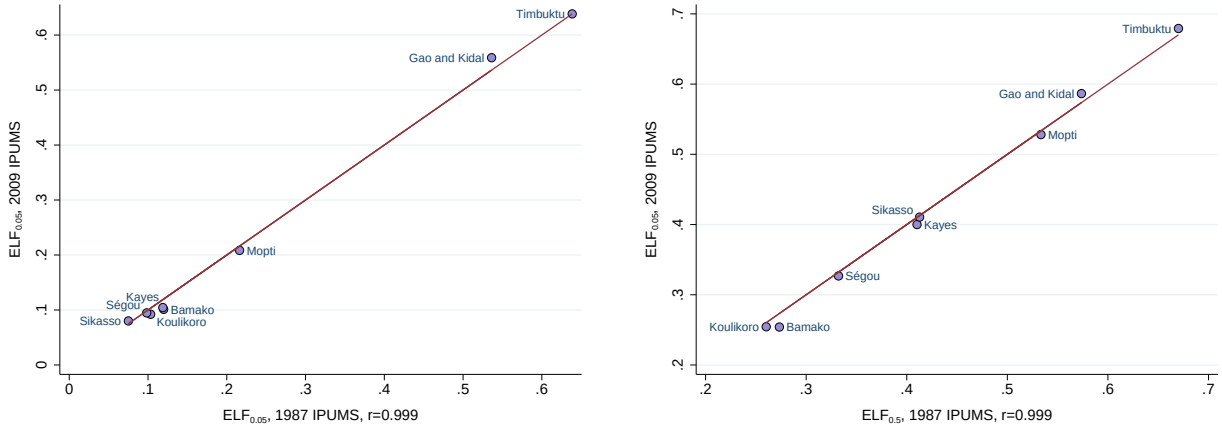


(f) Home births and ELF$_\delta$

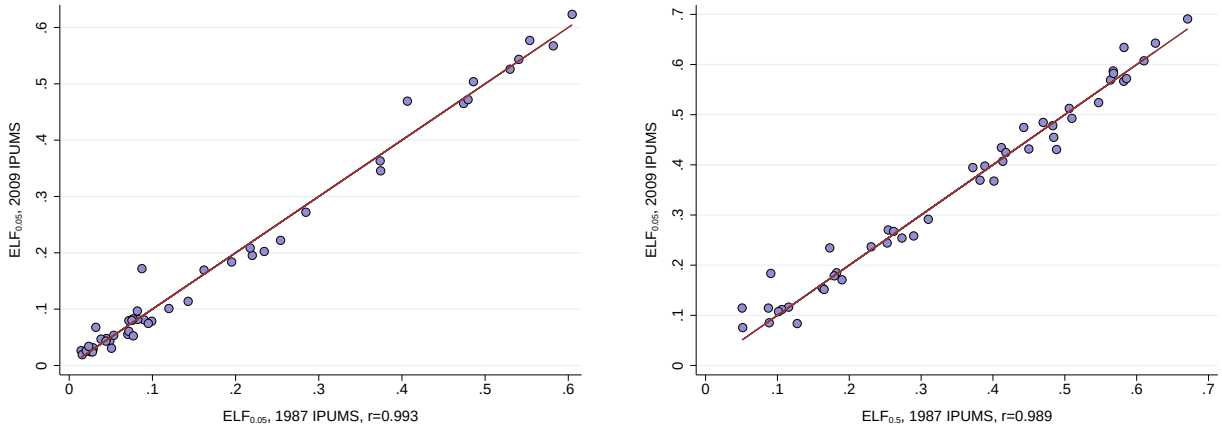
Figure E.1: Regression estimates with robust and clustered standard errors.

Notes. This figure compares the 95% confidence intervals based on robust and clustered (by country) standard errors for selected specifications from Figures 8 and 11.

F Persistence of subnational diversity: ELF_δ indices



(a) First-level subnational regions



(b) Second-level subnational regions

Figure F.1: Persistence of regional diversity in Mali, 1987–2009.

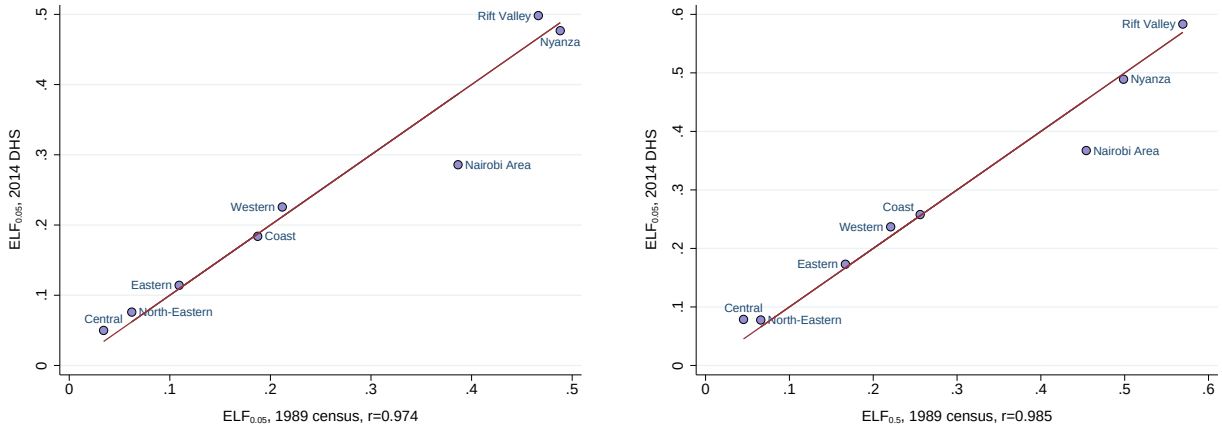


Figure F.2: Persistence of regional diversity in Kenya, 1989–2014.

G Description of variables

Diversity variables

Construction of all diversity indices is described in detail in section 2. The sources of original data on regional ethnolinguistic and religious composition are listed in Table A.2.

Outcome variables

The primary sources for all outcome variables except nighttime lights and GRP per capita are indicated in Table A.2.

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.

Net secondary school attendance ratio. The share of children of official secondary school age attending secondary school.

Share of home births. The share of births delivered at home rather than at a specialized medical facility.

Child malnutrition. Share of moderately underweight children under age 5. A child is considered underweight if her weight-for-age score is two standard deviations below the median value in the reference population.

Electricity access. Share of region's households that have access to electricity.

Nighttime lights. Data on luminosity come from the Defense Meteorological Satellite Program's Operational Linescan System that reports stable images of Earth at night captured between 20:00 and 21:30. The measure ranges from 0 to 63 and is available for cells at 30 arc-second resolution, see Henderson et al. (2012) for technical details. We calculate average luminosity for each region in 2010 and 2011 and then average across these two years. *Source:* `http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html`.

Gross regional product per capita. GRP per capita in 2005 measured in current PPP dollars. Most data come from Gennaioli et al. (2013), with Mitton (2016) serving as a source for Gambia, Mozambique, and Zimbabwe. The data for Nigeria come from the Canback Global Income Distribution Database (C-GIDD).

International wealth index (IWI). Wealth index, as proposed by Smits and Steendijk (2015), averaged across households within relevant regions. The original household surveys used for the construction of IWI in each country are listed in Table A.2. *Source*: `https://globaldatalab.org/iwi` and personal communication with Jeroen Smits.

Control variables

Absolute latitude. Absolute latitude of region's centroid. *Source:* own calculations.

Area. Surface area of the region in square kilometers. *Source:* own calculations.

Mean suitability of land for agriculture. Index of land suitability for rain-fed agriculture (maximizing technology mix). Coded on the scale from 1 (very high suitability) to 8 (not suitable) for cells at 5 arc-minute resolution. The variable used in the analysis is the average value of the suitability index across cells in each region. *Source:* FAO GAEZ dataset (plate 46) downloaded at `http://webarchive.iiasa.ac.at/Research/LUC/GAEZ/index.htm` and own calculations.

Spatial variability of land suitability for agriculture. Based on the same underlying data as the mean suitability index. Calculated as the standard deviation of cell values for each region.

Distance to capital city. Great circle distance to the country's capital city from the region's centroid measured in kilometers. *Source*: own calculations.

Capital city indicator. A dummy variable equal to one, if the region contains the country's capital city, and zero, otherwise. *Source*: own calculations.

Landlocked indicator. A dummy variable equal to one, if the region is landlocked, and zero, otherwise. *Source*: own calculations.

Ruggedness index. Index of terrain ruggedness as constructed by Nunn and Puga (2012) for cells at 30 arc-second resolution. The variable used in the analysis is the average value of the index across cells in each region. *Source:* `http://diegopuga.org/data/rugged/#grid`.

Urbanization rate. Share of region's households that live in urban areas. *Source*: see the "PG outcomes" column in Table A.2. For Equatorial Guinea, regional urbanization data come from the 2015 census report.

References

Alesina, Alberto and Ekaterina Zhuravskaya, "Segregation and the Quality of Government in a Cross Section of Countries," *American Economic Review*, August 2011, *101* (5), 1872–1911.

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

___ **, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg**, "Fractionalization," *Journal of Economic Growth*, June 2003, *8* (2), 155–194.

___ **, Caterina Genaiolli, and Stefania Lovo**, "Public Goods and Ethnic Diversity: Evidence from Deforestation in Indonesia," December 2015. Working Paper, Department of Economics, Harvard University.

___ **, Reza Baqir, and William Easterly**, "Public Goods and Ethnic Divisions," *Quarterly Journal of Economics*, November 1999, *114* (4), 1243–1284.

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

Algan, Yann, Camille Hémet, and David D. Laitin, "The Social Effects of Ethnic Diversity at the Local Level: A Natural Experiment with Exogenous Residential Allocation," *Journal of Political Economy*, June 2016, *124* (3), 696–733.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, February 2005, *113* (1), 151–184.

Ashraf, Quamrul and Oded Galor, "The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development," *American Economic Review*, February 2013, *103* (1), 1–46.

Baldwin, Kate and John D. Huber, "Economic versus Cultural Differences: Forms of Ethnic Diversity and Public Goods Provision," *American Political Science Review*, November 2010, *104*, 644–662.

Beugelsdijk, Sjoerd, Mariko Klasing, and Petros Milionis, "Value Diversity and Regional Economic Development," *Scandinavian Journal of Economics*, 2018, *forthcoming.*

Bjørnskov, Christian, "Determinants of Generalized Trust: A Cross-Country Comparison," *Public Choice*, January 2007, *130* (1), 1–21.

Collier, Paul, "Ethnicity, Politics and Economic Performance," *Economics & Politics*, November 2000, *12* (3), 225–245.

Conley, Timothy G., "GMM estimation with cross sectional dependence," *Journal of Econometrics*, September 1999, *92* (1), 1–45.

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.

Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg, "The Political Economy of Linguistic Cleavages," *Journal of Development Economics*, March 2012, *97* (2), 322–338.

___ , ___ , and ___ , "Culture, Ethnicity and Diversity," *American Economic Review*, September 2017, *107* (9), 2479–2513.

___ , Joseph Gomes, and Ignacio Ortuño-Ortín, "The Geography of Linguistic Diversity and the Provision of Public Goods," December 2016. CEPR Discussion Paper 11683.

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

Easterly, William and Ross Levine, "Africa's Growth Tragedy: Policies and Ethnic Divisions," *Quarterly Journal of Economics*, November 1997, *112* (4), 1203–1250.

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

___ and ___ , "Linking Conflict to Inequality and Polarization," *American Economic Review*, June 2011, *101* (4), 1345–74.

___ , **Laura Mayoral, and Debraj Ray**, "Ethnicity and Conflict: An Empirical Study," *American Economic Review*, June 2012, *102* (4), 1310–1342.

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*, February 2003, *97* (1), 75–90.

Gennaioli, Nicola, Rafael La Porta, Florencio Lopez de Silanes, and Andrei Shleifer, "Human Capital and Regional Development," *Quarterly Journal of Economics*, February 2013, *128* (1), 105–164.

Gerring, John, Strom C. Thacker, Yuan Lu, and Wei Huang, "Does Diversity Impair Human Development? A Multi-Level Test of the Diversity Debit Hypothesis," *World Development*, February 2015, *66*, 166–188.

Gershman, Boris and Diego Rivera, "Ethnicity, Religion, and Conflict: Evidence from African Regions," May 2017. Working Paper, American University.

___ **and** ___ , "Measuring Regional Ethnolinguistic Diversity in Sub-Saharan Africa: Surveys vs. GIS," January 2018. Working Paper, American University.

Ginsburgh, Victor and Shlomo Weber, "Linguistic Distances and Ethnolinguistic Fractionalization and Disenfranchisement Indices," in Victor Ginsburgh and Shlomo Weber, eds., *The Palgrave Handbook of Economics and Language*, Palgrave Macmillan UK, 2016, chapter 5, pp. 137–173.

Gisselquist, Rachel M., Stefan Leiderer, and Miguel Niño-Zarazúa, "Ethnic Heterogeneity and Public Goods Provision in Zambia: Evidence of a Subnational "Diversity Dividend"," *World Development*, February 2016, *78*, 308–323.

Glennerster, Rachel, Edward Miguel, and Alexander D. Rothenberg, "Collective Action in Diverse Sierra Leone Communities," *Economic Journal*, May 2013, *123* (568), 285–316.

Greenberg, Joseph H., "The Measurement of Linguistic Diversity," *Language*, January–March 1956, *32* (1), 109–115.

Habyarimana, James, Macartan Humphreys, Daniel N. Posner, and Jeremy M. Weinstein, "Why Does Ethnic Diversity Undermine Public Goods Provision?," *American Political Science Review*, November 2007, *101*, 709–725.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil, "Measuring Economic Growth from Outer Space," *American Economic Review*, April 2012, *102* (2), 994–1028.

Hodler, Roland and Paul A. Raschky, "Regional Favoritism," *Quarterly Journal of Economics*, May 2014, *129* (2), 995–1033.

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

Jerven, Morton, *Poor Numbers: How We Are Misled by African Development Statistics and What to Do about It*, Ithaca, NY: Cornell University Press, 2013.

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

La Porta, Rafael, Florencio Lopez de Silanes, Andrei Shleifer, and Robert Vishny, "The Quality of Government," *Journal of Law, Economics, and Organization*, March 1999, *15* (1), 222–279.

Laitin, David D., "What Is a Language Community?," *American Journal of Political Science*, January 2000, *44* (1), 142–155.

Michalopoulos, Stelios, "The Origins of Ethnolinguistic Diversity," *American Economic Review*, June 2012, *102* (4), 1508–1539.

___ **and Elias Papaioannou**, "Pre-Colonial Ethnic Institutions and Contemporary African Development," *Econometrica*, January 2013, *81* (1), 113–152.

___ **and** ___ , "National Institutions and Subnational Development in Africa," *Quarterly Journal of Economics*, February 2014, *129* (1), 151–213.

Miguel, Edward, "Tribe or Nation? Nation Building and Public Goods in Kenya versus Tanzania," *World Politics*, April 2004, *56* (3), 327–362.

___ **and Mary Kay Gugerty**, "Ethnic Diversity, Social Sanctions, and Public Goods in Kenya," *Journal of Public Economics*, December 2005, *89* (11–12), 2325–2368.

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

Montagu, Dominic, Gavin Yamey, Adam Visconti, April Harding, and Joanne Yoong, "Where Do Poor Women in Developing Countries Give Birth? A Multi-Country Analysis of Demographic and Health Survey Data," *PLoS ONE*, February 2011, *6* (2), 1–8.

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

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

___ **and Leonard Wantchekon**, "The Slave Trade and the Origins of Mistrust in Africa," *American Economic Review*, December 2011, *101* (7), 3221–3252.

Oster, Emily, "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 2018, *forthcoming.*

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

Reynal-Querol, Marta, "Ethnicity, Political Systems, and Civil Wars," *Journal of Conflict Resolution*, February 2002, *46* (1), 29–54.

Robinson, Amanda Lea, "Ethnic Diversity, Segregation, and Ethnocentric Trust in Africa," *British Journal of Political Science*, 2018, *forthcoming.*

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

Spolaore, Enrico and Romain Wacziarg, "Ancestry, Language and Culture," in Victor Ginsburgh and Shlomo Weber, eds., *The Palgrave Handbook of Economics and Language*, Palgrave Macmillan UK, 2016, chapter 6, pp. 174–211.

___ **and** ___ , "War and Relatedness," *Review of Economics and Statistics*, December 2016, *98* (5), 925–939.

Young, Alwyn, "The African Growth Miracle," *Journal of Political Economy*, August 2012, *120* (4), 696–739.