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

Boris Gershman    Diego Rivera

American University

ABCDE 2016: Data and Development Economics
The World Bank
June 20, 2016
Motivation

- Debates on ethnic diversity and development
  - Theoretically ambiguous relationship
  - Easterly and Levine (1997): source of Africa’s “growth tragedy”
  - Vibrant cross-country literature
Motivation

- Debates on ethnic diversity and development
  - Theoretically ambiguous relationship
  - Easterly and Levine (1997): source of Africa’s “growth tragedy”
  - Vibrant cross-country literature
- Research shifts to the subnational level
  - Inconclusive evidence
  - Miguel and Gugerty (2005) on villages in Kenya
  - Glennerster et al. (2013) on chiefdoms in Sierra Leone
  - Giselquist et al. (2016) on districts in Zambia

- Recurring measurement issues
  - Availability and quality of data on ethnolinguistic composition
  - Choice of appropriate diversity metrics
Motivation

- Debates on ethnic diversity and development
  - Theoretically ambiguous relationship
  - Easterly and Levine (1997): source of Africa’s “growth tragedy”
  - Vibrant cross-country literature

- Research shifts to the subnational level
  - Inconclusive evidence
  - Miguel and Gugerty (2005) on villages in Kenya
  - Glennerster et al. (2013) on chiefdoms in Sierra Leone
  - Gisselquist et al. (2016) on districts in Zambia

- Recurring measurement issues
  - Availability and quality of data on ethnolinguistic composition
  - Choice of appropriate diversity metrics
Contribution

- New high-quality dataset on subnational diversity in SSA
  - 36 countries, 398 regions, 732 unique ethnolinguistic groups
  - Regions are official first-level subnational administrative units
  - Based on censuses and regionally representative surveys
Contribution

- New high-quality dataset on subnational diversity in SSA
  - 36 countries, 398 regions, 732 unique ethnolinguistic groups
  - Regions are official first-level subnational administrative units
  - Based on censuses and regionally representative surveys
- New subnational diversity metrics
  - Account for linguistic relatedness between groups
Contribution

- New high-quality dataset on subnational diversity in SSA
  - 36 countries, 398 regions, 732 unique ethnolinguistic groups
  - Regions are official first-level subnational administrative units
  - Based on censuses and regionally representative surveys
- New subnational diversity metrics
  - Account for linguistic relatedness between groups
- Subnational analysis of diversity and development outcomes
  - Education, health, electricity access, nighttime lights
  - Standard diversity indices mostly irrelevant
  - Indices adjusted for linguistic similarity negative and significant
  - Indices of deep-rooted diversity negative and significant
The New Dataset: Sources

- Over 150 surveys and reports reviewed
- Selection criteria
  - Best regional coverage (representativeness)
  - Detailed list of well-defined ethnolinguistic groups
  - Small share of unidentified “other” ethnicities
The New Dataset: Sources

- Over 150 surveys and reports reviewed
- Selection criteria
  - Best regional coverage (representativeness)
  - Detailed list of well-defined ethnolinguistic groups
  - Small share of unidentified “other” ethnicities
- Almost 50% of the sample covered by census data
- The rest based on DHS, MICS, or similar surveys
- No reliable subnational data for certain countries
  - Most notably, DRC, Sudan, South Sudan
The New Dataset: Metrics

- Consider a region populated by $N$ ethnolinguistic groups
- Let $s_i$ be the regional population share of group $i$
- Standard fractionalization and polarization indices

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

- Problem: all groups are treated as equally distinct
The New Dataset: Metrics

- Consider a region populated by $N$ ethnolinguistic groups
- Let $s_i$ be the regional population share of group $i$
- Standard fractionalization and polarization indices
  
  \[
  \text{ELF} = \sum_{i=1}^{N} s_i(1 - s_i) \quad \text{ELP} = 4 \sum_{i=1}^{N} s_i^2(1 - s_i)
  \]
- Problem: all groups are treated as equally distinct
- Take into account linguistic relatedness
  - Greenberg (1956), Fearon (2003), Desmet et al. (2009)
Example: Gash-Barka Region of Eritrea

- Raw data from DHS
  - Seven distinct ethnic groups in the region
    - Bilen (0.004), Hedareb (0.084), Kunama (0.087), Nara (0.116), Saho (0.011), Tigre (0.363), and Tigrinya (0.335)
  - ELF = 0.728
Example: Gash-Barka Region of Eritrea

- Raw data from DHS
  - Seven distinct ethnic groups in the region
    - Bilen (0.004), Hedareb (0.084), Kunama (0.087), Nara (0.116), Saho (0.011), Tigre (0.363), and Tigrinya (0.335)
  - ELF = 0.728

- Take into account linguistic relatedness between groups
  - Match them to corresponding Ethnologue codes
  - Look at their relative locations on the linguistic tree
Example: Gash-Barka Region of Eritrea

“Proto-Human”

Afro-Asiatic

Cushitic

North

East

Central

Saho-Afar

Nada-Afar

Northern

Semitic

South

Ethiopian

E. Sudanic

Eastern

Nara

Kunama

Nilo-Saharan

North

Tigre

Tigrigna

Central

Bilen

Kunama
Distance-adjusted Diversity Measures

- Account for linguistic similarities/distances

\[
\text{ELF}_\delta = \sum_{i=1}^{N} \sum_{j=1}^{N} s_i s_j \tau_{ij}, \quad \tau_{ij} = 1 - \frac{l}{m} \delta
\]

- Weight \( \tau_{ij} \) captures linguistic distance between groups \( i \) and \( j \)
  
  - \( l \) is the number of shared branches of the linguistic tree
  - \( m = 13 \) is the maximum possible number of shared branches
  - \( \delta \) captures the importance of shared branches

Fearon (2003) picks \( \delta = 0.5 \), Desmet et al. (2009) set \( \delta = 0.05 \).
Distance-adjusted Diversity Measures

- Account for linguistic similarities/distances

\[
ELF_\delta = \sum_{i=1}^{N} \sum_{j=1}^{N} s_i s_j \tau_{ij}, \quad \tau_{ij} = 1 - (l/m)^\delta
\]

- Weight \( \tau_{ij} \) captures linguistic distance between groups \( i \) and \( j \)
  - \( l \) is the number of shared branches of the linguistic tree
  - \( m = 13 \) is the maximum possible number of shared branches
  - \( \delta \) captures the importance of shared branches

- We try different values of \( \delta \in [0.01, 1] \)
  - Fearon (2003) picks \( \delta = 0.5 \), Desmet et al. (2009) set \( \delta = 0.05 \)

- Similar adjustment for ELP
Deep and Recent Cleavages

- Desmet et al. (2012) suggested an alternative set of metrics
- Aggregate groups at different tiers of the linguistic tree
- Calculate diversity at each level of aggregation $k$
Desmet et al. (2012) suggested an alternative set of metrics

Aggregate groups at different tiers of the linguistic tree

Calculate diversity at each level of aggregation $k$

$k = 13$ corresponds to the finest definition of language

- 732 in our sample

$k = 1$ corresponds to major language families

- Just 6 in our sample: Afro-Asiatic, Niger-Congo, Nilo-Saharan, Khoisan, Creole, and Indo-European

In total, 13 ELF($k$) and 13 ELP($k$) indices
Example: Gash-Barka Region of Eritrea

- **Afro-Asiatic**
  - **Cushitic**
    - North (0.084)
    - East (0.011)
    - Central (0.004)
  - Semitic
    - South (0.698)
- **Nilo-Saharan**
  - **E. Sudanic**
    - Eastern (0.116)
  - **Kunama**
    - Kunama\(^1\) (0.087)

- **Semitic**
  - **E. Sudanic**
    - Ethiopian (0.698)
  - Kunama
    - Kunama\(^2\) (0.087)
- **Nilo-Saharan**
  - **Kunama**
    - Kunama\(^3\) (0.087)

- **Bedawiyet**
  - Bedawiyet\(^1\) (0.084)
  - Bedawiyet\(^2\) (0.084)
- **Saho-Afar**
  - Saho\(^1\) (0.011)
- **Northern**
  - Bilen\(^1\) (0.004)
- **Ethiopian**
  - North (0.698)
- **Nara**
  - Nara\(^1\) (0.011)
- **Kunama**
  - Kunama\(^3\) (0.087)
Example: Gash-Barka Region of Eritrea

- **Standard unadjusted index**
  - ELF = 0.728

- **Distance-adjusted indices**
  - ELF\(_{0.01}\) = 0.330
  - ELF\(_{0.05}\) = 0.354
  - ELF\(_{0.5}\) = 0.532
  - ELF\(_{1}\) = 0.622

- **Indices for different aggregation levels**
  - ELF(1) = 0.324
  - ELF(2) = 0.482
  - ELF(3) = ELF(4) = ELF(5) = 0.485
  - ELF(6) = ELF(7) = \cdots = ELF(13) = 0.728
Subnational Diversity: ELF(12)
Subnational Diversity: ELF(7)
Subnational Diversity: ELF(6)
Subnational Diversity: ELF(5)
Subnational Diversity: ELF(3)
Subnational Diversity: ELF(2)
Subnational Diversity: ELF(1)
Empirical Setup

- Estimating equation
  \[ y_i = \alpha_c + \beta D_i + X_i' \Gamma + \varepsilon_i \]

- Development outcomes \( y_i \) in region \( i \)
  - Education (literacy rate, school attendance)
  - Health (home births, child malnutrition)
  - Basic infrastructure (electricity access)
  - Nighttime luminosity

- Diversity measures \( D_i \)
  - 19 ELF_\delta and 13 ELF(k) indices, same for ELP

- Regional controls \( X_i \)
  - Geography, distance to capital city, urbanization rate

- Full set of country fixed effects \( \alpha_c \)
Subnational ELF and Literacy Rate

![Graph showing standardized regression coefficients across different aggregation levels. The x-axis represents aggregation levels, while the y-axis shows the standardized regression coefficients. There are error bars for each data point, indicating variability. The coefficients range from -0.25 to 0.05, with some levels showing a positive trend.]
Subnational ELF and Literacy Rate

![Graph showing standardized regression coefficient vs. x (in percentage)]
Other Outcomes and Robustness

- Qualitatively similar results for
  - School attendance, educational attainment
  - Prevalence of home births, child malnutrition
  - Access to electricity, nighttime lights (a bit less pronounced)

- Similar results for ELF and ELP indices

- Robustness
  - Exclusion of regions with large shares of “others”
  - Exclusion of regions containing capital cities
  - Rural subsample
  - Omission of geographic controls
Conclusions

- New high-quality dataset on subnational diversity in SSA
  - Wide coverage and best available data sources
  - Standardized definition of an ethnolinguistic group
  - Metrics accounting for the structure of linguistic cleavages
- New results on subnational diversity and development
  - Standard ELF and ELP indices mostly irrelevant
  - Indices adjusted for linguistic similarity negative and significant
  - Indices of deep-rooted diversity negative and significant
- Research in progress
  - Further diversity metrics (cross-cuttingness, ethnic inequality)
  - Further outcomes (conflict)
Total Number of Unique Ethnic Groups

![Graph showing the total number of unique ethnic groups against aggregation level. The graph indicates a steady increase in the number of groups as the aggregation level rises.]
The Niger-Congo Family and the Bantu

Number of ethnolinguistic groups

Niger-Congo (1)
Atlantic-Congo (2)
Volta-Congo (3)
Benue-Congo (4)
Bantoid (5)
Southern (6)
Narrow Bantu (7)
Central (8)
J (9)
ELF($k$): Descriptive Statistics
ELF($k$): Kernel Density, $k = 3, 4, 5, 7, 9, 13
ELF_δ: Descriptive Statistics
ELF$_\delta$: Kernel Density, $\delta = 0.01, 0.05, 0.1, 0.3, 0.5, 1$
ELF(1) vs. ELF$_{0.05}$
$\text{ELF}(1)$ vs. $\text{ELF}_{0.5}$
ELF(1) vs. ELF₁
ELF(5) vs. ELF_{0.05}
ELF(5) vs. ELF\textsubscript{0.5}
ELF(5) vs. ELF₁
\textbf{ELF(13) vs. ELF}_{0.05}
ELF(13) vs. ELF$_{0.5}$
ELF(13) vs. ELF₁
ELF(1) vs. ELP(1)
ELF(3) vs. ELP(3)
ELF(13) vs. ELP(13)
$\text{ELF}_{0.05}$ vs. $\text{ELP}_{0.05}$
ELF_{0.5} vs. ELP_{0.5}
$\text{ELF}_1$ vs. $\text{ELP}_1$
Subnational ELF and Literacy Rate

![Graph showing the relationship between ELF and female literacy rate. The x-axis represents ELF residual values, ranging from -2 to 3. The y-axis represents female literacy rate residual values, ranging from -2 to 2. The graph includes a scatter plot with data points and a linear trend line.](image-url)
Subnational ELF and Literacy Rate
Subnational ELF and Literacy Rate

![Graph showing the relationship between female literacy rate and ELF(2), residual. The graph includes a scatter plot with data points and a regression line.]
Subnational ELF and Literacy Rate

![Graph showing the relationship between Female literacy rate and ELF(5), residual.](image-url)
Subnational ELF and Literacy Rate

![Graph showing the relationship between female literacy rate, residual and ELF(9), residual.](image-url)
Subnational ELF and Literacy Rate

Female literacy rate, residual vs. ELF(13), residual
Subnational ELF and Net Secondary Attendance

The graph shows the standardized regression coefficient for different aggregation levels. The x-axis represents the aggregation level, while the y-axis shows the standardized regression coefficient. The coefficients range from approximately -0.25 to 0.05, with a trend indicating a decrease with increasing aggregation level.
Subnational ELF and Home Births

![Graph showing standardized regression coefficient against aggregation level.](image-url)
Subnational ELF and Home Births

![Graph showing standardized regression coefficient vs. x 100]
Subnational ELF and Underweight Children
Subnational ELF and Underweight Children
Subnational ELF and Access to Electricity

![Graph showing standardized regression coefficients at different aggregation levels. The x-axis represents aggregation level, and the y-axis represents standardized regression coefficients. The graph shows a trend where the coefficients fluctuate between -0.15 and 0.05 across different levels of aggregation.](image-url)
Subnational ELF and Access to Electricity

![Graph showing standardized regression coefficient](image-url)
Subnational ELP and Literacy Rate

![Graph showing the relationship between standardized regression coefficient and aggregation level. The x-axis represents aggregation level, and the y-axis represents the standardized regression coefficient. The graph includes error bars and a trend line.]
Subnational ELF and Nighttime Lights

The graph shows the standardized regression coefficient for different aggregation levels. The x-axis represents the aggregation level, ranging from 1 to 13. The y-axis represents the standardized regression coefficient, ranging from -0.25 to 0.

The data points indicate fluctuations in the standardized regression coefficient across the aggregation levels, with some levels showing a higher impact than others. The trend suggests a diminishing effect as the aggregation level increases.

The graph also includes error bars, indicating the variability or uncertainty in the standardized regression coefficients at each aggregation level.
Subnational ELF and Nighttime Lights
Persistence of Subnational Diversity: Mali-I

ELF(13), 2009 census

ELF(13), 1987 census, r=0.998

Bamako Gao and Kidal Kayes Koulikoro Mopti Sikasso Ségou Tombouctou
Persistence of Subnational Diversity: Mali-II

ELF(13), 1987 census, r=0.984
Persistence of Subnational Diversity: Zambia

ELF(13), 2010 census

ELF(13), 1990 census, r=0.982
Persistence of Subnational Diversity: Liberia

ELF(13), 2008 census
ELF(13), 1974 census, r=0.918

Bong
Grand Bassa and Rivercess
Grand Cape Mount
Grand Gedeh and River Gee
Lofa and Gbarpolu
Maryland and Grand Kru
Montserrado, Bomi, and Margibi
Nimba
Sinoe
Persistence of Subnational Diversity: Kenya

ELF(13), 2014 DHS

ELF(13), 1989 census, r=0.987
Persistence of Subnational Diversity: Gabon

The diagram shows the persistence of subnational diversity in Gabon, comparing data from the 1993 census and the 2012 DHS survey. The correlation coefficient, r, is 0.955. The regions are labeled as follows: Estuaire, Haut-Ogooué, Moyen-Ogooué, Ngounié, Nyanga, Ogooué-Ivindo, Ogooué-Lolo, Ogooué-Maritime, and Woleu-Ntem.
The GIS Approach: Murdock vs. ELF(13)
The GIS Approach: GREG vs. ELF(13)
The GIS Approach: WLMS vs. ELF(13)