Published in final edited form as:
Am Econ Rev. 2012 June ; 102(4): 1508–1539. doi:10.1257/aer.102.4.1508.

The Origins of Ethnolinguistic Diversity

Stelios Michalopoulos*

Stelios Michalopoulos: stelios.michalopoulos@tufts.edu

*Department of Economics, Tufts University, Braker Hall, 8 Upper Campus Rd, Medford, MA 02155 and the Institute for Advanced Study, Princeton, NJ, 08540

Abstract

This study explores the determinants of ethnolinguistic diversity within as well as across countries shedding light on its geographic origins. The empirical analysis conducted across countries, virtual countries and pairs of contiguous regions establishes that geographic variability, captured by variation in regional land quality and elevation, is a fundamental determinant of contemporary linguistic diversity. The findings are consistent with the proposed hypothesis that differences in land endowments gave rise to location-specific human capital, leading to the formation of localized ethnicities.

Keywords

Ethnic Diversity; Geography; Growth; Languages; Human Capital

Ethnicity has been widely viewed in the realm of social sciences as instrumental for the understanding of socioeconomic processes. Within economics, there has been a large and growing literature beginning with Mauro (1995), that uses indexes of ethnolinguistic diversity either as an instrument or as an explanatory variable for various economic indicators (e.g., Easterly and Levine (1997), Alesina et al. (2003) and Banerjee and Somanathan (2006), among others). Similarly, within sociology, anthropology, political science, psychology and history, the volume of work investigating the causes and effects of ethnicity attests to its paramount importance.[1]

The goal of this study is to empirically uncover the common exogenous features that produce the observed diversity at a global scale. It does so by bringing into the foreground the role of geographic heterogeneity in separating populations leading to the formation of distinct linguistic groups. The empirical investigation, conducted at various levels of spatial aggregation, establishes that geographic variability, captured by variation in regional land quality and elevation, is a fundamental determinant of contemporary ethnic diversity. It also shows that the link between geography and ethnicity is particularly strong across territories whose historically indigenous populations constitute a significant fraction of today's inhabitants.

[1]See Alesina and Ferrara (2005) and Hale (2004) for a survey within economics and social sciences, respectively.

Linking ethnicity to geographic variability has implications for the large empirical literature that uses the former as an explanatory variable. Specifically, the finding that geographic heterogeneity partially shapes contemporary ethnic diversity provides a justification for treating the latter as predetermined with respect to contemporary economic policies. However, geographic characteristics are likely to have an independent effect on economic indicators beyond their effect on cultural diversity.

In the empirical section I employ new data on agricultural suitability at a resolution of 0.5 by 0.5 decimal degrees to construct the distribution of land quality at a regional and country level. Such disaggregated data allow for the econometric analysis to be conducted at various levels of aggregation. First, I show the importance of geography in shaping ethnic diversity across countries. Nations characterized by more diverse land and elevation attributes exhibit higher levels of linguistic diversity. This highlights the fundamental role that heterogeneous regional land endowments have played in the formation of ethnically diverse societies. The results are robust to accounting for alternative hypotheses regarding the formation of ethnolinguistic groups.

Second, to mitigate concerns related to the endogeneity of contemporary political boundaries, inherent to the literature on cross-country regressions, I arbitrarily divide the world into geographic entities, called virtual countries. Ethnic diversity, measured by the number of languages spoken in each virtual country, is systematically related to the underlying heterogeneity in land quality for agriculture. At the same time, the empirical analysis reveals that regions with more variable terrain sustain more ethnically fragmented societies. Overall, geographically diverse territories give rise to more ethnic groups. Arguably, modern states in their quest to build unitary nation-states have largely affected local languages and identities via political centralization and state-sponsored education. As such it is crucial to account for these state-specific histories. Unlike a cross-country analysis, this is feasible in the context of virtual-country regressions, where I show that the results are robust to the inclusion of country fixed effects.

A third noteworthy feature of the empirical analysis is that it focuses on pairs of contiguous regions, by taking further advantage of information on the traditional location of ethnic groups. Pairing each 0.5 by 0.5 decimal degrees cell with its immediate neighbors, I find that differences in land quality and elevation within dyads of adjacent regions negatively affect ethnic similarity. The latter is proxied by the percentage of common languages spoken within a regional pair. The dyadic analysis allows for the inclusion of region fixed effects neutralizing any bias induced by local, within-country unobserved heterogeneity, such as differences in regional integration along the process of nation building. The evidence demonstrates that (i) local differences in land endowments systematically increase regional ethnic diversity and (ii) the spatial arrangement of a heterogeneous geography matters in determining the degree of overall cultural heterogeneity.

Historically, major population movements have been reshuffling the distribution of ethnicities across as well as within countries. In the last 500 years, in particular, several regions have experienced dramatic changes in their linguistic endowments via conquests, slavery, migrations, and colonization. In absence of estimates on historical ethnic diversity

the impact of these events is hard to quantify. Nevertheless, to the extent that geographic variability increases ethnic diversity by reducing mobility among groups, eventually causing a cultural drift, one would expect geography to be less important in places where native populations have been dramatically reduced and displaced. This is shown to be the case across all levels of aggregation. Linguistic diversity no longer exhibits a systematic link with the underlying geography in virtual countries, pairs of adjacent regions, and real countries where the majority of contemporary inhabitants cannot trace their ancestry in 1500 AD to that same geographic area. This is an important finding because it shows that unlike most countries where a large fraction of ethnic diversity may be treated as predetermined by geography, in places where recent settlers constitute the majority, ethnic diversity can no longer be treated as such.

Though the major contribution of this paper is to empirically identify the geographic origins of ethnolinguistic diversity, section *IV* discusses the possible mechanisms via which geographic heterogeneity may operate on the formation of ethnolinguistic groups. It suggests that differences in land endowments across regions gave rise to location-specific human capital, diminishing population mobility and leading to the formation of localized ethnicities. A prediction that follows this view is that groups found in multiple separate territories should display similarities in their subsistence pattern. Consistent with this prediction I show that within Africa there is a strong correlation of subsistence practices across non-adjacent partitions of the same language group.

The rest of the paper is organized as follows. Section *I* relates the empirical findings to the literature in economics as well as anthropology, linguistics, ecology and sociology. Section *II* presents the data. Section *III* constitutes the main part of the empirical analysis. This is conducted in a (i) cross-country, (ii) cross-virtual country, and (iii) cross-pair of adjacent regions framework. It includes the robustness checks and concludes by showing how a significant presence of recent immigrants affects the link between geographic and ethnic diversity. In section *IV* I describe the possible mechanisms via which geographic variability may lead to linguistic differentiation and I provide evidence consistent with the specific human capital hypothesis. Section *V* summarizes the key findings and concludes.

I. Related Literature

Within economics, Ahlerup and Olsson (2008) is the only other paper that investigates the roots of ethnic diversity at a global scale. The authors provide a theory where ethnic groups endogenously emerge over time among peripheral populations in response to an insufficient supply of public goods. Using data on the duration of human settlements since prehistoric times, they show that countries where modern humans settled earlier sustain higher ethnic diversity today. In my empirical analysis migratory distance from Addis Ababa, which systematically predicts early human settlements and is shown by Ashraf and Galor (2008) to be a strong predictor of genetic diversity across countries, enters with the expected sign, i.e. countries with a large migratory distance from East Africa sustain fewer ethnic groups. To understand the quantitative importance of this channel, note that variation in geography and migratory distance from East Africa have comparable effects on linguistic diversity across countries.

The findings of the current study provide evidence bridging the two opposing strands of thought regarding the formation of ethnic identity. On the one hand, establishing that deep geographic determinants explain a significant fraction of the observed linguistic diversity is consistent with the primordial view qualifying ethnic groups as deeply rooted, clearly-drawn entities (Geertz (1967)). On the other hand, showing that geography alone cannot account for all the variation in contemporary ethnic diversity, and in particular showing that there is no link between geographic and ethnic diversity in those places where the historically indigenous populations today constitute a minority provides evidence in favor of the constructivist or instrumentalist school of thought. This view highlights the contingent and situational character of ethnicity with modern states' policies being a significant determinant of the observed diversity (Barth (1969)).

Outside economics, studies in the fields of linguistics, anthropology and sociology have examined the relationship between the environment and cultural diversity. For example, Mace and Pagel (1995) use historical maps of native populations in North America at the time of contact by Europeans and show that more diverse habitats sustained more native groups. Also, Nettle (1998, 1999) looks at the global distribution of languages and finds that countries characterized by low rainfall and short growing seasons sustain fewer languages.

The link between variable land endowments and ethnic diversity established here has a striking parallel to the relationship between biodiversity and variation within species. Darwin's observations that ecologically diverse places would bring about and sustain variation within finches is of particular relevance.[2] Along the same lines, this study shows that geographic heterogeneity across regions fosters ethnic diversity.

Other studies have shown the relationship between species diversity and linguistic diversity. For example, Harmon (1996) notes that countries with larger endemism in vertebrate and flowering plant species also sustain higher linguistic diversity. Along the same lines, Moore et al. (2002) show that vertebrate species diversity and cultural diversity have similar spatial distribution within Sub-Saharan Africa, both increasing in rainfall and net primary productivity and Sutherland (2003) shows that countries with high linguistic diversity also have high bird and mammal diversity (see Maffi (2005) for a thorough literature review). The consistent finding in the literature regarding the negative relationship between absolute latitude and linguistic diversity has been linked to the very strong correlation between latitude and both temperature and climatic variability. According to Nettle (1998) where climate is variable populations are forced to use wider ecological niches homogenizing them linguistically. Also, regions further away from the equator have lower temperatures, thus lower net primary productivity, and are also subject to lower ultraviolet radiation (UVR).

By analyzing the origins of ethnic diversity this paper belongs to an emerging strand of literature within economics that examines the deep-rooted determinants of economic performance. Starting with Diamond (1997), who highlights the crucial role of geography in affecting long-run development via the differential timing of the transition to agriculture,

[2]Darwin (Originally 1839, Reprinted in 2006) observed that a certain ecological niche was giving rise to an optimal shape of the finches' beaks.

economists are becoming increasingly interested in uncovering (pre)historic forces that have shaped contemporary income levels. Galor and Moav (2002) and Galor and Michalopoulos (2011), for example, argue that the Neolithic Revolution triggered an evolutionary process that affected comparative development. Comin, Easterly and Gong (2010) find that very old history of technology adoption is a significant determinant of today's economic outcomes. Others are investigating the role of genetics, including Spolaore and Wacziarg (2009), who show the effect of genetic distance on the pairwise income differences between countries, and the recent study by Ashraf and Galor (2008) which documents a non-monotonic effect of genetic diversity on comparative development. This paper adds to this literature by showing that geographic diversity is an important determinant of linguistic diversity, a societal attribute that has been extensively found to correlate with various measures of economic performance.

II. Empirical section

A. The Data Sources

The study constructs and uses a variety of geographic indicators. For the cross-country and cross-virtual country analysis standard geographic measures like elevation, temperature and precipitation are derived using the G-Econ (2006) database. For the dyadic regressions alternative data sets are used in order to match the resolution of the land quality dataset.

The global data on land quality for agriculture were assembled by Ramankutty et al. (2002) to investigate the effect of future climate change on contemporary agricultural suitability. This dataset provides information on land quality at a resolution of 0.5 by 0.5 decimal degrees. In total there are 64,004 observations.

Each observation takes a value between 0 and 1 and represents the probability that a particular grid cell may be cultivated. In order to construct this index, the authors first empirically estimate the probability density function of the percentage of croplands around 1990 with respect to climate and soil characteristics. They then combine the derived probability with data on climate and soil quality to predict regional suitability for agriculture at the resolution of 0.5 degrees latitude by 0.5 degrees longitude worldwide. The climatic characteristics are based on mean-monthly climate conditions for the 1961–1990 period and capture (i) temperature (ii) precipitation and (iii) potential sunshine hours. All the climatic conditions, monotonically though weakly increase the suitability of land for agriculture. Regarding the soil suitability the traits considered are a measure of the total organic content (carbon density) and the nutrient availability (soil pH). The relationship of these indexes with agricultural suitability is non-monotonic. Low and high values of pH limit cultivation potential, since these values signal that soils are too acidic or too alkaline, respectively. In the web Appendix details on the data sources and the exact formulas used in the construction of the land quality index are presented.

This detailed dataset provides an accurate description of the global distribution of land quality for agriculture. Map 1 in the Appendix shows the worldwide distribution of land quality. Using these raw global data I construct the distribution of land quality at the desired level of aggregation.

With respect to the cross-country, cross-virtual country and cross-pair of adjacent regions analysis, ethnic diversity is constructed using information on the location of linguistic groups. In the case of virtual and real-country regressions the number of languages within each geographic unit provides a measure of the overall ethnolinguistic diversity. In the adjacent region analysis, an index of ethnic similarity is constructed by calculating the percentage of common languages, that is, the number of common languages over the total number of languages spoken within a pair of adjacent regions. Data on the location of linguistic groups' homelands are obtained from the WLMS (2006) database. This dataset covers most of the world and is accurate for the years between 1990 and 1995. Languages are based on the 15th edition of the Ethnologue (2005) database. To identify which languages are spoken within each unit of analysis I use the information on the location of language polygons. Each of these polygons delineates a traditional linguistic homeland; populations away from their homelands (e.g., in cities, refugee populations, etc.) are not mapped. Linguistic groups of unknown location, widespread languages,[3] and extinct languages, are not mapped and thus not considered in the empirical analysis. The only exception for not mapping widespread languages is the case of the English language, which is mapped for the United States. Note that the Ethnologue (2005) database also records the population of a language group. However, this information is available only at the country level and is listed for widely varying census years, so it cannot be used as the basis for constructing a fractionalization measure at the virtual country level.

B. The Properties of Geographic Variability

The distribution of land quality varies considerably across regions and countries. For example, Figure 1 in the Appendix plots the distribution of regional land quality for Greece and Nepal. In Greece the quality of land is concentrated around high values with an average quality of 0.81, and a variation of 0.12. On the other hand, land quality in Nepal averages 0.45 and differs significantly across regions with a variation of 0.36, featuring an almost uniform distribution across all levels of land quality. Similarly, Nepal exhibits a much larger variation in altitudes compared to Greece.

The variation in land quality within the respective unit of analysis and the standard deviation of elevation are the statistics used to capture the degree of geographic heterogeneity. These measures proxy for the frictions that geography exerts on populations' interactions with larger variance of land quality and elevation limiting mobility among groups leading over time to linguistic differentiation. Indeed, going back to the example of Greece and Nepal, according to the Ethnologue, there are 14 linguistic groups in Greece compared to the highly linguistically fragmented society of Nepal with 107 languages. Note that apart from their very different geographic endowments Nepal and Greece are similar along several dimensions. First, they have similar land mass. Second, they had comparable levels of early economic development as proxied by the population density in 1500 AD. Third, neither country had a European colonizer. Fourth, in both countries 100 percent of the current population can trace its ancestors within the same political boundaries as early as 1500 AD,

[3]According to the WLMS (2006) widespread languages are national or regional trade languages which cannot be properly mapped as polygons since they overlap a large number of other language areas.

and finally Nepal is located twice further from East Africa than Greece, if anything biasing the comparison in favor of finding more linguistic groups in Greece.

Comprehensive data on the timing of the formation of language groups are not available, with few studies documenting the evolution of specific language families. For example, Gray and Atkinson (2003) demonstrate that Indo-European languages expanded with the spread of agriculture from Anatolia around 8000–9500 years BP (before present). The language tree they construct provides information on the timing of linguistic divergence within the Indo-European group. According to this, at 7000 years BP Greek and Armenian languages diverged. At 5000 years BP, Italic, Germanic, Celtic and Indo-Iranian families diverged, and at 1750 years BP the Germanic languages split between West Germanic (German, Dutch, English) and North Germanic (Danish and Swedish). Similarly, according to Nurse and Philippson (2003) the spread of the Bantu languages within Africa took place between 3000 BC and 1000 AD. In general, according to Cavalli-Sforza and Cavalli-Sforza (1996) if a population with a common language is split into two groups, it takes about 1,000 to 1,500 years for the development of a mutually unintelligible language.

Hence, using contemporary geographic data to proxy for historical variability in physical geography presents its own potential pitfalls, which merit further discussion. For example, a potential concern is how representative these geographical characteristics are of the period when linguistic groups were being formed. Regarding the elevation index, despite some local natural events and human interventions at a very local scale, overall altitudes have not changed significantly since the retreat of the last Ice Age. Things are more complicated regarding the land quality index. This is because precipitation, temperature and soil properties, which are the basis of this index, may have changed regionally over the last 5,000 years. Thus the measure of land quality is a noisy index of what might have been the true distribution of agricultural quality in the past. On the one hand, this measurement error may be white noise, making it harder to detect a relationship between variation in land quality and linguistic heterogeneity. On the other hand, this measurement error could be systematic; the same forces that reduce ethnic diversity (centralized modern states) may also be associated with human interventions that have a homogenizing effect on the landscape, generating a spurious relationship between variability in land quality and ethnic diversity. This possibility underscores the need for the analysis to be conducted at a level of aggregation where country fixed effects can be accounted for. This is done in the virtual-country regressions whereby the introduction of country fixed effects accounts for any country-level unobserved forces that may have affected both land heterogeneity and linguistic diversity. At the same time, one might argue that such unobserved forces might vary even across regions within a country. In this respect, the inclusion of region fixed effects in the dyadic regressions neutralizes any bias induced by local, within-country unobserved heterogeneity, allowing for a sharper identification of the causal impact of geographic heterogeneity on linguistic diversity.

It is certainly the case that humans can alter land quality via technological innovations, for example the introduction of the heavy plough may make environments more favorable to agriculture. The issue is whether variation in land quality is affected by linguistic diversity. Although one cannot rule out entirely the possibility of reverse causality running from

exogenous group-specific subsistence practices to the characteristics of land quality, this would likely have a small effect. For example, Diamond (2005) describes several cases of cultures whose subsistence practices were not sustainable given the underlying agricultural capabilities, eventually triggering an environmental collapse tantamount to these cultures' own demise. Nevertheless, to alleviate concerns related to the possible endogeneity of the soil characteristics and to the extent that climate is less prone to human interventions, in the cross-country and cross-virtual country analysis I show that results are similar when I use variation in the climatic suitability to capture heterogeneity in agricultural endowments.

Having discussed the properties of geographic variability I now turn to the main empirical results.

III. Evidence

A. Cross-Country Analysis

I start the empirical analysis with the investigation of the geographic determinants of contemporary linguistic diversity across countries. To maintain consistency across different levels of aggregation I use the language data from the WLMS (2006) and derive the number of languages spoken within each country. The respective country-specific geographic measures are calculated focusing on regions with linguistic coverage according to the WLMS (2006). Note that once the geographic data are intersected with the language data, i.e. spatially clipped along the observed distribution of languages, several of the underlying cells obtain different shapes following the spatial pattern of the language coverage. All cells that overlap with a language irrespective of the magnitude of the overlap are considered valid. Results are similar if I discard cells with an overlap of less than 10 or 100 square kilometers.

The number of regional observations on agricultural suitability within each country ranges from a single observation for Monaco to 9,415 for Russia. The median number of cells per country is 73. To construct the measure of linguistic diversity I focus on languages with at least 1,000 speakers as recorded in the Ethnologue dataset.[4] The resulting kernel density estimate of the distribution of linguistic groups across countries is shown in Figure 2*a* in the Appendix. Note that the distribution of the number of languages is skewed, so the natural log of languages is used in the analysis (Appendix Figure 2*b*). In Tables 2*a* and 2*b* below I show that results are robust to using alternative estimation techniques, indexes of linguistic diversity and measures of land quality.

To make sure there are enough regional observations per country, only those with at least 10 cells of 0.5 by 0.5 decimal degrees with information on land quality and language coverage are included. This limits the sample size to 156 countries. Descriptive statistics and the raw correlations are presented in Tables 1*a* and 1*b* in the web Appendix.

For the cross-country regressions the following specification is adopted:

[4]Results are similar if I include all linguistic groups in the construction of the dependent variable or language groups with at least 3,000 recorded speakers.

$$Ln(Number\ of\ Languages_i) = a_0 + a_1\ Absolute\ Latitude_i + a_2\ Variation\ in\ Elevation_i + a_3\ Variation\ in\ Land\ Quality_i + \boldsymbol{\alpha}_4\mathbf{X}_i + \eta_i, \quad (1)$$

where X_i is a vector of other geographic and political controls for country *i*.

Given that distance from the equator has been found to be an important predictor of ethnic diversity, in the first column of Table 1 I include it as the only regressor. As expected it enters negatively and is precisely estimated. Absolute latitude alone explains 23 percent of the variation in contemporary linguistic diversity across countries, with those further from the equator displaying consistently fewer languages. In column 2 the measures of geographic variability are introduced and enter highly significant with the expected sign. Countries with more diverse soil and climatic characteristics as well as more variable terrain display higher linguistic diversity. Introducing geographic variability significantly increases the explanatory power of the model.

To facilitate comparison of the quantitative effect across different specifications and across regressors I report standardized coefficients. A one-standard deviation increase in the variation of elevation and a similar increase in the variation of land quality augments linguistic diversity by 0.31 and 0.34 standard deviations increasing the log number of languages by 0.52 and 0.47, respectively. Likewise, a 1 standard deviation increase in absolute latitude increases the log number of languages by 0.84. Average land quality and average elevation are negative and statistically significant; however they become insignificant, once additional geographic controls are introduced.

In the third column of Table 1, other geographic characteristics are accounted for. Consistent with Nettle (1998), average precipitation enters positively and is highly significant, whereas average temperature is insignificant. Smaller countries located at a greater migratory distance from Addis Ababa sustain fewer linguistic groups. Note that by including average temperature in the regression, absolute latitude becomes insignificant. This is because distance from the equator is very highly correlated with average temperature, as shown in web Appendix Table 1*b*. Not surprisingly, controlling for these additional geographic characteristics decreases the magnitude of the coefficients on variation in land quality and variation in elevation, however, their effect remains precisely estimated. The standardized coefficients suggest that the effect of variation in elevation and migratory distance from East Africa are comparable. The distance to the nearest coastline though insignificant, weakly increases diversity. This is consistent with the view that areas that are increasingly isolated from the sea have been experiencing limited population mixing and thus should on average display higher ethnolinguistic fractionalization. It should be noted, however, that mean distance from the coast also captures the vulnerability of different areas to both the incidence and the intensity of invasion and colonization. Thus, the coefficient should be cautiously interpreted.

In column 4 continental fixed effects are included.[5] In this regression I also add the log population density as of 1995 to ensure that the results are not driven by differential

[5]Specifically, dummies for countries in the Americas, Africa, Europe and East Asia and the Pacific are included.

population density across countries.[6] The partial scatter plots of the variation in land quality and the variation in elevation against the log number of languages from column 4 are presented in Figures 4 and 5 in the Appendix. In column 5 of Table 1, additional controls that may be correlated both with geographic factors and ethnic diversity are added. To capture variation in historical contingencies across countries the population density in 1500 *AD* and the country's year of independence are added. The former enters negatively and is statistically significant. This finding provides evidence that conditional on geographic characteristics, contemporary ethnic diversity may have been influenced by a country's historical levels of development as represented by the population density in 1500 *AD*. The year when each country gained its independence enters insignificantly. Finally, the timing of the transition to agriculture across countries which might affect both the suitability of land for agriculture and ethnic diversity is not significant.

Table 2*a* presents a series of robustness checks. Using the count of languages as the dependent variable, column 1 presents the estimates on geographic variability using the negative binomial model (the Poisson model is not appropriate because of overdispersion in the number of languages across countries). In column 2 of Table 2*a*, geographic diversity is proxied using the dispersion in land quality and the dispersion on elevation within a country. In the last two columns I experiment with alternative indexes of land quality. In column 3 I use the climatic component of the land quality index. Likewise, mean land quality is proxied by the mean climatic suitability for agriculture whereas variation in land quality is constructed using the variation in the climatic suitability across regions within a country. Similarly, column 4 uses the soil suitability component to derive the statistics of interest. In both cases variation in either climatic or soil suitability for agriculture along with more variable terrain systematically increase linguistic diversity across countries.

Table 2*b* reports results using as a dependent variable various measures of linguistic fractionalization and employs variation in climatic suitability to capture heterogeneity in agricultural endowments. In this case the geographical statistics are constructed using information across all cells within each country. In columns 1–3 the dependent variable is a widely used measure of ethnolinguistic fractionalization, *ELF*, based on data from a Soviet ethnographic source, Atlas Narodov Mira (1964), and augmented by Fearon and Laitin (2003). In column 1 I introduce distance from the equator which enters negatively and highly significant. In column 2 I add the measures of geographic variability. Variation in land quality is positive and significant however the variation in elevation is insignificant. This is mainly driven by continental differences since Africa which is the most ethnically fractionalized continent also features a much less variable topography compared to the rest of the world. In column 3 when I include continental fixed effects and further geographical features variation in elevation regains its economic and statistical significance. In the rest of the columns of Table 2*b* I employ measures of linguistic fractionalization constructed by Desmet, Ortuño-Ortín and Wacziarg (2011). The authors use information on language trees from the Ethnologue (2005) to produce estimates of ethnolinguistic fractionalization at different levels of linguistic aggregation depending on how finely or coarsely groups are

[6]One has to be careful in interpreting the coefficient on current population density as it may be affected by ethnic diversity.

defined. Columns 4–7 use fractionalization indexes based on progressively finer classifications of linguistic groups to investigate the role of geographic heterogeneity at various levels of linguistic aggregation. Across all specifications variation in elevation and variation in climatic suitability for agriculture are systematic determinants of contemporary linguistic fractionalization. Similarly mean precipitation and distance from the sea coast systematically increase linguistic fractionalization. Migratory distance from East Africa and absolute latitude on the other hand, enter with the expected negative sign however they are less precisely estimated.

The findings uncover the fundamental role that geographic endowments play in the formation of ethnically diverse countries. Nevertheless, the endogeneity of contemporary political boundaries and the fact that modern states in their quest to build unitary nation states have largely shaped local languages call for the investigation to be conducted at finer levels of aggregation where such country-specific characteristics may be properly accounted for. This is the task pursued below.

B. Cross-Virtual Country Analysis

This section focuses on virtual countries. It does so in order to investigate whether the relationship between geography and ethnic diversity holds true at an arbitrary level of aggregation. The virtual countries are constructed in the following way: I generate a global grid of 2.5 by 2.5 decimal degrees that extends from −180 to 180 degrees longitude and from 85 degrees latitude to −65 degrees latitude. This global grid is intersected with the territories that are covered linguistically by the WLMS (2006) database. As a result each and every part of a virtual country that remains after the spatial intersection has complete linguistic coverage and it is across these territories that geographic and population statistics for each virtual country are constructed. For example, in order to derive the area under water for a virtual country, I focus on the regions where a language is spoken and across these regions I sum up the areas of all water bodies found there. To identify water bodies I use the "Inland Water Area Features" dataset from Global Mapping International which offers comprehensive global coverage of all rivers and lakes. Out of the 64,004 cells in the land quality dataset, 18,941 contain no information on languages and are dropped from the analysis. This is mostly due to the incomplete mapping of regions in the Americas and Australia, (see Map 1 in the web Appendix illustrating the resulting virtual countries).[7]

For the construction of the linguistic diversity index I consider languages with at least 1,000 speakers as recorded in the Ethnologue dataset. The results remain robust whether one includes all languages or only focuses on linguistic groups with at least 3,000 speakers within a country. In the regression analysis, virtual countries of at least 3,000 inhabitants are included. The distribution of the number of languages spoken across virtual countries is skewed so, instead of the levels, the natural log of languages is used in the analysis (see

[7]Similar to the cross-country analysis partial geographic cells may arise after the intersection with the linguistic maps. Excluding geographic cells with less than 10 or 100 square kilometers of linguistic coverage delivers similar results. Moreover, column 1 in Table 5*b* shows that focusing on virtual countries with complete linguistic information on each of the 25 intact underlying cells delivers if anything stronger results.

Figures 3*a* and 3*b* in the Appendix). Table 5*a* shows that results are robust to using alternative estimation techniques and different population cutoffs for virtual countries.

Similar to the cross-country analysis, I focus on virtual countries with at least 10 cells of 0.5 by 0.5 decimal degrees with information on land quality. The resulting sample size is 1,663 observations with a mean of 22 regional land quality observations per virtual country. Descriptive statistics and the pairwise correlations are presented in Tables 2*a* and 2*b* in the web Appendix. The median virtual country has 3 languages spoken with the most linguistically fragmented virtual country located in the western part of Cameroon.

Map 2 below shows one example of a virtual country. The circles, which are the centroids of the original 0.5 by 0.5 cells, represent the regional land quality for agriculture. The differently colored polygons represent the locations of the linguistic groups. The virtual country of map 2 falls between two real countries with the squiggly line delineating the current borders between Iran on the east and Iraq on the west. There are in total 8 languages spoken in this area.[8] It is characterized by large variation in elevation equal to 0.56, and a significant variation in land quality equal to 0.26, ranging from places that are totally inhospitable to agriculture to areas where the climate and the soil conditions are highly conducive to cultivation.

Excluding regions without linguistic coverage raises the question whether the linguistically mapped territories reported in WLMS (2006) are systematically different from places for which there is no linguistic coverage. To address this concern I do the following: From the 1,663 virtual countries there are 1,206 that do not have complete language coverage. In order to determine whether there is selection into mapping I calculate for each virtual country the geographic statistics for the regions without linguistic coverage and compare them to the respective statistics of the regions with language information. As shown in Table 3 there are no statistically significant differences between linguistically mapped and non-mapped territories with respect to mean land quality, mean elevation and mean temperature. The only statistically significant difference (at 10 percent level) is found for average precipitation. Regions without language coverage receive slightly higher rainfall of 0.71 millimeters per month. Overall, these findings suggest that places with language coverage do not differ significantly from the non-mapped regions in their geographic endowments, so focusing on the former does not cause a selection bias in the estimation. To further illustrate that selection into mapping is not driving the results, in column 1 of Table 5*b* I focus only on the 452 virtual countries for which there is complete linguistic coverage in each of the 25 underlying grid cells.

For the cross-virtual country regressions the following specification is adopted:

$$Ln(Number\ of\ Languages_i) = \beta_0 + \beta_1 \underset{Latitude_i}{Absolute} + \beta_2 \underset{Elevation_i}{Variation\ in} + \beta_3 \underset{Land\ Quality_i}{Variation\ in} + \boldsymbol{\beta}_4 \mathbf{X}_i + \xi_i, \quad (2)$$

[8]Namely these are: Central Kurdish, Gurani, Koy Sanjaq Surat, North Mesopotamian Spoken Arabic, Sangisari, South Azerbaijani, Southern Kurdish and Northern Kurdish. Languages' traditional homelands may overlap. For example, in this particular grid, regions where Gurani is spoken also have Northern Kurdish speakers.

where X_i is a vector of other geographic traits of virtual country i.

The analysis in Table 4 is presented in a fashion similar to the cross-country case by starting with absolute latitude as the only regressor.[9] As expected it enters negatively and is highly significant. In the second column the measures of geographic variability are introduced.

Because I report standardized coefficients one can directly compare the relative magnitudes of the variables of interest. A one-standard deviation increase in the variation of elevation and a similar increase in the variation of land quality augment linguistic diversity by 0.12 and 0.14 standard deviations adding 0.12 and 0.13 log number of languages respectively. These economically important findings reveal the geographic origins of contemporary ethnolinguistic diversity. The effect of geographic variability has half the effect of distance from the equator, however, as soon as country fixed effects are introduced, distance from the equator becomes insignificant. In the same specification I control for average land quality and average elevation but they are statistically insignificant.

In column 3 of Table 4 an array of additional geographic controls is introduced. Average precipitation is positive and highly significant whereas average temperature is insignificant. As expected, larger territories have more languages but the effect is not precisely estimated. Distance from the shoreline of an artificial country does not systematically affect linguistic diversity. The variable capturing areas of water like rivers and lakes, enters positively and it is statistically insignificant.[10] The more countries a virtual country falls into the more languages it sustains. This finding has a dual interpretation. It may be suggestive of the effect of state formation on ethnic diversity and/or an artifact of modern states having drawn political borders along ethnic boundaries. Finally, migratory distance from East Africa enters negatively and is highly significant. These geographic characteristics decrease the estimated coefficients of geographic variability which nevertheless remain economically and statistically significant. Overall, such geographic features capture 53 percent of the variation in linguistic diversity across virtual countries.

Taking advantage of the arbitrarily drawn borders of these geographic units one may explicitly control for country fixed effects. Territories falling into more than one country are assigned to the country in which their centroid falls. For example, the centroid of the virtual country in Map 2 belongs to Iraq. This is done in all subsequent specifications. Such inclusion of powerful controls, not possible in a cross-country framework, allows me to explicitly take into account any systematic elements related to the nation-building process of current states and thus produce reliable estimates of the effect of geographic heterogeneity on ethnic diversity. Note that as soon as country fixed effects are introduced in column 4 of Table 4, both migratory distance from East Africa and distance from the equator become insignificant but this is largely due an increase in the standard errors. In this regression I also add the log population density as of 1995 which is insignificant.

[9]The results presented here are OLS estimates with the standard errors clustered at the country level. Note that in case of virtual countries split by international boundaries they are assigned to the country in which their centroid falls.

[10]Note that the bivariate relationship between water area and linguistic diversity is negative but insignificant. This obtains because water area and surface area are positively correlated and the latter is positively related to linguistic diversity. In fact, controlling for surface area in this bivariate regression the coefficient on water area is negative and precisely estimated.

Columns 5 and 6 of Table 4 examine whether the identified effect of geographic variability is driven by any inherent differences between regions in the tropics and the rest of the climatic zones. In column 5 the sample is restricted to virtual countries in the tropics which extend from 23.5 latitude degrees south to 23.5 latitude degrees north. The estimated coefficients on both variation in land quality and variation in elevation remain stable and the effect of average land quality now becomes positive and significant. The latter is consistent with the finding of Moore et al. (2002) that cultural diversity within Sub-Saharan Africa increases in net primary productivity. Also, within the tropics larger territories receiving more precipitation and those located further inland display systematically larger linguistic diversity. On the contrary, across virtual countries out of the tropics in column 6, geographic heterogeneity is the only significant predictor of ethnic diversity. Finally, column 7 of Table 4 focuses on virtual countries that belong entirely to a single existing country. This specification allows one to investigate whether the estimated strong positive relationship between geographic variability and ethnic diversity holds true across regions within countries. According to the estimates a one standard deviation increase in both variation in land quality and variation in elevation increases by 0.22 the log number of languages contributing significantly to the formation of ethnically diverse regions within countries.

Tables 5*a* and 5*b* present a series of robustness checks. For brevity I only report the estimates on geographic variability but all specifications include the same controls as in column 4 of Table 4. Column 1 of Table 5*a* reports the effect of geographic variability on the number of languages within a virtual country estimated by a negative binomial model. In the second column all virtual countries are included in the regression irrespective of the total population, whereas in column 3 only those with at least 50,000 inhabitants as of 1995 are included. Finally, in the last two columns the dependent variable is constructed using alternative population thresholds for the linguistic groups. In particular, in column 4 all linguistic groups that fall within a virtual country are counted irrespective of their population, whereas in column 5 only language groups with at least 3,000 recorded speakers are included.

Across all these alternative specifications geographic variability remains both quantitatively and qualitatively significant demonstrating the robustness of the findings to alternative estimation techniques and indexes of linguistic diversity.

In Table 5*b* I perform additional robustness checks. In column 1 I focus on virtual countries with complete linguistic coverage across all 25 underlying intact geographic cells. The coefficients on geographic diversity increase in magnitude and remain precisely estimated. In column 2 I control for the size of each virtual country in a flexible way. Specifically, I add a fixed effect for each percentile of the size distribution. Doing so, I am exploiting variation within states, across virtual countries of similar sizes. The results show that the estimated coefficients on geographic variability remain unaffected. In columns 3–5 I use alternative measures of geographic variability. In Column 3 geographic diversity is captured by the dispersion in land quality and the dispersion of elevation within a virtual country. In columns 4 and 5 I use the climatic and soil components of land quality, respectively, to capture variation in agricultural suitability. Across these alternative indexes of geographic

diversity the main finding remains unaltered. Territories characterized by variable geographic endowments have more linguistic groups.

This section establishes that heterogeneity in land quality and elevation across virtual countries is a fundamental determinant of contemporary ethnic diversity. The fact that these results obtain at an arbitrary level of aggregation, in and out of the tropics and after controlling for country fixed effects brings into light the geographic origins of ethnic diversity.

C. Pairwise Analysis of Adjacent Regions

There is a large literature in education, urban and trade economics that concentrates on bordering regions to investigate the effect of various policies, (Black (1999), McCallum (1995) among others). The rationale is that focusing on adjacent regions, while accounting for observable characteristics, should neutralize any local unobservable differences that would otherwise contaminate inference. Similarly, this section investigates the effect of local geographic differences in determining the degree of linguistic similarity within pairs of adjacent regions. To implement this test I identify the neighboring regions of each 0.5 by 0.5 degrees cell. The immediate neighbors of each cell are those directly to the north, south, east and west, as well as those that are diagonally contiguous at a distance of 0.71 degrees (i.e., to the northwest, southwest, northeast and southeast). In total, a single region may belong to at most eight pairs. In Map 3 in the Appendix the dots of regional land qualities are the centroids of the individual regions and the arrows point to their regional neighbors. Pairs with a total area of less than 1,000 square kilometers are excluded, resulting in a total of 156,570 unique dyads that constitute the units of analysis.

For the pairwise regressions of adjacent regions the following specification is adopted:

$$\begin{matrix}\textit{Percentage of}\\ \textit{Common Languages}_{ij}\end{matrix} = \gamma_0 + \gamma_1 \begin{matrix}\textit{Difference in}\\ \textit{Land Quality}_{ij}\end{matrix} + \gamma_2 \begin{matrix}\textit{Difference in}\\ \textit{Elevation}_{ij}\end{matrix} + \gamma_3 \mathbf{X}_{ij} + \xi_{ij}, \quad (3)$$

where the percentage of common languages$_{ij}$ is the number of common languages divided by the total number of languages spoken within the regional pair *ij* capturing the degree of ethnic similarity of any two adjacent regions. The absolute difference in land quality and elevation between regions *i* and *j* provide a measure of how dissimilar are the local geographic attributes. Tables 3*a* and 3*b* in the web Appendix present the summary statistics and the correlation table. Note that the mean of the dependent variable implies that adjacent regions by virtue of proximity have on average 77 percent of the total number of languages in common.

The first column in Table 6 produces evidence on the role of regional geographic differences in shaping local ethnolinguistic diversity. Specifically, a one standard deviation increase in the difference in land quality and a similar increase in difference in elevation decreases local linguistic similarity by 3.5 percentage points, contributing to the formation of ethnically distinct neighbors. The combined estimated effect is a modest 12 percent of the dependent variable's standard deviation, however; it is quantitatively similar to the effect of physical distance within pairs. A one standard deviation increase in the geodesic distance (that is,

approximately 14 kilometers) decreases ethnic similarity by 3.4 percentage points. The significant effect of physical proximity conditional on geographic factors is consistent with a migration mechanism where physical distance between populations leads over time to linguistic differentiation among groups.

At the same time, regional pairs located further from the equator and with large migratory distance from East Africa are more ethnically similar, whereas those characterized by higher levels of precipitation have fewer languages in common. Finally, pairs that fall within a country are systematically more similar, whereas pairs more densely populated in 1995 tend to be more linguistically diverse but the effect is insignificant at conventional levels.

The dyadic structure of the observations allows me to control for region fixed effects by including 42,644 regional dummies. This demanding specification explicitly takes into account any systematic elements related to the regional histories. As is evident from column 2, the inclusion of region fixed effects decreases the magnitude of the coefficients on pairwise differences in elevation and land quality considerably, but they remain highly statistically significant. Also, unlike the geodesic distance within the pair and distance from the equator which remain significant, migratory distance from East Africa and average precipitation are no longer systematically related to local ethnic diversity.

The last 3 columns of Table 6 consider different continents. Column 3 focuses on pairs within Africa, whereas column 4 focuses on pairs within Europe. Comparing the two samples it is clear that geographic variability is quantitatively equally important, whereas the effect of physical distance within dyads in Africa is twice as large as the effect within regional pairs in Europe. Additionally, dyads within countries in Africa have 13.6 percent more common languages compared to regional pairs that are partitioned by country borders. On the other hand, any two adjacent regions within Europe located in the same country have 34.4 percent more common languages compared to pairs split by borders. Finally, within Asia (column 5), more geographically distinct dyads and those located further away from the coast are more linguistically dissimilar.

These findings demonstrate the following: first, the differences in land quality and elevation between adjacent regions shape local ethnic diversity. Second, the spatial arrangement of a given heterogeneous land endowment matters in determining the degree of cultural heterogeneity, that is, regions with similar land endowments in closer proximity would result in lower ethnic diversity.

D. Recent Migrations and Ethnic Diversity

In the last 500 years several regions have experienced dramatic changes in their ethnic and linguistic endowments because of conquests, slavery, migrations, and colonization. Although distinguishing the relative impact of these historical events on shaping contemporary ethnic diversity is beyond the scope of the current study, in this section I show that across territories where indigenous groups as of 1500 AD constitute a minority today, the current ethnic composition of the population is no longer systematically related to the underlying geography.

Two explanations may be offered for this broken link. One is that the new settlers have spent only limited time under these geographic endowments, so geography has not yet made an impact upon the formation of distinct linguistic traits. The second observation is related to the fact that these migration flows have taken place mainly over the last two centuries when geography has become less important in shaping mobility between groups. This is largely due to industrialization, that is, the declining importance of land in the production process, and the ease of communication and transportation thanks to technological advances.

Table 7 investigates whether the effect of geographic variability depends on the presence of native populations. Specifically, at every level of aggregation results are presented for territories where more than 40 percent of the current population was indigenous in these countries as of 1500 AD. These are compared to the point estimates derived for regions where *less* than 40 percent of the population can trace ancestry back to 1500 AD.[11] Columns 1 and 2 present the results across virtual countries. The controls are similar to those in column 4 of Table 4. Out of a total of 1,663 virtual countries 1,302 belong to countries with significant representation of indigenous people. Geographic variability in this subset of virtual countries is a powerful predictor of linguistic diversity. On the contrary, in the remaining 311 regions with greatly diminished indigenous populations, geographic heterogeneity is no longer systematically related to contemporary linguistic diversity. If anything, as column 2 suggests, more variable elevation leads to lower linguistic fragmentation. The same pattern holds true when focusing across pairs of adjacent regions and across countries, shown in columns 3, 4 and 5, 6 respectively.

Although geographic variability is not related to contemporary ethnic diversity in regions with significant population reshuffling in the last 500 years, variations in the intensity of colonization and slavery have been linked to the geographic conditions of the colonized countries. For example, Engerman and Sokoloff (1997) argue that slavery became so predominant in certain colonies like the Caribbean or Brazil, because of the climatic and soil conditions that were well suited for growing specific valuable crops, like sugar, which were most efficiently produced on large slave plantations. On the other hand, the limited presence of slavery in the North American mainland was due to land endowments favoring an economy based on grains and livestock, whose production required few slaves due to the limited economies of scale. Also, Acemoglu, Johnson and Robinson (2001) argue that ecological conditions shaped the intensity of European colonization with regions characterized by unfavorable disease environments experiencing little European migration. Evidence consistent with the geographic determinants of the post-1500 migration flows is produced by Ertan and Putterman (2007). First, the authors show that Sub-Saharan Africans ended up in larger numbers in places where endowments were more suitable for sugarcane production (to be used as slaves in sugarcane plantations) and less suitable for wheat cultivation and second, European populations were more likely to migrate in places characterized by low disease loads further away from the equator.

Overall, the findings in Table 7 suggest that in places with large population changes over the last 500 years, the advent of new groups coupled with the replacement and displacement of

[11]Using any cut-off below 40% of indigenous population the results are similar.

the indigenous populations have severed the link between geographic variability and ethnic diversity. This is an important finding because it highlights that unlike most countries where a significant fraction of contemporary ethnic diversity may be treated as predetermined by geography; ethnic diversity in places where recent settlers constitute the majority, is neither geographically predetermined nor may be treated as exogenous. In addition to their languages, recent settlers have brought along institutions and other country-specific characteristics of their native lands. These have been shown by Putterman and Weil (2010) to directly affect contemporary economic performance.

IV. Channels Linking Geography to Ethnic Diversity

In this section I discuss the possible mechanisms that may give rise to the observed relationship between geographic and ethnic diversity.

The empirical findings are consistent with several channels. First, one might argue that groups of people form an ethnic identity along a homogeneous land endowment in order to defend it against intruders and enforce property rights over it. If this is the case, places characterized by more diverse land endowments would naturally sustain more linguistic groups. Second, to the extent that homogeneous territories are easier to conquer and invasions have a homogenizing effect then a similar empirical relationship would hold true. Third, geographic differences, by magnifying migration costs, may increase isolation between groups leading via a process of cultural drift to the formation of distinct cultural and linguistic traits (Boyd and Richerson (1985)). In fact, the analysis on pairs of contiguous regions produces evidence on the role of pure migration costs by showing that the physical distance within a pair of adjacent regions significantly decreases linguistic similarity. Fourth, in a stage of development when land dominates production decisions, people working on different types of land acquired location-specific skills not easily transferable to a different natural environment. Therefore, geographic diversity reduced the mobility of people in a given area, increasing isolation, leading to the formation of distinct languages. This last mechanism highlights that ethnic groups may be thought of as bearers of location-specific human capital.[12]

Disentangling the importance of each channel is difficult. A heterogeneous geography may directly increase migration costs and/or lead groups of people to accumulate skills specific to their local environment. However, the specific human capital mechanism generates an auxiliary prediction not borne out by other potential explanations. Specifically, to the extent that a language group is characterized by specific skills then non-adjacent partitions of the same language group would exhibit similar modes of subsistence. In an attempt to provide empirical evidence on this channel I examine African linguistic groups.

Figure 4*a* maps the linguistic homeland of the Silt'e. This group is located in Ethiopia and according to the WLMS (2006) its traditional location extends over 2 non-adjacent regions. To obtain a proxy for the dominant subsistence activity across linguistic partitions I use data on the actual allocation of land between pasture and farming and within farming among the

[12]Region-specific human capital should be thought of as encompassing both the technical knowledge necessary to be productive in a given region and the capacity of the immune system to adapt to the local disease vectors.

different crops in 1992 AD. Specifically, I use information at the grid level (0.5 by 0.5 decimal degrees) on how land is allocated between pasture and agriculture to identify the dominant activity within each partition. For example, the northern part of Silt'e in Figure 4*b* has 61 percent of its land under some form of cultivation compared to 38 percent used for pasture. Hence, it is classified as an agricultural community. Similarly, the southern partition of Silt'e uses most land for farming. Naturally, one may argue that both partitions being farmers is not a manifestation of common human capital within the group, but rather an overall characteristic of this part of Ethiopia. Therefore, I generate for each partition a buffer zone of 0.5 decimal degrees radius (see Figure 4*a*) and calculate the dominant land use in these neighboring regions. Figure 4*b* shows that in the case of Silt'e the neighboring regions of both partitions are predominantly pastoral. For example, the regions found in the buffer zone of the northern partition allocate on average 50 percent of the land towards pasture compared to 39 percent towards agriculture.

To explore the correlation structure of the subsistence activities within groups located in multiple territories I focus on African linguistic groups which according to the WLMS (2006) have homelands that span over 2 non-contiguous regions.[13] In total there are 209 groups with 2 partitions each. For the cross-language partition analysis the following specification is adopted:[14]

$$Specialization_{i,g}=\delta_0+\delta_1 Buffer\ Specialization_{i,g}+\delta_2 Specialization_{j,g}+\delta_3 Land\ Quality+\varphi_{i,g}, \quad (4)$$

where $Specialization_{i,g}$ is the dominant activity in partition i of group g. Buffer $Specialization_{i,g}$ is the primary activity in the regions within a radius of 0.5 decimal degrees around partition i of group g and $Specialization_{j,g}$ captures the dominant activity in partition j that belongs to the same ethnic group g as partition i. I consider three types of activities: first, whether pasture is dominant, second whether maize is the most common crop when agriculture is present and third whether sorghum is the most cultivated crop. I focus on these two crops because they are the most widespread in Africa.

In columns 1 and 2 of table 8 the dependent variable is an indicator that equals 1 if partition i is predominantly pastoral. Column 1 shows that if the neighboring regions of a partition are pastoral this increases by 75 percent the likelihood that the partition itself is pastoral. Also, partitions with higher suitability for agriculture are more likely to be farmers. In column 2 I add whether partition j of the same linguistic group g is pastoral. Conditional on the specialization pattern of the surrounding regions of partition i, if partition j is pastoral this increases significantly the likelihood that i is also a pastoral community.

Similar is the pattern when I look at specific crops. Columns 3 and 4 focus on whether maize is the most cultivated crop and columns 5 and 6 on sorghum. In the case of maize, column 4 reveals that if partition j has maize as its dominant crop, this increases by 54 percent the probability that maize will also be the most cultivated crop in partition i.

[13]To capture significant partitions I exclude those with less than 100 square kilometers. Results are similar if I include all partitions irrespective of size and/or groups that are partitioned in more than 2 non-adjacent territories.

[14]The results presented here are Probit maximum likelihood estimates with the standard errors clustered at the ethnicity level.

Similarly, column 6 shows that conditional on the subsistence pattern of the neighboring regions if sorghum is the crop with the largest share of land under cultivation in partition *j*, this increases by 37 percent the probability that sorghum will also be the dominant crop in partition *i*.

Overall, the uncovered correlation in the subsistence practices across partitions of the same linguistic group is supportive of the hypothesis that African ethnic groups are bearers of specific skills. Although the findings highlight that the human capital mechanism is a relevant dimension for understanding language group formation within Africa they do not shed light on the relative importance of other channels, a task left for future research.

V. Concluding Remarks

This study examines the geographic origins of ethnic diversity. I construct detailed data on the distribution of land quality and elevation across regions and countries and show that geographic variability systematically gives rise to linguistic diversity. I examine both cross-virtual country and cross-country specifications. The former is of particular significance since the relationship between geographic and ethnic diversity holds true at an arbitrary level of aggregation, explicitly avoiding the endogeneity of current countries' borders and after controlling for country fixed effects. These results are corroborated by examining how local differences in land quality and elevation shape the degree of ethnic similarity within pairs of adjacent regions. Accounting for region-specific effects, contiguous cells sharing similar land features are ethnically more homogeneous compared to pairs characterized by different land endowments. Overall, the importance of the distribution of land quality and elevation in determining ethnic diversity is a recurrent finding across different levels of aggregation and remains robust to alternative specifications and indexes of geographic and linguistic heterogeneity.

The empirical evidence may be used to explain the pattern of technology diffusion within and across countries as well as across ethnic groups. Technology would diffuse more quickly over places characterized by homogeneous geographic endowments, whereas in relatively heterogeneous places - which, according to the evidence, are also more ethnically diverse, the diffusion would be less rapid, leading to the emergence of inequality across countries as well as ethnic groups (Diamond (1997)).

Furthermore, this research suggests that differences in land endowments across regions gave rise to location-specific human capital, diminishing population mobility and leading to the formation of localized ethnicities. On the other hand, homogeneous land endowments facilitated population mixing resulting eventually in the formation of a common ethnolinguistic identity. Viewing languages as agents of specific human capital implies that speakers of the same language found in separate territories, should be observed to undertake similar activities. Consistent with the hypothesis, I show that within Africa the subsistence patterns across non-contiguous partitions of the same language group are systematically correlated.

Acknowledgments

Comments from 3 anonymous referees improved the manuscript substantially. Daron Acemoglu, Roland Benabou, Matteo Cervellati, James Fearon, Andrew Foster, Oded Galor, Ioanna Grypari, Peter Howitt, Yannis Ioannides, Masayuki Kudamatsu, Nippe Lagerlof, David Laitin, Ashley Lester, Ross Levine, Glenn Loury, Ignacio Palacios-Huerta, Elias Papaioannou, Stephen Ross, Yona Rubinstein, Francesco Trebbi and David Weil provided valuable suggestions. I would like, also, to thank the participants at the 2007 NEUDC Conference, the 2007 LAMES Meetings in Bogotá, the 2007 NBER Summer Institute on Income Inequality and Growth and the 2008 Ethnicity Conference in Budapest, as well as the seminar participants at Brown University, Chicago GSB, Collegio Carlo Alberto, Dartmouth College, EIEF, IIES, Princeton University, Stanford GSB, Tufts University, UCL, University of Bologna, University of Copenhagen, University of Connecticut, University of Cyprus, University of Gothenburg, University of Houston, Warwick University and Yale University for the useful discussions. Lynn Carlsson's ArcGIS expertise proved of invaluable assistance. Financial support from the Watson Institute's research project "Income Distribution Across and Within Countries" at Brown University is gratefully acknowledged.

References

Acemoglu, Daron; Johnson, Simon; Robinson, James A. The Colonial Origins of Comparative Development: An Empirical Investigation. American Economic Review. 2001; 91(5):1369–1401.

Ahlerup, Pelle; Olsson, Ola. Working Papers. University of Gothenburg; 2008. The Roots of Ethnic Diversity.

Alesina, Alberto; Ferrara, Eliana La. Ethnic Diversity and Economic Performance. Journal of Economic Literature. 2005; 43:762–800.

Alesina, Alberto; Devleeschauwer, Arnaud; Easterly, William; Kurlat, Sergio; Wacziarg, Romain. Fractionalization. Journal of Economic Growth. 2003; 8:155–194.

Ashraf, Quamrul; Galor, Oded. Working Paper. Brown University; 2008. Human Genetic Diversity and Comparative Economic Development. mimeo

Atlas Narodov Mira, Atlas of the People of the World. 1964. Moscow:Glavnoe Up-ravlenie Geodezii i Kartograi, Bruck, S.I., and V.S. Apenchenko.

Banerjee, Abhijit; Somanathan, Rohini. The Political Economy of Public Goods: Some Evidence from India. Journal of Development Economics. 2006; 82:287–314.

Barth, Frederik. Ethnic Groups and Boundaries: The Social Organization of Cultural Difference. Boston: Little, Brown; 1969.

Black, Sandra E. Do Better Schools Matter? Parental Valuation of Elementary Education. The Quarterly Journal of Economics. 1999; 114(2):577–599.

Boyd, Robert; Richerson, Peter J. Culture and the Evolutionary Process. Chicago, IL: University of Chicago Press; 1985.

Cavalli-Sforza, Luigi Luca; Cavalli-Sforza, Francesco. The Great Human Diasporas: The History of Diversity and Evolution. Cambridge: Perseus Books; 1996.

Comin, Diego; Easterly, Bill; Gong, Erick. Was the Wealth of Nations Determined in 1000 B.C.? American Economic Journal: Macroeconomics. 2010; 2:65–97.

Darwin, Charles. The Voyage of the Beagle. Originally 1839, Reprinted in 2006. Reprinted by Black Dog and Leventhal Publishers

Desmet, Klaus; Ortuño-Ortín, Ignacio; Wacziarg, Romain. The Political Economy of Ethnolinguistic Cleavages. Journal of Development Economics. Forthcoming.

Diamond, Jared. Guns, Germs, and Steel: The Fates of Human Societies. New York, NY: W. W. Norton & Co; 1997.

Diamond, Jared. Collapse: How Societies Choose to Fail or Succeed. New York, NY: Viking Press; 2005.

Easterly, William; Levine, Ross. Africa's Growth Tragedy: Policies and Ethnic Divisions. Quarterly Journal of Economics. 1997; 112(4):1203–1250.

Engerman, Stanley L.; Sokoloff, Kenneth L. Factor Endowments, Institutions, and Differential Paths of Growth among New World Economies. In: Haber, Stephen, editor. How Latin America Fell Behind. Stanford, CA: Stanford University Press; 1997. p. 260-304.

Ertan, Arhan; Putterman, Louis. mimeo. Brown University; 2007. Determinants and Economic Consequences of Colonization: A Global Analysis.

Ethnologue. Languages of the World. 15. SIL International; 2005.

Fearon, James; Laitin, David. Ethnicity, Insurgency and Civil War. American Political Science Review. 2003; 97:75–90.

Galor, Oded; Moav, Omer. Natural Selection and the Origin of Economic Growth. Quarterly Journal of Economics. 2002; 117(4):1133–1191.

Galor, Oded; Michalopoulos, Stelios. Evolution and the Growth Process: Natural Selection of Entrepreneurial Traits. Journal of Economic Theory. Forthcoming.

G-Econ. 2006. Available on-line at http://gecon.yale.edu

Geertz, Clifford. Old Societies and New States: The Quest for Modernity in Asia and Africa. New York: Free Press; 1967.

Gray, Russell D.; Atkinson, Quentin D. Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin. Nature. 2003; 426:435–439. [PubMed: 14647380]

Hale, Henry E. Explaining Ethnicity. Comparative Political Studies. 2004; 37(4):458–485.

Harmon, David. Losing Species, Losing Languages: Connections Between Biological and Linguistic Diversity. Southwest Journal of Linguistics. 1996; 15:89–108.

Mace, Ruth; Pagel, Mark. A Latitudinal Gradient in the Density of Human Languages in North America. Proceedings of the Royal Society; London. 1995. p. 117-121.

Maffi, Luisa. Linguistic, Cultural and Biological Diversity. Annual Review of Anthropology. 2005; 29:599–617.

Mauro, Paolo. Corruption and Growth. The Quarterly Journal of Economics. 1995; 110:681–712.

McCallum, John. National Borders Matter: Canada-U.S. Regional Trade Patterns. American Economic Review. 1995; 85:615–623.

Moore, Joslin L.; Manne, Lisa; Brooks, Thomas; Burgess, Neil D.; Davies, Robert; Rahbek, Carsten; Williams, Paul; Balmford, Andrew. The Distribution of Cultural and Biological Diversity in Africa. Proceedings of the Royal Society B. 2002; 269:1645–1653. [PubMed: 12204124]

Nettle, Daniel. Explaining Global Patterns of Language Diversity. Journal of anthropological archaeology. 1998; 17:354–374.

Nettle, Daniel. Linguistic Diversity. Oxford: Oxford University Press; 1999.

Nurse, Derek; Philippson, Gérard. The Bantu Languages. London, UK: Routledge; 2003.

Putterman, Louis; Weil, David N. Post-1500 Population Flows and the Long Run Determinants of Economic Growth and Inequality. Quarterly Journal of Economics. 2010; 125:1627–1682. [PubMed: 24478530]

Ramankutty, Navin; Foley, Jonathan A.; Norman, John; McSweeney, Kevin. The Global Distribution of Cultivable Lands: Current Patterns and Sensitivity to Possible Climate Change. Global Ecology and Biogeography. 2002; 11:377–392.

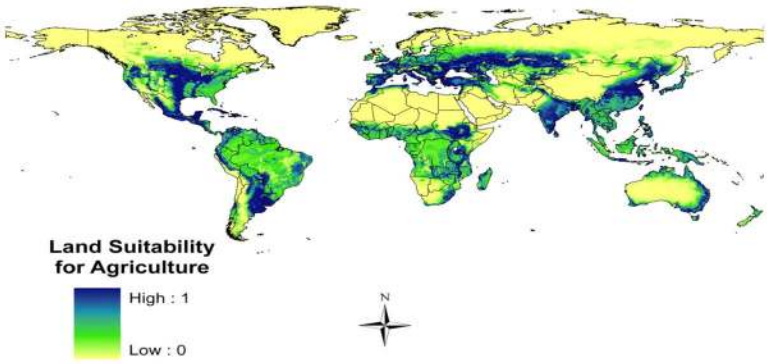
Spolaore, Enrico; Wacziarg, Romain. The Diffusion of Development. Quarterly Journal of Economics. 2009; 124(2)

Sutherland, William J. Parallel Extinction Risk and Global Distribution of Languages and Species. Nature. 2003; 423:276–279. [PubMed: 12748639]

WLMS. World Language Mapping System. 2006. Version 3.2 Available on-line at http://www.gmi.org/wlms/

1 Appendix

Global Maps



Map 1.
Land Quality Across Countries

Kernel Density Estimates

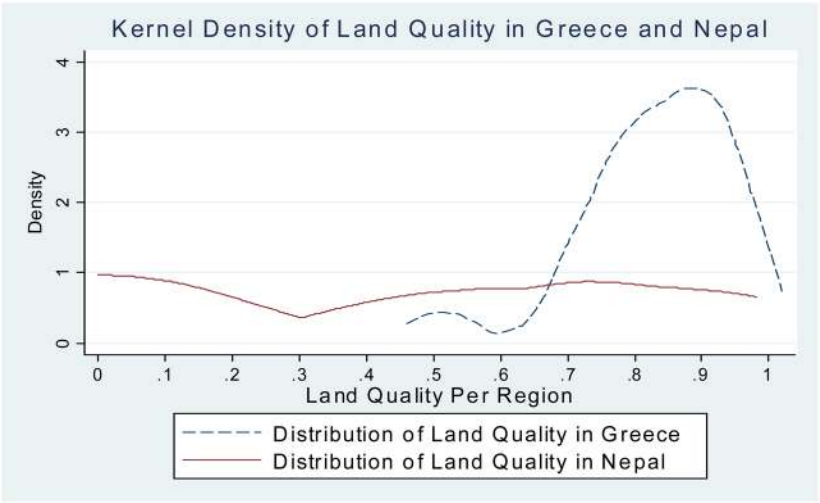


Figure 1[15]

[15]All density estimates presented are weighted by the Epanechnikov kernel.

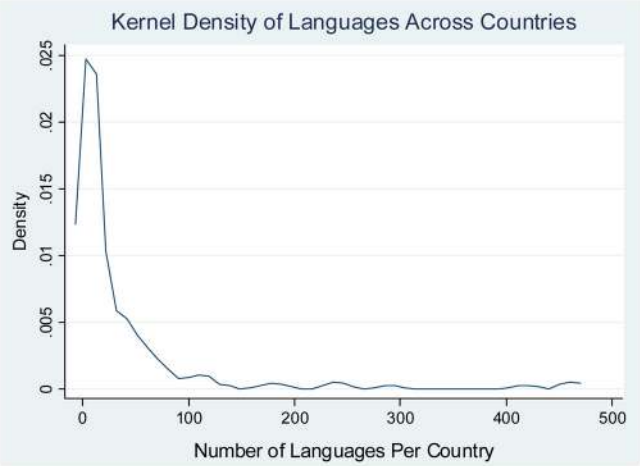


Figure 2*a*

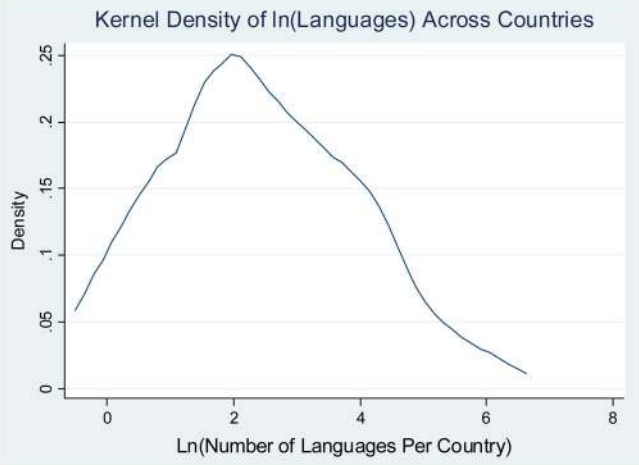


Figure 2*b*

Figure 2.

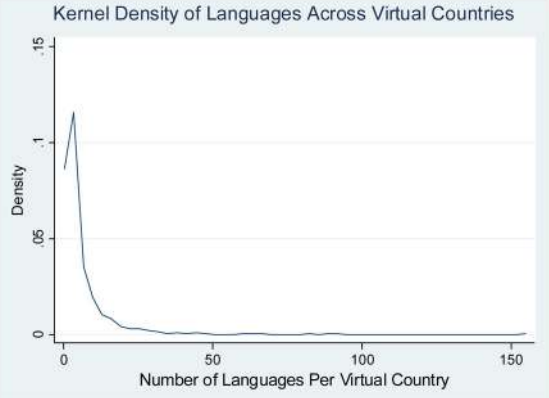


Figure 3*a*

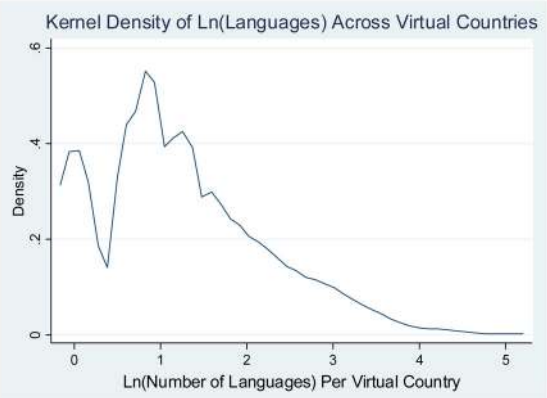


Figure 3*b*

Figure 3.

Real Country Analysis

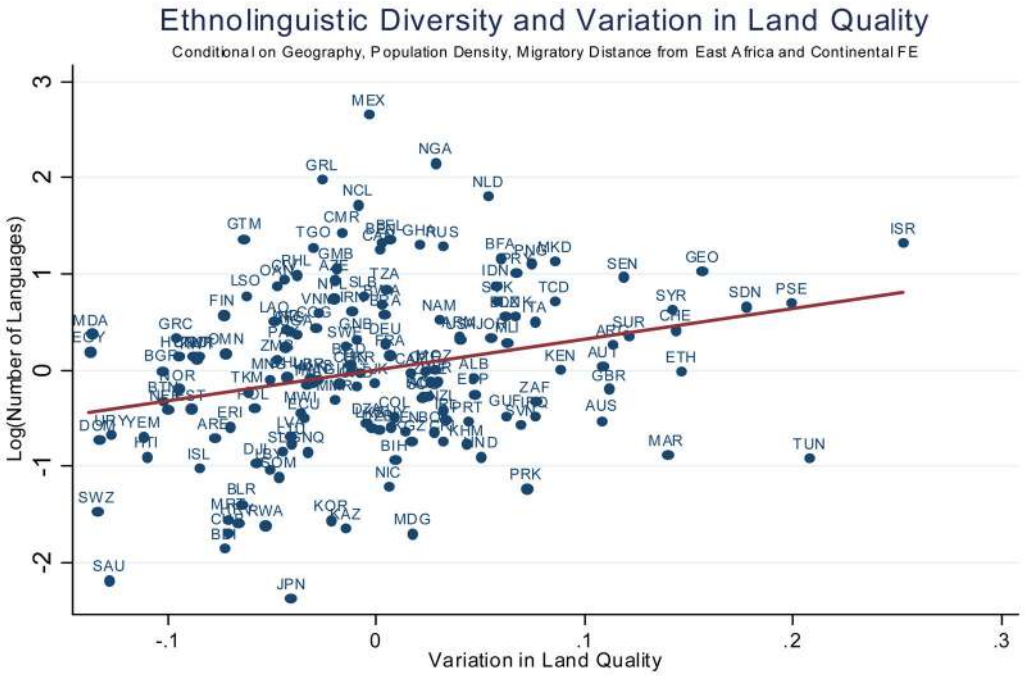


Figure 4.

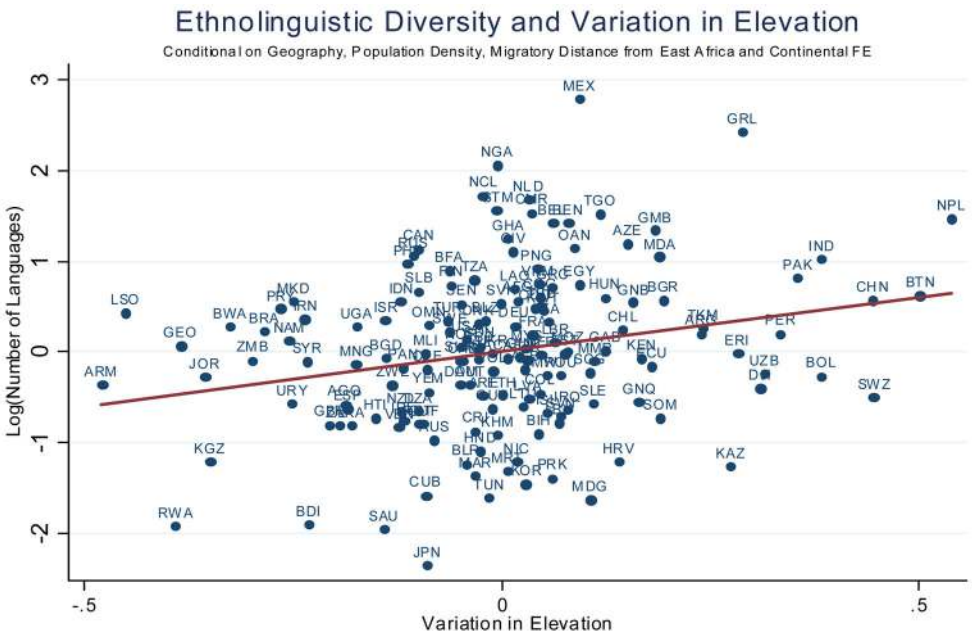
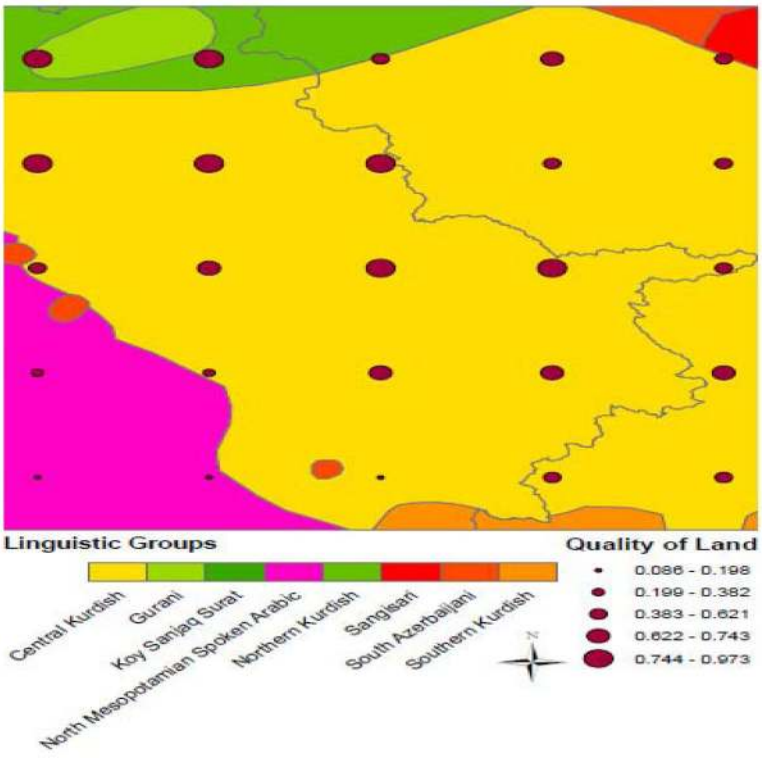
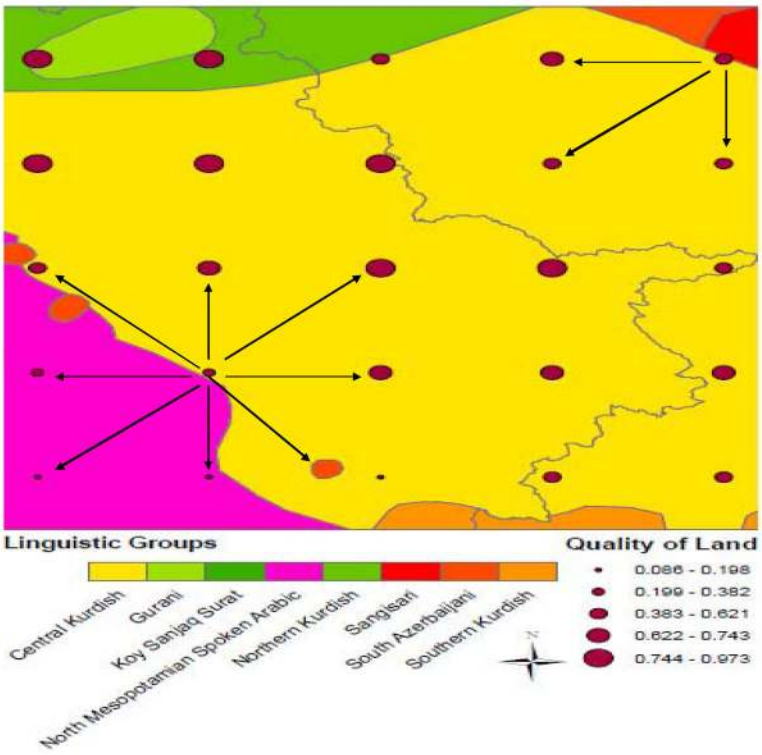


Figure 5.

Pairwise Analysis of Adjacent Regions



Map 3.
Examples of Pairs of Adjacent Regions



Map 2.
Example of a Virtual Country

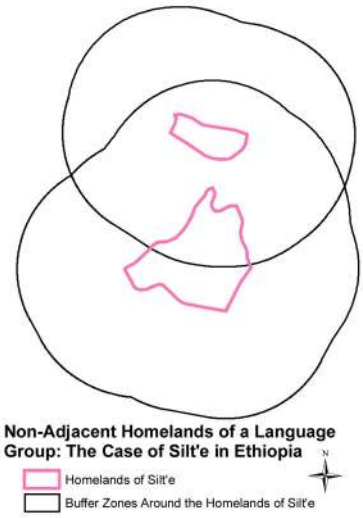


Figure 4*a*

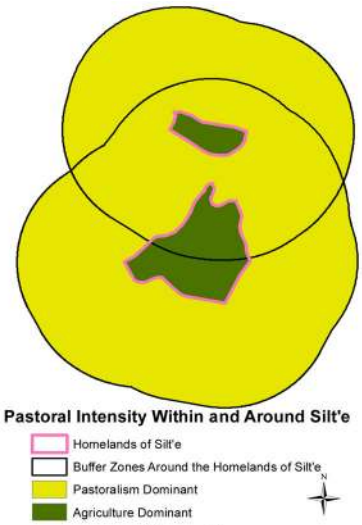


Figure 4*b*

Figure 4.

Table 1

Main Specification for the Cross-Country Analysis

VARIABLES	(1)	(2)	(3)	(4)	(5)
	Ln Number of Languages				
Variation in Elevation		0.310*** (0.113)	0.256*** (0.079)	0.291*** (0.089)	0.275*** (0.101)
Variation in Land Quality		0.340*** (0.084)	0.177*** (0.061)	0.208*** (0.058)	0.211*** (0.060)
Mean Elevation		−0.249** (0.113)	−0.111 (0.106)	−0.104 (0.118)	−0.085 (0.113)
Mean Land Quality		−0.179** (0.069)	−0.069 (0.065)	−0.029 (0.068)	0.006 (0.064)
Absolute Latitude	−0.479*** (0.070)	−0.547*** (0.061)	−0.058 (0.192)	−0.033 (0.214)	−0.131 (0.201)
Mean Precipitation			0.468*** (0.086)	0.447*** (0.088)	0.479*** (0.088)
Mean Temperature			0.270 (0.197)	0.385* (0.213)	0.404** (0.183)
Ln(Area)			0.517*** (0.067)	0.482*** (0.073)	0.464*** (0.074)
Distance from the Sea			0.053 (0.065)	0.063 (0.062)	0.073 (0.064)
Migratory Distance from East Africa			−0.281*** (0.063)	−0.518** (0.199)	−0.513** (0.218)
Ln(Population density in 1995)				−0.118 (0.087)	0.023 (0.072)
Ln(Population density in 1500)					−0.235** (0.105)
Year of Independence					−0.108 (0.066)
Timing of Transition to Agriculture					0.134 (0.094)
Continental FE	N	N	N	Y	Y
Observations	156	156	156	156	142
R^2	0.23	0.40	0.67	0.69	0.73

Standardized beta coefficients are reported, with robust standard errors in parentheses. ***$p < 0.01$,

**$p < 0.05$,

*$p < 0.1$; See web Appendix for variables' sources and definitions.

Table 2a

Robustness Checks for the Cross-Country Analysis

	(1)	(2)	(3)	(4)
VARIABLES	**Negative Binomial**	**OLS**	**OLS**	**OLS**
	Number of Languages		**Log Number of Languages**	
Variation in Elevation	1.313*** (0.395)		1.045** (0.473)	1.309*** (0.445)
Variation in Land Quality	3.252*** (0.952)			
Dispersion of Elevation		0.429*** (0.130)		
Dispersion of Land Quality		1.315*** (0.399)		
Variation in Climatic Suitability			2.505*** (0.749)	
Mean Climatic Suitability			0.661* (0.349)	
Variation in Soil Suitability				3.785*** (1.324)
Mean Soil Suitability				0.653 (0.474)
Continental FE	Y	Y	Y	Y
Observations	142	142	142	142
R^2	.	0.73	0.73	0.73
Log pseudolikelihood	−536.365	.	.	.

Nonstandardized coefficients reported, with robust standard errors in parentheses

*** $p < 0.01$,

** $p < 0.05$,

* $p < 0.1$

All specifications include the same controls as those of column 5 in Table 1. Column (1) is estimated using a negative binomial and the dependent variable is the number of languages. Columns (2)–(4) use as dependent variable the ln Number of Languages. Column (2) uses the dispersion in elevation and land quality to capture geographic variability. (3) and (4) use the climatic and soil components of land quality, respectively.

Table 2b

Linguistic Fractionalization Across Countries

VARIABLES	(1) ELF	(2) ELF	(3) ELF	(4) ELF3	(5) ELF5	(6) ELF7	(7) ELF9
Variation in Elevation		−0.111 (0.124)	0.363** (0.142)	0.356** (0.169)	0.472*** (0.160)	0.426*** (0.149)	0.413*** (0.156)
Variation in Climatic Suitability		0.294*** (0.094)	0.231*** (0.084)	0.291*** (0.103)	0.293*** (0.103)	0.215** (0.086)	0.156* (0.089)
Mean Elevation		0.093 (0.127)	−0.301** (0.148)	−0.352** (0.171)	−0.475*** (0.169)	−0.462*** (0.166)	−0.367** (0.167)
Mean Climatic Suitability		0.053 (0.083)	0.218** (0.108)	−0.141 (0.128)	−0.062 (0.121)	−0.213** (0.108)	−0.024 (0.110)
Absolute Latitude	−0.369*** (0.080)	−0.434*** (0.081)	−0.397 (0.331)	−0.116 (0.358)	−0.185 (0.326)	−0.064 (0.295)	−0.124 (0.311)
Mean Precipitation			0.180 (0.151)	0.455*** (0.174)	0.404** (0.167)	0.487*** (0.142)	0.375*** (0.140)
Mean Temperature			−0.030 (0.266)	0.248 (0.296)	0.181 (0.281)	0.316 (0.274)	0.302 (0.280)
Ln(Area)			0.030 (0.125)	−0.247* (0.146)	−0.186 (0.153)	−0.174 (0.132)	−0.015 (0.130)
Distance from the Sea			0.281*** (0.086)	0.452*** (0.118)	0.414*** (0.109)	0.326*** (0.098)	0.229** (0.093)
Migratory Distance from East Africa			−0.122 (0.256)	−0.535* (0.289)	−0.280 (0.296)	−0.359 (0.287)	−0.540* (0.287)
Ln(Population Density in 1995)			−0.022 (0.103)	−0.093 (0.141)	0.006 (0.141)	0.007 (0.137)	−0.090 (0.132)
Ln(Population Density in 1500)			−0.268** (0.123)	−0.239* (0.144)	−0.214 (0.150)	−0.211 (0.135)	−0.166 (0.137)
Year of Independence			0.146 (0.096)	0.058 (0.122)	0.111 (0.118)	0.061 (0.108)	0.046 (0.104)
Timing of Transition to Agriculture			−0.080 (0.131)	0.154 (0.168)	0.196 (0.155)	0.338** (0.135)	0.155 (0.135)
Continental FE	N	N	Y	Y	Y	Y	Y
Observations	143	143	143	143	143	143	143
R^2	0.14	0.20	0.53	0.32	0.35	0.44	0.48

Standardized beta coefficients reported, with robust standard errors in parentheses.

*** $p < 0.01$,

** $p < 0.05$,

* $p < 0.1$; See web Appendix for variables' sources and definitions.

Table 3

Selection into Language Mapping

	(1)	(2)	(3)
	Regions in Virtual Countries with Language Coverage	**Regions in Virtual Countries without Language Coverage**	**Standard Error**
Mean Elevation	0.6794	0.6581	(0.0150)
Mean Land Quality	0.3403	0.3390	(0.0030)
Mean Precipitation	0.0729	0.0736	(0.0004)*
Mean Temperature	13.2809	13.3542	(0.0780)
Observations	1206	1206	

Standard errors in parentheses clustered at the country level.

*** $p < 0.01$;

** $p < 0.05$;

* $p < 0.1$

See web Appendix for variables' sources and definitions.

Table 4

Main Specification for the Virtual Country Analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	**Log Number of Languages**						
Variation in Elevation		0.124*** (0.044)	0.114*** (0.033)	0.082*** (0.030)	0.118** (0.057)	0.093** (0.043)	0.091** (0.035)
Variation in Land Quality		0.144** (0.058)	0.088** (0.043)	0.116*** (0.033)	0.103** (0.048)	0.173*** (0.055)	0.134** (0.052)
Mean Elevation		−0.024 (0.033)	0.028 (0.053)	0.032 (0.072)	0.008 (0.154)	−0.019 (0.147)	−0.034 (0.107)
Mean Land Quality		0.043 (0.060)	0.042 (0.055)	0.065 (0.074)	0.262*** (0.075)	−0.026 (0.112)	−0.006 (0.098)
Absolute Latitude	−0.557*** (0.054)	−0.515*** (0.066)	−0.390*** (0.134)	−0.317 (0.270)	−0.216 (0.131)	−0.353 (0.272)	−0.402 (0.335)
Mean Precipitation			0.354*** (0.061)	0.191** (0.087)	0.222*** (0.081)	0.076 (0.114)	0.111 (0.124)
Mean Temperature			−0.118 (0.131)	−0.140 (0.202)	−0.054 (0.173)	−0.129 (0.219)	−0.209 (0.247)
Ln(Area)			0.059 (0.055)	0.038 (0.038)	0.197*** (0.067)	−0.012 (0.042)	0.046 (0.055)
Sea Distance			0.026 (0.039)	0.021 (0.042)	0.198*** (0.071)	0.008 (0.043)	−0.000 (0.036)
Water Area			0.037 (0.027)	0.005 (0.024)	0.026 (0.027)	0.013 (0.037)	−0.017 (0.036)
Within Country			−0.041 (0.034)	−0.041 (0.035)	−0.036 (0.057)	−0.087 (0.066)	
Number of Countries			0.197*** (0.033)	0.191*** (0.038)	0.121** (0.058)	0.288*** (0.063)	
Migratory Distance from East Africa			−0.284*** (0.050)	−0.087 (0.215)	1.000 (0.710)	−0.268 (0.272)	−0.234 (0.288)
Ln(Population Density in 1995)				0.023 (0.053)	0.169 (0.111)	0.006 (0.067)	0.039 (0.079)
Country FE	N	N	N	Y	Y	Y	Y
Observations	1663	1663	1663	1663	536	1127	994
R^2	0.31	0.36	0.53	0.70	0.73	0.56	0.66

Standardized beta coefficients are reported, with standard errors in parentheses clustered at the country level.

*** p < 0.01,

** p < 0.05,

* p < 0.1

(5) focuses on virtual countries in the tropics, (6) on virtual countries out of the tropics and (7) on virtual countries belonging entirely within an existing real country.

See web Appendix for variables' sources and definitions.

Table 5a

Robustness Checks for the Virtual Country Analysis

	(1)	(2)	(3)	(4)	(5)
VARIABLES	**Negative Binomial**	**OLS**	**OLS**	**OLS**	**OLS**
	Number of Languages	**Ln Number of Languages**			
Variation in Elevation	0.425*** (0.149)	0.383*** (0.129)	0.455*** (0.142)	0.492*** (0.126)	0.361*** (0.121)
Variation in Land Quality	1.074*** (0.387)	1.211*** (0.357)	1.128*** (0.363)	1.081*** (0.337)	1.108*** (0.350)
Country FE	Y	Y	Y	Y	Y
Observations	1663	1888	1324	1709	1651
R^2	.	0.71	0.73	0.69	0.71
Log-Pseudolikelihood	−3586.497	.	.	.	.

Nonstandardized coefficients reported, standard errors in parentheses clustered at the country level.

*** $p < 0.01$;

** $p < 0.05$;

* $p < 0.1$

All specifications include the same controls as those of column 4 in Table 4. Column (1) is estimated using a negative binomial and the dependent variable is the Number of Languages.

The dependent variable in (2)–(5) is the Ln Number of Languages. Column (2) includes all virtual countries irrespective of the population. (3) includes virtual countries with at least 50,000 inhabitants. (4) includes in the Number of Languages all linguistic groups of a virtual country irrespective of their population. (5) includes in the Number of Languages groups with at least 3,000 recorded speakers. See web Appendix for variables' sources and definitions.

Table 5b

Robustness Checks for the Virtual Country Analysis

VARIABLES	(1)	(2)	(3)	(4)	(5)
	Ln # of Languages				
Variation in Elevation	1.003*** (0.327)	0.360*** (0.129)		0.407*** (0.139)	0.428*** (0.135)
Variation in Land Quality	1.583*** (0.475)	1.194*** (0.295)			
Dispersion of Elevation			0.145*** (0.050)		
Dispersion of Land Quality			0.403*** (0.116)		
Variation in Climatic Suitability				0.661** (0.321)	
Mean Climatic Suitability				0.532** (0.212)	
Variation in Soil Suitability					0.875*** (0.304)
Mean Soil Suitability					(0.021) (0.183)
Country FE	Y	Y	Y	Y	Y
Observations	452	1663	1663	1663	1663
R^2	0.63	0.74	0.71	0.71	0.70

Nonstandardized coefficients reported, standard errors in parentheses clustered at the country level.

*** $p < 0.01$;

** $p < 0.05$;

* $p < 0.1$

All specifications include the same controls as those of column 4 in Table 4.

(1) focuses on virtual countries with linguistic information across all 25 complete underlying cells, (2) adds 100 fixed effects one for each percentile of the size distribution of virtual countries. Column (3) uses the dispersion in elevation and land quality, respectively. (4) uses the variation in climatic suitability to capture the heterogeneity in the suitability for agriculture. In this case mean land quality is proxied by mean climatic suitability for agriculture. (5) uses the variation in soil suitability to capture variation in the suitability for agriculture. In this case mean land quality is proxied by mean soil suitability for agriculture. See web Appendix for variables' sources and definitions.

Table 6

Main Specification for the Pairwise Analysis of Adjacent Regions

	(1)	(2)	(3)	(4)	(5)
VARIABLES	**Percentage of Common Languages**				
Difference in Land Quality	−0.111*** (0.032)	−0.038*** (0.012)	−0.054*** (0.018)	−0.048** (0.021)	−0.056*** (0.016)
Difference in Elevation	−0.091*** (0.014)	−0.051*** (0.006)	−0.050*** (0.016)	−0.046** (0.022)	−0.053*** (0.009)
Geodesic Distance Within the Pair	−0.244*** (0.050)	−0.303*** (0.039)	−0.382*** (0.041)	−0.199** (0.074)	−0.274*** (0.051)
Migratory Distance from East Africa	0.005*** (0.002)	−0.010 (0.078)	−0.021 (0.146)	−0.040 (0.192)	0.232** (0.093)
Absolute Latitude	0.005*** (0.001)	0.077*** (0.026)	0.006 (0.025)	−0.008 (0.049)	0.093** (0.042)
Mean Elevation	−0.006 (0.013)	−0.006 (0.028)	0.011 (0.044)	0.040 (0.051)	0.010 (0.046)
Mean Land Quality	−0.004 (0.024)	−0.003 (0.014)	0.021 (0.037)	0.018 (0.061)	−0.026 (0.034)
Mean Precipitation	−1.270*** (0.130)	0.089 (0.152)	−0.639 (0.661)	0.116 (0.646)	−0.004 (0.194)
Mean Temperature	0.001 (0.002)	−0.002 (0.007)	−0.008 (0.007)	−0.008 (0.010)	0.004 (0.010)
Sea Distance	−0.015 (0.014)	−0.138 (0.114)	−0.034 (0.120)	−0.039 (0.212)	−0.375*** (0.073)
Water Area	−0.012 (0.018)	−0.007 (0.010)	−0.009 (0.012)	−0.009 (0.019)	0.004 (0.011)
Ln(Population Density in 1995)	−0.005 (0.005)	−0.003* (0.002)	−0.005 (0.004)	−0.000 (0.008)	0.003 (0.002)
Ln(Area)	0.106*** (0.016)	0.034*** (0.010)	0.068*** (0.020)	−0.029 (0.026)	0.014 (0.012)
Within Country Indicator	0.144*** (0.010)	0.155*** (0.009)	0.136*** (0.012)	0.344*** (0.015)	0.145*** (0.012)
Region FE	N	Y	Y	Y	Y
Observations	156,570	156,570	35,305	11,975	74,830
R^2	0.28	0.73	0.74	0.70	0.72

Nonstandardized coefficients reported, standard errors in parentheses clustered at the country level.

*** $p < 0.01$;

** $p < 0.05$;

* $p < 0.1$

Specification (3) focuses on pairs within Africa, (4) on pairs within Europe, (5) on pairs within Asia

Table 7

Recent Migrations and Linguistic Diversity

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	**Ln Number of Languages**		**Percentage of Common Languages**		**Ln Number of Languages**	
Variation in Elevation	0.470*** (0.132)	−0.352** (0.168)			1.312*** (0.500)	0.5694 (2.4711)
Variation in Land Quality	1.220*** (0.406)	0.256 (0.361)			3.144*** (1.121)	1.465 (6.969)
Difference in Land Quality			−0.048*** (0.011)	−0.001 (0.008)		
Difference in Elevation			−0.052*** (0.007)	−0.038 (0.023)		
Country FE	Y	Y	N	N	N	N
Region FE	N	N	Y	Y	N	N
Continental FE	N	N	N	N	Y	Y
Observations	1352	311	124522	32048	121	21
R^2	0.71	0.56	0.74	0.69	0.72	0.97

Nonstandardized coefficients reported, standard errors clustered at the country level in (1), (2), (3), (4) and robust standard errors are reported in (5) and (6).

*** $p < 0.01$;

** $p < 0.05$;

* $p < 0.1$

(1), (3) and (5): countries where at least 40% of the current population can trace ancestry to 1500AD.

(2), (4) and (6): countries where less than 40% of the current population can trace ancestry to 1500AD.

(1) and (2): virtual country sample including all controls of column 4 in Table 4.

(3) and (4): pairs of contiguous regions sample including all controls of column 2 in Table 6.

(5) and (6): real countries sample including all controls of column 5 in Table 1.

Table 8

Specialization Pattern Across Non-Contiguous Partitions of the Same Language Group

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	**Pasture**		**Maize**		**Sorghum**	
Pasture Across Buffer Regions	0.749*** (0.094)	0.420*** (0.137)				
Pasture in the non-Adjacent Partition of the Same Group		0.307*** (0.161)				
Maize Across Buffer Regions			0.945*** (0.018)	0.881*** (0.050)		
Maize in the non-Adjacent Partition of the Same Group				0.540*** (0.144)		
Sorghum Across Buffer Regions					0.926*** (0.023)	0.853*** (0.059)
Sorghum in the non-Adjacent Partition of the Same Group						0.367*** (0.135)
Mean Land Quality	−0.102** (0.039)	−0.059** (0.028)	0.064 (0.206)	0.055 (0.182)	0.154 (0.155)	0.120 (0.149)
Observations	418	418	372	372	372	372
Log pseudolikelihood	−40.581	−30.356	−54.667	−43.563	−65.098	−57.339

The unit of analysis is a partition of a language group. Standard errors in parenthesis clustered at the language group level. All columns are estimated by maximum-likelihood Probit models.

*** $p < 0.01$;

** $p < 0.05$;

* $p < 0.1$

The coefficients report the change in the probability of the dependent variable induced by a change in the respective regressor. In (1) and (2) the dep. var. is an indicator that takes the value 1 if pasture dominates land use in the partition. In (3) and (4) the dep. var. is an indicator that takes the value 1 if maize is the dominant crop in the partition. In (5) and (6) the dep. var is an indicator that takes the value 1 if sorghum is the dominant crop in the partition. See web Appendix for variable's definitions.