Journal of Development Economics 133 (2018) 231–263

Contents lists available at ScienceDirect

Journal of Development Economics

journal homepage: www.elsevier.com/locate/devec

Subnational diversity in Sub-Saharan Africa: Insights from a new dataset☆

Boris Gershman *, Diego Rivera

Department of Economics, American University, USA

ARTICLE INFO

JEL classification numbers:
H41
O10
O15
Z12
Z13

Keywords:
African development
Ethnolinguistic diversity
Public goods provision
Religious diversity
Subnational analysis

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.

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

☆ 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, USA.
E-mail address: boris.gershman@american.edu (B. Gershman).

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

https://doi.org/10.1016/j.jdeveco.2018.01.003
Received 8 June 2017; Received in revised form 10 December 2017; Accepted 12 January 2018
Available online 31 January 2018
0304-3878/

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.

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 economic 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

[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.

[4] See Easterly and Levine (1997), La Porta et al. (1999), Collier (2000), Alesina et al. (2003), Alesina and La Ferrara (2005), and 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.