Measuring Regional Ethnolinguistic Diversity in Sub-Saharan Africa: Surveys vs. GIS

Boris Gershman[*]
American University

Diego Rivera
American University

January 2018

Abstract

This paper compares two approaches to measuring subnational ethnolinguistic diversity in Sub-Saharan Africa, one based on censuses and large-scale population surveys and the other relying on the use of geographic information systems (GIS). The two approaches yield sets of regional fractionalization indices that are moderately positively correlated, with a stronger association across rural areas. These differences matter for empirical analysis: in a common sample of regions, survey-based indices of deep-rooted diversity are more strongly negatively associated with a range of development indicators relative to their highest-quality GIS-based counterparts.

Keywords: African development, ethnolinguistic diversity, GIS, subnational analysis

JEL Classification Numbers: O10, O15, Z13

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

1 Introduction

Ethnolinguistic diversity has long been viewed as an important determinant of various development outcomes including income per capita, public goods provision, quality of governance, and violent conflict. Given the extraordinary diversity of its population, the political salience of ethnicity, and the relatively low rates of economic growth, the African continent occupies a special place in this line of research.

Naturally, reliable data are the key ingredient in any empirical study. While there are several datasets on ethnic and linguistic diversity commonly used in cross-country analyses, there are no similarly established sources at the subnational level. This paper examines two approaches to constructing such a dataset for a large sample of regions within Sub-Saharan African countries. First, we briefly discuss the contribution of Gershman and Rivera (2018) who employ censuses and large-scale household surveys to measure subnational ethnolinguistic diversity. Second, we show how the GIS approach, based on combining digital maps of ethnolinguistic groups with disaggregated population data, may be used to accomplish the same goal. We next compare the sets of fractionalization indices resulting from the two strategies, discuss the likely sources of measurement error, and demonstrate that the choice of approach matters for the empirical analysis of the relationship between regional diversity and development outcomes.

2 The survey-based approach

A standard approach to constructing any dataset on regional sociodemographic characteristics is to compile the information from relevant population surveys. In Gershman and Rivera (2018), we introduce and describe the product of a comprehensive effort to compile the data on subnational ethnolinguistic composition of Sub-Saharan Africa based on this approach. Our dataset contains 36 countries, almost 400 first-level administrative units, and 750 unique ethnolinguistic groups. For each country, we picked the best available data source according to the following main criteria: breadth of coverage and regional representativeness, the number of listed well-defined ethnolinguistic groups, and the population share of unidentified "other" ethnicities. As a result, about half of the dataset is based on national censuses while the other half relies on large-scale household surveys, mostly various waves of the Demographic and Health Surveys and Multiple Indicator Cluster Surveys.

Defining ethnicity is a notoriously difficult task. We standardize this notion by linking ethnicities to the corresponding unique spoken languages, which is rather straightforward in the context of Sub-Saharan Africa. Specifically, we match all ethnic groups in the original survey data to the *Ethnologue* database, a comprehensive compendium of worlds languages. Thus, the basic elements of our dataset are ethnolinguistic groups as captured by *Ethnologue*'s three-letter coding system.

The matching procedure has an additional benefit of incorporating original groups into *Ethnologue*'s linguistic family tree model which allows to construct diversity metrics for coarser language divisions, as in the country-level analysis of Desmet et al. (2012). This enables the exploration of diversity based on both recent and deeper ethnolinguistic cleavages since upper tiers of the linguistic tree roughly correspond to the distant hypothetical common ancestors of modern languages. There are thirteen possible levels of linguistic aggregation in our sample, where level 13 represents the original groups, 750 in total, and level 1 corresponds to the six major families to which all languages listed in the dataset belong. After properly aggregating the relevant regional population shares, fractionalization indices can be readily calculated for each of the thirteen levels. We denote them as ELF(k), where $k = 1, \ldots, 13$ is the level of linguistic aggregation and ELF(13) is the standard index based on the most disaggregated survey data.

3 The GIS approach

An alternative approach to measuring subnational diversity relies on a straightforward application of GIS. The idea is to combine one of the available digital maps of ethnolinguistic groups with a high-resolution population grid in order to reconstruct the ethnolinguistic composition of any given geographic area and calculate corresponding diversity measures. Two basic conditions must hold for this approach to yield reliable metrics: first, the digital maps must accurately represent the actual territorial boundaries within which various groups reside and, second, the population grid must provide accurate counts of people located in each cell of the map. Neither of these conditions is expected to hold perfectly since both key inputs of the exercise are only approximations. However, given the increasing popularity of GIS applications in social sciences and the simplicity of implementing this approach, it is instructive to compare its outcomes to those of the survey-based method in Gershman and Rivera (2018).

We apply the GIS approach to three different digital maps of ethnolinguistic groups actively used in recent research. The first one originates from Murdock (1959) and is meant to roughly capture the precolonial boundaries of ethnic groups in Africa. The second database, commonly referred to as GREG (Geo-Referencing of Ethnic Groups), is the digitized version of *Atlas Narodov Mira*, a global map of ethnic "homelands" put together by Soviet ethnographers in 1960s (Weidmann et al., 2010). Finally, the third and by far the most detailed available source is the *World Language Mapping System* (WLMS) which depicts "traditional linguistic homelands" for each entry in the *Ethnologue* database (16th edition in our analysis). Unlike the Murdock map, both GREG and WLMS contain multiple areas populated by two or more groups. In the absence of information on the relative population shares of coexisting groups, we assume that in all such cases the total population of a mixed area is equally split among relevant groups.

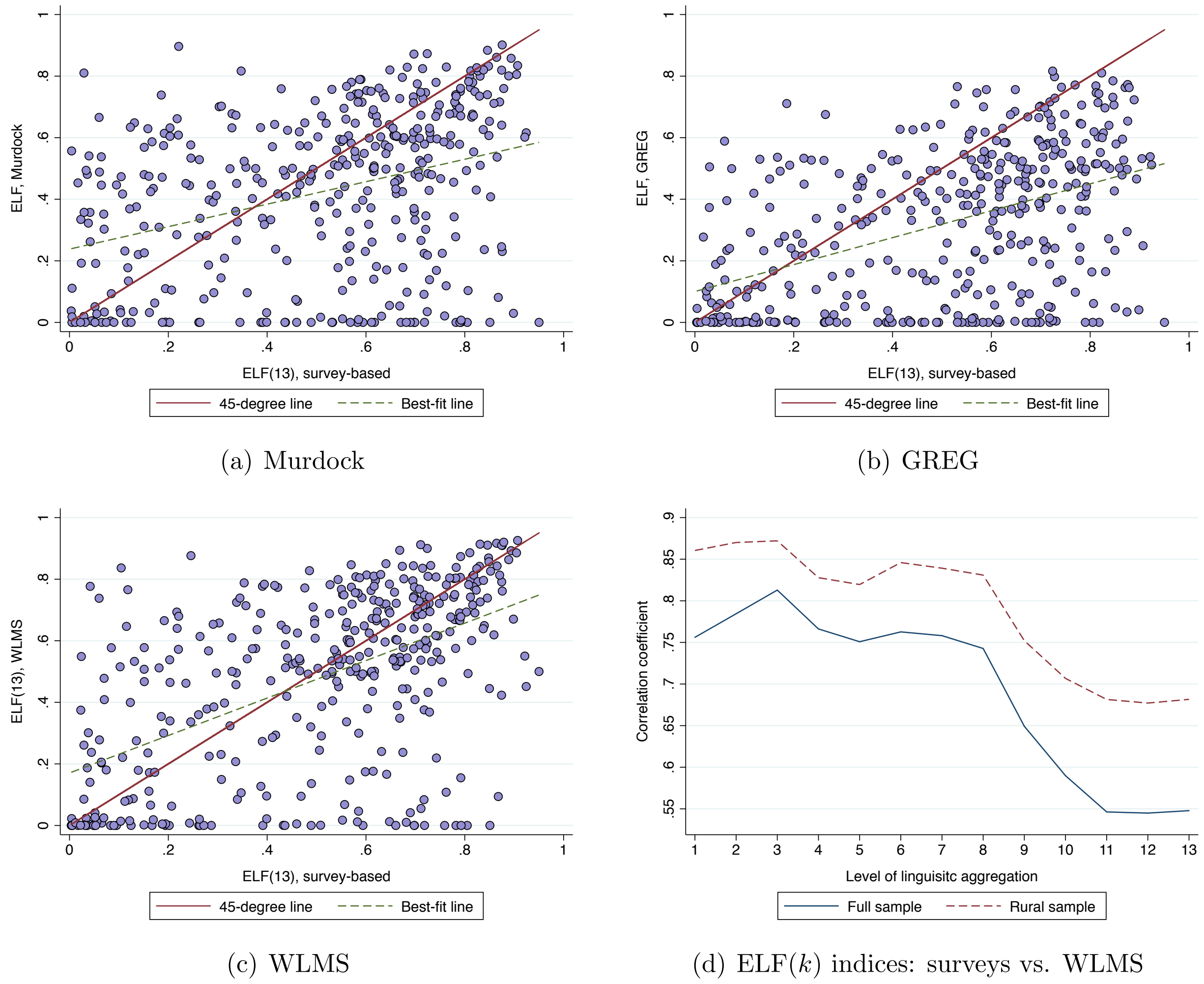
To determine the population counts for ethnolinguistic groups in each region we use LandScan 2013, a high-resolution raster database developed by Oak Ridge National Laboratory. The final population grid composed of 30 arc-second cells (approximately 1 square kilometer near the equator) is produced by disaggregating census data within administrative boundaries via a "smart interpolation" technique that employs spatial data and imagery analysis. For consistency, we use the same set of administrative boundaries as in Gershman and Rivera (2018).

4 Comparison of subnational ELF indices

We start by directly comparing the regional ELF indices generated using the GIS approach to the ELF(13) index from our survey-based dataset. Table 1 and Figure 1 show that the match between various indices is far from perfect. The metrics based on the Murdock, GREG, and WLMS maps are all positively associated with the survey-based index, with rather low correlation coefficients of 0.35 and 0.46 in the former two cases and a higher value of 0.55 in the latter.

Table 1: Pairwise correlations between regional ELF indices

	Full sample (396 regions)				Rural sample (301 regions)			
	Murdock	GREG	WLMS	Surveys	Murdock	GREG	WLMS	Surveys
Murdock	1.00				1.00			
GREG	0.58	1.00			0.55	1.00		
WLMS	0.62	0.54	1.00		0.61	0.52	1.00	
Surveys	0.35	0.46	0.55	1.00	0.51	0.52	0.69	1.00



(a) Murdock

(b) GREG

(c) WLMS

(d) ELF(k) indices: surveys vs. WLMS

Figure 1: Regional ELF indices: surveys vs. GIS

We start by directly comparing the regional ELF indices generated using the GIS approach to the ELF(13) index in the high-quality subsample of our survey-based dataset (that is, after excluding the regions with information on ethnicity missing for more than 20% of the population). Table 1 and Figure 1 show that the match between various indices is far from perfect. The metrics based on the Murdock, GREG, and WLMS maps are all positively associated with the survey-based index, with rather low correlation coefficients of 0.35 and 0.46 in the former two cases and a higher value of 0.55 in the latter.

There are multiple sources of measurement error inherent in the GIS approach. Setting aside the noise in disaggregated population data introduced by the LandScan interpolation technique and the arbitrary equal split of the population in mixed areas, important distortions stem from inaccuracies in spatial representation of ethnolinguistic groups on the available maps. First, they depict territories "traditionally" inhabited by certain groups and thus cannot capture any migration-driven changes in population composition and, most importantly, higher diversity of urban areas. In extreme cases, when a city or an urban area constitute a region of their own (e.g., Nairobi, Addis Ababa, Brazzaville), the GIS approach almost surely yields an ELF index close or equal to zero. Not surprisingly, as the right panel of Table 1 reveals, in the "rural" sample, limited to regions with urbanization rates below 50%, the correlation between GIS-based indices and the survey benchmark becomes stronger reaching 0.69 in the case of WLMS.

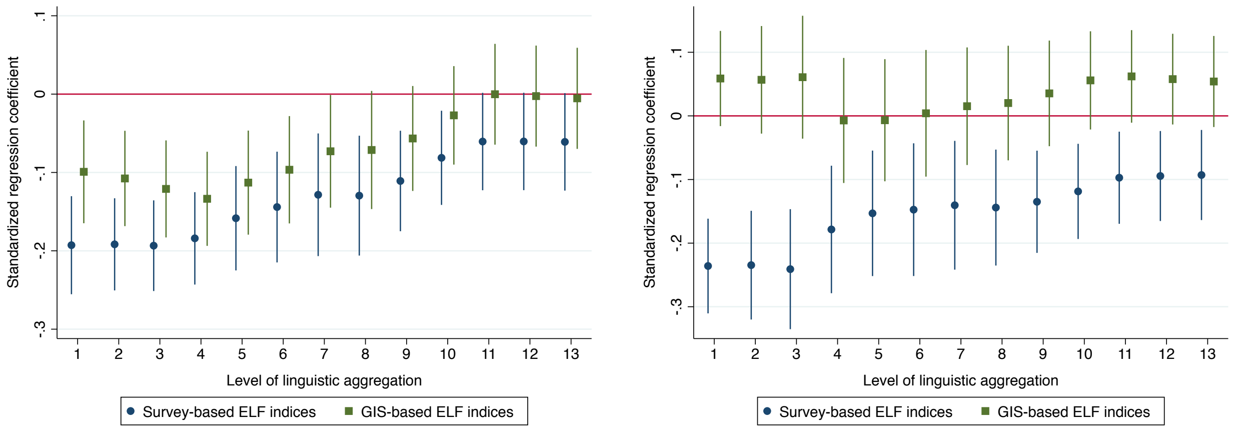
The second important issue is the level of detail and other properties of the chosen classification of ethnolinguistic groups. Panels (a) and (b) of Figure 1 demonstrate that the metrics based on Murdock and GREG maps, especially the latter, tend to underestimate diversity relative to the survey-based index. In fact, GREG has only 191 groups, some of them clearly representing a whole family of ethnicities or languages. For example, according to this source, the Maryland county of Liberia is populated entirely by the "Bantu-speaking Pygmy tribes," whereas the 2008 census lists 17 distinct groups for the same region. On the other hand, WLMS is by far the most detailed and standardized database, with over 1700 languages in our sample of regions (many of them spoken by very few people). Unfortunately, this level of detail, in combination with outdated or imprecise boundaries, does not always improve the accuracy of diversity estimates. To give an extreme, but not unique, example, according to *Ethnologue*, as of the year 2000, the Vili language had a total of 7310 speakers in the Republic of the Congo. Yet, its "traditional" homeland happens to intersect the urban region of Pointe-Noire containing the countrys second largest city. As a result, the GIS approach yields a count of more than 0.92 *million* Vili speakers. In general, inaccurate boundaries on ethnolinguistic maps may result in estimates of contemporary regional diversity that are very different from the best available survey-based measures. Unfortunately, existing documentation lacks in detail and transparency making it hard to identify all of the likely weak spots of the original maps.

Due to their direct link to *Ethnologue*'s classification system, the WLMS languages are readily comparable to those in Gershman and Rivera (2018) and can be similarly aggregated at the more basic levels of the linguistic family tree, as discussed above. Interestingly, as shown in panel (d) of Figure 2, the survey- and WLMS-based ELF indices are much more strongly connected at higher levels of linguistic aggregation, with the correlation coefficient hovering around 0.85 for levels 18 in the rural sample. This is expected: as languages are aggregated into more basic families, the scope for measurement error is diminished since neighboring groups are likely to be merged together under the umbrella of their common ancestors.

5 Data matter: regional diversity and development

In Gershman and Rivera (2018), we show that deep-rooted diversity, based on linguistic cleavages formed in the distant past, is a much stronger predictor of contemporary local public goods provision in Sub-Saharan Africa. To establish this pattern, we regress a variety of indicators, including metrics of education, health, and electricity access, on the ELF(k) indices (one at a time), a battery of regional characteristics, and country fixed effects. Our findings for literacy rate, largely representative of the results for other outcome variables, are reproduced in panel (a) of Figure 2. It shows the standardized coefficient estimates on ELF(k) indices from thirteen regressions, along with respective robust 95% confidence intervals (round markers with spikes). It is clear that both economic and statistical significance increase for ELF indices calculated at higher levels of linguistic aggregation.

Here, we repeat the same exercise but substitute the original survey-based ELF indices with the ones constructed using WLMS, the highest-quality GIS metrics. As can be seen from comparing the new estimates (square markers with spikes) to the baseline, despite the similarity of the overall pattern, the coefficient estimates are substantially lower in magnitude and less statistically significant for the GIS-based indices of deep-rooted diversity. This is consistent with the effect of attenuation bias due to measurement error inherent in the GIS approach. In fact, when pairs of indices are included in "horse-race" type of regressions, the coefficient estimates for relevant survey-based measures remain very strong whereas their GIS counterparts decrease in absolute value and lose statistical significance, as shown in panel (b) of Figure 2.



(a) Regional literacy and ELF(k)

(b) "Horse-race" regressions

Figure 2: Deep-rooted diversity and regional literacy rate

6 Concluding remarks

While the GIS approach to measuring subnational diversity is straightforward to implement and is universally applicable, one has to be mindful of the biases associated with this method. Our analysis compares its outcomes to the high-quality survey-based dataset on Sub-Saharan Africa and shows that the distortions introduced by the GIS approach are smaller for rural regions. Furthermore, the measurement error appears to be less severe for ELF indices based on the WLMS map of languages, especially when the latter are aggregated into more basic language families to capture deep-rooted diversity.

References

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

Gershman, Boris and Diego Rivera, "Subnational Diversity in Sub-Saharan Africa: Insights from a New Dataset," *Journal of Development Economics*, 2018, *forthcoming.*

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

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