# Ethno-Linguistic Diversity and Urban Agglomeration

**Ulrich J. Eberle**[a,b,1,2], **J. Vernon Henderson**[a,1,2], **Dominic Rohner**[b,1,2], **and Kurt Schmidheiny**[c,1,2]

[a]London School of Economics, Centre for Economic Performance, Houghton Street, London WC2A2AE, UK.; [b]University of Lausanne, Department of Economics, Internef, 1015 Lausanne, Switzerland.; [c]University of Basel, Faculty of Business and Economics, Peter Merian-Weg 6, 4002 Basel, Switzerland.

This manuscript was compiled on May 14, 2020

**This article shows that higher ethno-linguistic diversity is associated with a greater risk of social tensions and conflict, which in turn is a dispersion force lowering urbanization and the incentives to move to big cities. We construct a novel worldwide data set at a fine-grained level on urban settlement patterns and ethno-linguistic population composition. For 3,540 provinces of 170 countries, we find that increased ethno-linguistic fractionalization and polarization are associated with lower urbanization and an increased role for secondary cities relative to the primate city of a province. These striking associations are quantitatively important and robust to various changes in variables and specifications. We find that democratic institutions affect the impact of ethno-linguistic diversity on urbanization patterns.**

Ethno-Linguistic Diversity | Fractionalization | Polarization | Urbanization | Urban Agglomeration | Primacy | Conflict | Democracy

The conflict literature has found that ethnic diversity within a region can induce tensions and raise the potential for conflict (1–3). Existing game-theoretic models of spatial distributions of ethnic groups and social tensions (4) predict that, in the presence of tensions between groups, conflicts are more costly when bigger numbers of members of different groups live at close range. To avoid such conflict costs caused by inter-group hostility, members of ethnic groups have an incentive to remain dispersed in the countryside as opposed to moving to cities to live in close quarters. Further, when they do urbanize, instead of agglomerating into one giant regional "melting pot" megapolis, they may spread over smaller cities.

This paper presents what is, to the best of our knowledge, the first global empirical investigation of the nexus between ethno-linguistic diversity and major patterns of where people live within countries. We show that initial ethnic diversity reduces urban agglomeration. This has important consequences as policies which inhibit urbanization and urban concentration can strongly restrict economic growth (5, 6). Yet, economists have largely ignored the role of ethno-linguistic cleavages when studying agglomeration benefits, urbanization and development, the size distribution of cities, and policies which impact concentration (7–14).

Many anecdotal examples of the impact of ethno-linguistic diversity on urbanization patterns may come to mind. One example is the archetypical bilingual city of Montreal which has stagnated in size since the 1960s, while nearby predominantly English-speaking cities like Toronto or French-speaking cities like Quebec-Ville have typically grown by at least 50% over the same time period (15). As a more structured example we pick the two Indian states with the highest degree of ethno-linguistic diversity in India as measured by fractionalization, a common measure of diversity in the literature which we define later. These states, Nagaland and Himachal Pradesh, are also in the top 3% of degree of diversity by provinces worldwide and Nagaland is at the center of India's well known on-going conflict in its Northeast. These highly fractionalized states rank in the top 6% and 3%, respectively, of provinces worldwide in incidence of conflict for 1975-2015 (defined below). In terms of the resulting urban concentration, we develop two measures below: share of the population that is urbanized, and primacy (fraction of the urban population in the biggest city in the province). These two Indian states both rank in the bottom 30% worldwide of provinces in terms of urban share and in the bottom 1% in terms of primacy share. In other words, their high degree of ethnic fractionalization and conflict is closely associated with people staying in the countryside and avoiding agglomerating into one main city by spreading urban population across cities.

To comprehensively assess these relationships, we created a novel, fine-grained data set of geographical population distribution and language use. For 233 countries around the world, these data allow us to compute indices of urban concentration in the year 2015, as well as ethnolinguistic diversity at the province level in 1975. Provinces are the first-level administrative boundaries within countries such as U.S. states or German Bundesländer (see the SI Appendix, p.5 for details). We identify the effects of ethno-linguistic diversity on urban concentration from within country variation in urban concentration at the provincial level for 3,540 provinces in the 170 countries with more than one province, controlling for the 1975 levels of the variables of interest. Drawing on data of the *Global Human Settlement Layer* (GHSL) and the *GHS Settlement Model* (GHS-SMOD, (16)) on geo-localised population and urban boundaries, we first establish a data set

## Significance Statement

Urbanization and agglomeration of economic activity are key drivers of economic development. Many factors underlying city sizes and locations continue to be well studied. However, a key factor has so far been generally ignored: the role of the ethno-linguistic composition of local populations. We address this gap, drawing on a novel, very detailed dataset on local urban agglomeration and ethno-linguistic diversity. We find that, in multi-ethnic areas, social tensions arise more easily, discouraging the move to bigger cities. Ethno-linguistically diverse regions feature less urbanization and agglomeration, with potentially profound economic consequences.

[1]U.J.E., J.V.H, D.R. and K.S. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: u.eberle@lse.ac.uk, J.V.Henderson@lse.ac.uk, dominic.rohner@unil.ch, kurt.schmidheiny@unibas.ch.

DRAFT

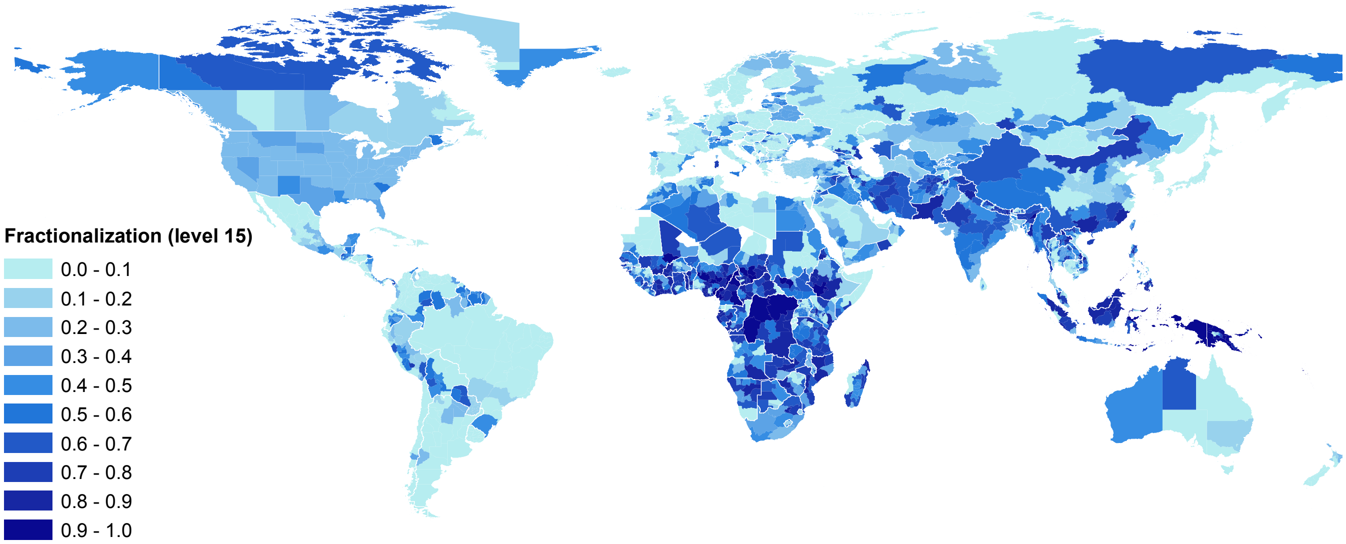


**Fig. 1.** Global Map of Ethno-linguistic Fractionalization at the Province Level. Fractionalization is calculated at language tree level 15. See text for data sources and construction.

at the 1 km grid level, which distinguishes between *city cores*, *dense towns*, *semi-dense towns*, *suburbs* and *rural areas* for 2015. The GHS project for the first time defines areas such as cities, based solely on population and population density measures consistently across the world, with no regards to local administrative borders and to census bureau qualitative views on what defines urban areas and cities. This consistency in definition across and within countries is an important feature of our contribution.[*]

In this paper, we first match the grid cells with fine-grained language information, drawing on the *World Language Mapping System* (WLMS) data capturing the traditional languages (as defined by Ethnologue (18)) present in the early 1990s. Ethnologue contains the number of speakers of all languages in a given country and WMLS maps the information of the Ethnologue into the geographic location of ethno-linguistic groups. All details of the data construction are relegated to the "Data and Methods" Section below.

In Figure 1, the average ethno-linguistic fractionalization at the province level is displayed graphically for all countries for level 15 (which is the most disaggregated level of language distinction, as detailed below). In the map, darker colours indicate higher levels of ethnic fractionalization. The map illustrates the fine-grained data structure and one reason why we study our research question at the provincial rather than national level. Figure 1 shows that large countries have enormous within country variation across provinces. Taking the province rather than the country as the unit of observation allows us to exploit this variation. Moreover, in robustness checks, we will show that our results in fact hold for small-province countries as well as large-province countries. Another key factor is that, given the high inter-provincial migration costs in many countries, with evidence for China (19) and Indonesia (20), and the role of provinces in governance, the province seems a natural way to study our phenomena. In addition, in statistical work, province-level data allow us to control through country fixed effects for unobservable confounding country characteristics (like national governance) which also influence the urban structure.

Next, using fractionalization as a measure of ethno-linguisitic diversity, we graph three motivating sets of associations. Figure 2 displays the association between a conflict measure and ethno-linguistic fractionalization, as well as between the two urban concentration measures and ethno-linguistic fractionalization, for all provinces across the world.

In panel A of Figure 2 we show with a non-linear regression that ethnic fractionalization correlates positively with the count of conflict incidents in each province from 1975 to 2015 (based on data from "Geographical Research on War, United Platform", GrowUP (21)), as postulated at the beginning of the article. This is in line with our premise that ethnic diversity may go in hand with heightened ethnic tensions and conflict. As argued above, this risk of unrest may be a dispersion force, leading to less urbanization and less urban concentration.

Hence, panel B of Figure 2 illustrates the correlation at the province level between ethnic fractionalization in 1975 and urban population share in 2015, while panel C displays the relationship of ethnic fractionalization in 1975 and primary city share in 2015. In both cases we detect – at least for intermediate and high levels of ethnic fractionalization – a clear association between ethnic diversity and both urbanization and primacy.

Taken together, the correlations suggest that places with greater fractionalization have less urbanization with more people staying in the countryside and a smaller share of urban population in the primate city of the province, so a bigger share is found in smaller cities. It appears that fractionalization strongly impacts where people live and the degree of urban concentration. Of course there will be heterogeneity in these relationships. As one example at the end of the paper, we consider a policy question of how democratization may influence outcomes, because the extent of democratization may influence the tensions associated with any degree of ethno-linguistic fractionalization.

While the associations in Figure 2 are intriguing, below

[*]There are also country specific efforts to measure urban area sizes based on density of buildings (e.g. delineating urban areas with building density for France, see (17)), but our outcomes involve population measures, so we need population data as well as worldwide coverage.

DRAFT

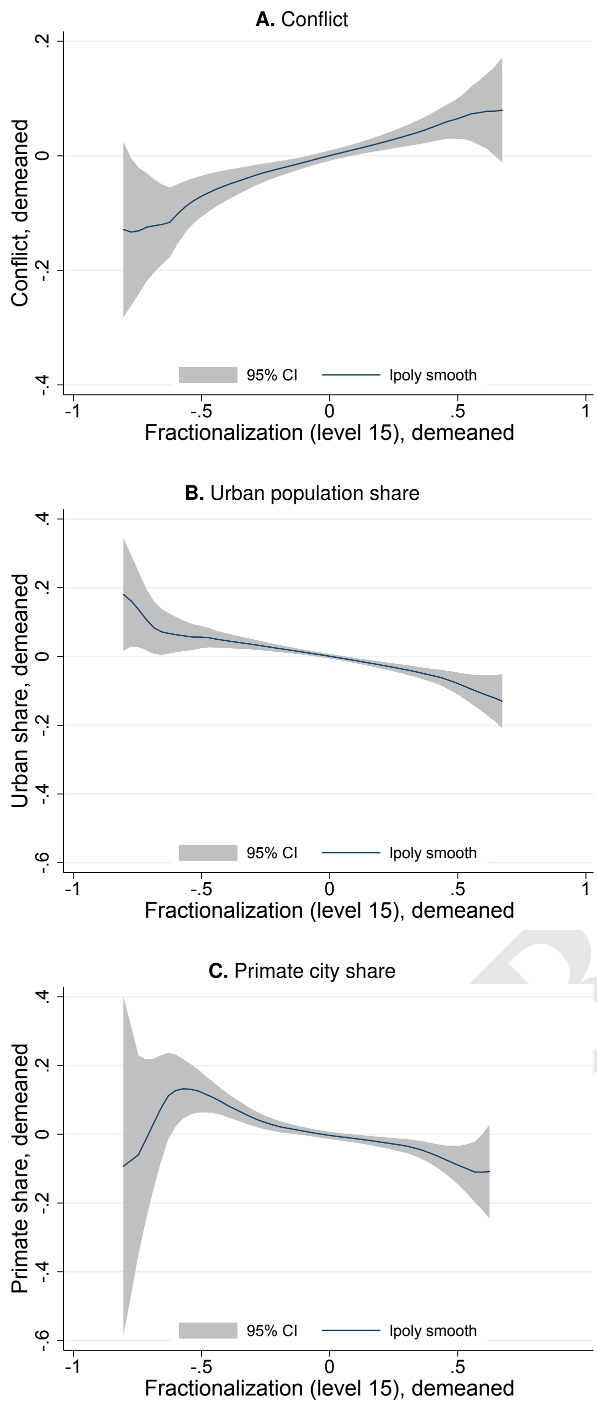


**Fig. 2.** Distributions and Regressions: Ethno-linguistic Fractionalization, Conflict and Urban Concentration. The unit of observation is a province. The sample includes 3,540 provinces worldwide. The graphs depict kernel-weighted local polynomial regressions of $1^{st}$ degree. The plots show the association between different outcome variables on the vertical axis and fractionalization on the horizontal axis. Each variable's country mean is subtracted. Fractionalization is calculated at language tree level 15 for the year 1975. Panel A: Conflict is reported for 3169 provinces in 154 countries. The outcome variable indicates provinces with at least one ethnic group involved in a conflict incident (implying at least 25 deaths) during the period 1975-2015, with data from the Geographical Research on War United Platform. Panels B and C: Urbanization indices for the year 2015 calculated with data from the Global Human Settlement Layer. Panel B. Urban share is the share of urban population of a province divided by the total population; Panel C: Primate share is the population of the largest city in a province divided by the total population of all other cities in the province.

we turn to a more full-fledged statistical analysis. For this purpose, we now first discuss in some detail the data and methods before studying these relationships in more depth in a regression analysis, controlling for a variety of potential confounders.

**Data and Methods.** Our urban concentration measures capture the extent to which provincial populations concentrate into cities ($Urban$), and the extent to which that urbanized population is found in just one city ($Primate$). To construct them, we classify each grid cell in the categories city cores ($core$), dense towns ($dense$), semi-dense towns ($semi$), suburban ($sub$) and rural area (see SI Appendix for a detailed description of definitions and algorithms). Given this classification, our dependent variables are defined as:

$$Urban_i = \frac{Pop_i^{core} + Pop_i^{dense} + Pop_i^{semi} + Pop_i^{sub}}{Pop_i}, \qquad [1]$$

$$Primate_i = \frac{Pop_i^{1st}}{Pop_i^{core} + Pop_i^{dense}}, \qquad [2]$$

where $Pop_i$ is the total population of province $i$ in 2015, $Pop_i^{1st}$ is the population in the largest city core in province $i$ and $Pop_i^{core}$, $Pop_i^{dense}$, $Pop_i^{semi}$, and $Pop_i^{sub}$ correspond to the total population of all grid cells in province $i$ of the respective category. For the urban share equation, we note that urban in the numerator is broadly defined. The GHS project has a low density threshold as part of its urban definitions of semi-dense towns and suburbs (300 per sq km) meaning that, in general, it reports higher urban shares worldwide than the UN World Urbanization Prospects data. However, we are only interested in relative comparisons across provinces within countries. For the primate share equation, we note that, for any specific city, the GHS project only identifies the dense $Pop_i^{core}$ population; suburban populations are not assigned to specific core cites. Thus to have a denominator consistent with the numerator in eqn (2), for all cities in a province, we include only dense urban populations, $Pop_i^{core}$ and $Pop_i^{dense}$. Later, as robustness checks, we will employ a stricter definition of urban share limited to core cities and dense towns in the numerator of eqn (1); and we will use a measure of primate city size that attempts to incorporate commuting zones around cities in eqn (2).

As noted above, we match the grid cells with fine-grained language information. Our language data from the *World Language Mapping System* (WLMS) is arguably the most precise source currently available, and has recently been used by (22), (23) and (24). The need to disentangle subtle differences in urbanization patterns has required us to construct our data at a more fine-grained level (1 km grid cells) than previous publications. Moreover, we apply the algorithm pioneered by (24) for allocating languages to population in multi-linguistic areas, which further increases precision. These features and the use of consistent definitions and data sources for urbanization and linguistic measures account for our dataset being the most precise of its kind currently available.

To compute measures of ethno-linguistic diversity we use the *Fractionalization* measure capturing the degree to which the population is segmented into many different groups at a provincial level. We also show in the appendix results for the *Polarization* index capturing the extent to which the population is divided into two equal sized and potentially opposing groups.

The reason we focus on ethnic *Fractionalization* as main measure is that it has been linked to both small scale frictions in public good provision (25, 26) as well as to large scale social conflict and civil wars (2, 27, 28), whereas the use of ethnic *Polarization* has been more confined to the study of large-scale wars (e.g. in (1, 2, 28)), making the concept arguably narrower and in our view slightly less relevant than *Fractionalization* for studying urbanization outcomes. Thus, we use *Polarization* as alternative measure and relegate it to the appendix. Formally, the two measures are defined in the literature (1) as:

$$Fractionalization_i = 1 - \sum_{m=1}^{M_i} (\pi_i^m)^2 \ , \qquad [3]$$

$$Polarization_i = 1 - \sum_{m=1}^{M_i} ((0.5 - \pi_i^m)/0.5)^2 \, \pi_i^m \ , \qquad [4]$$

where $M_i$ designates the total number of groups $m = 1, ..., M_i$ in province $i$ and $\pi_i^m$ corresponds to the population share of a group $m$ in the province's total population.

We populate the language map with 1975 GHS population numbers (29), so as to represent language diversity historically. Ethnologue has up to 15 levels of distinction yielding 6208 country-language pairs (e.g. "French-Canada" and "French-Switzerland" are two country-language pairs) when applying the finest level of language distinction. The information of Ethnologue and WMLS allows us to distinguish ethno-linguistic groups at different levels of language affinity; and these indices can be computed at any of the 15 levels. High levels of aggregation distinguish only major language families while low levels of aggregation, e.g. level 15, result in distinguishing very fine-grained differences between similar languages. Some countries such as India have enormous diversity, with 391 languages distinguished at the most disaggregated level and 18 already at level 2.

As an example, in Figure 3 we graphed the language structure for Himachal Pradesh, the above-mentioned province of about 7.5 million in northwest India. The figure illustrates the branches of its language tree, showing for each branch the highest level of disaggregation. The province starts on level 1 with 2 languages and then proceeds down to its finest division at level 8 with 18 final languages and ethnic groups.

In the main analysis, as in (24), we shall focus on level 15, the highest disaggregation level worldwide. For most states in India like Himachal Pradesh, the branches of the tree end at levels 6 through 8 (denoted by the underlining end language). When looking at level 15, branches ending sooner (say level 6 or 8) are accounted level 15 language affinity. In Figure S2 in the SI Appendix, we show a similar graph for Switzerland. In the regression analysis, we demonstrate robustness at more aggregated levels, where related languages in the tree are lumped together.

**Baseline Results.** This section systematically studies the association between ethno-linguistic factors and urbanization patterns by regressing contemporary measures of urban concentration on historical measures of ethno-linguistic diversity, as well as initial urban concentration levels from four decades ago, using data from provinces across the world.

Table 1 displays our results. It is divided into two panels: the top panel A is a cross sectional analysis while the bottom panel B is longitudinal by additionally controlling for the past (1975) value of the dependent variable. Columns are in pairs for different samples and outcomes; and, within each pair, columns are distinguished by the set of controls.

Column 1, Panel A regresses the *Urban share* in a given province in 2015 on pre-sample ethno-linguistic fractionalization in 1975, yielding a coefficient of -0.126 that is statistically significant at the 1 % level. To give perspective, this means that moving from a perfectly ethno-linguistically homogeneous province (i.e. with ethno-linguistic fractionalization of 0) to a perfectly diverse one (i.e. with ethno-linguistic fractionalization of 1) would be associated with a 13 percentage points lower share of the urban population in the province. This change in urbanization corresponds to about half a standard deviation of the *Urban share* measure, or the difference between the very urbanized Netherlands and the less urbanized United States which contains more rural area population. Note that this specification controls for country fixed effects, which means that the estimation is based solely on within-country comparisons of provinces, filtering out unobserved between-country heterogeneity. There is a concern however that estimates in Panel A could be biased because of omitted variables and reverse causality. For example, urbanization over long periods of time could influence fractionalization.

To deal with this, we move in column (1), Panel B to a more demanding specification where we also control for 1975 values of urban share, in which we investigate the impact of fractionalization on the evolution of urbanization over the following 4 decades. A control for the 1975 urban share also controls for the influence of omitted variables at least on historical urbanization, a topic we return to below. Of course, it also sweeps up any impact of ethnolinguistic fractionalization on historical urban share, leading us to potentially understate the total effect of fractionalization on urban share in 2015. However, conditioning on base period urbanization tells us more unambiguously how subsequent urbanization is influenced by baseline fractionalization. When controlling at the province level for urban share in 1975 in Panel B, we still find a statistically significant negative coefficient, albeit its magnitude is reduced by half compared to Panel A. Of note, the coefficient of past urban share is sizeable and highly significant, pointing towards a large persistence of urbanization patterns over time. Overall, it is reassuring that in Panel B we continue to find evidence of ethnic fractionalization slowing down the pace of urbanization, after controlling for pre-sample urbanization.

In column 2, Panels A and B, we estimate the analogous specifications as in column 1, Panels A and B, but controlling in addition for terrain ruggedness and population density in 1975 (see SI Appendix p. 5 for a detailed description of these control variables and Table S2 for all estimated coefficients). The results remain very similar and the coefficients of interest remain statistically significant at the 1 percent level.

With regard to the measure of urban concentration, we estimate the same specifications for the share of the primate city in total urban population (*Primate*). Note that unlike the 1975 urban share, the past primate share from 1975 is only observable for a restricted sample, since some provinces in 1975 did not have a core city ($Pop_i^{core}$). Hence, we run the regressions of primate share in Panel A on fractionalization first

DRAFT

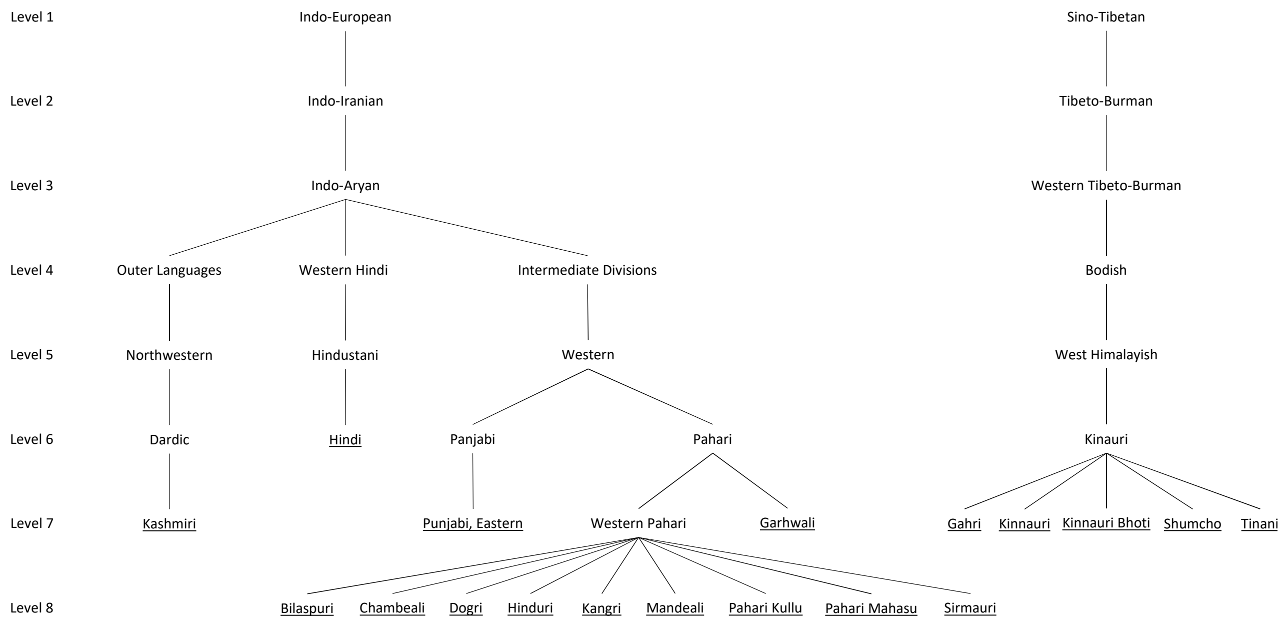


**Fig. 3.** The Use of Ethnologue Language Trees: Illustration for the Indian Province Himachal Pradesh. The graph depicts the language tree of Himachal Pradesh. The languages of Himachal Pradesh are divided in up to 8 levels, with level 1 being the most aggregated and level 8 being the least aggregated level. The endpoint (underlined) of each branch depicts the commonly-referred name of a language. The language tree is based on data by the Ethnologue. Four very minor languages at the extension of Western Pahari are omitted for presentation purposes.

on the full sample (columns 3 and 4) and then on the restricted sample (columns 5 and 6) to improve comparability. We find that the importance of the biggest city among urbanized areas is considerably reduced in the face of ethno-linguistic fractionalization. Put differently, ethno-linguistic diversity is associated with having several smaller cities instead of a single mega city. Quantitatively, moving from a fully homogeneous to a fully heterogeneous society (i.e. moving ethno-linguistic fractionalization from 0 to 1) would be associated with an at least 8 percentage points lower *Primate share* in columns 5 and 6 in Panel B, equal to about a quarter of a standard deviation of this variable.

Note that we also carry out a regression analysis linking ethnic diversity to conflict. In the interest of space, this investigation is relegated to the SI Appendix. In Table S8 we show that there is a strong and statistically significant association between ethno-linguistic fractionalization in 1975 at the province level and several measures of armed conflict between 1975 and 2015 at the province level.

How robust are our results to various considerations? The first concern is omitted variables. In SI Appendix Table S2 our results are robust to including further control variables that could potentially influence the spread of cities. In particular, we control for square and cubic terms of population density, for distance to coast, elevation, latitude, provincial GDP and for whether the national capital is located in the given province. We also control for the degree of historical conflict from 1946-1974 to address concerns that initial antagonism may have shaped diversity and urbanization in 1975. The SI Appendix p.5 contains a detailed description of these control variables. Note that these robustness checks can reduce sample size, as the additional information is not observed in all countries. Coefficients on ethno-lingusitic fractionalization move very little in response to varying the sets of controls. Finally, we assess the maximum potential remaining bias from omitted (unobserved) variables by performing a test following Altonji et al. (30) and Oster (31). In our specification with most controls for observables, i.e. Panel B of Table 1, we calculate an estimate of the extent of possible bias for the effect of fractionalization of +0.020 for urban share and +0.022 for primate share.[†] Hence our point estimates remain substantially below zero even allowing for such potential bias.

Next for robustness, we show that the overall stability of estimated coefficients remains when varying the threshold levels in the language tree for distinguishing different languages. As explained above, our data allow us to compute ethno-linguistic diversity measures for different definitions of what constitutes distinct languages. When using an aggregation level of 1, we only distinguish the most fundamental differences in the language tree, such as the difference between Indo-European and Sino-Tibetan language families, but lump together distinctions, such as Italian and German, into the Indo-European group. In contrast, as we move down the tree, the distinctions become more fine-grained, where local dialects are distinct such as Kangri, Hinduri, and Dogri as dialects of Western Paharai which in turn is related to Punjabi in Figure 3 above; or, say, Arpitan, Romansch, Lombard, and French in Non-German Switzerland (see SI Appendix Figure S2).

We graph the pattern of coefficients and their significance in Figure 4, linking ethnic diversity to urban share, primate share and conflict. Overall, the results of Figure 4 highlight the stability of estimated coefficients over a range of possible aggregation levels of the language data. In particular, we observe a statistically significant negative association between ethnic fractionalization, on the one hand, and urban and primate shares, on the other hand, across a wide range of

[†] We calculate the maximum bias with conservative assumptions for this context, i.e. $\delta = 1$ and $R^2_{max} = 0.9$. See the SI Appendix p.7 for more details and calculation.

DRAFT

**Table 1. Ethno-linguistic Fractionalization and Urbanization Patterns.**

| Dependent variable: | Urban share | | Primate share | | | |
|---|---|---|---|---|---|---|
| Sample: | Full sample | | Full sample | | Restricted sample | |
| Controls: | No | Yes | No | Yes | No | Yes |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *Panel A: Cross sectional* | | | | | | |
| Fractionalization | -0.126*** (0.024) | -0.107*** (0.023) | -0.144*** (0.025) | -0.115*** (0.023) | -0.212*** (0.031) | -0.175*** (0.028) |
| Adjusted $R^2$ | 0.467 | 0.515 | 0.360 | 0.462 | 0.342 | 0.459 |
| *Panel B: Longitudinal* | | | | | | |
| Fractionalization | -0.057*** (0.020) | -0.054*** (0.020) | | | -0.082*** (0.026) | -0.080*** (0.025) |
| Urban share (1975) | 0.612*** (0.049) | 0.591*** (0.048) | | | | |
| Primate share (1975) | | | | | 0.846*** (0.028) | 0.819*** (0.032) |
| Adjusted $R^2$ | 0.732 | 0.735 | | | 0.824 | 0.826 |
| Provinces | 3540 | 3540 | 2359 | 2359 | 1623 | 1623 |
| Countries | 170 | 170 | 154 | 154 | 138 | 138 |
| Country FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Ruggedness | | ✓ | | ✓ | | ✓ |
| Population density (1975) | | ✓ | | ✓ | | ✓ |

The unit of observation is a province. OLS estimates are reported in all columns. Robust standard errors clustered at the country level are reported in parentheses. "Restricted sample" refers to the set of provinces with data available on the outcome variable for 1975. The regressions control for country fixed-effects. Statistical significance is represented by * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

DRAFT

possible language aggregation levels. Moreover, the positive correlation between ethnic fractionalization and conflict is found across the board of different aggregation levels. We note that explanatory power of the regressions across all these graphed levels varies minimally.[‡]

Next we turn to our alternative measure of ethno-linguistic diversity. While the fractionalization measure takes high values for areas with a large number of groups, the main alternative diversity measure defined above, ethno-linguist polarization, reaches high values for situations closer to bi-modal distributions of a small number of sizeable groups. As discussed above, we prefer fractionalization – the arguably somewhat broader concept, fitting better the context of urbanization, and have relegated polarization to the SI Appendix.

The relationship in the data between our fractionalization and polarization measures is displayed in SI Appendix Figure S4. After filtering out country averages (Panel B), the two diversity measures are highly correlated though the correlation is far from perfect. It is therefore useful to replicate our baseline Table 1 using polarization measures instead of fractionalization. Studying the role of ethno-linguistic polarization also provides a different perspective on diversity – the effect of being more bimodal versus simply more diverse. The results of the baseline specification using polarization instead of fractionalization are displayed in SI Appendix Table S3 with very similar results for primacy and somewhat weaker results for urban share.

Further, we consider alternative measures for the outcome variables urban share and primate share reported in SI Appendix Table S4. First we consider in columns (1) and (2) a narrower measure of the degree of urbanization by only considering city cores and dense towns in eqn (1), leading to similar results for both fractionalization and polarization. Then we consider an alternative definition of primate share. We draw on data from a joint OECD/EC project described in (32) which offers a globally harmonized definition of commuting zones called functional urban areas (FUA). We measure primate city share as the FUA population divided by the broad definition of urban population in the numerator in eqn (1). We use the broad definition since FUA's contain population in less dense areas. Using this definition for primate city share in columns (3) and (4) again yields very similar results for both fractionalization and polarization.

Last, we explore the "modifiable areal unit problem (MAUP)" (33, 34) and ecological correlations (35), which could arise if results at the levels of (large) provinces do not carry over to smaller spatial units. Put differently, our results could be sensitive to the definition and scale of units for which data are collected. One way to investigate this is to split our provincial sample in two, according to the scales of provinces (area or population); and then check whether the findings hold similarly for the samples of countries with smaller versus larger provinces. This is what we do in SI Appendix Tables S5 and S6. In the former table we split the sample according to average population area (unweighted and population-weighted), while in the latter we split according to average province population and the number of provinces in a country. For both small and large province samples, in all cases, we continue to find large negative effects of ethnic fractionalization on urban share and primate share, with no clear pattern of whether results are stronger for the small or large province samples. We conclude that the modifiable areal unit problem is not driving our results.

**Discussion and Role of Policies.** The above results tell a stark story of ethno-linguistic diversity slowing down urbanization and urban concentration, hence potentially affecting economic development. Still, there may be room for policies to dampen the extent of this relationship. One natural candidate for a

[‡] For the three outcomes the ranges are respectively 0.734-0.735, 0.824-0.826, and 0.615-0.618.

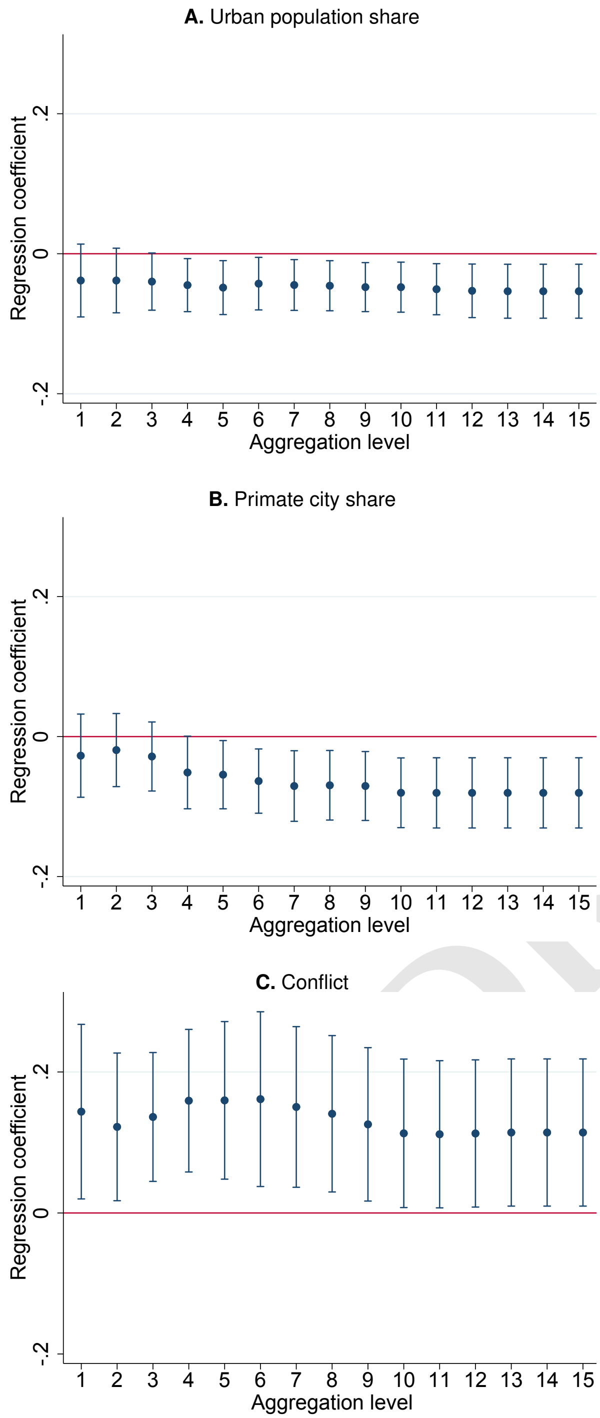


**Fig. 4.** Ethno-linguistic Fractionalization, Conflict and Urban Concentration: Results for Different Aggregation Levels. Regression results of the two measures of urban concentration and conflict incident on ethno-linguistic fractionalization, at all 15 linguistic aggregation levels. Panel A and B: the regressions performed control for country fixed effects, ruggedness and 1975 population density and 1975 outcome variables, as specified in columns (2) and (6) of the lower panel of Table 1. Panel C: the regressions performed are as specified in column (3) of SI Appendix Table S8. Point estimates are shown as dots and confidence intervals at the 95% level as bars.

policy dimension that may be able to modulate ethnic tensions is democracy. In particular, there exists evidence that while full, consolidated democracy reduces the risk of ethnic tensions and conflict, nascent or fragile/intermediate democracies may bear higher risks of political violence than autocracies (3, 36).[§] Hence, in what follows we shall investigate whether the impact of ethnic fractionalization is magnified in countries with intermediate democracy levels.

In particular, we interact our fractionalization measure with three regime types: full democracy, intermediate regime, full autocracy. We control for the full set of fixed effects and other baseline controls (ruggedness, population density), including the 1975 levels of the urban variables in panel B. Results are reported in Table 2. In the first columns (1)–(2) the democracy measure is taken from the Polity IV project (38), while in columns (3)–(4) we rely on democracy scores from Freedom House (39). The overall picture emerging from Table 2 is that indeed the impact of ethnic diversity on urban share and primate share tends to be distinctly magnified in intermediate regimes. However, the differences in coefficient magnitudes in many cases are statistically weak and stronger for primacy than for urban share (see tests at the bottom of the panels for details). Hence, these results need to be interpreted with caution. We find similar patterns in SI Appendix Table S7 for ethnic polarization, as for fractionalization.

**Data Availability.** All data used in this study are from public and commercial data sources as described in the SI Appendix. Upon publication, generated data and code to generate variables and results will be published in a publicly available repository allowing to replicate all tables and figures of the current paper. Before publication, data and code are available from the corresponding authors upon request.

**ACKNOWLEDGMENTS.** We are grateful for helpful comments from participants at the 13th Meeting of the Urban Economics Association and the 9th European Meeting of the Urban Economics Association. We thank Yannis Ioannides for comments on a formative version of the project and the suggestion that urban share should be a measure of urban concentration. We thank the editor and two referees for very helpful comments. Ulrich Eberle gratefully acknowledges financial support from the Swiss National Science Foundation Doc.Mobility fellowship P1LAP1_181253 and especially thanks the Centre for Economic Performance (CEP) at the LSE for hosting him during 2018-2020. Dominic Rohner gratefully acknowledges financial support from the ERC Starting Grant POLICIES FOR PEACE-677595 and warmly thanks UBC for the hospitality during 2018-2019 when the first draft of this paper was written. Kurt Schmidheiny is grateful to the UC Berkeley for the hospitality during the Academic Year 2017-2018 when this project took shape.

Authors are listed in alphabetical order with equal weight; all contributed equally to the project and paper.

1. JG Montalvo, M Reynal-Querol, Ethnic polarization, potential conflict, and civil wars. *Am. Econ. Rev.* **95**, 796–816 (2005).
2. J Esteban, L Mayoral, D Ray, Ethnicity and conflict: An empirical study. *Am. Econ. Rev.* **102**, 1310–42 (2012).
3. J Esteban, M Morelli, D Rohner, Strategic mass killings. *J. Polit. Econ.* **123**, 1087–1132 (2015).
4. HF Mueller, D Rohner, D Schönholzer, The peace dividend of distance: violence as interaction across space. CEPR Discussion Paper No. 11897 (2017).
5. V Henderson, The urbanization process and economic growth: The so-what question. *J. Econ. growth* **8**, 47–71 (2003).
6. JV Henderson, Urbanization and growth in *Handbook of economic growth.* (Elsevier) Vol. 1, pp. 1543–1591 (2005).
7. GK Zipf, *Human behavior and the principle of least effort.* (Addison-Wesley Press), (1949).

[§] In particular, democracy is a double-edged knife in terms of political stability, as better accountability and governance reduce the motives for revolt, but freedom of assembly and speech can be exploited by extremists (37).

**Table 2. Policy Implications: The Role of Democracy.**

| Data source: | Polity | | Freedom | |
|---|---|---|---|---|
| Dependent variable: | Urban share | Primate share | Urban share | Primate share |
| | (1) | (2) | (3) | (4) |
| *Panel A: Cross sectional* | | | | |
| Fractionalization × Democracy | -0.196** (0.082) | -0.009 (0.052) | -0.281*** (0.084) | -0.035 (0.067) |
| Fractionalization × Intermediate regime | -0.162** (0.070) | -0.368*** (0.090) | -0.079*** (0.028) | -0.198*** (0.041) |
| Fractionalization × Autocracy | -0.085*** (0.026) | -0.178*** (0.037) | -0.083** (0.032) | -0.242*** (0.055) |
| Adjusted $R^2$ | 0.530 | 0.477 | 0.515 | 0.466 |
| P(Test: Democracy = Int. regime) | .756 | .001 | .025 | .041 |
| P(Test: Int. regime = Autocracy ) | .305 | .054 | .922 | .52 |
| P(Test: Democracy = Autocracy) | .2 | .01 | .029 | .018 |
| *Panel B: Longitudinal* | | | | |
| Fractionalization × Democracy | -0.047 (0.039) | -0.029 (0.031) | -0.095* (0.057) | -0.028 (0.042) |
| Fractionalization × Intermediate regime | -0.107** (0.043) | -0.198*** (0.061) | -0.059** (0.026) | -0.140*** (0.044) |
| Fractionalization × Autocracy | -0.056** (0.027) | -0.102** (0.039) | -0.056* (0.033) | -0.074* (0.041) |
| Urban share (1975) | 0.548*** (0.065) | | 0.571*** (0.059) | |
| Primate share (1975) | | 0.809*** (0.041) | | 0.811*** (0.037) |
| Adjusted $R^2$ | 0.728 | 0.824 | 0.727 | 0.822 |
| P(Test: Democracy = Int. regime) | .297 | .001 | .559 | .071 |
| P(Test: Int. regime = Autocracy ) | .288 | .18 | .935 | .255 |
| P(Test: Democracy = Autocracy) | .847 | .012 | .519 | .449 |
| Provinces | 2627 | 1245 | 2776 | 1313 |
| Countries | 117 | 103 | 131 | 110 |
| Country FE / Base controls | ✓ | ✓ | ✓ | ✓ |

The unit of observation is a province. OLS estimates are reported in all columns. Robust standard errors clustered at the country level are reported in parentheses. Fractionalization is interacted with variables capturing the degree of democratization in countries in 1975. Columns 1-2: Data on democracy is derived from the variable "Polity" by the Polity IV Project (38). Democracy refers to the third of countries with the highest Polity score. Autocracy refers to the third of countries with the lowest Polity score. Intermediate refers to the remaining third of countries with an intermediate Polity score. Columns 3-4: Data on democracy is derived from the variable "Freedom Status" by Freedom House (39) evaluating political rights and civil liberties (accessed via the Quality of Government data catalogue). Democracy refers to countries classified as "Free". Autocracy refers to countries classified as "Not Free". Intermediate refers to countries classified as "Partly Free". The regressions control for country fixed-effects. Statistical significance is represented by * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

DRAFT


8. X Gabaix, Zipf's law and the growth of cities. *Am. Econ. Rev.* **89**, 129–132 (1999).
9. D Black, V Henderson, A theory of urban growth. *J. political economy* **107**, 252–284 (1999).
10. J Eeckhout, Gibrat's law for (all) cities. *Am. Econ. Rev.* **94**, 1429–1451 (2004).
11. K Schmidheiny, J Suedekum, The pan-european population distribution across consistently defined functional urban areas. *Econ. Lett.* **133**, 10–13 (2015).
12. AF Ades, EL Glaeser, Trade and circuses: explaining urban giants. *The Q. J. Econ.* **110**, 195–227 (1995).
13. K Desmet, JV Henderson, The geography of development within countries in *Handbook of regional and urban economics.* (Elsevier) Vol. 5, pp. 1457–1517 (2015).
14. SS Rosenthal, WC Strange, Evidence on the nature and sources of agglomeration economies in *Handbook of regional and urban economics.* (Elsevier) Vol. 4, pp. 2119–2171 (2004).
15. Statistics Canada, Census data canada (2019) Dataset (url).
16. M Pesaresi, A Florczyk, M Schiavina, M Melchiorri, L Maffenini, GHS settlement grid, updated and refined regio model 2014 in application to ghs-built r2018a and ghs-pop r2019a, multitemporal (1975-1990-2000-2015), r2019a. European Commission, Joint Research Centre (JRC) (2019) Dataset (url).
17. MP De Bellefon, PP Combes, G Duranton, L Gobillon, C Gorin, Delineating urban areas using building density. *J. Urban Econ.*, 103226 (2019).
18. MP Lewis, GF Simons, CD Fennig, Ethnologue: Languages of the world, nineteenth ed. SIL International, Dallas. (2018).
19. T Tombe, X Zhu, Trade, migration, and productivity: A quantitative analysis of china. *Am. Econ. Rev.* **109**, 1843–72 (2019).
20. G Bryan, M Morten, The aggregate productivity effects of internal migration: Evidence from indonesia. *J. Polit. Econ.* **127**, 2229–2268 (2019).
21. L Girardin, P Hunziker, LE Cederman, NC Bormann, M Vogt, Growup-geographical research on war, unified platform. ETH Zurich (2015) Dataset (url).
22. A Alesina, S Michalopoulos, E Papaioannou, Ethnic inequality. *J. Polit. Econ.* **124**, 428–488 (2016).
23. R Hodler, M Valsecchi, A Vesperoni, Ethnic geography: Measurement and evidence. CEPR Discussion Paper No. 12378 (2017).
24. K Desmet, J Gomes, I Ortuño-Ortín, The geography of linguistic diversity and the provision of public goods, (National Bureau of Economic Research), Technical report (2018).
25. A Alesina, R Baqir, W Easterly, Public goods and ethnic divisions. *The Q. journal economics* **114**, 1243–1284 (1999).
26. A Alesina, EL Ferrara, Ethnic diversity and economic performance. *J. economic literature* **43**, 762–800 (2005).
27. D Rohner, Reputation, group structure and social tensions. *J. Dev. Econ.* **96**, 188–199 (2011).
28. J Esteban, L Mayoral, D Ray, Ethnicity and conflict: Theory and facts. *science* **336**, 858–865 (2012).
29. M Schiavina, S Freire, K MacManus, GHS population grid multitemporal (1975, 1990, 2000, 2015) r2019a. European Commission, Joint Research Centre (JRC) (2019) Dataset (url).
30. JG Altonji, TE Elder, CR Taber, Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *J. political economy* **113**, 151–184 (2005).
31. E Oster, Unobservable selection and coefficient stability: Theory and evidence. *J. Bus. & Econ. Stat.* **37**, 187–204 (2019).
32. AI Moreno-Monroy, M Schiavina, P Veneri, Metropolitan areas in the world. delineation and population trends. *J. Urban Econ.*, 103242 (2020).
33. CE Gehlke, K Biehl, Certain effects of grouping upon the size of the correlation coefficient in census tract material. *J. Am. Stat. Assoc.* **29**, 169–170 (1934).
34. AS Fotheringham, DW Wong, The modifiable areal unit problem in multivariate statistical analysis. *Environ. planning A* **23**, 1025–1044 (1991).
35. DA Freedman, Ecological inference and the ecological fallacy. *Int. Encycl. social & Behav. sciences* **6**, 1–7 (1999).
36. H Hegre, Toward a democratic civil peace? democracy, political change, and civil war, 1816–1992. *Am. political science review* **95**, 33–48 (2001).
37. P Collier, D Rohner, Democracy, development, and conflict. *J. Eur. Econ. Assoc.* **6**, 531–540 (2008).
38. MG Marshall, TR Gurr, K Jaggers, Political regime characteristics and transitions, 1800-2008, Polity IV Project (2012) Dataset (url).
39. Freedom House, Freedom in the world - country and territory ratings and statuses, 1973-2018 (2019) Dataset (url).

PNAS
www.pnas.org

# Supplementary Information for

## Ethno-Linguistic Diversity and Urban Agglomeration

**Ulrich J. Eberle, J. Vernon Henderson, Dominic Rohner, and Kurt Schmidheiny**

**Ulrich J. Eberle**
**E-mail: u.eberle@lse.ac.uk**
**J. Vernon Henderson**
**E-mail: J.V.Henderson@lse.ac.uk**
**Dominic Rohner**
**E-mail: Rohner dominic.rohner@unil.ch**
**Kurt Schmidheiny**
**E-mail: kurt.schmidheiny@unibas.ch**

### This PDF file includes:

Supplementary text
Figs. S1 to S4
Tables S1 to S8
SI References

## Supporting Information Text

## Data

The main data sources are described in detail below.

**Population Data.** Calculating historical linguistic indices and contemporary population densities alike require some form of fine-grained population data, preferably available for multiple periods from the same source to ensure consistency over time. We use gridded population data by the Global Human Settlement (GHS) project's 1 sq km "population grid" (GHS-POP, (1)), available for the years 1975, 1990, 2000 and 2015, derived from GPW4, and provided by the European Commission, Joint Research Centre and Columbia University, Center for International Earth Science Information Network. GHS generates population counts per grid cell by dis-aggregating population data of administrative units (CIESIN GPWv4) into grid cells based on built-up cover (impermeable surface) as determined primarily from Landsat satellite imagery (Global Human Settlement Layer, GHSL) for the respective year. The Global Human Settlement Layer is an initiative of the European Commission's Joint Research Centre (JRC), the European Commission's Directorate General for Regional Development, and the GEO Human Planet Initiative which maps built-up cover from satellite imagery. We calculate ethno-linguistic indices based on population data for the year 1975 and the urban outcome measures on population data for the year 2015.

**Urban Boundary Data.** Our main dependent variables measure the fraction of urbanized population in provinces and the fraction of the largest primate city within the urban population. To define these variables as precisely as possible, a globally consistent definition of meaningful population density thresholds is required. We take information on the degree of urbanization from the GHS "Settlement Model layers" (GHS-SMOD, (2)). This data set defines seven population density groups and assigns a group to each 1 sq km grid cell: City cores (at least 50,000 inhabitants with cells having at least 1500 per sq km), dense towns (5,000 to 50,000 inhabitants meeting the 1500 density requirement per cell), semi-dense towns (5,000 to 50,000 inhabitants meeting a density requirement of 300 people per sq km and 2 km distance to the next city core or dense town), suburbs (accounting for the residual inhabitants of the urban cluster having density over 300 people per sq km) and three low-density, i.e. rural categories. This classification is based on the formation of contiguous areas of high-density cells and the total population within such areas (2). Our primate city in each province is based on these GHS core cities boundaries, summing the grid square populations within those boundaries.

The two panels of Figure S1 illustrate –for the regions of Northern France, Belgium, Netherlands and Southern UK– how the classification into settlement categories (right panel) allows for a clear distinction between urban and rural population clusters and gives us agglomerations such as cities and towns, while the left panel shows the underlying population densities. Note, in this part of Europe, only the center of Paris in the left panel shows really high population densities over 20,000 people per square km.

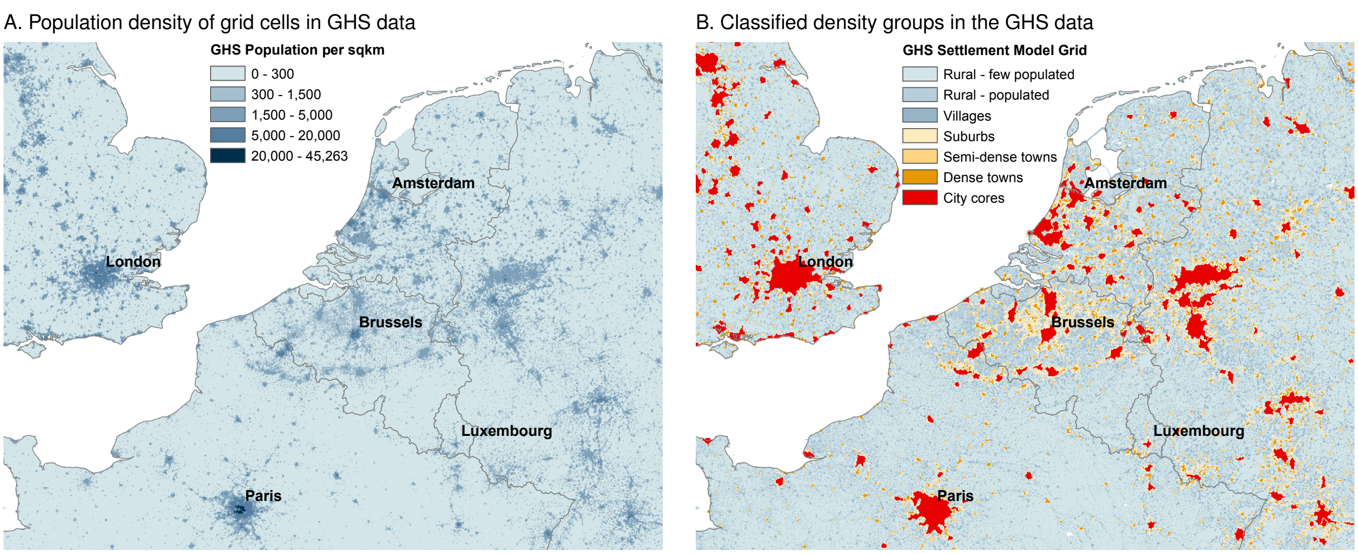


**Fig. S1.** Degree of Urbanization in Europe, 2015. The left panel depicts raw population data from the GHS population grid (GHS-POP), with a population density per km$^2$ ranging from 0 (uninhabited) to 45,263 (Central Paris). The right panel shows the seven urbanization classes from the GHS Settlement Model grid (GHS-SMOD), which are used to define urban and city core populations in our outcome variables.

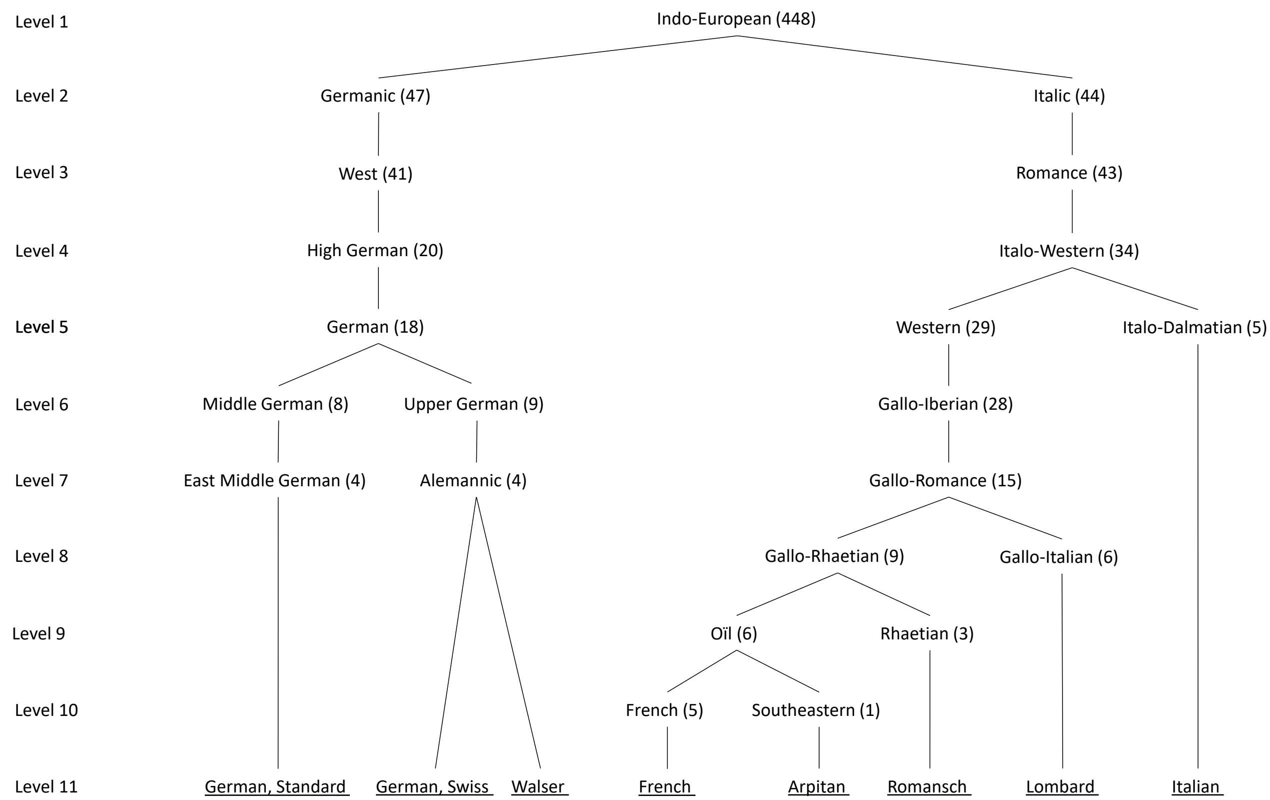


**Fig. S2.** Ethnologue Language Tree for Switzerland. The graph depicts the language tree of Switzerland. Swiss languages are divided in up to 11 levels, with level 1 being the most aggregated and level 11 being the least aggregated level. The endpoint (underlined) of each branch depicts the commonly-referred name of a language. The language tree is based on data by the Ethnologue.

**Language Data.** To calculate province-level ethno-linguistic fractionalization and polarization, we require information on the number of speakers per language in each province. We obtain information on the spatial distribution of languages from the $19^{th}$ edition of the World Language Mapping System (WLMS), the georeferenced counterpart of the Ethnologue (3).[*] This map covers most parts of the world with polygons, each depicting the extent of a traditional language, as it occurred in the early/mid 1990s. The data accounts for multilingual regions by letting language polygons overlap. The scale of reporting varies across regions: while languages in many parts of the Old World appear to be well-documented, the recording is limited in regions subject to past large-scale migration waves, such as Oceania and South America ((4)). After data cleaning, we identify 6,208 country-language pairs (when focusing on the finest level of language distinction, level 15), with the median (mean) country having 6 (26.64) languages. Linguistic diversity spans from mono-linguistic nations - mostly isolated island states such as Cuba, Iceland or Jamaica - to countries with complex linguistic nets - such as India, Indonesia and Papua New Guinea with 391, 435 and 467 languages, respectively.[†]

Not mapped are mostly unpopulated areas, such as deserts. Even though these areas are unlikely to have any impact on any of our variables, we interpolate missing data by assigning the closest language polygon. As discussed in the main text, various ways exist to compute ethno-linguistic diversity, depending on what threshold is used for distinguishing different languages. For very high levels of aggregation (Level 1) only mere major language families are considered different when computing diversity measures, while low levels of aggregation (Level 15) result in distinguishing very fine-grained differences between similar languages. Figure S2 illustrates the branches of the language tree for one country, Switzerland, to supplement the example of the province of Himachal Pradesh, India in the text.

To illustrate how the various levels of thresholds of the language tree map into diversity measures, compare Figure 1 in the main text which displayed ethno-linguistic fractionalization around the world for the most fine-grained level 15, with Figure S3 where we display the analogue world map of ethno-linguistic fractionalization at the province level for the most aggregate level 1. Unsurprisingly, using a higher level of disaggregation results in more clear-cut differences between areas and leads to higher

[*] WLMS has recently been used by e.g. (4–6). Note that alternative global georeferenced group-level data include "Geo-referencing of Ethnic Groups" and "Geo-referencing Ethnic Power Relations". Both are of high quality and frequently used in the related literature. Unlike WLMS, however, neither of the data sets reports all language speakers per country, which is crucial to adopt an iterative fitting process in areas with overlapping group coverage.

[†] We exclude a set of mostly minor languages, due to insufficient information necessary for data processing: languages with unknown location; point languages with a population share smaller .5%; languages without or unknown number of first language speakers; languages with insufficient linguistic tree information including "isolate languages" (no language trees available), "mixed languages" (hybrids without clearly defined language trees) and sign languages.

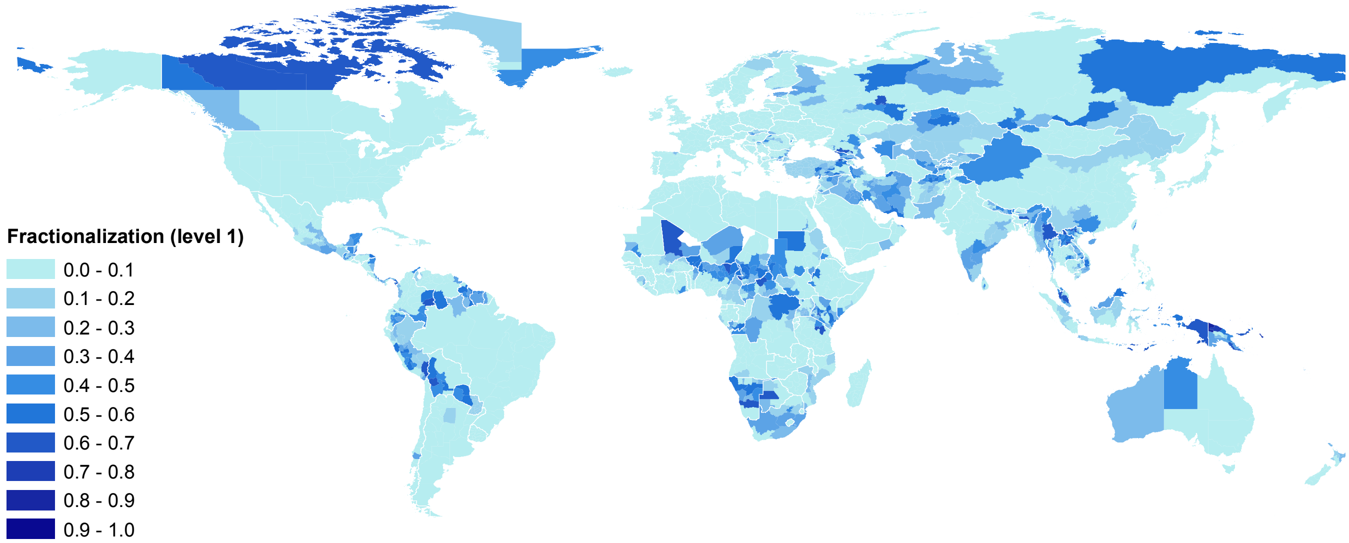


**Fig. S3.** Global Map of Ethno-Linguistic Fractionalization (Tree Level 1) at the Province Level. Fractionalization is calculated at language tree level 1. See text for data sources and construction.

computed diversity scores.

Note that some languages require special attention, for instance those not bound to a specific region, but spread throughout a country, known as "widespread languages", e.g. Russian speakers in Uzbekistan. We distribute widespread language speakers uniformly across a country, which is equivalent to spanning a polygon along a country's boarders. Further, a small number of languages are marked as a point in the WLMS raw data when the location of speakers within a country is known, but not the extent of their geographical spread. We then follow (6) and draw a circle around these points, proportional in size to the share of speakers in the country.[‡] The last unmapped language class in WLMS describes languages for which neither the location, nor the geographic extent is known. We choose to omit those languages, as we are unable to assign them to the correct province.[§] Combining all the above steps results in a fully polygonized ethno-linguistic map, making use of all available information.

In a next step, we allocate local populations to the languages spoken in each province. This task would be straightforward in the absence of spatial overlaps of languages or if province-level language speaker numbers were available in case of on overlap. Unfortunately, neither is the case, making the assignment of local populations to spoken language consequently more complex in multilingual regions. To address this issue, we employ an iterative proportional fitting algorithm, a statistical procedure that assigns people in a certain region to a language, conditional on the nation-wide share of speakers. This procedure has been recently applied in a similar context by (6), whose steps we follow.

We prepare the data by converting the linguistic map into $K$ 1 km grid cells. There are $M$ languages spoken in a country. The data can thus be organized in a $K \times M$ dimensional matrix $\mathcal{B}$, where each column represents a language and each row accounts for a single grid cell. Next, we assign the value 1 to element $b_{km}$ if language $m$ is spoken in cell $k$, 0.00000004 otherwise. The rationale behind assigning a small positive value rather than zero to languages not spoken in a cell is to account for intrastate migration. For instance, while it is highly likely that at least some Canadian French speakers moved to Vancouver at some point, the linguistic data does not map them accordingly. We address inconsistencies in the linguistic map, by distributing a small amount of Canadian French speakers across Canada.[¶] In addition, we define a $K \times 1$ matrix $\mathcal{N}$, with each cell's GHS population count. Finally, the $1 \times M$ dimensional matrix $\mathcal{L}$ contains the total number of speakers per language in that country (the data is obtained from the Ethnologue). The iterative proportional fitting process adjusts the elements of matrix $\mathcal{B}$ such that row and column totals sum up to the corresponding entry in matrix $\mathcal{N}$ and $\mathcal{L}$, respectively. The algorithm follows the steps below:

1. Proportionally adjust each row's sum to equal entries in matrix $\mathcal{N}$: Divide each row by its row-total, then multiply each column by $\mathcal{N}$.

2. Proportionally adjust each column's sum to equal entries in matrix $\mathcal{L}$: Divide each column by its column-total, then multiply each row by $\mathcal{L}$.

3. Repeat steps 1) and 2) until convergence is reached.

[‡] Languages representing less than .5 % of the country's population are omitted, because they would otherwise result in very small, and most likely imprecise circles.

[§] Omitted languages are relatively small, with speakers representing on average only a tenth of those from mapped languages. In addition, a third of these have insufficient information or are classified as sign languages, hence not revealing any information of ethnic affiliation.

[¶] (6) assign $0.000001$ to each 25 sq km (5 km $\times$ 5 km) grid cell. We proportionally adjust this value to our 1 sq km grid cells: $0.000001/25 = 0.00000004$. A small positive amount is further desirable because exclusively positive values in matrix $\mathcal{B}$ guarantees convergence, as shown in (7).

In other words, the process re-balances the values in matrix $\mathcal{B}$ until (i) the sum of all speakers in a cell equals the GHS cell population count, and at the same time (ii) the sum of each language's speakers across all cells equals the total of speakers in the Ethnologue.[‖]

**Polarization and Fractionalization Measures.** In Figure S4 we display graphically the relationship between our polarization versus fractionalization measures. While these diversity measures are obviously positively related, they are far from identical. This is intuitive given that polarization measures take high values for settings with two dominant groups of similar size, while fractionalization spikes when the number of groups is very large.

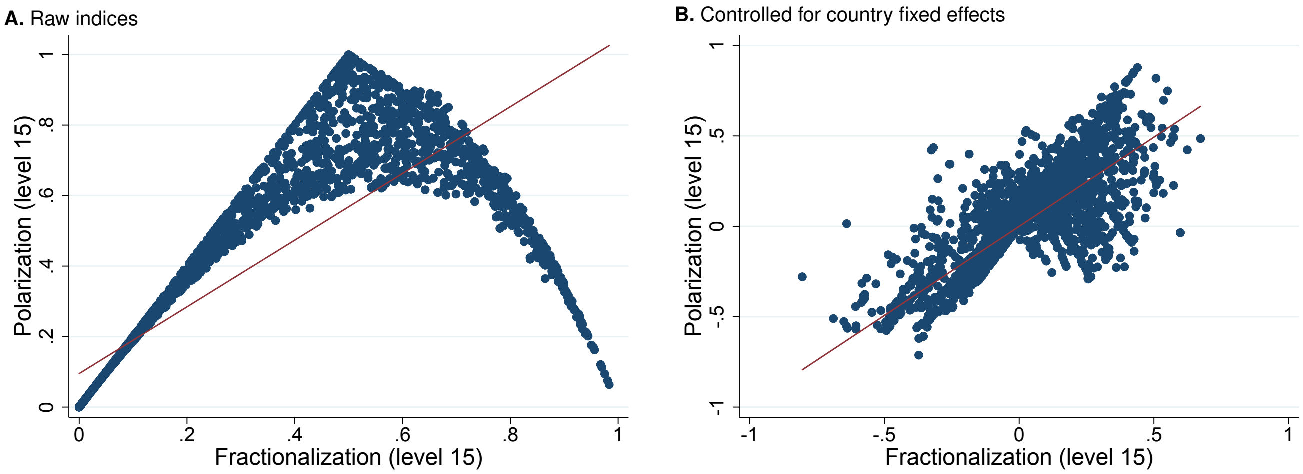


**Fig. S4.** Ethno-linguistic Polarization and Fractionalization. Scatter plots studying the relationship between fractionalization and polarization at level 15 of the language tree across 3,540 provinces in 170 countries. In the right panel, each the country mean of each variable is subtracted.

**Administrative boundaries.** National and first-level administrative boundaries are from the Digital Chart of the World "Province" data set for the year 2000. The median (mean) country has 27 (49.80) provinces. The unit of observation of the regression analysis is based on countries' first-level administrative boundaries. For instance, the unit of observation is a state for the United States and a Bundesland for Germany. We prefer this data set over alternative options, among others because WLMS is also based on Digital Chart of the World's national boundary data.

**Further variables.** Throughout the manuscript and appendix we also include a series of control variables which we shall describe in some detail in what follows. In particular, ruggedness data by (8) is used to calculate the province average of the "Terrain Ruggedness Index", an index measuring irregularities in the local terrain, based on elevation data and first defined by (9). The variable is measured in units of hundreds of metres and the granularity of the underlying elevation data is 30 arc-seconds. Population density (1975) is calculated by dividing province populations in 1975 by land area. Population numbers are derived from the GHS Population Grid (1) and land area is based on all land pixels defined in (2). Elevation depicts the province average altitude, based on data by (10) and in units of hundreds of metres, with a granularity of the underlying data of 30 arc-seconds. Latitude is measured at the province centroid and specifies the geographic north–south position in decimal degrees. Distance to coast measures the spherical distance between province centroids and the nearest coast line, based on data by the Digital Chart of the World 2000. The variable is reported in kilometres. Capital in province measures the spherical distance between province centroids and the according capital city, based on capital location data by (11). # Conflicts (1946-1974) depicts the number of conflict events between 1946 and 1974, derived from conflict data by the "Geographical Research on War, United Platform" (GrowUP, (12)). Provincial GDP (1990) measures the total GDP per province for the year 1990, based on "Gross Cell product" (purchasing power parity) data by (13), a data set globally available at the 1 by 1 decimal degree level. All distance-based variables are calculated in ArcGIS.

**Descriptive statistics.** Drawing on the aforementioned data sets, we are able to construct our main variables of interest used in the main text. The descriptive summary statistics of these measures are displayed below in Table S1.

[‖] For a more detailed discussion of this procedure, please consult (6).

**Table S1. Descriptive summary statistics of main variables**

| | Obs. | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| *Dependent variables:* | | | | | |
| Urban share (2015) | 3540 | 0.63 | 0.26 | 0.00 | 1.00 |
| Primate share (2015) | 2368 | 0.52 | 0.29 | 0.01 | 1.00 |
| *Ethnicity Indices:* | | | | | |
| Fractionalization (Level 1, 1975) | 3540 | 0.09 | 0.16 | 0.00 | 0.80 |
| Fractionalization (Level 8, 1975) | 3540 | 0.23 | 0.26 | 0.00 | 0.98 |
| Fractionalization (Level 15, 1975) | 3540 | 0.26 | 0.28 | 0.00 | 0.98 |
| Polarization - (Level 1, 1975) | 3540 | 0.16 | 0.27 | 0.00 | 1.00 |
| Polarization - (Level 8, 1975) | 3540 | 0.32 | 0.33 | 0.00 | 1.00 |
| Polarization - (Level 15, 1975) | 3540 | 0.34 | 0.32 | 0.00 | 1.00 |
| *Control variables:* | | | | | |
| Ruggedness (100m) | 3540 | 1.12 | 1.21 | 0.00 | 8.88 |
| Population density (population/km$^2$, 1975) | 3540 | 0.29 | 0.96 | 0.00 | 13.42 |
| Urban share (1975) | 3540 | 0.54 | 0.29 | 0.00 | 1.00 |
| Primate share (1975) | 1712 | 0.54 | 0.29 | 0.01 | 1.00 |

The unit of observation is a province.

## Selection on unobserved variables

The practical formula in (14) is $\beta^* = \beta_1 - \delta(\beta_0 - \beta_1)(R^2_{max} - R^2_1)/(R^2_1 - R^2_0)$. In this, $\beta^*$ is an estimator that converges to the true coefficient, $\beta_0$ the estimated coefficient without controls and $\beta_1$ the coefficient with controls; $R^2_0$ and $R^2_1$ are the corresponding $R^2$'s. $\delta$ has an upper bound of 1 under equal selection (unobservables and observables equally related to the treatment), which we assume. $R^2_{max}$ is the maximum explanatory power obtainable from included and omitted variables, excluding measurement error and purely idiosyncratic items. We think measurement error is fairly high for urban share given the controversies in defining urban, although perhaps less so for primacy which is a ratio where different measures of city populations affect both the numerator and denominator. We also note that, in our case, when we control for the lagged dependent variable, we are in essence controlling for the effect of all omitted variables on at least historical populations and we think of the R2 's in panel B of Table 1 as being pretty much at the maximum, before measurement error. We note in Table S2 in columns 4 and 8 when we add in the long list of controls with the lagged dependent variable present, the R2 relative to columns 1 and 5 changes by less than 1.5%. We could take a very conservative view by assuming $R^2_{max} = 1$, which would give a possible bias of +0.032 for urban share and +0.052 for primacy, which still leaves noticeable negative effects of fractionalization on both outcomes. However it is unreasonable to assume no measurement error or pure noise in these two cases. For the text, in both cases we set $R^2_{max} = 0.9$, which still may be conservative. We then estimate for fractionalization in Table 1, Panel B, col. (2) and (5) as the estimate with controls, $\beta_1$, for urban share and primate share, respectively. We compare this to the estimated effect of fractionalization in a bivariate regression without any control variables beyond fractionalization and with no country fixed effects. The estimated coefficient of these bivariate regressions are $\beta_0 = -0.140$ (standard error clustered for countries = 0.062, $R^2 = 0.022$, $N = 3540$) for urban share and $\beta_0 = -0.300$ (s.e. = 0.061, $R^2 = 0.081$, $N = 1623$) for primate share.

## Robustness checks

In what follows we shall display a series of robustness tables that we have discussed at length in the main text. In Table S2 we include a battery of further control variables capturing terrain, location, economic, political and past historical conflict characteristics that could potentially influence the potential growth of cities (see the detailed description of these control variables above).

Further, in Table S3 we replicate the results of the baseline specifications but focusing on ethnic polarization instead of fractionalization. Moreover, Table S4 estimates the baseline specifications, but focusing on alternative definitions of urban share and primate share. For urban share we apply a narrower measure of total urban population by only considering city cores and dense towns in eqn (1). For the primate share, the OECD has a project to define commuting zones of cities worldwide, which they call functional urban areas [FUA] (15). In Table S4 primacy is measured as the FUA population divided by the broad definition of urban population in the numerator in eqn (1). We use the broad definition since FUA's contain population in less dense areas.

To investigate the potential sensitivity of our results to the size of provinces, we split our sample according to the scales of provinces (area, population, etc) in Tables S5 and S6. While in the former we split the sample according to average population area (unweighted and population-weighted), in the latter the sample is split according to average province population and the number of provinces in a country. In each case the splits are intended to divide provinces into equal size groups. All provinces in a country are put in one or the other group, so the number of countries in each sample differs.

Finally in Table S7 we replicate findings of Table 2 in the main text on policy analysis, but focusing on ethic polarization rather than fractionalization.

As discussed in depth in the main text, for all the aforementioned sensitivity checks our findings continue to hold.

**Table S2. Robustness to Alternative Control Variables**

| Sample: | Restricted sample | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dependent variable: | Urban share | | | | Primate share | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| *Panel A: Cross sectional* | | | | | | | | |
| Fractionalization | -0.107*** | -0.109*** | -0.079*** | -0.080*** | -0.175*** | -0.173*** | -0.131*** | -0.149*** |
| | (0.023) | (0.023) | (0.023) | (0.024) | (0.028) | (0.028) | (0.030) | (0.029) |
| Ruggedness | -0.011** | | -0.004 | -0.004 | 0.006 | | 0.018** | 0.003 |
| | (0.004) | | (0.005) | (0.007) | (0.008) | | (0.007) | (0.009) |
| Population density (1975) | 0.063*** | 0.064*** | 0.353*** | 0.053*** | 0.083*** | 0.082*** | 0.372*** | 0.067*** |
| | (0.008) | (0.008) | (0.036) | (0.008) | (0.007) | (0.007) | (0.021) | (0.008) |
| (Population density, 1975)$^2$ | | | -0.070*** | | | | -0.066*** | |
| | | | (0.009) | | | | (0.006) | |
| (Population density, 1975)$^3$ | | | 0.003*** | | | | 0.003*** | |
| | | | (0.001) | | | | (0.000) | |
| Distance to coast | | | | -0.000* | | | | -0.000*** |
| | | | | (0.000) | | | | (0.000) |
| Elevation | | | | -0.002 | | | | 0.002 |
| | | | | (0.001) | | | | (0.002) |
| Latitude | | | | 0.000 | | | | 0.001 |
| | | | | (0.001) | | | | (0.001) |
| Capital in province | | | | 0.189*** | | | | 0.223*** |
| | | | | (0.019) | | | | (0.028) |
| # Conflicts (1946-1974) | | | | -0.000 | | | | -0.005** |
| | | | | (0.002) | | | | (0.002) |
| Provincial GDP (1990) | | | | 0.001*** | | | | -0.000* |
| | | | | (0.000) | | | | (0.000) |
| Adjusted R$^2$ | 0.515 | 0.514 | 0.585 | 0.549 | 0.459 | 0.459 | 0.556 | 0.511 |
| *Panel B: Longitudinal* | | | | | | | | |
| Fractionalization | -0.054*** | -0.055*** | -0.047** | -0.048** | -0.080*** | -0.080*** | -0.075*** | -0.072*** |
| | (0.020) | (0.020) | (0.020) | (0.021) | (0.025) | (0.026) | (0.026) | (0.027) |
| Ruggedness | -0.009** | | -0.007* | -0.001 | 0.001 | | 0.004 | -0.003 |
| | (0.004) | | (0.004) | (0.006) | (0.005) | | (0.005) | (0.006) |
| Population density (1975) | 0.014** | 0.015** | 0.113*** | 0.014*** | 0.013*** | 0.013*** | 0.088*** | 0.009*** |
| | (0.006) | (0.006) | (0.023) | (0.005) | (0.004) | (0.004) | (0.022) | (0.003) |
| (Population density, 1975)$^2$ | | | -0.020*** | | | | -0.017*** | |
| | | | (0.006) | | | | (0.005) | |
| (Population density, 1975)$^3$ | | | 0.001** | | | | 0.001*** | |
| | | | (0.000) | | | | (0.000) | |
| Distance to coast | | | | -0.000 | | | | -0.000 |
| | | | | (0.000) | | | | (0.000) |
| Elevation | | | | -0.002 | | | | 0.001 |
| | | | | (0.001) | | | | (0.001) |
| Latitude | | | | -0.001 | | | | -0.001 |
| | | | | (0.001) | | | | (0.001) |
| Capital in province | | | | 0.079*** | | | | 0.066*** |
| | | | | (0.019) | | | | (0.021) |
| # Conflicts (1946-1974) | | | | -0.000 | | | | -0.002** |
| | | | | (0.001) | | | | (0.001) |
| GDP (1990) | | | | 0.000*** | | | | -0.000** |
| | | | | (0.000) | | | | (0.000) |
| Urban share (1975) | 0.591*** | 0.592*** | 0.545*** | 0.568*** | | | | |
| | (0.048) | (0.048) | (0.049) | (0.058) | | | | |
| Primate share (1975) | | | | | 0.819*** | 0.819*** | 0.778*** | 0.807*** |
| | | | | | (0.032) | (0.032) | (0.036) | (0.039) |
| Adjusted R$^2$ | 0.735 | 0.734 | 0.744 | 0.746 | 0.826 | 0.827 | 0.831 | 0.838 |
| Provinces | 3540 | 3540 | 3540 | 3061 | 1623 | 1623 | 1623 | 1459 |
| Countries | 170 | 170 | 170 | 147 | 138 | 138 | 138 | 120 |
| Country FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

The unit of observation is a province. Variable definitions and sources are outlined above. OLS estimates are reported in all columns. Robust standard errors clustered at the country level are reported in parentheses. The regressions control for country fixed-effects. Statistical significance is represented by * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

**Table S3. Ethno-linguistic Polarization and Urbanization Patterns**

| Dependent variable: | Urban share | | Primate share | | | |
|---|---|---|---|---|---|---|
| Sample: | Full sample | | Full sample | | Restricted sample | |
| Controls: | No | Yes | No | Yes | No | Yes |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *Panel A: Cross sectional* | | | | | | |
| Polarization | -0.085*** | -0.071*** | -0.105*** | -0.082*** | -0.123*** | -0.094*** |
| | (0.023) | (0.020) | (0.022) | (0.021) | (0.028) | (0.025) |
| Adjusted $R^2$ | 0.465 | 0.513 | 0.358 | 0.460 | 0.334 | 0.453 |
| *Panel B: Longitudinal* | | | | | | |
| Polarization | -0.012 | -0.011 | -0.048** | -0.046** | -0.048** | -0.046** |
| | (0.015) | (0.015) | (0.019) | (0.018) | (0.019) | (0.018) |
| Urban share (1975) | 0.615*** | 0.594*** | | | | |
| | (0.049) | (0.048) | | | | |
| Primate share (1975) | | | 0.849*** | 0.822*** | 0.849*** | 0.822*** |
| | | | (0.027) | (0.031) | (0.027) | (0.031) |
| Adjusted $R^2$ | 0.731 | 0.734 | 0.823 | 0.825 | 0.823 | 0.825 |
| Provinces | 3540 | 3540 | 2359 | 2359 | 1623 | 1623 |
| Countries | 170 | 170 | 154 | 154 | 138 | 138 |
| Country FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Ruggedness | | ✓ | | ✓ | | ✓ |
| Population density (1975) | | ✓ | | ✓ | | ✓ |

The unit of observation is a province. OLS estimates are reported in all columns. Robust standard errors clustered at the country level are reported in parentheses. “Restricted sample” refers to the set of provinces with data available on the outcome variable for 1975. The regressions control for country fixed-effects. Statistical significance is represented by * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

**Table S4. Robustness to Alternative Urban Definitions**

| Sample: | Restricted sample | | | |
|---|---|---|---|---|
| Dependent variable: | Urban share (core and dense) | | Primate share (FUA) | |
| | (1) | (2) | (3) | (4) |
| *Panel A: Cross sectional* | | | | |
| Fractionalization | -0.089*** | | -0.149*** | |
| | (0.023) | | (0.029) | |
| Polarization | | -0.067*** | | -0.123*** |
| | | (0.021) | | (0.027) |
| Adjusted $R^2$ | 0.581 | 0.581 | 0.439 | 0.439 |
| *Panel B: Longitudinal* | | | | |
| Fractionalization | -0.061*** | | -0.104*** | |
| | (0.020) | | (0.028) | |
| Polarization | | -0.028* | | -0.073*** |
| | | (0.016) | | (0.027) |
| Urban share (core and dense, 1975) | 0.546*** | 0.546*** | | |
| | (0.049) | (0.049) | | |
| Urban share (1975) | | | 0.521*** | 0.518*** |
| | | | (0.072) | (0.073) |
| Adjusted $R^2$ | 0.742 | 0.741 | 0.513 | 0.512 |
| Provinces | 3540 | 3540 | 2407 | 2407 |
| Countries | 170 | 170 | 156 | 156 |
| Country FE | ✓ | ✓ | ✓ | ✓ |
| Ruggedness | ✓ | ✓ | ✓ | ✓ |
| Population density (1975) | ✓ | ✓ | ✓ | ✓ |

The unit of observation is a province. OLS estimates are reported in all columns. In columns 1-2, the dependent variable employs a narrower definition of urban population, with the numerator only considering the population located in city cores and dense towns and the denominator still capturing the whole province population. In columns 3-4, the dependent variable uses an alternative primate share definition, with the numerator capturing the province-wide population within "Functional Urban Areas", derived from data by GHS (15) and the denominator based on the baseline definition of the urban population (city cores, dense towns, semi-dense towns and suburbs). Robust standard errors clustered at the country level are reported in parentheses. "Restricted sample" refers to the set of provinces with data available on the outcome variable for 1975. The regressions control for country fixed-effects. Statistical significance is represented by * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

**Table S5. Robustness of Provinces as Unit of Observation (Area-based Sample Splits)**

| Sample: | Restricted sample | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Splitting criteria: | Average province area | | | | Average province area (population weighted) | | | |
| Dependent variable: | Urban share | | Primate share | | Urban share | | Primate share | |
| Sample split criteria: | <median | >median | <median | >median | <median | >median | <median | >median |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| *Panel A: Cross sectional* | | | | | | | | |
| Fractionalization | -0.166*** | -0.073*** | -0.107** | -0.113*** | -0.145*** | -0.081*** | -0.098** | -0.114*** |
| | (0.043) | (0.019) | (0.045) | (0.027) | (0.044) | (0.019) | (0.042) | (0.028) |
| Adjusted $R^2$ | 0.507 | 0.469 | 0.486 | 0.366 | 0.492 | 0.465 | 0.510 | 0.366 |
| *Panel B: Longitudinal* | | | | | | | | |
| Fractionalization | -0.101*** | -0.031* | -0.138 | -0.066*** | -0.092*** | -0.034** | -0.136 | -0.065*** |
| | (0.029) | (0.016) | (0.087) | (0.022) | (0.030) | (0.017) | (0.083) | (0.022) |
| Urban share (1975) | 0.667*** | 0.480*** | | | 0.659*** | 0.480*** | | |
| | (0.068) | (0.045) | | | (0.065) | (0.047) | | |
| Primate share (1975) | | | 0.821*** | 0.815*** | | | 0.805*** | 0.824*** |
| | | | (0.053) | (0.039) | | | (0.058) | (0.040) |
| Adjusted $R^2$ | 0.755 | 0.672 | 0.842 | 0.772 | 0.742 | 0.676 | 0.831 | 0.781 |
| Provinces | 1784 | 1756 | 615 | 1008 | 1833 | 1707 | 583 | 1040 |
| Countries | 73 | 97 | 52 | 86 | 77 | 93 | 55 | 83 |
| Country FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Ruggedness | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Population density (1975) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

The unit of observation is a province. OLS estimates are reported in all columns. The sample is split according to country-wide province features. In columns 1-4, odd (even) columns only consider provinces located in countries with a below (above)-median average province area, with the province area calculated in ArcGIS. In columns 5-8, odd (even) columns only consider provinces located in countries with a below (above)-median population weighted area, i.e. with the province area weighted by GHS population counts for 1975. Robust standard errors clustered at the country level are reported in parentheses. The regressions control for country fixed-effects. Statistical significance is represented by * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

**Table S6. Robustness of Provinces as Unit of Observation (Population-based Sample Splits and Number of Provinces per Country)**

| Sample: | Restricted sample | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Splitting criteria: | Average. province population (1975) | | | | Number of provinces | | | |
| Dependent variable: | Urban share | | Primate share | | Urban share | | Primate share | |
| Sample split criteria: | <median | >median | <median | >median | <median | >median | <median | >median |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| *Panel A: Cross sectional* | | | | | | | | |
| Fractionalization | -0.085** | -0.124*** | -0.077 | -0.125*** | -0.071*** | -0.138*** | -0.143*** | -0.087*** |
| | (0.042) | (0.026) | (0.054) | (0.025) | (0.024) | (0.036) | (0.034) | (0.030) |
| Adjusted $R^2$ | 0.491 | 0.486 | 0.506 | 0.357 | 0.519 | 0.519 | 0.402 | 0.518 |
| *Panel B: Longitudinal* | | | | | | | | |
| Fractionalization | -0.056* | -0.061*** | -0.132** | -0.076*** | -0.030 | -0.072** | -0.062** | -0.095** |
| | (0.033) | (0.022) | (0.065) | (0.027) | (0.020) | (0.031) | (0.026) | (0.040) |
| Urban share (1975) | 0.635*** | 0.501*** | | | 0.546*** | 0.626*** | | |
| | (0.062) | (0.052) | | | (0.040) | (0.081) | | |
| Primate share (1975) | | | 0.822*** | 0.816*** | | | 0.851*** | 0.785*** |
| | | | (0.062) | (0.037) | | | (0.030) | (0.064) |
| Adjusted $R^2$ | 0.743 | 0.674 | 0.831 | 0.780 | 0.719 | 0.752 | 0.817 | 0.834 |
| Provinces | 1790 | 1750 | 440 | 1183 | 1775 | 1765 | 815 | 808 |
| Countries | 84 | 86 | 53 | 85 | 138 | 32 | 106 | 32 |
| Country FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Ruggedness | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Population density (1975) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

The unit of observation is a province. OLS estimates are reported in all columns. The sample is split according to country-wide province features. In columns 1-4, odd (even) columns only consider provinces located in countries with a below (above)-median average province population, with population counts based on GHS data for 1975. In columns Columns 5-8, odd (even) columns only consider countries with a below (above)-median number of provinces. Robust standard errors clustered at the country level are reported in parentheses. The regressions control for country fixed-effects. Statistical significance is represented by * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

**Table S7. Policy Implications: The Role of Democracy (Polarization)**

| Sample: | Restricted sample | | | |
|---|---|---|---|---|
| Data source: | Polity | | Freedom | |
| Dependent variable: | Urban share | Primate share | Urban share | Primate share |
| | (1) | (2) | (3) | (4) |
| *Panel A: Cross sectional* | | | | |
| Polar. × Democracy | -0.108* | -0.011 | -0.157** | -0.021 |
| | (0.061) | (0.046) | (0.062) | (0.046) |
| Polar. × Intermediate regime | -0.114* | -0.239*** | -0.048* | -0.140*** |
| | (0.064) | (0.087) | (0.027) | (0.048) |
| Polar. × Autocracy | -0.052** | -0.109*** | -0.048* | -0.143*** |
| | (0.023) | (0.038) | (0.026) | (0.052) |
| Adjusted $R^2$ | 0.526 | 0.470 | 0.511 | 0.459 |
| P(Test: Democracy = Int. regime) | .947 | .022 | .111 | .078 |
| P(Test: Int. regime = Autocracy ) | .358 | .172 | .99 | .965 |
| P(Test: Democracy = Autocracy) | .386 | .107 | .107 | .083 |
| *Panel B: Longitudinal* | | | | |
| Polar. × Democracy | -0.013 | -0.003 | -0.044 | -0.014 |
| | (0.029) | (0.030) | (0.036) | (0.026) |
| Polar. × Intermediate regime | -0.069* | -0.143*** | -0.001 | -0.105** |
| | (0.039) | (0.047) | (0.019) | (0.044) |
| Polar. × Autocracy | -0.002 | -0.060 | -0.012 | -0.052 |
| | (0.022) | (0.036) | (0.024) | (0.038) |
| Urban share (1975) | 0.552*** | | 0.574*** | |
| | (0.065) | | (0.059) | |
| Primate share (1975) | | 0.813*** | | 0.814*** |
| | | (0.040) | | (0.037) |
| Adjusted $R^2$ | 0.726 | 0.823 | 0.725 | 0.820 |
| P(Test: Democracy = Int. regime) | .248 | .015 | .297 | .079 |
| P(Test: Int. regime = Autocracy ) | .121 | .166 | .698 | .354 |
| P(Test: Democracy = Autocracy) | .766 | .232 | .441 | .404 |
| Provinces | 2627 | 1245 | 2776 | 1313 |
| Countries | 117 | 103 | 131 | 110 |
| Country FE | ✓ | ✓ | ✓ | ✓ |
| Ruggedness | ✓ | ✓ | ✓ | ✓ |
| Population density (1975) | ✓ | ✓ | ✓ | ✓ |

The unit of observation is a province. OLS estimates are reported in all columns. Robust standard errors clustered at the country level are reported in parentheses. Polarization is interacted with variables capturing the degree of democratization in countries in 1975. Columns 1-2: Data on democracy is derived from the variable "Polity" by the Polity IV Project (16). Democracy refers to the third of countries with the highest Polity score. Autocracy refers to the third of countries with the lowest Polity score. Intermediate refers to the remaining third of countries with an intermediate Polity score. Columns 3-4: Data on democracy is derived from the variable "Freedom Status" by Freedom House (17) evaluating political rights and civil liberties (accessed via the Quality of Government data catalogue). Democracy refers to countries classified as "Free". Autocracy refers to countries classified as "Not Free". Intermediate refers to countries classified as "Partly Free". The regressions control for country fixed-effects. Statistical significance is represented by * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

## Supplementary results

In Table S8 we show results when regressing conflict measures on ethno-linguist fractionalization and polarization. In the table, we report the results of cross-sectional regressions at the province level, covering 3,170 provinces across 151 countries. Our ethno-linguistic diversity measures remain the same as throughout the paper. We have three dependent variables: count of conflict incidents in a province from 1975 to 2015 (estimated as a Poisson count model) in columns 1 and 2, conflict incidence (i.e. equals 1 if at least 1 conflict event present within 1975-2015) which is the extensive margin in columns 3 and 4, and count of incidents in a province conditional on there being a least one incident (also done as a Poisson) which is the intensive margin in columns 5 and 6. The Poisson overall count in columns 1 and 2 covers aspects of both the intensive and extensive margins– whether there are zero conflict incidents and, when positive, how many. To construct these variables, we draw on disaggregate data from "Geographical Research on War, United Platform" (GrowUP, (12)).

We generally find a strong and statistically significant association between our ethno-linguistic diversity measures and these armed conflict measures, especially for fractionalization. This table is consistent with the view that part of the costs of bigger cities in ethno-linguistically diverse areas could be related to higher risk of political tensions and violence.

**Table S8. Ethno-linguistic Diversity and Conflict**

| Sample | Full sample | | | | | |
|---|---|---|---|---|---|---|
| Dependent variable: | Overall | | Extensive margin | | Intensive margin | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Fractionalization | 0.757** | | 0.114** | | 0.339*** | |
| | (0.297) | | (0.053) | | (0.121) | |
| Polarization | | 0.410*** | | 0.050** | | 0.213* |
| | | (0.122) | | (0.025) | | (0.109) |
| Mean Dep. var. | 4.021 | 4.022 | .218 | .218 | 18.416 | 18.416 |
| Adjusted $R^2$ | | | 0.616 | 0.614 | | |
| Pseudo $R^2$ | .672 | .668 | | | .554 | .552 |
| Provinces | 3169 | 3169 | 3166 | 3166 | 691 | 691 |
| Countries | 154 | 154 | 151 | 151 | 87 | 87 |
| Country FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Ruggedness | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Population density (1975) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

The unit of observation is a province. In columns 1 and 2, the dependent variable is a count variable of the total number of events between 1975 and 2015. In columns 3 and 4, the dependent variable is a dummy indicating conflict incidence. Columns 5 and 6 restrict the sample to provinces with at least 1 conflict event. Poisson estimates are reported in columns 1, 2, 5, and 6, and OLS estimates in columns 3-4. The regressions control for country fixed-effects. Robust standard errors are reported in parentheses, clustered at the country level. Statistical significance is represented by * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

## References


1. M Schiavina, S Freire, K MacManus, GHS population grid multitemporal (1975, 1990, 2000, 2015) r2019a. European Commission, Joint Research Centre (JRC) (2019) Dataset (url).
2. M Pesaresi, A Florczyk, M Schiavina, M Melchiorri, L Maffenini, GHS settlement grid, updated and refined regio model 2014 in application to ghs-built r2018a and ghs-pop r2019a, multitemporal (1975-1990-2000-2015), r2019a. European Commission, Joint Research Centre (JRC) (2019) Dataset (url).
3. MP Lewis, GF Simons, CD Fennig, Ethnologue: Languages of the world, nineteenth ed. SIL International, Dallas. (2018).
4. A Alesina, S Michalopoulos, E Papaioannou, Ethnic inequality. *J. Polit. Econ.* **124**, 428–488 (2016).
5. R Hodler, M Valsecchi, A Vesperoni, Ethnic geography: Measurement and evidence. CEPR Discussion Paper No. 12378 (2017).
6. K Desmet, J Gomes, I Ortuño-Ortín, The geography of linguistic diversity and the provision of public goods (2018).
7. SE Fienberg, , et al., An iterative procedure for estimation in contingency tables. *The Annals Math. Stat.* **41**, 907–917 (1970).
8. N Nunn, D Puga, Ruggedness: The blessing of bad geography in africa. *Rev. Econ. Stat.* **94**, 20–36 (2012) Dataset (url).
9. SJ Riley, SD DeGloria, R Elliot, A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain J. sciences* **5**, 23–27 (1999).
10. B Lehner, G Grill, Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrol. Process.* **27**, 2171–2186 (2013).
11. NB Weidmann, D Kuse, KS Gleditsch, The geography of the international system: The Shapes dataset. *Int. Interactions* **36**, 86–106 (2010) Dataset (url).
12. L Girardin, P Hunziker, LE Cederman, NC Bormann, M Vogt, Growup-geographical research on war, unified platform. ETH Zurich (2015) Dataset (url).
13. WD Nordhaus, Geography and macroeconomics: New data and new findings. *Proc. Natl. Acad. Sci.* **103**, 3510–3517 (2006) Dataset (url).
14. E Oster, Unobservable selection and coefficient stability: Theory and evidence. *J. Bus. & Econ. Stat.* **37**, 187–204 (2019).
15. A Moreno-Monroy, L Maffenini, P Veneri, GHS-FUA r2019a - ghs functional urban areas, derived from ghs-ucdb r2019a, (2015), r2019a. European Commission, Joint Research Centre (JRC) (2019) Dataset (url).
16. MG Marshall, TR Gurr, K Jaggers, Political regime characteristics and transitions, 1800-2008, Polity IV Project (2012) Dataset (url).
17. Freedom House, Freedom in the world - country and territory ratings and statuses, 1973-2018 (2019) Dataset (url).