Ethnolinguistic 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]Centre for Economic Performance, London School of Economics, London WC2A2AE, United Kingdom; [b]Department of Economics, Faculty of Business and Economics (HEC Lausanne), University of Lausanne, 1015 Lausanne, Switzerland; and [c]Faculty of Business and Economics, University of Basel, 4002 Basel, Switzerland

Edited by Eric Sheppard, University of California, Los Angeles, CA and accepted by Editorial Board Member Susan Hanson May 19, 2020 (received for review February 5, 2020)

This article shows that higher ethnolinguistic 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 worldwide dataset at a fine-grained level on urban settlement patterns and ethnolinguistic population composition. For 3,540 provinces of 170 countries, we find that increased ethnolinguistic 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 ethnolinguistic diversity on urbanization patterns.

ethnolinguistic diversity | fractionalization | urbanization | 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 intergroup 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 a global empirical investigation of the nexus between ethnolinguistic 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 ethnolinguistic 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 ethnolinguistic 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 ethnolinguistic 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 ongoing 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 fine-grained dataset 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 US states or German Bundesländer (see *SI Appendix*, *Data* for details). We identify the effects of ethnolinguistic 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 (16) on geolocalized population and urban boundaries, we first establish a dataset at the 1-km grid level, which distinguishes between city cores, dense towns, semidense 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 regard to local administrative borders and to census bureau qualitative views on what defines urban areas and

Significance

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 ethnolinguistic composition of local populations. We address this gap, drawing on a very detailed dataset on local urban agglomeration and ethnolinguistic diversity. We find that, in multiethnic areas, social tensions arise more easily, discouraging the move to bigger cities. Ethnolinguistically diverse regions feature less urbanization and agglomeration, with potentially profound economic consequences.

Author contributions: U.J.E., J.V.H., D.R., and K.S. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. E.S. is a guest editor invited by the Editorial Board.

Published under the PNAS license.

Data deposition: Replication code and generated data have been deposited in the Harvard Dataverse (https://doi.org/10.7910/DVN/PLDXPD).

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

[2] To whom correspondence may be addressed. Email: j.v.henderson@lse.ac.uk, kurt.schmidheiny@unibas.ch, dominic.rohner@unil.ch, or ulrich.eberle@unil.ch.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2002148117/-/DCSupplemental.

First published June 29, 2020.

Downloaded at Palestinian Territory, occupied on December 30, 2021

المنارة للاستشارات
www.manaraa.com

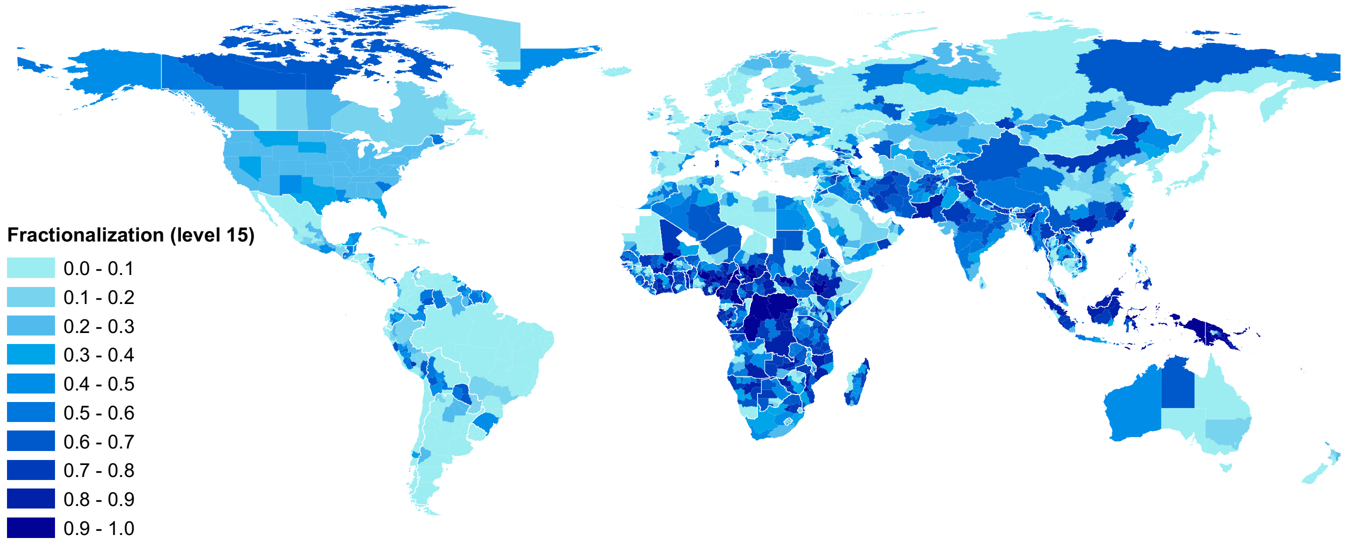

Fig. 1. Global map of ethnolinguistic fractionalization at the province level. Fractionalization is calculated at language tree level 15. See *Data and Methods* for data sources and construction.

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; ref. 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 ethnolinguistic groups. All details of the data construction are relegated to *Data and Methods*.

In Fig. 1, the average ethnolinguistic 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 colors 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. Fig. 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 interprovincial 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 ethnolinguistic diversity, we graph three motivating sets of associations. Fig. 2 displays the association between a conflict measure and ethnolinguistic fractionalization, as well as between the two urban concentration measures and ethnolinguistic fractionalization, for all provinces across the world.

In Fig. 2*A*, we show, with a nonlinear 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; ref. 21), as postulated at the beginning of the article. This is in line with our premise that ethnic diversity may go hand 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, Fig. 2*B* illustrates the correlation at the province level between ethnic fractionalization in 1975 and urban population share in 2015, while Fig. 2*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 ethnolinguistic fractionalization.

While the associations in Fig. 2 are intriguing, below, 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 ($dense$), semidense 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

*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 ref. 17), but our outcomes involve population measures, so we need population data as well as worldwide coverage.

Downloaded at Palestinian Territory, occupied on December 30, 2021

المنارة للاستشارات
www.manaraa.com

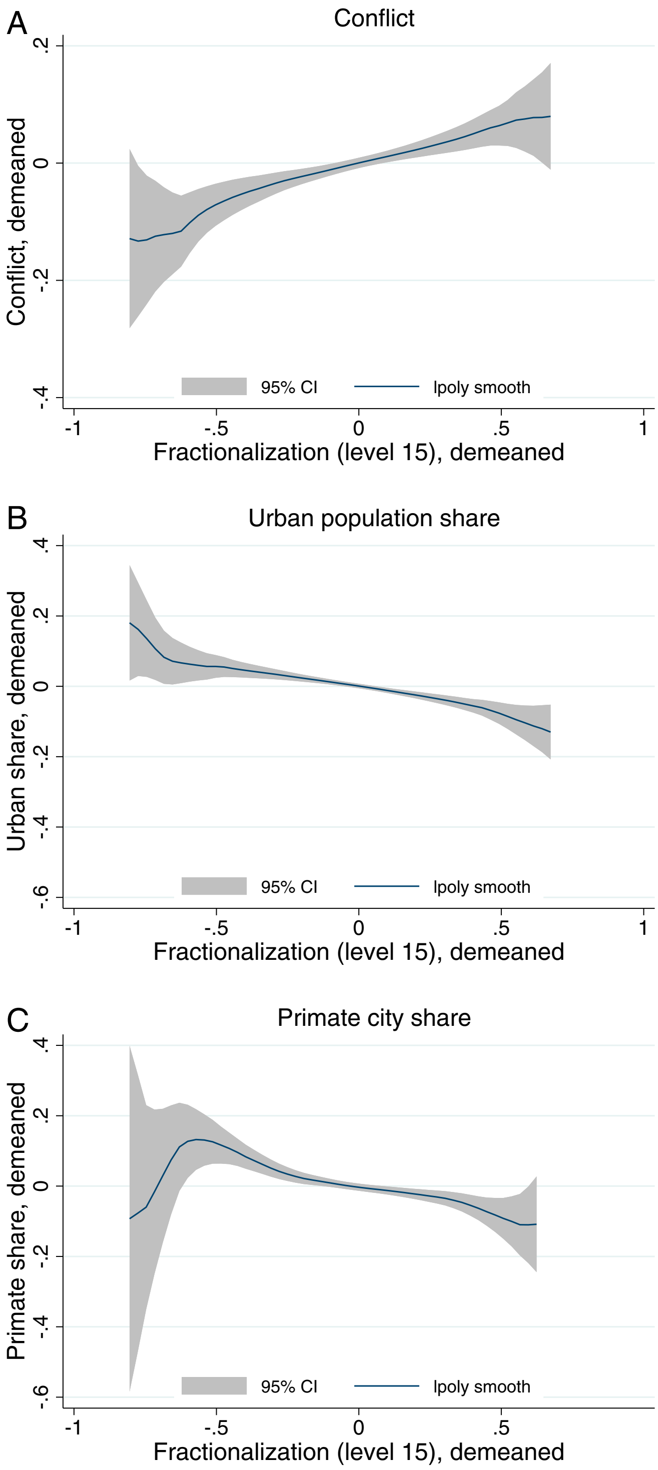

Fig. 2. Distributions and regressions: ethnolinguistic fractionalization, conflict, and urban concentration. The unit of observation is a province. The sample includes 3,540 provinces worldwide. The graphs depict kernel-weighted local polynomial regressions of first degree. The plots show the association 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. (*A*) Conflict is reported for 3,169 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. (*B* and *C*) Urbanization indices for the year 2015 calculated with data from the GHSL. (*B*) Urban share is the share of urban population of a province divided by the total population. (*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.

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

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

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 semidense towns and suburbs (300 people per square km), meaning that, in general, it reports higher urban shares worldwide than the United Nations 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 Eq. **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 Eq. **1**, and we will use a measure of primate city size that attempts to incorporate commuting zones around cities in Eq. **2**.

As noted above, we match the grid cells with fine-grained language information. Our language data from the WLMS are arguably the most precise source currently available, and have recently been used by refs. 22–24. 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 ref. 24 for allocating languages to population in multilinguistic 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 ethnolinguistic 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 *SI 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 the main measure is that it has been linked to both small-scale frictions in public good provision (25, 26) and 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 refs. 1, 2, and 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 *SI Appendix*. Formally, the two measures are defined in the literature (1) as

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

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

Downloaded at Palestinian Territory, occupied on December 30, 2021

المنارة للاستشارات
www.manaraa.com

where M_i designates the total number of groups $m = 1, \dots, M_i$ in province i, and π_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 6,208 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 ethnolinguistic 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, for example, 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 Fig. 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 two 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 ref. 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 as level 15 language affinity. In *SI Appendix*, Fig. S2, 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 ethnolinguistic factors and urbanization patterns by regressing contemporary measures of urban concentration on historical measures of ethnolinguistic 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 three parts: the top two rows are a cross-sectional analysis, the next four rows are longitudinal, by additionally controlling for the past (1975) value of the dependent variable, while the last five rows show the number of observations and the controls used in both types of analyses. Columns are in pairs for different samples and outcomes, and, within each pair, columns are distinguished by the set of controls.

Column 1 in Table 1 for the cross-sectional rows regresses the *Urban share* in a given province in 2015 on presample ethnolinguistic 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 ethnolinguistically homogeneous province (i.e., with ethnolinguistic fractionalization of 0) to a perfectly diverse one (i.e., with ethnolinguistic fractionalization of 1) would be associated with a 13% lower share of the urban population in the province. This change in urbanization corresponds to about half of 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 the cross-sectional regressions could be biased because of omitted variables and reverse causality. For example, urbanization over long periods of time could influence fractionalization.

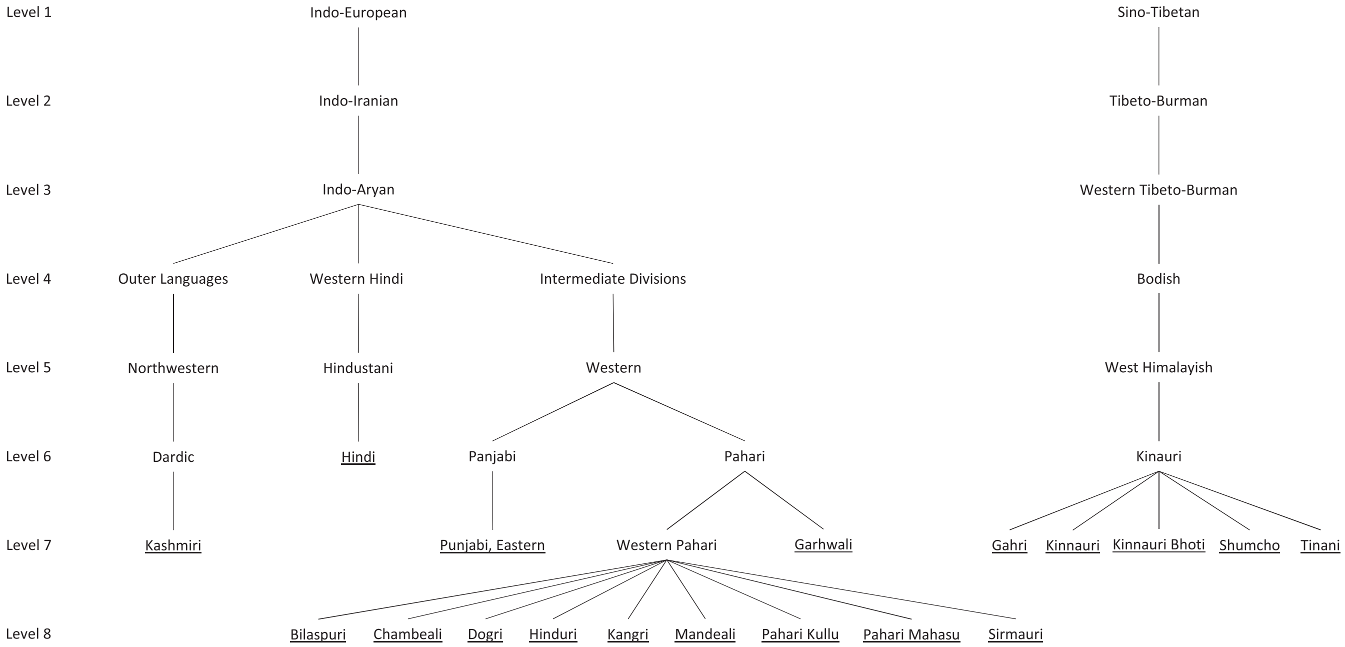

Fig. 3. The use of ethnologue language trees: illustration for the Indian province Himachal Pradesh. The graph depicts the language tree of Himachal Pradesh. The languages of Himachal Pradesh are divided in up to eight levels, with level 1 being the most aggregated and level 8 being the least aggregated level. The endpoint (underlined) of each branch depicts the commonly used name of a language. The language tree is based on data by Ethnologue. Four very minor languages at the extension of Western Pahari are omitted, for presentation purposes.

Downloaded at Palestinian Territory, occupied on December 30, 2021

www.manaraa.com

Table 1. **Ethnolinguistic fractionalization and urbanization patterns**

	Urban share		Primate share			
	Full sample		Full sample		Restricted sample	
	No control	Control	No control	Control	No control	Control
	(1)	(2)	(3)	(4)	(5)	(6)
Cross-sectional regressions						
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
Longitudinal regressions						
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
Observations and controls						
Provinces	3,540	3,540	2,359	2,359	1,623	1,623
Countries	170	170	154	154	138	138
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Ruggedness		Yes		Yes		Yes
Population density (1975)		Yes		Yes		Yes

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 $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

To deal with this, we move, in Table 1, column 1 of the longitudinal regressions, 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 four 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 the longitudinal regressions, we still find a statistically significant negative coefficient, albeit its magnitude is reduced by half compared to the cross-sectional regressions. Of note is that the coefficient of past urban share is sizeable and highly significant, pointing toward a large persistence of urbanization patterns over time. Overall, it is reassuring that, in the longitudinal rows, we continue to find evidence of ethnic fractionalization slowing down the pace of urbanization, after controlling for presample urbanization.

In column 2 of Table 1, we estimate the analogous specifications as in column 1, but controlling, in addition, for terrain ruggedness and population density in 1975 (see *SI Appendix, Data*, for a detailed description of these control variables, and see *SI Appendix*, Table S2 for all estimated coefficients). The results remain very similar, and the coefficients of interest remain statistically significant at the 1% 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 the cross-sectional rows of Table 1 on fractionalization first 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 ethnolinguistic fractionalization. Put differently, ethnolinguistic diversity is associated with having several smaller cities instead of a single megacity. Quantitatively, moving from a fully homogeneous to a fully heterogeneous society (i.e., moving ethnolinguistic fractionalization from 0 to 1) would be associated with an at least 8% lower $Primate\ share$ in columns 5 and 6 in the longitudinal rows, 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 *SI Appendix*. In *SI Appendix*, Table S8, we show that there is a strong and statistically significant association between ethnolinguistic 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, and 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. *SI Appendix, Data*, 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 ethnolinguistic 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, that is, the longitudinal rows in 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

[†] We calculate the maximum bias with conservative assumptions for this context, that is, $\delta = 1$ and $R^2_{max} = 0.9$. See *SI Appendix, Selection on unobserved variables* for more details and calculation.

Downloaded at Palestinian Territory, occupied on December 30, 2021

المنارة للاستشارات
www.manaraa.com

in the language tree for distinguishing different languages. As explained above, our data allow us to compute ethnolinguistic 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 Fig. 3 above, or, say, Arpitan, Romansch, Lombard, and French in non-German Switzerland (*SI Appendix*, Fig. S2).

We graph the pattern of coefficients and their significance in Fig. 4, linking ethnic diversity to urban share, primate share, and conflict. Overall, the results of Fig. 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 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 ethnolinguistic diversity. While the fractionalization measure takes high values for areas with a large number of groups, the main alternative diversity measure defined above, ethnolinguist polarization, reaches high values for situations closer to bimodal 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 *SI Appendix*.

The relationship in the data between our fractionalization and polarization measures is displayed in *SI Appendix*, Fig. S4. After filtering out country averages (*SI Appendix*, Fig. S4*B*), the two diversity measures are highly correlated, although 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 ethnolinguistic 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 Eq. **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 Organisation for Economic Co-operation and Development (OECD) and European Commission (EC) project described in ref. 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 Eq. **1**. We use the broad definition, since FUAs contain population in less dense areas. Using this definition for primate city share in columns 3

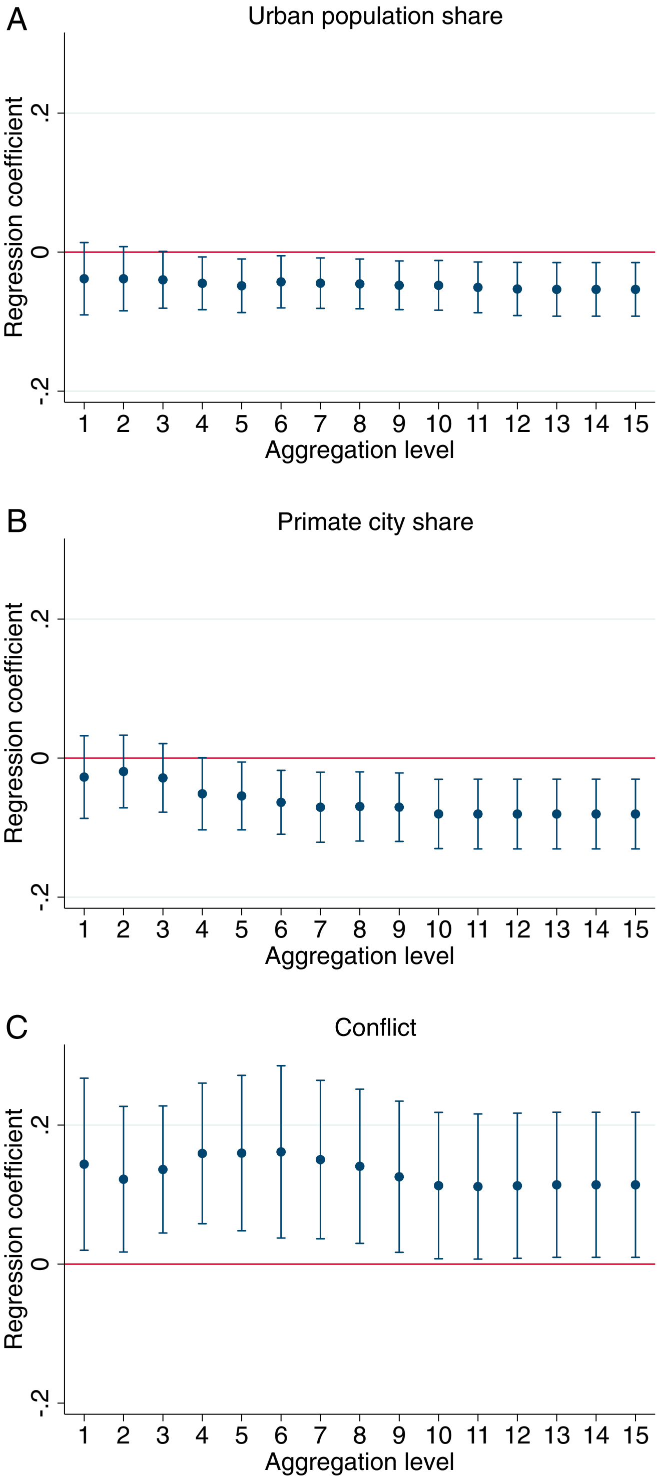

Fig. 4. Ethnolinguistic fractionalization, conflict, and urban concentration: results for different aggregation levels. Regression results of the two measures of urban concentration and conflict incident on ethnolinguistic fractionalization, at all 15 linguistic aggregation levels. (*A* and *B*) The regressions performed control for country fixed effects, ruggedness, and 1975 population density and 1975 outcome variables, as specified in the longitudinal regressions in columns 2 and 6 of Table 1 for aggregation level 15. (*C*) The regressions performed are as specified in column 3 of *SI Appendix*, Table S8. Point estimates are shown as dots, and CIs at the 95% level are shown as bars.

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

Downloaded at Palestinian Territory, occupied on December 30, 2021

المنارة للاستشارات

www.manaraa.com

Table 2. Policy implications: The role of democracy

	Polity		Freedom	
	Urban share	Primate share	Urban share	Primate share
	(1)	(2)	(3)	(4)
Cross-sectional regressions				
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)	0.756	0.001	0.025	0.041
P(Test: Int. regime = Autocracy)	0.305	0.054	0.922	0.52
P(Test: Democracy = Autocracy)	0.2	0.01	0.029	0.018
Longitudinal regressions				
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)	0.297	0.001	0.559	0.071
P(Test: Int. regime = Autocracy)	0.288	0.18	0.935	0.255
P(Test: Democracy = Autocracy)	0.847	0.012	0.519	0.449
Observations and controls				
Provinces	2,627	1,245	2,776	1,313
Countries	117	103	131	110
Country FE/Base controls	Yes	Yes	Yes	Yes

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 and 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 and 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 catalog). Democracy refers to countries classified as "Free." Autocracy refers to countries classified as "Not Free." Intermediate (Int.) refers to countries classified as "Partly Free." The regressions control for country fixed effects. Statistical significance is represented by *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

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 MAUP is not driving our results.

Discussion and Role of Policies

The above results tell a stark story of ethnolinguistic diversity slowing down urbanization and urban concentration, and hence potentially affecting economic development. Still, there may be room for policies to dampen the extent of this relationship. One natural candidate for a 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, and 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 the longitudinal regressions. Results are reported in Table 2. In columns 1 and 2, the democracy measure is taken from the Polity IV project (38), while, in columns 3 and 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 in the bottom three rows of the cross-sectional and the longitudinal regressions 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 *SI Appendix*. Generated data

[§] 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).

Downloaded at Palestinian Territory, occupied on December 30, 2021

المنارة للاستشارات
www.manaraa.com

and code to generate variables and results are publicly available at the Harvard Dataverse (https://doi.org/10.7910/DVN/PLDXPD) (40).

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. U.J.E. gratefully acknowledges financial support from the Swiss National Science Foundation Doc.Mobility Fellowship P1LAP1_181253 and especially thanks the Centre for Economic Performance at the London School of Economics and Political Science for hosting him during 2018–2020. D.R. gratefully acknowledges financial support from the European Research Council Starting Grant POLICIES FOR PEACE-677595 and warmly thanks University of British Columbia for the hospitality during 2018–2019 when the first draft of this paper was written. K.S. is grateful to University of California, Berkeley for the hospitality during the academic year 2017–2018 when this project took shape.

1. J. G. 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**, 1310–1342 (2012).
3. J. Esteban, M. Morelli, D. Rohner, Strategic mass killings. *J. Polit. Econ.* **123**, 1087–1132 (2015).
4. H. F. Mueller, D. Rohner, D. Schönholzer, "The peace dividend of distance: Violence as interaction across space" (Discuss. Pap. 11897, Centre for Economic Policy Research, 2017).
5. V. Henderson, The urbanization process and economic growth: The so-what question. *J. Econ. Growth* **8**, 47–71 (2003).
6. J. V. Henderson, "Urbanization and growth" in *Handbook of Economic Growth*, P. Aghion, S. Durlauf, Eds. (Elsevier, 2005), vol. 1, pp. 1543–1591.
7. G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, 1949).
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. Polit. Econ.* **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. A. F. Ades, E. L. Glaeser, Trade and circuses: Explaining urban giants. *Q. J. Econ.* **110**, 195–227 (1995).
13. K. Desmet, J. V. Henderson, "The geography of development within countries" in *Handbook of Regional and Urban Economics*, G. Duranton, V. Henderson, W. Strange, Eds. (Elsevier, 2015), vol. 5, pp. 1457–1517.
14. S. S. Rosenthal, W. C. Strange, "Evidence on the nature and sources of agglomeration economies" in *Handbook of Regional and Urban Economics*, V. Henderson, J. F. Thisse, Eds. (Elsevier, 2004), vol. 4, pp. 2119–2171.
15. Statistics Canada, Census data Canada. https://www12.statcan.gc.ca/census-recensement/index-eng.cfm. Accessed 4 April 2020.
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)*. Dataset, http://data.europa.eu/89h/42e8be89-54ff-464e-be7b-bf9e64da5218. Accessed 4 April 2020.
17. M. P. De Bellefon, P. P. Combes, G. Duranton, L. Gobillon, C. Gorin, Delineating urban areas using building density. *J. Urban Econ.*, 10.1016/j.jue.2019.103226 (2019).
18. M. P. Lewis, G. F. Simons, C. D. Fennig, *Ethnologue: Languages of the World* (SIL International, Dallas, TX, ed. 19, 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, L. E. Cederman, N. C. Bormann, M. Vogt, *GROWup-Geographical Research on War, Unified Platform* (ETH Zurich, 2015). Dataset, http://growup.ethz.ch/. Accessed 4 April 2020.
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" (Discuss. Pap. 12378, Centre for Economic Policy Research, 2017).
24. K. Desmet, J. F. Gomes, I. Ortuño-Ortín, The geography of linguistic diversity and the provision of public goods. *J. Dev. Econ.* **143**, 102384 (2020).
25. A. Alesina, R. Baqir, W. Easterly, Public goods and ethnic divisions. *Q. J. Econ.* **114**, 1243–1284 (1999).
26. A. Alesina, E. L. Ferrara, Ethnic diversity and economic performance. *J. Econ. Lit.* **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*. Dataset, http://data.europa.eu/89h/0c6b9751-a71f-4062-830b-43c9f432370f. Accessed 4 April 2020.
30. J. G. Altonji, T. E. Elder, C. R. Taber, Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *J. Polit. Econ.* **113**, 151–184 (2005).
31. E. Oster, Unobservable selection and coefficient stability: Theory and evidence. *J. Bus. Econ. Stat.* **37**, 187–204 (2019).
32. A. I. Moreno-Monroy, M. Schiavina, P. Veneri, Metropolitan areas in the world. Delineation and population trends. *J. Urban Econ.*, 10.1016/j.jue.2020.103242 (2020).
33. C. E. 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. A. S. Fotheringham, D. W. Wong, The modifiable areal unit problem in multivariate statistical analysis. *Environ. Plann. A* **23**, 1025–1044 (1991).
35. D. A. Freedman, Ecological inference and the ecological fallacy. *Int. Encycl. Soc. Behav. Sci.* **6**, 1–7 (1999).
36. H. Hegre, Toward a democratic civil peace? Democracy, political change, and civil war, 1816–1992. *Am. Polit. Sci. Rev.* **95**, 33–48 (2001).
37. P. Collier, D. Rohner, Democracy, development, and conflict. *J. Eur. Econ. Assoc.* **6**, 531–540 (2008).
38. M. G. Marshall, T. R. Gurr, K. Jaggers, *Polity IV Project: Political Regime Characteristics and Transitions, 1800-2008* (2012). Dataset, https://www.systemicpeace.org/inscrdata.html. Accessed 4 April 2020.
39. Freedom House, *Freedom in the World–Country and Territory Ratings and Statuses, 1973-2018* (2019). Dataset, https://freedomhouse.org/content/freedom-world-data-and-resources. Accessed 4 April 2020.
40. U. J. Ulrich, J. V. Henderson, D. Rohner, K. Schmidheiny, Replication data for: Ethnolinguistic diversity and urban agglomeration. Harvard Dataverse, V1. https://doi.org/10.7910/DVN/PLDXPD. Deposited 21 May 2020.

Downloaded at Palestinian Territory, occupied on December 30, 2021

المنارة للاستشارات
www.manaraa.com