The Origins of Ethnolinguistic Diversity: Theory and Evidence

Stelios Michalopoulos*
Tufts University

October 19, 2008

Abstract

This research examines theoretically and empirically the economic origins of ethnolinguistic diversity. The empirical analysis constructs detailed data on the distribution of land quality and elevation across contiguous regions, virtual and real countries, and shows that variation in elevation and land quality has contributed significantly to the emergence and persistence of ethnic fractionalization. The empirical and historical evidence support the theoretical analysis, according to which heterogeneous land endowments generated region specific human capital, limiting population mobility and leading to the formation of localized ethnicities and languages. The research contributes to the understanding of the emergence of ethnicities and their spatial distribution and offers a distinction between the natural, geographically driven, versus the artificial, man-made, components of contemporary ethnic diversity.
Keywords: Ethnic Diversity, Geography, Technological Progress, Human Capital, Colonization.
JEL classification Numbers: O11, O12, O15, O33, O40, J20, J24.

*Part of this research circulated earlier under the title "Ethnolinguistic Diversity: Origins and Implications." I am indebted to Oded Galor for his constant advice and mentorship. Daron Acemoglu, Roland Benabou, James Fearon, Andrew Foster, Ioanna Grypari, Peter Howitt, Masayuki Kudamatsu, Nippe Lagerlof, David Laitin, Ashley Lester, Ross Levine, Glenn Loury, Ignacio Palacios-Huerta, Stephen Ross, Yona Rubinstein, Francesco Trebbi and David Weil provided valuable comments. 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 Copenhagen, University of Connecticut, University of Cyprus, University of Gothenburg, University of Houston, Warwick University, 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.

1 Introduction

Ethnicity has been widely viewed in the realm of social sciences as instrumental for the understanding of socioeconomic processes. A rich literature in the fields of economics, political science, psychology, sociology, anthropology and history attests to this.[1] Nevertheless, the economic origins of ethnic diversity have not been identified, limiting our understanding of the phenomenon and its implications for comparative economic development.

This research examines theoretically and empirically the economic origins of ethnic diversity. The empirical investigation, conducted at various levels of aggregation, establishes that geographic variability, captured by the variation in regional land quality and elevation, is a fundamental determinant of ethnic diversity. In particular, the analysis shows that contemporary ethnic diversity displays a natural component and a man-made one. The natural component is driven by the diversity in land quality and elevation across regions, whereas the man-made one captures the idiosyncratic state histories of existing countries, reflecting primarily their colonial experience. The evidence supports the proposed theory according to which, heterogeneous land endowments generated region specific human capital, limiting population mobility and leading to the formation of localized ethnicities and languages.[2]

The identification of the geographical origins of ethnic group formation produces a wide range of applications. For example, the proposed distinction between the natural versus the man-made components of contemporary ethnic diversity raises the question of whether the well documented negative relationship between ethnolinguistic fractionalization and countries' economic performance, (see e.g., Easterly and Levine (1997), Fearon and Laitin (2003), Alesina et al. (2003) and Banerjee and Somanathan (2006) among others) reflects the direct effect of divergent state histories across countries, rather than a true effect of ethnic diversity on economic outcomes.[3] Additionally, the results 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 land endowments, whereas in relatively heterogeneous ones, and according to the evidence more ethnically diverse, the diffusion would be less rapid leading to the emergence of inequality across countries as well as ethnic groups.

This research argues that ethnicities and languages were formed in a stage of development when land was the single most important factor of production. Particularly, the theory suggests that differences in land endowments across regions gave rise to location specific human capital,

[1] See Hale (2004).

[2] Languages and ethnicities are arguably related but distinct dimensions of cultural heterogeneity. Nevertheless, indexes of ethnic and linguistic diversity are highly correlated. Henceforth, I will be using these terms interchangeably.

[3] Michalopoulos (2008) employes the proposed framework to uncover the causal impact of ethnolinguistic diversity on economic performance across regions and countries.

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.

The link between variable land endowments and ethnic diversity 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.[4] Along the same lines, this study argues that variation in elevation and land qualities across regions is the ultimate cause of the emergence and persistence of ethnic diversity.

The model uses a two-region overlapping generations framework. Human capital is specific to each area, accumulates over time through learning by doing and is available to the region's population.[5] In the beginning of each period, individuals compare the expected income that can be earned in their place of origin to that in case of moving. The incentive to move stems from regional productivity shocks. Transferring region specific know-how across places, however, is costly in the sense that the human capital of those who relocate may not be perfectly applicable to the production structure of the receiving place. According to the theory, these differences in the transferability of region specific knowledge gave rise to regional variation in population mixing and ultimately to distinct ethnolinguistic traits.[6]

In the empirical section I employ new data on land's agricultural suitability at a resolution of 0.5 degrees latitude by 0.5 degrees longitude to construct the distribution of land quality at a regional and country level. Such disaggregated level data, never before used in an economic application, allow for the econometric analysis to be conducted at various levels of aggregation. Specifically, to mitigate the problem of endogenous borders, inherent to the literature on cross-country regressions, I arbitrarily divide the world into geographical entities of a fixed size, called virtual countries. As predicted by the theory, I find that ethnic diversity, measured

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

[5] 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. The latter is bound to accumulate more slowly over time.

[6] One could argue that the intensity of trade between regions could be an independent force leading to a convergence in the regional cultural traits. However, one would expect that trade would be more intense between regions with distinct factor endowments, i.e. with different land characteristics. Such a prediction, nevertheless, is at odds with the empirical findings suggesting that any trade induced force towards ethnic homogenization is quantitatively dominated by the elements identified in the theory. An additional reason why the quantitative importance of trade appears to be limited may stem from the fact that whenever there are gains from trade to be made, customarily this is accompanied by the emergence of a class within a society specializing in the relevant activities rather than a uniform participation in trade across individuals. Similarly, the pursuit of economic diversification through marrying across regions of different productive endowments would also operate against finding a systematic positive relationship between ethnic diversity and heterogeneity in regional land qualities and elevation.

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 diverse societies. Overall, geographically diverse territories, that is places characterized by a wide spectrum of land qualities and variable altitudes, give rise and support more ethnic groups. The findings are robust to the inclusion of continental and country fixed effects which effectively capture any systematic elements related to the state and continental histories of these geographical units.

Taking further advantage of the information on where ethnic groups are located, a more demanding test of the theory's predictions is conducted in a novel empirical setting. In particular, focusing on pairs of adjacent regions I find that the difference in land quality and elevation between any two adjacent areas negatively affects ethnic similarity, as reflected in the percentage of common languages spoken within the regional pair. This finding demonstrates that (i) the difference in land quality and elevation between adjacent regions is a significant determinant of local ethnic diversity and (ii) the spatial arrangement of a given heterogeneous land endowment matters in determining the degree of the overall cultural heterogeneity.

Moving into a cross-country framework, the empirical findings obtained at the alternative levels of spatial aggregation are further validated. Countries characterized by more diverse land attributes exhibit higher levels of ethnolinguistic fractionalization. This highlights the fundamental role that regional land endowments have played in the formation of more or less ethnically diverse societies. Testing alternative hypotheses regarding the formation of ethnolinguistic diversity, focusing on differential historical paths and additional geographical characteristics, the qualitative predictions remain intact.[7]

Historical accidents have also influenced contemporary fractionalization outcomes. The European colonization after the 15th century, for example, is an obvious candidate. Europeans substantially affected the ethnolinguistic spectrum of the places they colonized. In particular, their active manipulation of the original ethnolinguistic endowment, including the introduction of their own ethnicities and the replacement of the indigenous populations, introduced a man-made component of contemporary ethnic fractionalization, tipping the balance in favor of an ethnic spectrum whose identity and size is not a natural consequence of the primitive land characteristics. This decomposition of contemporary ethnic fractionalization into a natural component, driven by the geographic variability, and a man-made one, offers new insights

[7] According to the theory, places experiencing persistent productivity shocks would be less ethnically diverse due to the resulting population mixing. Although the empirical focus of this study is not on testing this prediction, I find consistent results. Specifically, distance from the equator has a significant negative impact on ethnic diversity. This interpretation derives from the observation that distance from the equator correlates with more variable climates and, thus, more frequent productivity shocks. Note also that biodiversity generally decreases further away from the equator (Rosenzweig, 1995) effectively allowing for fewer productive niches along which groups of people may specialize.

regarding the origins and implications of ethnic diversity.

The results of this study are directly related to the literature on state formation, see Alesina and Spolaore (1997). In this literature, preference heterogeneity is a key determinant of the optimal size of a state. Taking into account that heterogeneous land endowments may be associated with distinct needs for public goods,[8] and establishing that these differences in land endowments are behind ethnic fragmentation, generate new insights about the relationship between state formation and ethnic diversity.

Another line of research, to which the findings are relevant, is a recent study by Spolaore and Wacziarg (2009). The authors document empirically the effect of genetic distance, a measure associated with the time elapsed since two populations' last common ancestors, on the pairwise income differences between countries. Larger genetic distance is associated with larger income differences. According to the proposed theory, population mixing, which affects genetic distance between two countries, is endogenous to the transferability of country specific human capital within the pair. The more similar the geographic endowments between two countries, the smaller should their genetic distance be, ceteris paribus. Therefore, the theory predicts that the uneven diffusion of technology across countries may be an outcome of the differences in society's specific human capital. By introducing the pair-wise country differences in the distributions of land quality and elevation, one can decisively improve upon the interpretation of the existing results.

The proposed theory also bridges the divide in the literature regarding the formation of ethnicities, by identifying the economic mechanism at work. There are two main strands of thought. The primordial one qualifies ethnic groups as deeply rooted clearly drawn entities, see Geertz (1967), whereas the constructivists or instrumentalists, see Barth (1969), highlight the contingent and situational character of ethnicity. In the current framework, it is the heterogeneity in regional land endowments that initially gives rise to relatively stable ethnic diversity, an element of primordialism. However, as the process of development renders land increasingly unimportant ethnic identity is ultimately bound to become less attached to a certain set of region specific skills and, thus, more situational and ambiguous in character. For example, Miguel and Posner (2006) provide evidence that ethnic identification in Africa becomes more pronounced as political and economic competition increases. Similarly, Rao and Ban (2007) provide evidence on the man-made component of ethnic diversity in India by showing how state policies and local politics have had an important impact on shaping caste structures over the last fifty years.[9]

[8] Irrigation projects, for example, would be much more complementary to farmers' needs than herders.

[9] In another recent study Caselli and Coleman (2006) provide a theory where ethnic traits provide a dimension along which voluntary coalitions may be formed and Esteban and Ray (2007) investigate the salience of ethnic

According to the theory, to the extent that ethnolinguistic groups are bearers of region specific human capital and land is a significant productive input, ethnicities would tend to disperse over territories of similar productive endowments. This prediction generates new insights for understanding the pattern of population movements like the spread of the first agriculturalists and herders following the Neolithic Revolution, the settlement intensity of colonizers across the colonized world as well as the contemporary spatial distribution of ethnic groups in general.

This study is a stepping stone for further research. Equipped with a more substantive understanding of the origins and determinants of ethnolinguistic diversity, long standing questions among development and growth economists, in which ethnic diversity plays a significant role, may be readdressed.

The rest of the paper is organized as follows. In section 2, historical evidence on the building blocks of the theory is presented. Section 3 advances the theory and its predictions. Section 4 discusses the data and shows empirically how geographic variability shapes production decisions. Section 5 presents the main part of the empirical analysis. This is conducted in a (i) cross-virtual country (ii) cross-pair of adjacent regions and (iii) cross-country framework. It includes the various robustness checks and concludes by focusing on the impact of the European colonizers on the ethnolinguistic endowment of the colonized world. Finally, section 6 summarizes the key findings and concludes.

2 Evidence on Migrations and Language Spreads

The theory rests upon three fundamental building blocks: (*i*) population movements influence the ethnolinguistic identity of the places involved (*ii*) ethnic groups and languages tend to disperse along places with similar productive endowments (*iii*) regional productivity shocks generate the incentive to relocate from one place to another.

Linguists have long recognized the role of population mixing in producing common linguistic elements between places. As Nichols (1997a) points outs "almost all literature on language spreads focuses on either demographic expansion or migration as the basic mechanism."[10] Both instances are a result of population movements towards territories previously unoccupied by their ancestors. As an outcome of population mixing, the regional populations experience a language shift either to or from the immigrants' language. Similarly, languages long in contact come to resemble each other in several dimensions like sound structure, lexicon, and grammar. This resultant structural approximation is called convergence. To the

identity on the eruption of civil conflict.

[10] Nichols (1997a) defines a spread zone as "an area of low density where a single language or family of languages occupies a large range."

extent that recurrent contact between regional populations may occur through repetitive cross-migrations, the modeling of the long run emergence of common ethnolinguistic characteristics as an increasing function of the intensity of population mixing between places is, thus, justified.

There are several examples showing that migrations have been occurring between places of similar productive characteristics. Linguistic research, in particular, has identified several regions of the world which are called "spread zones" of languages, that is, regions sustaining low linguistic diversity. These regions, in fact, are typically characterized by relatively homogeneous land endowments, as is the case for the grasslands of central Eurasia.

Examples of groups that migrated along areas that were similar to their region of origin include Austronesians and speakers of Eskimoan languages, who are coastally adapted peoples, and have accordingly spread along coasts rather than inland. Along similar lines, Bellwood (2001) argues that the spread zones of agriculturalists and their languages following the Neolithic Revolution trace closely land qualities that were amenable to agricultural activities. Considering languages of the Indo-European family, their expansion after the Neolithic revolution is embedded to the notion of "spread" and "friction" or "mosaic" zones.[11] Spread regions are characterized by similar land qualities where the early agriculturalists could easily apply their own specific knowledge. Friction zones on the other hand, are areas less conducive to such activities. In these places the populations maintained their distinct ethnolinguistic behavior. Examples of the latter include regions like Melanesia, Northern Europe and Northern India, see Renfrew (2000) for a comprehensive review. Early agriculturalists and pastoralists, perhaps not surprisingly, targeted and expanded into areas where their specific human capital would best apply, homogenizing them linguistically.[12]

In general, as long as land dominates the production process, ethnic human capital is bound to be tied to a set of regional productive activities and consequently the ethnic groups would target and disperse into territories similar to the region of origin, minimizing, thus, erosion of their human capital endowment.

Lastly, evidence suggests that climatic shocks, which in the context of the theory proxy

[11] Gray and Atkinson (2003) produce evidence demonstrating that Indo-European languages indeed expanded with the spread of agriculture from Anatolia around 8,000–9,500 years BP. The language tree constructed by the authors provides information about the timing of linguistic divergence within the Indo-European group. For example, at 7000 years BP (before present) Greek and Armenian diverge. At 5000 years BP, Italic, Germanic, Celtic, Indo-Iranian families diverge and at 1750 years BP the Germanic languages split between West Germanic (German, Dutch, English) and North Germanic (Danish and Swedish).

[12] Other relatively more recent examples of ethnic groups that consistently migrated to places where they could utilize their ethnic human capital, include the Greeks and the Jews, among others, who belong to the historic trade diasporas (Curtin, 1984). In this case, it is the knowledge of how to conduct commerce that allowed these groups to spread into areas where merchandising was both possible and profitable. Botticini and Eckstein (2005), for example, document the religiously driven transformation of the Jewish ethnic human capital towards literacy and the resulting urban expansion.

for productivity shocks, were indeed an important factor in generating movements of people.[13] For example, Nichols (1997) suggests that at least since the advent of the Little Ice Age in the late middle ages, highland economies have been precarious, whereas the lowlands, with their longer growing seasons, were relatively prosperous offering winter employment for the essentially transhumant male population of the highlands. This caused lowland dialects to spread uphill. Prior to the global cooling, however, lowlands were dry and uplands moist and warm. Under these conditions, with highlands being relatively more economically secure, upland dialects spread downhill, through a similar process. The linguistic patterns found in regions like central Caucasus and the highland spread of Quechua fall in this category.

3 The Basic Structure of the Model

Consider an overlapping-generations economy in which economic activity extends over infinite discrete time. In every period, the economy produces a single homogeneous good using land, labor and region specific technology as inputs to the production process. The supply of land is exogenous and fixed over time. There are two regions i and j. The regional labor supply is governed by the evolution of the region specific know-how, its transferability between the places and the state of the relative temporary productivity shock.

Each individual lives two periods and population size is fixed. In the first period, agents are economically idle, passively accumulating the specific know-how of the place they are born to. In the second period, they supply inelastically their unit of labor in one of the two regions and consume the earnings. Individuals' preferences are defined over consumption in the second period of their lives,[14] c_{t+1}, and are represented by a strongly monotone and strictly quasi-concave utility function, $U = u(c_{t+1})$.

3.1 Production of Final Output

Production in each area displays constant-returns-to-scale with respect to land and labor. The output produced at time t in region r, is $Y_t^r = (z_t^r h_t^r)(L_t^r)^{\alpha}(m^r X^r)^{1-\alpha}$; $\alpha \in (0,1)$, $r \in \{i, j\}$. The productivity shock in period t in region r is denoted z_t^r, the level of knowledge, h_t^r, in period t relevant to region r evolves over time through learning by doing - it is the region r specific human capital - L_t^r is the total labor employed in period t in region r, m^r represents the land quality and X^r is the size of land used in production, normalized to 1 for all r.

Suppose that there are no property rights over land.[15] The return to land in every period

[13] The independent role of regional climatic fluctuations in generating the differential timing of the transition to agriculture across places has been proposed by Ashraf and Michalopoulos (2007).

[14] Allowing both for endogenous fertility and intergenerational altruism the predictions would not be reversed.

[15] The modeling of the production side is based upon two simplifying assumptions. First, capital is not an

is therefore zero, and the wage rate in period t is equal to the output per worker produced at time t, y_t^r, where

$$y_t^r = (z_t^r h_t^r)\,(m^r/L_t^r)^{1-\alpha} \tag{1}$$

3.2 Accumulation of region specific technology

The level of regional technology available to the indigenous population at time t in region r advances as a result of learning by doing $h_{t+1}^r = \psi\,(h_t^r)\,,\ r \in \{i,\ j\}$ with $h_0^r = 1$, $\psi_{h_t^r} > 0$ and $\psi_{h_t^r h_t^r} < 0$. Since both region specific technologies start from the same initial level and follow the same law of motion, the technology available to the indigenous in each region is identical in every period, i.e. $h_t^i = h_t^j = h_t$. Differences in the accumulation rate of region specific technology would not alter the predictions of the model. As it will become apparent, it would in principle make people of the region enjoy a higher technological growth rate and less willing to move, ceteris paribus. Furthermore, it is not a priori clear which places should enjoy higher technological accumulation rates. The literature has stressed both the role of pure population density, which is proportional to the productivity of the land, see Galor and Weil (2000), and the "necessity as the mother of invention" in promoting technological progress. For the latter see Boserup (1965).

As adults, individuals may move freely from one region to the other.[16] However, this comes at a cost arising from differences in the region specific human capital. In particular, since the level of technology, h_t^r, is region r specific, relocation renders obsolete part of the knowledge the individual may apply as a worker in the receiving place. This erosion increases as places become increasingly different in the set of productive activities.

The following equation captures how the know-how of the region of origin is converted into units of know-how relevant to the receiving place:

$$k_t^r = (h_t^q)^{1-\varepsilon}\ \ \forall\ r,\ q \in \{i,j\},\ r \neq q, \quad 0 \leq \varepsilon \leq 1, \quad h_t^q \geq 1 \tag{2}$$

where k_t^r are the units of knowledge that a migrant may apply should she move to region r and ε captures the degree of erosion within a regional pair. Those characterized by more heterogeneous productive endowments score higher along this dimension. In the empirical

input in the production function, and second the return to land is zero. Allowing for capital accumulation and private property rights over land would complicate the model to the point of intractability, but would not affect the qualitative results. Specifically, if property rights were preassigned to the indigenous then the rental price of land would adjust as a result of the demand from migrants. Alternatively, property rights could be endogenized in a conflict model sharing the same basic properties as the current set up leading to qualitatively similar predictions.

[16] Including additional costs associated with moving, either as a result of time expended on relocating or in the form of a transfer to the indigenous in the receiving area would not change the results. It would, however, add an additional dimension along which places might differ.

section these differences in regional productive characteristics will be captured by differences in land endowments. Note that within a regional pair erosion of region-specific knowledge is symmetric. The properties of transferring region-specific technology across places, follow directly by differentiating (2). In particular, the migrant's know-how relevant to the receiving place decreases in the level of erosion between the regions, $\frac{\partial k_t^r}{\partial \varepsilon} < 0$, $\forall\ r \in \{i,j\}$. Second, the migrant's know-how relevant to the receiving place increases in the human capital of the place of origin, $\frac{\partial k_t^r}{\partial h_t^q} > 0, \forall\ r,\ q \in \{i,j\},\ r \neq q$. Third, there exist diminishing returns to the transferability of the know-how of the place of origin, $\frac{\partial^2 k_t^r}{\partial^2 h_t^q} < 0$, $\forall\ r,q \in \{i,j\}$, $r \neq q$. This captures that the accumulation of technology becomes increasingly region specific and, as a result, less useful in case of relocation.[17] Lastly, the transferability of region-specific knowledge decreases with the level of erosion, $\frac{\partial^2 k_t^r}{\partial h_t^q \partial \varepsilon} < 0, \forall\ r, q \in \{i,j\}, r \neq q$. In other words, an additional unit of domestic know-how is less applicable to the receiving region in pairs characterized by higher erosion.

Taking into account the common evolution of region specific human capital and the preceding discussion, it follows that the indigenous population of region r, that is individuals who work in the same region they are born to, have higher level of know-how compared to that of the migrants during the period the migrants arrive, that is the output per worker is higher for the indigenous population.[18] Specifically, using (1)

$$y_t^r = (z_t^r h_t^r)\,(m^r/L_t^r)^{1-\alpha} \quad and \quad y_t^{q\to r} = (z_t^r k_t^r)\,(m^r/L_t^r)^{1-\alpha} \tag{3}$$

$\forall\ r,\ q \in \{i,j\}, r \neq q$, where y_t^r is the output per indigenous worker of region r and $y_t^{q\to r}$ is the output per migrant-worker from region q working in region r.

3.3 Defining Common Ethnicity

A probabilistic framework regarding the formation of shared ethnolinguistic elements is adopted. Particularly, it is conjectured that the probability that individuals from regions i and j will share common traits increases in the intensity of population mixing between the two regions over time.[19] As individuals cross-migrate, they add their cultural traits from the place of origin

[17] Such diminishing returns could be conceived as an outcome of increasing specialization in the set of activities relevant for each region. At any given level of heterogeneity within a regional pair, further specialization in the respective activities diminishes the transferability of the additional know-how.

[18] It is useful to note that migrants' offspring have the same level of region specific human capital as the offspring of non-migrants. Gradual accumulation of the region specific technology for the offspring of immigrants would not alter the results. It could, however, create selection into reverse migration of the people whose ancestors were immigrants.

[19] Assuming that regions in the beginning are either ethnolinguistically fragmented or homogeneous does not affect the pattern of ethnolinguistic assimilation. Should the latter be the case, then distinct cultural practices would form regionally over time due to cultural drift, see Boyd and Richerson (1985).

to the cultural pool of the indigenous population. This addition may be an outcome of the pure interaction in everyday activities between the locals and the contemporary immigrants or may take the form of intermarrying. Although we do not explicitly model the household formation decision, the probability of mixed households would increase in the intensity of cross migration. Should this process occur repeatedly over time, then the respective regions would share an increasingly larger set of common practices. On the other hand, pairs of regions characterized by few cross–migrations would evolve to exhibit distinct ethnolinguistic characteristics.

Formally, let f_T denote the probability that places, i and j, observed at the end of period T will exhibit common ethnolinguistic elements:

$$f_T = \frac{\sum_{t=1}^{T} I_t}{T} \tag{4}$$

where I_t is an indicator function that takes the value of 1 if migration occurs in period t between regions i and j, irrespective of the direction, and 0 otherwise. Such formulation could alternatively be interpreted as an inverse measure of ethnic distance between the two regions. Note that this relationship applies in the long-run, so T should be thought as relatively large.[20] According to this definition pairs of places whose populations never mixed until period T would have zero probability of sharing common ethnic traits, or alternatively put, maximal ethnolinguistic distance. Alternative specifications of (4) could accommodate a potential "founder" effect, in which case earlier migrations have a larger impact than later ones in the formation of common ethnicity. Including both the occurrence and the actual size of migration in every period would reinforce the qualitative predictions.

Variations in the intensity of population mixing between regions are according to the theory the main determinant of ethnic diversity across places. The analysis below establishes how this intensity is shaped by the forces of the environment.

3.4 Labor Allocation Across Regions

Individuals in each period t maximize earnings. In the beginning of every period t, regional productivity shocks, z_t^r, which last for one period, are realized. Adults observe the realization of the shock and decide whether or not to migrate by comparing the respective incomes in (3).[21] Erosion of region-specific technology decreases potential income in case of relocation,

[20] Indeed, in the short run population mixing may increase diversity in the receiving place, see Willliamson (2006).

[21] Migration in this framework lasts for at least one generation. It would be straightforward to incorporate short term migration by allowing for several productivity shocks per generation per region. Accounting for seasonality in the climatic fluctuations, would strengthen the theoretical predictions. Conditional on the similarity of productive endowments, places characterized by higher seasonality would exhibit larger and more frequent short-term migration movements.

whereas a relatively higher productivity shock in the host area acts as an incentive for an agent to migrate. This is the fundamental trade-off created by the forces in the economy.

Consequently, in period t, after the realization of regional productivity shocks and before any migration movement, individuals in each region compare the potential income of either migrating or staying in the region of origin. Let $\{\lambda_t\}_{t=0}^{T}$ denote the sequence of the ratios of productivity shocks of region i relative to region j, that is $\lambda_t = \frac{z_t^i}{z_t^j}$. It follows that $\lambda_t > 0$ and $\lambda_t \gtreqless 1 \; iff \; z_t^i \gtreqless z_t^j$. Using (3) and substituting L_t^i, L_t^j with their values from the preceding period, individuals from region i have an incentive to move to region j in the beginning of period $t \; iff$:

$$y_t^{i\to j} > y_t^i \Rightarrow \lambda_t < \left(h_t^i\right)^{-\varepsilon} \left(\frac{m^j}{m^i}\frac{L_{t-1}^i}{L_{t-1}^j}\right)^{1-\alpha} \tag{5}$$

Similarly, individuals from region j are willing to migrate to region i in the beginning of period $t \; iff$:

$$y_t^{j\to i} > y_t^j \Rightarrow \lambda_t > \left(h_t^j\right)^{\varepsilon} \left(\frac{m^j}{m^i}\frac{L_{t-1}^i}{L_{t-1}^j}\right)^{1-\alpha} \tag{6}$$

It is obvious from (5) and (6) that the incentive to move depends on the relative size of the regional productivity shocks, the level of the specific human capital of the region of origin, the erosion that such a migration entails and the ratio of the population densities relative to the ratio of land qualities. Simple inspection of (5) and (6) shows that when individuals in one region strictly prefer to migrate then individuals in the other region strictly prefer not to.

Given the absence of mobility barriers, as long as either (5) or (6) obtains in the beginning of period t, population movement will be observed.

Let $M_t^{i\to j}$, $M_t^{j\to i}$denote the size of the population that migrates from region i to j and j to i, respectively, in period t. The size of the realized migration makes the marginal individual from the place of origin, indifferent between moving and staying where she was born. In particular, when in the beginning of the period t the incentive to migrate is from region i to region j, then once migration, $M_t^{i\to j}$, has taken place, (5) should hold with equality. Adding the size of the migration $M_t^{i\to j}$ to the population of the receiving region, j, subtracting it from the region of origin, i, and manipulating (5) the level of population movement may be explicitly derived as

$$M_t^{i\to j} = \frac{L_{t-1}^i - \left(\lambda_t \left(h_t^i\right)^{\epsilon}\right)^{\frac{1}{1-\alpha}} \frac{m^i}{m^j} L_{t-1}^j}{1 + \left(\lambda_t \left(h_t^i\right)^{\epsilon}\right)^{\frac{1}{1-\alpha}} \frac{m^i}{m^j}} \tag{7}$$

Note that the numerator of (7) is strictly positive, as a long as (5) holds in the beginning of period t. Similar reasoning applies to deriving the size of the labor movement from region j to region i. Specifically,

$$M_t^{j\to i} = \frac{\left(\lambda_t \left(h_t^j\right)^{-\epsilon}\right)^{\frac{1}{1-\alpha}} \frac{m^i}{m^j} L_{t-1}^j - L_{t-1}^i}{1 + \left(\lambda_t \left(h_t^j\right)^{-\epsilon}\right)^{\frac{1}{1-\alpha}} \frac{m^i}{m^j}} \tag{8}$$

Again, note that the numerator in (8) is strictly positive, as long as (6) holds in the beginning of period t.

3.5 The M^iM^j and M^jM^i loci

Given the definition of common ethnicity in (4) it is necessary to explore how the environment, captured by the degree of erosion, the regional population densities, the contemporary level of regional know-how and productivity shocks, determines the occurrence of population mixing in any period t.

The M^iM^j locus is the geometric locus of all tuples $\left(h_t^i, \lambda_t, \frac{L_{t-1}^i}{L_{t-1}^j}, \varepsilon\right)$ such that the marginal individual in region i is indifferent between moving, that is, $y_t^{i\to j} = y_t^i$. In particular, $M^iM^j \equiv \left\{\left(h_t^i, \lambda_t, \frac{L_{t-1}^i}{L_{t-1}^j}, \varepsilon\right) : y_t^{i\to j} = y_t^i\right\}$. Solving explicitly for the level of the relative productivity shock in period t, $\lambda_t|_{M^iM^j}$, that makes people in region i indifferent to moving I get:

$$y_t^{i\to j} = y_t^i \Rightarrow \lambda_t|_{M^iM^j} = \left(\frac{L_{t-1}^i}{L_{t-1}^j}\frac{m_j}{m_i}\right)^{1-\alpha} \left(h_t^i\right)^{-\varepsilon} \tag{9}$$

Similarly,

$$y_t^{j\to i} = y_t^j \Rightarrow \lambda_t|_{M^jM^i} = \left(\frac{L_{t-1}^i}{L_{t-1}^j}\frac{m_j}{m_i}\right)^{1-\alpha} \left(h_t^j\right)^{\varepsilon} \tag{10}$$

As it is evident in (9) and (10) the ratio of the regional population densities from the last period is important in determining the no-migration loci. In Appendix A equations (A1) and (A2) show that the ratio of regional population densities in period $t-1$ is a function of the population densities generated by the last population movement across places in period s. The following lemma summarizes the properties of the migration indifference curves.

Lemma 1 *The properties of the non-migration loci:*

The M^iM^j locus			*The M^jM^i locus*		
$\frac{\partial\lambda_t}{\partial h_t^i}\Big\vert_{M^iM^j} < 0$	&	$\frac{\partial^2\lambda_t}{\partial^2 h_t^i}\Big\vert_{M^iM^j} > 0$	$\frac{\partial\lambda_t}{\partial h_t^j}\Big\vert_{M^jM^i} > 0$	&	$\frac{\partial^2\lambda_t}{\partial^2 h_t^j}\Big\vert_{M^jM^i} < 0$
$\frac{\partial\lambda_t}{\partial\varepsilon}\Big\vert_{M^iM^j} < 0$	&	$\frac{\partial^2\lambda_t}{\partial^2\varepsilon}\Big\vert_{M^iM^j} > 0$	$\frac{\partial\lambda_t}{\partial\varepsilon}\Big\vert_{M^jM^i} > 0$	&	$\frac{\partial^2\lambda_t}{\partial^2\varepsilon}\Big\vert_{M^jM^i} > 0$

Proof. First, substitute in (9) the two possible realizations of the past population densities, either (A1) or (A2), and differentiate accordingly. Repeat the same process for (10). □

Figure 1*a* shows the effect of the erosion, ε, on the occurrence of migration. As it follows from Lemma 1, conditional on the past that is on λ_s, h_s^j, and h_s^i, the distance between the no-migration loci, M^jM^i and M^iM^j, increases with the level of erosion. Given the contemporary relative productivity shock, λ_t, pairs of regions i and j with more dissimilar productive structures, i.e. higher ε, experience infrequent population mixing limiting the formation of common ethnolinguistic traits. Figure 1*b* is drawn with a higher level of region specific technology than 1*a* to exemplify the adverse effect of the accumulation of region specific human capital on migration outcomes. Note that in the absence of erosion, i.e. at $\varepsilon = 0$, regional knowledge is perfectly applicable across areas, as it is effectively general. In this case, the migration loci coincide and all it matters for migration is the relative size of the current ratio of regional productivity shocks, λ_t, with respect to λ_s, where s is the last period cross-migration occurred.

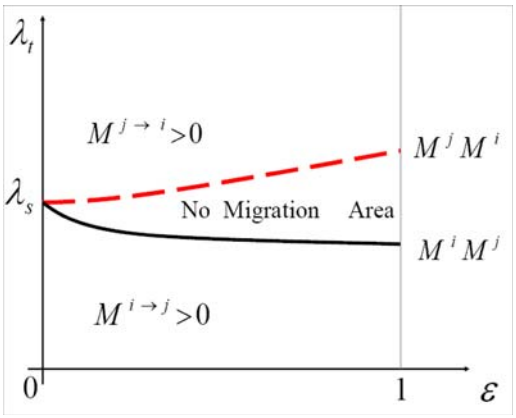


Figure 1a

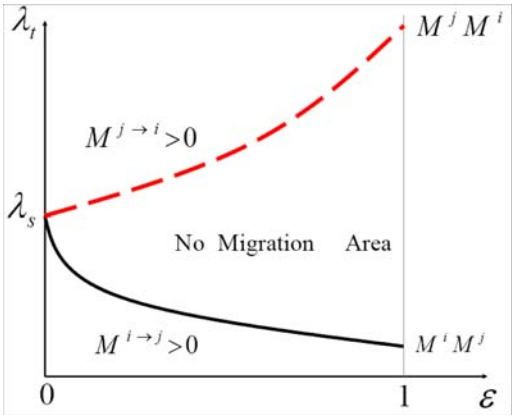


Figure 1b

In the set of figures above, it is evident the role of the temporal variation in regional productivity shocks in inciting or inhibiting migration patterns. Conditional on any level of erosion and region specific technology, which jointly determine the no migration area, the larger the difference between the temporary shock λ_t, and λ_s, the more probable is the occurrence of migration. Lemma 2 in Appendix A summarizes the cases of migration occurrences.

3.6 The Formation of Common Traits Over Time

Having established how the environment shapes population mixing, the formation of common ethnolinguistic elements may be traced over time. In period $t = 0$, the region specific technology is at its minimum, $h_0^i = h_0^j = 1$, since no accumulation has occurred yet, and individuals distribute themselves in places i and j such that the output per capita at time $t = 0$ is the

same across regions. It is assumed that the relative productivity shock, λ_t, is a discrete random variable independently and identically distributed over time. In particular,

$$\lambda_t = \begin{cases} \lambda_{\min} & \text{with probability} \quad p \\ \\ \lambda_{\max} & \text{with probability} \quad 1-p \end{cases} \tag{B1}$$

with $\lambda_{\min} < \lambda_{\max}$.[22] The following Proposition shows how erosion, ε, the ratio of the relative productivity shocks, λ_t/λ_s, and the level of region specific technology determine the probability that two regions will share common cultural elements.

Proposition 1 *Under (B1)*

1. *The probability that regions i and j share common ethnolinguistic traits as observed in period T, weakly decreases in the size of the erosion, ε,*

$$\frac{\partial f_T\left(\varepsilon;\lambda_t,\lambda_s,h_T\right)}{\partial \varepsilon} \leq 0$$

2. *The probability that regions i and j share common ethnolinguistic traits as observed in period T, weakly increases in the variance of the regional productivity shock, λ_t,*

$$\frac{\partial f_T\left(\lambda_t;\varepsilon,\lambda_s,h_T\right)}{\partial var\left(\lambda_t\right)} \geqslant 0$$

3. *The probability that regions i and j share common ethnolinguistic traits as observed in period T, weakly decreases in the level of region specific human capital in period T, h_T,*

$$\frac{\partial f_T\left(h_T;\varepsilon,\lambda_t,\lambda_s\right)}{\partial h_T} \leq 0$$

Proof. See Appendix A. □

Proposition 1 underlines the key role geographic conditions play in the formation of common ethnolinguistic traits. The adverse effect of an increase in the region specific know-how on the formation of common cultural elements stems from diminishing returns in the transformation of regional knowledge to units of knowledge relevant to the host region.[23] In Appendix A it is shown that the probability that two regions share common elements weakly increases both when productivity shocks differ intertemporally, i.e. $\lambda_t/\lambda_s \neq 1$, and by the

[22] This distributional assumption allows to explicitly follow the occurrence of migration pattern over time. Specifically, as it will become evident it disallows for successive migrations to occur towards the same region, reducing, thus, the cases to consider at any point in time. Different distributions of temporary productivity shocks would not affect the qualitative results.

[23] To the extent that the duration of human settlements is a proxy of the level of region specific human capital, the empirical finding of Ahlerup and Olsson (2007) that the former positively affects ethnic diversity is consistent with the third prediction of Proposition 1.

absolute distance between shocks, $|\lambda_t - \lambda_s|$. The variance of the regional productivity shocks, $var(\lambda_t)$, is a sufficient statistic that captures both dimensions. Ultimately, and perhaps more importantly, more heterogeneous productive structures across places summarized by ε, hinder population mixing. Consequently, low transferability of region specific human capital resulted in increasing inertia across regional populations, leading eventually to entrenched ethnicities tied to each locality. This will be the focus of the empirical analysis.[24]

The following section presents the data and the empirical strategy.

4 Empirical section

4.1 The Data Sources

To test the main theoretical prediction, an index of the transferability of region specific human capital is needed. The ideal index could be derived by examining the distribution of productive activities across regions, in a period of human history when the formation of cultural traits was taking place. Such quest for detailed data is bound to be an overwhelming endeavor. To overcome this issue I employ an alternative strategy. Given that ethnicities were formed at a point in time when land was the single most important input in the production process and in absence of historical data, I use contemporary disaggregated data on the suitability of land for agriculture and data on elevation, to proxy for the regional productive characteristics.

The intuition for using differences in land quality and elevation as the ultimate determinants of the differences in productive activities across regions is the following. Farming would be the dominant form of production in places characterized by high land quality, with the regions possibly differing in the optimal mix of plants and crops under cultivation. That is, even within agriculture, the specificity of human capital derives from the different crops produced regionally. However, herding/pastoralism is bound to be more widespread at intermediate and low levels of land quality, exactly because agriculture is less suitable in such areas. At very low levels of land quality being a middleman has been perhaps the most widespread activity as the case for cultures residing along trade routes suggests.[25] Along similar lines, different

[24] The predictions of the theory are consistent with the pre(historic) evidence about the formation of homogeneous linguistic areas across regions of common productive endowments. Also, the increased linguistic diversity in climates characterized by low climatic volatility, coupled with the low linguistic diversity at higher latitudes where regions are subject to seasonal fluctuations support the theoretical prediction that pairs of regions characterized by recurrent productivity shocks are bound to form homogeneous ethnolinguistic traits. This prediction is in line with the finding of Nettle (1996) that countries facing higher ecological risk sustain lower linguistic diversity.

[25] A famous example includes the trading routes of West Africa from the $5th$ - $15th$ century AD. These routes ran north and south through the Sahara and traded commodities like gold from the African rivers, salt, ivory, ostrich feathers and the cola nut. In absence of these trading routes, such places would hardly maintain any other activity, and this is a prime example where the regional knowledge, of how to transfer goods safely through a certain passage, is entirely location specific and thus almost impossible to transfer to other places.

altitudes are known to impose limits on the extent of agriculture as well as on the very choice of cultivated crops, see Grigg (1995). The next section provides empirical evidence which shows that geographic variability, as captured by the heterogeneity in land suitability for agriculture and elevation, is a significant determinant of actual crop diversity. Note that differences in elevation are likely also to be associated with higher transportation costs in case of relocation, further deterring population mobility.

The global data on agricultural suitability were assembled by Ramankutty et al. (2002) to investigate the effect of the future climate change on contemporary agricultural suitability.[26] This dataset provides information on land quality characteristics at a disaggregated level. 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 (i) empirically fit a relationship between the percentage of croplands around 1990 and both climate and soil characteristics and (ii) use the derived relationship to generate the regional suitability for agriculture across the globe.

The climatic characteristics are based on mean-monthly climate conditions for the 1961–1990 period and capture (i) monthly temperature (ii) precipitation and (iii) potential sunshine hours. All these measures weakly monotonically increase the suitability of land for agriculture. Regarding the soil suitability the traits taken into account are a measure of the total organic content of the soil (carbon density) and the nutrient availability (soil pH). The relationship of these indexes with agricultural suitability is non monotonic. In particular, low and high values of pH limit cultivation since this is a sign of soils being too acidic or alkaline respectively. Note that the derived measure does not capture topography and irrigation.

The resolution is 0.5 degrees latitude by 0.5 degrees longitude, thus the average cell has a size of about 55 km by 35 km. In total there are 58920 observations.

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

With respect to the cross-virtual country and cross-pair of adjacent regions analysis, ethnic diversity is captured using information on the location of linguistic groups. In the case of virtual country regressions the number of languages within each geographical 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 within each pair of adjacent regions. Data on the location of linguistic groups' homelands are obtained

[26] Appendix H provides a summary of the data sources used in this study.

from the Global Mapping International's World Language Mapping System. This dataset is covering most of the world and is accurate for the years between 1990 and 1995. Languages are based on the 15th edition of the Ethnologue database on languages around the world.[27]

In the cross-real country analysis a wealth of alternative measures of ethnic diversity is available. The measure of fractionalization widely used is the probability that two *individuals* randomly chosen from the overall population will differ in the characteristic under consideration, i.e. ethnicity, language, religion. The results presented below use the index most widely employed in the literature which is the ethnolinguistic fractionalization index, ELF, based on data from a Soviet ethnographic source, *Atlas Narodov Mira (Atlas of the People of the World) (1964)*, and augmented by Fearon and Laitin (2003). This index represents for each country the probability that two individuals randomly drawn from the overall population will belong to different ethnolinguistic groups. Using the linguistic, ethnic and religious fractionalization indexes constructed by Alesina et al. (2003), the absolute number of ethnic or linguistic groups derived by Fearon (2003) or the ethnic fractionalization measure proposed by Montalvo and Reynal-Querol (2005), the qualitative results are similar.[28]

4.2 The Properties of Geographic Variability and Productive Decisions

The distribution of land quality varies considerably across regions and across countries. For example, the following graph plots the distribution of regional land quality for Swaziland and Bhutan. In Swaziland the quality of land is concentrated around high values with average quality, $avg = 0.69$, and a *range* (this is the difference between the region with the highest land quality from that with the lowest) of 0.29.[29] On the other hand, land quality in Bhutan averages 0.30 and it spans a much larger spectrum. In fact, $range_{Bhutan} = 0.69$. The difference in elevation between these two countries is similar with Bhutan exhibiting a much larger diversity in altitudes.

[27] The data are available at www.gmi.org. To identify which languages are spoken within the unit of analysis I use the information on the location of language polygons. Each of these polygons delineate a traditional linguistic homeland; populations away from their homelands (e.g. in cities, refugee populations, etc.) are not mapped. Also, the World Language Mapping System does not attempt to map immigrant languages. Finally, linguistic groups of unknown location, widespread languages i.e. languages whose boundaries coincide with a country's boundaries and extinct languages are not mapped and, thus, not considered in the empirical analysis.

[28] Modifying the current framework to uncover the determinants of ethnic *polarization* is a topic for future research.

[29] The figure shows the kernel density estimate (weighted by the Epanechnikov kernel) of regional land qualities for each country.

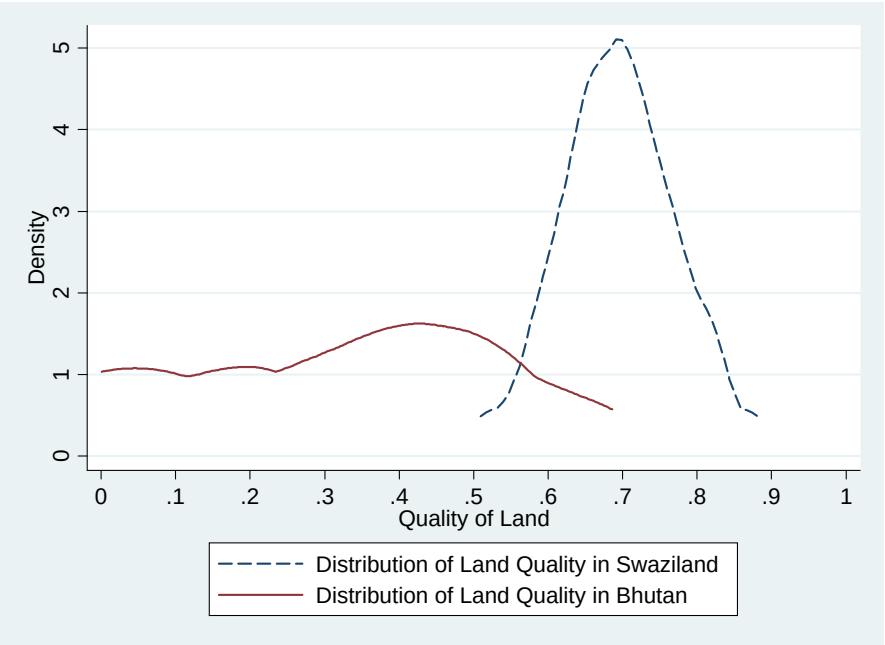


Figure 2

The range of land quality, i.e. the support of the distribution within the respective unit of analysis, and the standard deviation of elevation, *elev_sd*, are the statistics used to capture the degree of geographical heterogeneity.[30] These capture, albeit imperfectly, how readily location specific knowledge may be transferred across places. Intuitively, a larger range and/or a more variable topography implies that the geographical unit is composed of territories with increasingly different underlying productive characteristics, effectively enlarging the set of activities along which groups may specialize. The larger the spectrum of land qualities and the variation in elevation, the less transferable is the regional know-how. Thus, according to the theory, higher geographic diversity would increase the probability of ethnically distinct regions, ceteris paribus.[31,32] Indeed, going back to the example of Swaziland and Bhutan, ethnolinguistic

[30] The standard deviation of regional land quality is an alternative measure of a country's productive heterogeneity. Such proxy inherently captures variation both in the extensive, that is, in the extremes of the distribution of the land endowment, and the intensive margin. Conditional on the range, however, increases in the standard deviation of the endowment increase the weight towards the fixed extremes of the land quality distribution. This effectively results in fewer distinct land qualities along which groups may specialize. A further consequence of such an increase is that it causes a more unequal distribution of population across regions and since by construction the fractionalization indexes at the real country level are affected by the distribution of the population across ethnic groups (see below) an increase in the intensive margin may decrease fractionalization. Results not shown, indeed suggest that controlling simultaneously for the range and the standard deviation of land quality both enter significantly, the range with a positive sign and the standard deviation with a negative one. It should be noted, nevertheless, that the results, although quantitatively smaller for the reasons mentioned here, remain qualitatively intact when we use only the standard deviation.

[31] Dividing land quality into different categories according to the degree of suitability and calculating a measure of land quality fractionalization similar to how ethnic fractionalization is constructed, delivers results very similar to the ones presented here.

[32] The average quality of land, *avg*, according to the theory, should not directly effect ethnic diversity, because

fractionalization in Swaziland is only 0.38 compared to the highly ethnolinguistically fragmented society of Bhutan with $ELF_{Bhutan} = 0.69$.

The narrative so far suggests that geographic variability should manifest itself into different productive choices. Appendices $C1$ and $C2$ provide evidence on this direction. Appendix $C1$, in particular, demonstrates how different land qualities dictate the choice between pastoralism versus agriculture across ethnic groups in Kenya.

Appendix $C2$ shows how geographic diversity shapes farming decisions. Specifically, using data on the global distribution of major crops cultivated around 1990, I calculate the number of crops across countries, $nmbr\_crops$. The regression results in Table 1 show that countries endowed with larger variation in elevation, $elev\_sd$, and more diverse land qualities, $range$, systematically cultivate a larger number of major crops. Figures $5a$ and $5b$ present the partial scatter plots as generated by the regression in Table 1 of the number of crops cultivated against the variation in elevation and diversity in land quality respectively. Regarding the rest of the controls included in the regression in Table 1, the average level of land quality, avg, is not significantly related to crop variety. As expected larger countries appear to grow more crops. Further, countries in Western Europe, denoted by reg_we, cultivate systematically fewer crops whereas a typical country in Sub-Saharan Africa, denoted by reg_ssa, exhibits systematically larger crop diversity. These results strengthen the claim that variation in elevation and land quality diversity are the primitive elements behind productive choices.[33]

Using contemporary geographic data to proxy for differences in productive activities several centuries back in time presents its own potential pitfalls which merit further discussion. For example, a potential concern is how representative these geographical characteristics are of a period when ethnic groups were being formed. Regarding the elevation index, despite some local natural events and human interventions at a very local level, overall altitudes have not changed significantly since the retreat of the last Ice Age. Things are slightly more complicated regarding the land quality index. This is because precipitation, temperature and soil properties may have changed regionally over the last 5000 years. Hence, this measure of land quality is a noisy index of what might have been the true distribution of the land's agricultural quality in the past. This makes the task of identifying a relationship between land quality heterogeneity

if places are productively homogeneous then the regional know-how is perfectly applicable across all pockets of land, i.e. erosion is zero, irrespective of the level of land quality. Nevertheless, a higher land quality by sustaining denser populations may affect the path of a country's economic development, indirectly influencing ethnic diversity. I return to this point in the regression analysis.

[33] It should be noted that using the actual crop diversity to explain ethnic diversity is not an appealing approach for several reasons. Crop choice is endogenous to a host of things like the level of economic development, among others, so if ethnic diversity affects economic development and development affects crops cultivated then in that case causality would run from ethnic diversity to crop diversity. Also, the number of crops grown around 1990 is a limited measure of productive diversity since it captures heterogeneity only *within* farming. These considerations advise against using the crop diversity as a predictor of ethnic diversity.

and ethnic group formation harder.

Another concern is whether the results are subject to reverse causality. Variation in elevation is plausibly exogenous and not subject to human intervention at the regional scales the study investigates. However, diversity in land quality may be endogenous to human activities. In particular, the part of the index that depends on soil characteristics. This makes land quality possibly endogenous to the duration of agriculture and herding. Reassuringly, controlling for the timing of the rise of agriculture does not affect the results. Also, it is important to note that soil quality is itself endogenous to the regional climate. Comparing the global distribution of annual precipitation with the distribution of soil pH, it is evident that regions receiving a lot of precipitation are characterized by highly acidic soils, whereas in places with low precipitation the soil becomes alkaline.[34]

Although one cannot rule out entirely the possibility of reverse causality running from exogenous group specific subsistence practises to soil diversity, this would only be operative at small changes in soil quality. It would seem unlikely to posit that herders in Kenya, for example, transformed their lands into semi-deserts because of their herding cattle and camels and that agriculturalists transformed their own territories into fertile lands by systematically planting certain crops. If anything it would be the agricultural practises leading to a deterioration of the land's soil properties.

Having discussed the properties of geographic variability and established how it shapes production decisions we are ready to turn to the main empirical results.

5 Empirical Results

5.1 Cross-Virtual Country Analysis

Before going into the cross-country analysis, it is important to investigate whether the predictions of the theory obtain at an arbitrary level of aggregation. Finding that geographical diversity leads to higher ethnic diversity, irrespective of country borders, will greatly enhance the validity of the proposed theory and alleviate any concerns related to border and country formation inherent to any cross-country analysis.

The way that the artificial countries are constructed is the following. First, I generate a global grid where each regional unit is 2.5 degrees longitude by 2.5 degrees latitude and then I intersect it with the global data on land quality and elevation (see map $1b$ in Appendix B with the resulting artificial countries which constitute the unit of analysis). Using alternative dimensions like 4 by 4 or 5 by 5 degrees does not change the results.

[34] These maps are available at http://www.sage.wisc.edu/atlas/maps/anntotprecip/atl_anntotprecip.jpg and http://www.sage.wisc.edu/atlas/maps/soilph/atl_soilph.jpg respectively.

For each virtual country, I construct the distribution of land quality and elevation and calculate the number of unique languages spoken. In particular, I focus on languages with at least 1% area coverage within an artificial country. The latter captures the level of ethnic diversity, denoted *nmbr_lang*. Including all languages irrespective of their spatial extent or only focusing on those languages with at least 2% of area coverage within a virtual country, the results remain qualitatively intact.

In the regression analysis the sample of virtual countries is restricted in the following way. Territories for which there are at least 3 regions with information on land quality, elevation and languages are included. Also, to ensure that the findings are not driven by including in the regressions regions with negligible population density, only virtual countries whose individual regions have at least one person per sq km are considered.[35] Given these considerations a kernel density estimate of the distribution of the number of languages spoken across virtual countries is shown in Figure 3:[36]

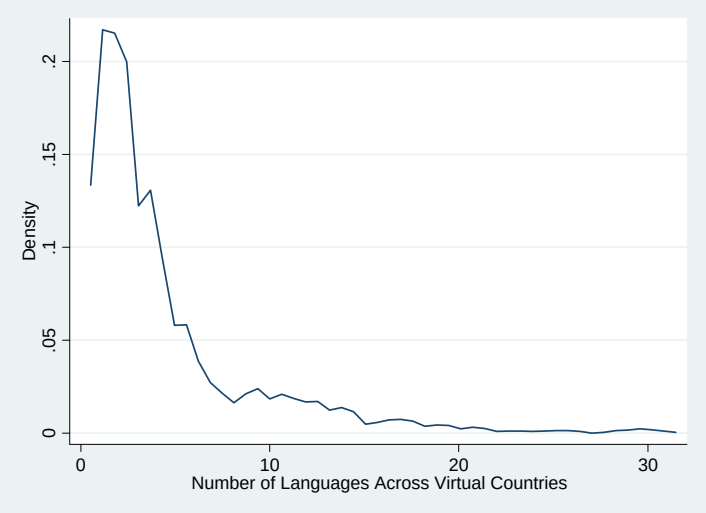


Figure 3

The resulting sample size is 1373 observations with a median of 25 regional land quality observations per virtual country. Descriptive statistics and the raw correlation between the variables used in the regressions are presented in Tables 2*a* and 2*b*. As one might expect, diversity in land quality, denoted by *range*, is higher in larger virtual countries, where *areakm*2 denotes the area of a virtual country, as well as in virtual countries characterized by more variable elevation, *elev_sd*.

[35] The population density data come from the Center for International Earth Science Information Network (CIESIN), Columbia University (2005) and were aggregated at the resolution level of the land quality data.

[36] Note that the distribution of the number of languages is skewed so instead of the levels the log of languages, *lnnmbr_lang*, is used in the regressions below. Excluding the extremely linguistically fragmented artificial countries, i.e. those with more than 20 languages spoken, the qualitative results are similar.

In each artificial country, there are on average 3.03 languages spoken and the pairwise correlations of both the spectrum of land qualities, *range*, and *elev_sd* with the number of languages are positive and large, 0.27. Map 2 in Appendix *D* shows one example of a virtual country. The circles, which are located in the centroids of the original cells, represent the regional land quality for agriculture. The different colored polygons represent the locations of the different linguistic groups. The virtual country in 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[37] and the spectrum of land qualities is 0.89, ranging from places that are totally inhospitable to agriculture to areas where the climate and the soil are highly conducive to cultivation.

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

$$\ln nmbr\_lang_i = \beta_0 + \beta_1 range_i + \beta_2 elev\_sd_i + \beta_3 X_i + \xi_i \tag{11}$$

where $\ln nmbr\_lang_i$ is the log number of languages spoken in virtual country i, $range_i$ is the support of the distribution of land quality, $elev\_sd_i$ is the variation in elevation and X_i is a vector of other geographical and political controls. The key prediction of the theory is that the greater the geographic variability across regions within virtual countries, the higher is the probability that these regions will bring forward and sustain more ethnically diverse societies.

This main prediction is corroborated across all alternative specifications of Table 3.[38] In the first regression of Table 3 both *elev_sd* and the *range* have a large and significant positive impact on linguistic diversity. A two-standard deviation increase in *range* increases linguistic diversity by 24% adding on average 0.72 languages to an average virtual country whereas a two-standard deviation increase in *elev_sd* increases linguistic diversity by 20%, adding on average 0.61 languages to an average virtual country. These are novel and economically important findings that reveal the geographic origins of contemporary ethnolinguistic diversity.

In the same specification, an array of additional geographical features are simultaneously accounted for. In particular, the size of each artificial country, *areakm2*, the average land quality, *avg*, the latitudinal distance from the equator, *abs_lat*, the number of real countries a virtual country falls into, *nmbr_cntry*, a dummy for the units that belong as a whole to an existing country, *in_cntry*, the area under water, *waterarea*, as well as the distance from the

[37] Namely these are: Central Kurdish, Gurani, Koy Sanjaq Surat, North Mesopotamian Spoken Arabic, Sangisari, South Azerbaijani and Southern Kurdish. Languages' traditional homelands may overlap. In this particular grid, for example, places that speak Gurani also speak Northern Kurdish.

[38] The results presented here are OLS estimates with the standard errors adjusted for spatial correlation following Conley (1999). This correction requires the choice of a cutoff distance, beyond which artificial countries do not influence each other. After projecting the world into the euclidean space using the Plate Carrée projection I use a cutoff distance of 2500 km. Results are similar using 1000 km, 3000 km, and 6000 km.

coastline, *sea_dist*, are controlled for. Larger artificial units sustain more languages. Areas that entirely belong to a single country display systematically lower ethnic fragmentation, whereas the more real countries a virtual country falls into, the more languages it sustains. This evidence points towards the effect of state formation on ethnic diversity. The distance from the equator itself enters negatively and significantly, consistent with the prediction that more climatically variable environments lead to lower ethnic diversity. Average land quality does not seem to affect linguistic diversity significantly. The variable capturing under water areas, *waterarea*, enters negatively and is marginally statistically significant losing significance in the rest of the specifications. This raises the issue of whether water bodies are a barrier or a facilitator of population mobility. Finally, the distance from the shoreline of an artificial country, *sea_dist*, does not systematically affect linguistic diversity. Overall, these geographical characteristics capture 45% of the variation in linguistic diversity across virtual countries.

The statistical and quantitative importance of geographic diversity is robust to alternative specifications. In particular, taking advantage of the arbitrarily drawn borders of these geographical units one may explicitly control for real country and continental fixed effects.[39] This is done in all subsequent specifications. Such inclusion of powerful controls, not possible in a cross-country framework, allows to explicitly take into account any systematic elements related to the state histories of existing real countries and, thus, produce reliable estimates of the effect of diversity in land quality on ethnic diversity. The inclusion of country and continental fixed effects in the second column of Table 3 only slightly changes the coefficients on *range* and *elev_sd*.

Columns 3 and 4 of Table 3 investigate whether the identified effect of geographic variability is driven by the inherent differences between regions in the tropics and the rest of the climatic zones. In column 3, the sample is restricted to virtual countries out of the tropics.[40] The estimated coefficient on *range* remains largely unchanged whereas the coefficient on the variation in elevation increases by almost 50%. This implies that out of the tropics variation in elevation is quantitatively a relatively more important determinant of linguistic diversity. This pattern reverses, however, when one examines the impact of geographic variability on ethnic diversity in the tropics, see column 4 of table 3. Across virtual countries in the tropics the coefficient of variation in elevation becomes less precisely estimated whereas diversity in land quality remains qualitatively and quantitatively significant. Within the tropics virtual countries with higher average land quality are characterized by larger linguistic diversity whereas

[39] For artificial countries falling into more than one real countries they are assigned the value of zero across the real country dummies. Alternatively, for these virtual countries one could assign as country dummies instead of zeros the fraction of the virtual country's area that falls into each real country. Doing so does not change the results.

[40] The tropics extent from 23.5 latitude degrees south to 23.5 latitude degrees north.

the opposite is true for virtual countries out the tropics. Also, within the tropics distance from the coast line enters significantly with a positive sign.

In column 5 of Table 3 the main specification (11) is estimated focusing on artificial units that entirely belong to a single existing country. This robustness check allows to investigate whether the estimated strong positive relationship between geographic variability and ethnic diversity obtains across regions within existing countries. Reassuringly, the variation of land quality across regions within countries systematically shapes ethnolinguistic diversity. Namely, territories within countries that display more heterogeneous land endowments give rise and sustain more ethnic and linguistic groups. A one standard deviation increase in both land quality diversity and variation in elevation increases by 30% the number of languages within an artificial country contributing significantly to the formation of ethnically diverse societies.

This section establishes that heterogeneity in land quality and elevation across virtual countries are both significant determinants 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 and continental fixed effects brings into light the, so far neglected, geographical origins of ethnic diversity.

5.2 Pairwise Analysis of Adjacent Regions

The theoretical framework has focused on how differences in the productive structure between *two* regions contribute or deter the formation of common ethnic traits. Hence, a direct test of the theory naturally dictates pairs of regions as the unit of analysis. In this setting, the empirically relevant question becomes how differences in land quality and elevation within a regional pair affects the degree of ethnic similarity between the two places. The information provided in the language dataset on the location of linguistic groups allows for such detailed investigation. To implement such a test I identify the neighboring regions of each grid. The neighbors of each area are those who are adjacent at a distance of 0.5 degrees, i.e. directly to the: north, south, east, west as well as those that are immediately and 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 (see map 2 in Appendix D where the dots of regional land qualities are centroids of the individual regions). Out of the 58920 regions in the land quality dataset 15982 contain no information on languages and are dropped from the analysis. I also exclude pairs whose individual regions belong to different countries focusing on pairs of adjacent regions that fall entirely within a single country. There are 134657 unique regional pairs within countries.

For the pairwise regressions of adjacent regions the following specification is adopted:[41]

$$pct\_comlang_{ij} = \beta_0 + \beta_1 lqdiff_{ij} + \beta_2 eldiff_{ij} + \beta_2 X_{ij} + \xi_{ij} \tag{12}$$

where $pct\_comlang_{ij}$ is the percentage of common languages, i.e. the number of common languages divided by the total number of unique languages spoken in pair i, j, and captures the degree of ethnic similarity between any two adjacent regions.[42] The variables $lqdiff_{ij}$ and $eldiff_{ij}$ stand for the absolute difference in land quality and elevation respectively between regions i and j and both are an inverse measure of how similar the primitive productive characteristics of any two adjacent regions are. Tables 4*a* and 4*b* present the summary statistics and the raw correlation of the variables used in the analysis. Note that the mean of $pct\_comlang$ has an interesting economic interpretation: adjacent regions within countries, by virtue of proximity, have on average 80% of the total number of languages in common.

According to the theory, regions characterized by large differences in their productive characteristics, would hinder regional population mixing, eventually giving rise to ethnically distinct populations. The first column in Table 5 supports this focal prediction. The difference in land quality and elevation within a regional pair both have a strong negative effect on the formation of common ethnic traits. In particular, a two standard deviation increase in the difference in land quality, $lqdiff_{ij}$, decreases the percentage of common languages by 3.5 points and a similar increase in the difference in elevation, $eldiff_{ij}$, decreases the percentage of common languages by 5.5 points contributing significantly to the formation of ethnically distinct neighbors. In the same specification several geographical characteristics are taken into account. Distance from the equator, abs_lat, systematically produces more linguistically homogeneous neighbors, whereas average elevation, $elev$, and the average land quality of the regional pair, avg, are not significantly affecting local ethnic diversity. Similarly, distance from the shoreline of a regional pair, sea_dist, the area under water within a pair, $waterarea$, and the difference in population density within the pair, $popdiff$, do not systematically affect local ethnic diversity. Finally, a control for the difference in the area of language coverage between the regional neighbors is included. Pairs whose individual regions differ in the spatial extent of their languages' coverage show lower linguistic similarity. Overall, these geographical characteristics capture 21% of the variation in local ethnic diversity.

In column 2 of Table 5, I take advantage of the relatively small size of the regional pairs to control for country and continental fixed effects. Regarding the country fixed effects each pair is assigned the dummy of the country it belongs to. This specification explicitly takes into

[41] Standard errors are clustered at the country level.

[42] Using as an inverse measure of local ethnic similarity, the number of languages spoken within each pair of regions, the results are unchanged.

account any systematic elements related to the state histories of each individual pair of regions which might have independently affected the formation of common ethnic traits. Despite the inclusion of such powerful controls the point estimates of $lqdiff$ and $eldiff$ remain largely unaffected.

In column 3 of Table 5, I allow for the effect of the pairwise difference in regional land quality and elevation to vary across continents. The marginal effects of both $lqdiff_{ij}$ and $eldiff_{ij}$ differ significantly across continents.[43] Within Africa and Asia changes in regional land quality have the greatest impact on local ethnic diversity, whereas changes in regional elevation are qualitative and quantitatively less important. On the other hand, elevation differences are relatively more important in shaping ethnic diversity across regional pairs in Europe and North America. Within the Pacific which includes Australia, New Zealand and parts of Papua New Guinea, both $lqdiff$ and $eldiff$ are quantitatively strong. For regional pairs in South America the extremely poor language coverage may be responsible for the insignificant findings.

Focusing on specific countries to investigate the impact of local geographic variability on ethnic diversity is possible thanks to the high resolution of the data. Column 4 in Table 5, for example, includes regional pairs that belong entirely to China. Within China a two-standard deviation increase in $lqdiff$ decreases local ethnic similarity by 5.4% and a similar magnitude change in local elevation decreases ethnic similarity by 2.5%. Looking in the constant of the regressions reveals that conditional on geographical characteristics adjacent regions within China are much more homogeneous being 89% ethnically similar compared to a 71% of ethnic similarity which is the case for an average regional pair across the globe. This difference in local ethnic similarity may well be an outcome of China's thousand years long and uninterrupted experience with statehood acting as a homogenizing force.

Considering that the data on language location is accurate for the period around the $1990's$ one would expect that the better transportation means and the lesser role of land in the production process would facilitate population mobility and eventually lead to the spatial dispersion of ethnic groups. Despite these reasonable factors weighing against finding any systematic relationship between local ethnic diversity and differences in land endowments, this novel empirical setting uncovers the importance of geographic variability, as captured by the local differences in land quality and elevation, in determining the degree of ethnic similarity within pairs of adjacent regions.

[43] See Table 5 for a complete description of the marginal effects by continent.

5.3 Cross-Real Country Analysis

Having established that the differences in land quality and elevation, between adjacent regions and within virtual countries affect systematically the local ethnic endowment, I now proceed into investigating the relationship between geographic variability and ethnolinguistic fractionalization across existing countries. In particular, using the global data on suitability of land for agriculture and elevation I construct the desired measures of geographic variability for each country. The number of regional observations per country range from a single observation for Monaco to 11937 for Russia. The median number of data points per country is 98.

Existing countries vary widely in the distribution of land qualities. Figures 6*a* and 6*b* in Appendix F, map the regional land qualities for Lesotho and Malawi respectively. A visual inspection of these maps reveals the homogeneity of land quality in Lesotho, $range_{Lesotho} = 0.40$ compared to the apparent heterogeneity inherent to the land quality of Malawi, $range_{Malawi} = 0.61$. Note that these two countries have nonetheless comparable overall levels of land quality, i.e. $avg_{Lesotho} = 0.67$ and $avg_{Malawi} = 0.73$. Mapping the languages spoken in Lesotho and Malawi a striking parallel emerges. The ethnically fragmented society of Malawi, $ELF_{Malawi} = 0.62$, reflects the large underlying spectrum of land qualities compared to the ethnically homogeneous Lesotho, $ELF_{Lesotho} = 0.22$.

As mentioned earlier the index of ethnolinguistic fractionalization, ELF, represents the probability that two individuals randomly drawn from a country's overall population will belong to different ethnolinguistic groups. This implies, that the way people are distributed across places affects measured fractionalization.[44] For example, consider a two-region framework. It is straightforward to manipulate (4) to elucidate how population density across regions affects measured fractionalization. The expected fractionalization, $E(ELF)$, for a pair of places reads:

$$E(ELF) = (1-f)\left(1 - \left(\frac{L^i}{L^j + L^i}\right)^2 - \left(\frac{L^j}{L^j + L^j}\right)^2\right) \tag{13}$$

where $(1-f)$ is the probability that the two regions i and j will have different ethnic traits and $\left(1 - \left(\frac{L^i}{L^j+L^i}\right)^2 - \left(\frac{L^j}{L^j+L^j}\right)^2\right)$ is the probability that two randomly chosen individuals will belong to *different regions.* It is evident from (13) that the more unequally the population is distributed across places, the lower the fractionalization, ceteris paribus. In Appendix A, the regional population densities are expressed as a function of the regional land qualities. It is shown that in the two-region case, conditional on the probability that two places will have different ethnolinguistic elements, $(1-f)$, a more unequal distribution of land quality decreases

[44] This is less of a concern in the preceding empirical sections given that the dependent variable is either the count of languages spoken or the percentage of common languages, rather than a transformation of the count of people speaking these languages.

fractionalization. Consequently, the gini coefficient of land quality for each country, denoted by *lqgini*, is constructed. As expected the gini of land quality is highly correlated (0.62) with how unequally population density is distributed across regions within a country in 1990.[45, 46]

Given the preceding discussion the following main specification is adopted:

$$ELF_i = a_0 + a_1 range_i + a_2 elev\_sd_i + a_3 avg_i + a_4 lqgini_i + a_5 X_i + \eta_i \tag{14}$$

where ELF_i is the level of ethnolinguistic fractionalization in country i, range$_i$ is the support of the distribution of land quality within a country, $elev\_sd_i$ is the variation in elevation, avg_i stands for the average land quality in country i, and $lqgini_i$ is the gini coefficient measuring how unequally land quality is distributed among regions of country i.

In the regression analysis the sample is restricted in the following way. Only countries for which there are at least 4 regions with information on land quality and elevation are included. This limits the sample size to 146 countries.[47] Descriptive statistics and the raw correlation between the variables of interest are presented in Tables 6*a* and 6*b*.

The results of the main specification (14) are presented in column 1 of Table 7. A two standard deviation increase in the dispersion of land quality, *range*, increases ethnolinguistic fractionalization by 22%. To better understand the magnitude of the effect note that the average difference in ethnolinguistic fractionalization between a Sub-Saharan and a non Sub-Saharan country is 0.33. The non-significant effect of variation in elevation on fractionalization in column 1, is driven mainly by the fact that although Sub-Saharan Africa is the most ethnically diverse region, it has an average standard deviation of elevation of 0.28 km, whereas for a non Sub-Saharan country the average is 0.46 km. Indeed, controlling for continental fixed effects, see column 2, a more variable topography increases ethnic diversity significantly. The gini of land quality, *lqgini*, as expected, enters with a negative sign. Average land quality enters also negatively and statistically significant, it turns insignificant, though, once I control for population density in 1500 *AD*.This shows that average land quality by sustaining denser popu-

[45] To measure the latter, I construct a gini index of population density for each country. The population density data come from the Center for International Earth Science Information Network (CIESIN), Columbia University (2005) and were aggregated at the resolution level of the land quality data in order to make the inequality indexes comparable.

[46] Results not shown also suggest that the gini coefficient of land quality is strongly correlated (the correlation is 0.55) with how clustered is land quality within a country, computed by the Moran's I index, a commonly used measure of spatial autocorrelation. That is, in countries with more unequal distribution of land quality, contiguous regions are on average of similar land characteristics. Consequently, the adjacency of productively similar regions would facilitate cross migration, due to low relocation costs, leading to lower fractionalization. Indeed, directly including in the regressions the level of clustering it enters negatively and decreases the coefficient of *lqgini*. However, it is significant only in the regressions using as dependent variable the ethnic fractionalization index derived by Alesina et al. (2003).

[47] Using alternative thresholds for the minimum number of observations per country and constructing the geographical indexes excluding regions with low population density the qualitative results are similar.

lation densities historically may have indirectly influenced contemporary ethnic diversity These purely geographical features account for 15% of the variation in contemporary ethnolinguistic fractionalization across countries.

In the second column of Table 7, dummies for Sub-Saharan Africa, *reg_ssa*, Latin America and Caribbean, *reg_lac*, and Western Europe, *reg_we*, the Americas, *americas*, and East Asia and Pacific, *reg_eap*, are introduced, in order to make sure that the results are not driven by a particular region. The coefficients of interest (except for *elev_sd*) generally decrease remain, though, both economically and statistically significant. Repeating the analysis excluding all the countries of Sub-Saharan Africa or focusing only within the latter produces qualitatively similar results.

In the last column of table 7 geographic and historical controls that could potentially affect fractionalization are accounted for. The pure size of a country, denoted by *areakm2*, enters positively but insignificantly. The mean distance to the nearest coastline or sea-navigable river, denoted by *distcr*, though insignificant, weakly increases fractionalization. This is conforming with the view that places which are increasingly isolated from water passages 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 sea, also captures the vulnerability of places to both the incidence and the intensity of colonization. Thus, the coefficient should be cautiously interpreted. The distance from the equator, denoted by *abs_lat*, has a strong negative effect on ethnolinguistic fractionalization.

The population density in 1500 *AD* and a country's year of independence are added to capture variation in historical contingencies across countries. The log of the population density in 1500 *AD*, *lpd*1500,[48] enters negatively and significantly. This finding is evidence that contemporary ethnic diversity may have been influenced by a country's historical levels of development as represented by the population density in 1500. Also, the year when each country gained independence, *yrentry*, is negatively correlated with fractionalization. Specifically, the later the year of independence, the higher the level of fractionalization. This is consistent with the historical evidence suggesting that since their inception modern states systematically attempted to homogenize their populations along ethnolinguistic dimensions. The expansion of public schooling, for example, had exactly such an impact on linguistic diversity. However, the causality may run in both directions. More fractionalized regions may cause a later emergence of modern states either because of being colonized or because of having a slower statehood

[48] This measure is highly correlated, around 0.56, with the index of state antiquity constructed by Bockstette et al. (2002). Including both makes them insignificant. Consequently, I only include in the regressions the log of the population density in 1500. It may be useful to note that the term "state history" used throughout this study is distinct from the state antiquity index.

formation. Figures $7a$ and $7b$ provide the partial scatter plots of the dispersion in land quality and the variation in elevation against ELF, as generated by the specification 3 in Table 7.

These robustness checks, on the one hand, highlight the fundamental role of the distribution of land quality and elevation in the formation of ethnically diverse societies and on the other hand, hint towards the endogeneity of the contemporary ethnolinguistic endowment to the divergent state histories across countries. In the next section we explore the latter in more detail.

5.4 Colonization and Ethnic Diversity

This section investigates an issue that has received particular attention within economics: the European colonization after the $15th$ century. Ample historical evidence suggests that colonizers impacted the indigenous populations. The way they affected the locals varied widely: from almost entirely eliminating the indigenous populations as in United States, Australia, Argentina and Brazil, to settling at very low levels in other places, such as Congo for example. In several instances, they actively influenced preexisting groups by giving territories to those that were not the initial claimants and politically favoring some groups over others, see Herbst (2002). Generally, the European colonization created an imbalance in the mix of the indigenous populations, directly affecting the preexisting ethnic spectrum.

Consequently, ethnic diversity across countries colonized by Europeans is itself endogenous to their colonial experience, the identity of the colonizers and how intensely the colonizers settled, among other things. Column 1 in Table 8 presents several correlations between ethnic diversity and the identity of the colonizers. Conditional on geographical characteristics, countries colonized by Germans, French, Dutch, British and Portuguese display consistently higher levels of contemporary ethnic fractionalization compared to places where the Italians, Belgians and Spaniards landed.[49]

The role of geography in shaping the endowment of ethnicities across space is predicated on the assumption that the indigenous groups have not been severely disrupted. However, in reality, there is great variation in the percentage of indigenous people across countries. For example, there are several countries whose ethnic mix is a relatively recent phenomenon. The United States, Brazil, Australia and Canada all fall into this category. According to the theory, in such countries geographic variability should no longer be a determinant of ethnic diversity because the indigenous element was severely affected by the advent of the colonizers whose arrival coincided with the economic take-off into industrialization and the beginning of land's

[49] An alternative reading of these correlations is that colonizers differed in the way they chose which places to colonize depending on the level of preexisting ethnic diversity. In absence of time series data on ethnic diversity before and after colonization one cannot disentangle between these two hypotheses.

declining importance in the production process.

In column 2 of Table 8, the sample is restricted into countries whose percentage of indigenous population as of 1500 *AD* still comprises at most 50% of the current population mix.[50] The coefficients of the variables of interest decrease substantially in magnitude and even change sign in the case of *elev_sd*, becoming insignificant. Overall and as expected, within this subset of countries geographic variability cannot account for the observed ethnic diversity emphasizing the power of historical events in dramatically altering the spectrum of ethnic diversity.

The last column of Table 8, investigates the effects of the European colonizers' identity on the percentage of indigenous people living in the colonized countries today. Countries colonized by the Spaniards lost 51.2% of their indigenous population, 26.1% was lost across countries colonized by the British and 16.6% was lost across French colonies.

Combining the findings across columns in table 8 suggests that European colonizers substantially affected the ethnolinguistic spectrum of the places they colonized. The introduction of their own ethnicities and the replacement of the indigenous populations, in particular, introduced a man-made component of contemporary ethnic fractionalization tipping the balance in favor of an ethnic spectrum whose identity and size is not a natural consequence of the primitive land characteristics.

These results suggest that contemporary fractionalization may be decomposed into two parts a natural and a man-made one. The natural component is driven by the geographic variability across regions, whereas the man-made one reflects the history dependent nature of contemporary ethnic diversity as exemplified by the experience of European colonization.

6 Concluding Remarks

This research examines the economic origins of ethnic diversity. It argues that the differences in geographical characteristics shaped the intensity of population mixing. Places exhibiting homogeneous land endowments were characterized by high transferability of region specific human capital. This facilitated population mobility leading to the formation of a common ethnolinguistic identity. On the contrary, among regions characterized by distinct land attributes, population mixing would be limited leading to the formation of local ethnicities and languages giving rise to a wider cultural spectrum.

Constructing detailed data on the distribution of land quality and elevation across regions and countries, I find that geographic variability systematically brings forward and sustains higher ethnic diversity. Both cross-virtual country and cross-country regressions are examined.

[50] A special thanks to Louis Putterman for providing his data set.

The former is of particular significance since the proposed relationship obtains at an arbitrary level of aggregation, explicitly avoiding the endogeneity of current countries' borders and after controlling for continental and country fixed effects. These results are further corroborated by looking into how differences in land quality and elevation shape the degree of ethnic similarity within pairs of adjacent regions. Regional neighbors, sharing common land features, are ethnically more similar than pairs of adjacent regions with different land endowments. Overall, the importance of the distribution of land quality and elevation in determining the natural component of ethnic diversity is a recurrent finding which obtains across different levels of aggregation and remains robust to alternative specifications.

The evidence is also suggestive of the role of state history in shaping contemporary ethnic diversity. In particular, it shows that across countries with a low representation of indigenous people, contemporary ethnic diversity is no longer related to the underlying geography. This is an outcome of the widespread European interference with the indigenous populations along the process of colonization which eventually tipped the balance in favor of a contemporary ethnic spectrum whose identity and size is not a natural consequence of the primitive land characteristics.

The findings provide a stepping stone for further research. Equipped with a more substantive understanding of the origins of ethnic diversity, long standing questions among development and growth economists in which ethnic diversity plays a significant role, may be readdressed. Specifically, the distinction between the natural versus the man-made components of contemporary ethnic diversity calls for a careful reinterpretation of the documented negative relationship between ethnic diversity and economic outcomes.

Additionally, the proposed way of thinking about ethnicities as bearers of specific human capital may be used to understand how and why inequality emerges across ethnic groups. Along the process of development the advent of new technologies, being differentially complementary to the specific human capital of each ethnicity, would lead to differential rates of technology adoption and thus inequality across groups. This notion of specific human capital, driven by the underlying distribution of land endowments, could also be applied at a societal level generating new insights about the diffusion of development both within and across countries.

Furthermore, establishing that diversity in land endowments drives ethnic diversity has profound implications for understanding why preferences about public goods provision might differ across groups. This geographically driven component of preference heterogeneity may be used to explain the differential timing of the emergence of politically centralized societies along the process of development and provide a new way of thinking about the optimal size of nations.

7 Appendix

Appendix A - Proofs

Past Migrations

As it is evident from (7) and (8) the size of the migration movement in period t depends on the level of regional population densities in period $t-1$. The latter is a function of past migration movements. In particular, in the beginning of any period t, and before any labor movement occurs (if any), the ratio of the regional population densities equals $\frac{L_{t-1}^i}{L_{t-1}^j}$. Depending on the direction of the last migration either (5) or (6) should hold with equality when evaluated at the regional population densities after the last occurrence of migration, in period, s. Solving for the ratio of regional population in period s, $\frac{L_s^i}{L_s^j}$, the following two cases obtain:

1. The last migration occurred in period s, $0 \leq s \leq t-1$ from region i to region j

$$\frac{L_{t-1}^i}{L_{t-1}^j} = \frac{L_s^i}{L_s^j} = \left(\lambda_s \left(h_s^{\,i}\right)^{\epsilon}\right)^{\frac{1}{1-\alpha}} \frac{m_i}{m_j} \quad if \quad M_s^{i\to j} > 0 \tag{A1}$$

2. The last migration occurred in period s, $0 \leq s \leq t-1$ from region j to region i

$$\frac{L_{t-1}^i}{L_{t-1}^j} = \frac{L_s^i}{L_s^j} = \left(\lambda_s \left(h_s^{\,j}\right)^{-\epsilon}\right)^{\frac{1}{1-\alpha}} \frac{m_i}{m_j} \quad if \quad M_s^{j\to i} > 0 \tag{A2}$$

Lemma 2 *In any period t there are the following cases regarding the occurrence of migration or not.*

1. If the last migration occurred in period s, $0 \leq s < t-1$, from region i to region j then

$$M_t^{i\to j} > 0 \qquad iff \qquad \lambda_t < \lambda_s \left(\frac{h_s^i}{h_t^i}\right)^{\varepsilon}$$

$$M_t^{j\to i} > 0 \qquad iff \qquad \lambda_t > \lambda_s \left(h_t^j h_s^i\right)^{\varepsilon}$$

$$M_t^{i\to j} = M_t^{j\to i} = 0 \quad iff \quad \lambda_s \left(\frac{h_s^i}{h_t^i}\right)^{\varepsilon} \leq \lambda_t \leq \lambda_s \left(h_t^j h_s^i\right)^{\varepsilon}$$

2. If the last migration occurred in period s, $0 \leq s < t-1$, from region j to region i then

$$M_t^{i\to j} > 0 \qquad iff \qquad \lambda_t < \lambda_s \left(h_s^j h_t^i\right)^{-\varepsilon}$$

$$M_t^{j\to i} > 0 \qquad iff \qquad \lambda_t > \lambda_s \left(\frac{h_t^j}{h_s^j}\right)^{\varepsilon}$$

$$M_t^{i\to j} = M_t^{j\to i} = 0 \quad iff \quad \lambda_s \left(h_s^j h_t^i\right)^{-\varepsilon} \leq \lambda_t \leq \lambda_s \left(\frac{h_t^j}{h_s^j}\right)^{\varepsilon}$$

Proof. Substituting the relevant ratio of the past population densities, either (A1) or (A2) depending on the direction of the last migration, in both (7) and (8) and solving for the required inequalities completes the proof. □

Proof of Proposition 1.

Under Assumption $(B1)$ the ratio λ_t/λ_s may take three unique values either $\lambda_{\min}/\lambda_{\max}$, $\lambda_{\max}/\lambda_{\min}$ or 1. Obviously, $\lambda_{\min}/\lambda_{\max} < 1 < \lambda_{\max}/\lambda_{\min}$. In this case there will be no successive migrations towards the same region. For example, for migration to occur in period t from j to i it is necessary (though not sufficient, see Lemma 2) that $\lambda_t > \lambda_s$. This implies that $\lambda_t = \lambda_{\max}$ and $\lambda_s = \lambda_{\min}$. Consequently, it follows that since in period s migration also occurred, the direction of this last migration could have only taken place from region i towards region j, i.e. $\lambda_s = \lambda_{\min}$ and $\lambda_{s-b} = \lambda_{\max}$. Similar reasoning rules out successive migration towards region i. This simplifies the analysis considerably since one may focus only on the cases of Lemma 2 where a current migration, should it take place, is always in the opposite direction of the last one. If $\lambda_t/\lambda_s = \lambda_{\min}/\lambda_{\max} < \left(h_s^j h_t^i\right)^{-\varepsilon}$ migration occurs towards region j. So, conditional on $\lambda_{\min}/\lambda_{\max}$, any regional pair characterized by higher ε and higher region specific technology, h_t^i, will experience fewer migrations towards region j. Similarly, migration occurs towards region i in period t iff $\lambda_t/\lambda_s = \lambda_{\max}/\lambda_{\min} > \left(h_s^j h_t^i\right)^{\varepsilon}$. It is evident that the right hand-side increases as erosion increases. Once it becomes sufficiently large no more population movements towards region i take place.

Conditional on $(B1)$ the probability that productivity shocks differ intertemporally, that is $\lambda_t/\lambda_s = \lambda_{\max}/\lambda_{\min}$ or $\lambda_t/\lambda_s = \lambda_{\min}/\lambda_{\max}$ equals $2p(1-p)$. This is maximized at $p = 1/2$. It is also obvious from Lemma 2 that the larger $\lambda_{\max}/\lambda_{\min}$ is (equivalently the smaller $\lambda_{\min}/\lambda_{\max}$ is), the more probable population mixing is. Consequently, increases in the variance of relative productivity shocks $var(\lambda_t) = p(1-p)(\lambda_{\max} - \lambda_{\min})^2$ increases the probability that the two regions will share common cultural traits.

These observations taken together provide a sketch of the proof.

Interpreting expected fractionalization, (13), in terms of regional land qualities:

Manipulating (13) may be rewritten as:

$$E(ELF) = (1 - f_T)\left(\frac{L_T^i}{2L_T^j} + \frac{L_T^j}{2L_T^i} + 1\right)^{-1}$$

Evaluating (A2) at $h_s^j = 1$, for example, and substituting the ratio of regional population densities, $E(ELF)$ may be rewritten as:

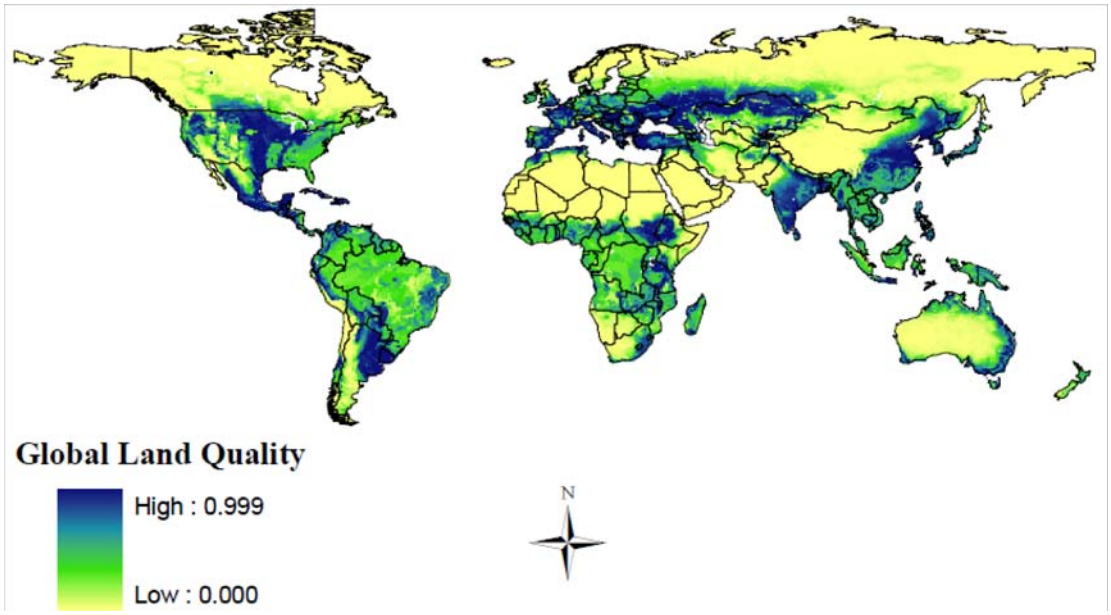
$$E(ELF) = (1 - f_T)\left(\frac{m^i}{2m^j} + \frac{m^j}{2m^i} + 1\right)^{-1} \tag{15}$$

It is easy to show that conditional on the probability that two places will not share the same cultural traits, $(1 - f_T)$, a more unequal distribution of the quality of land will decrease measured fractionalization. For example, let $m^i > m^j$, then an increase in m^i and/or a decrease in m^j will decrease $E(ELF)$. This obtains by differentiating (15) with respect to m^i and m^j accordingly.

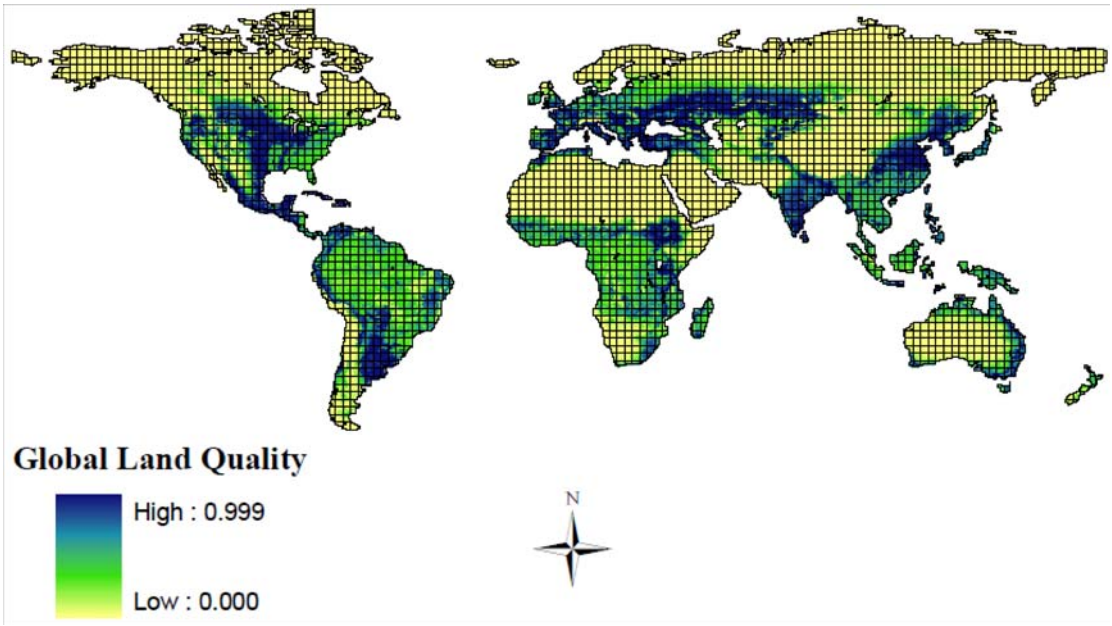
This derivation highlights the fact that conditional on the probability that individuals from two regions will have different ethnicities, an increase in the inequality of population density between these places, which is function of how unequally land quality itself is distributed affects negatively fractionalization outcomes.

Appendix B - Global Maps

Map 1*a*: Land Quality Across Countries



Map 1*b*: Land Quality Across Virtual Countries



Appendix C1 - Ethnic Groups and Land Quality in Kenya

The theoretical premise of this study is that ethnic groups are bearers of specific human capital and this specificity derives from the land quality in which an ethnic group resides. This section presents anecdotal evidence in support of the hypothesis.

The graph below plots the distribution of land quality within ethnic groups in Kenya, with similar spatial extent (a group of those examined here spans on average 35 regions of 0.5 degrees latitude by 0.5 degrees longitude). Land suitability for agriculture (described in the empirical section) is in the horizontal axis, whereas the vertical axis displays the name of each group. The boxes map the interquartile range of land quality with the dots representing regions with land quality more than three standard deviations further from the mean.

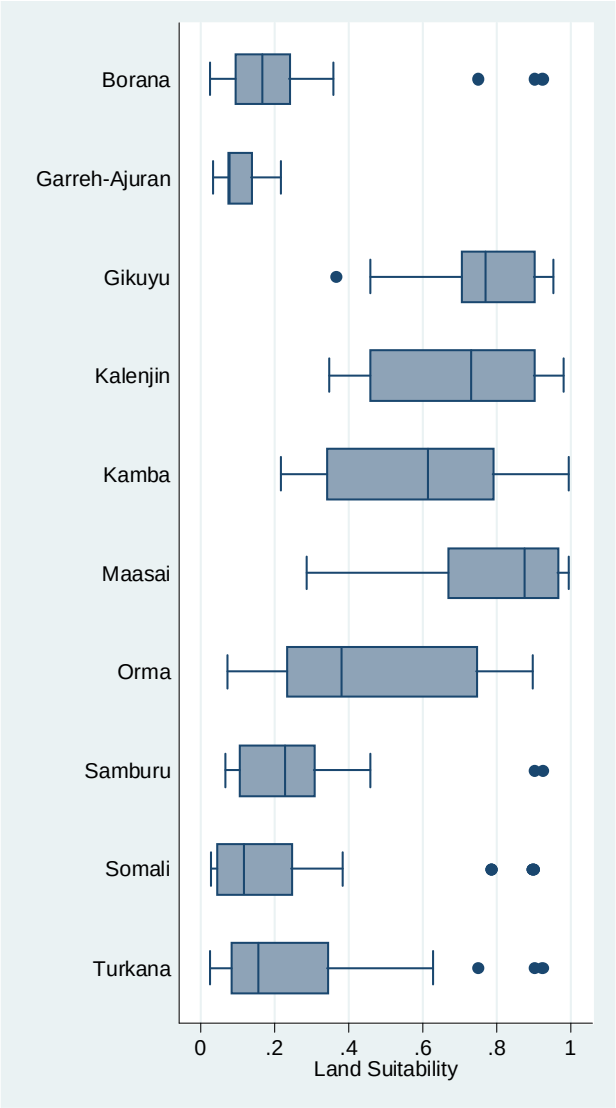


Figure 4: Land Quality within ethnic groups in Kenya

A cursory inspection of the box plots reveals that ethnic groups are not randomly dispersed across regional land qualities within Kenya. In fact, they seem to cluster in territories of distinct and homogenous land endowments. The Borana, the Garreh-Ajuran, the Samburu, the Somali, the Turkana and the Orma people are all located at relatively low levels of land quality where agriculture is almost impossible to maintain.[51] The Samburu, the Borana, the Turkana are semi-nomadic pastoralists who herd mainly cattle but also keep sheep, goats and camels, see Pavitt (2001). The Garreh-Ajuran and the Somali are semi-nomadic shepherds. These groups have the human capital to undertake the productive activities which are optimal for the places in which they are located. On the other hand, the Gikuyu and the Kalenjin are concentrated in territories of high land quality and they are mainly engaged in agriculture, producing: sorghum, millet, beans, sweet potatoes, maize, potatoes, cassava, bananas, sugarcane, yams, fruit, tobacco and coffee. The Kamba people are often found in different professions; some are agriculturalists others hunters, and a large number are pastoralists. This, according to the theory, is an outcome of the fact that Kamba reside in intermediate levels of land quality which may sustain different optimal activities. The Orma people are mainly pastoralists who herd cattle, sheep and goats however, people within the Orma group who speak the dialect of Munyo are agriculturalists. This would explain the spread out distribution of the Orma people.

An interesting example is the case of the Maasai people. As it is evident from the map they are located at regions endowed with climatic and soil characteristics very favorable to farming. Nevertheless, the Maasai are semi-nomadic pastoralists with the herding of cattle being the dominant activity. At first, this observation may seem at odds with the theory which posits that groups should develop human capital optimal and specific to their region. The history of Maasai, however, sheds important light on this issue, see Olson (1990). Upon the arrival of the British colonizers two treaties, one in 1904 and another in 1911, reduced the Maasai lands in Kenya by 60%. The eviction took place in order for the British to make room for settler ranches, subsequently confining Maasai to their present-day territories. It was exactly in these ancestral grazing areas where the Maasai's human capital, i.e. herding cattle was optimal. The very fact that today this group essentially practises and uses its ancestral human capital in territories that are mostly conducive to agriculture is itself a manifestation that ethnic human capital may be a very persistent factor in the economic choices of ethnic groups.

[51] The description of the main productive activities of each ethnic group, unless otherwise noted, comes from the entries found in the Ethnologue website, (http://www.ethnologue.com/).

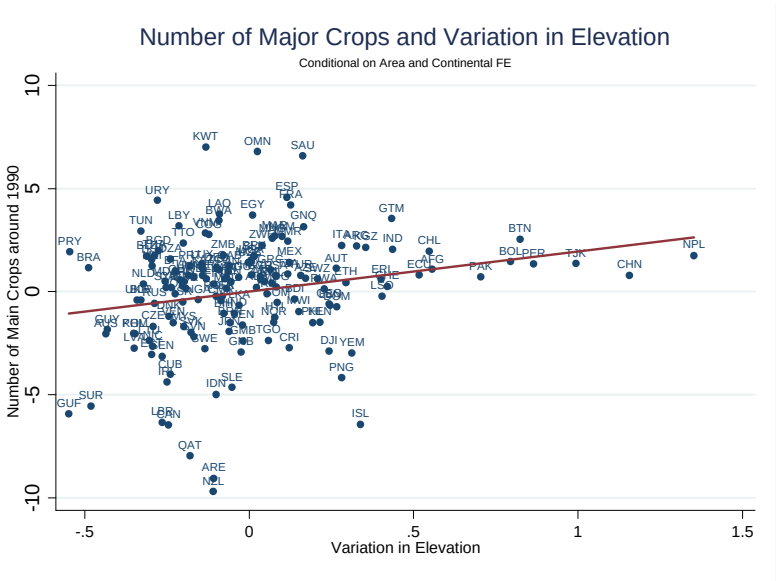
Appendix C2 - Crops and Geographic Variability

Table 1: Geographic Diversity and Cultivated Crops

Dep. Var. **nmbr_crops**	*range*	*elev_sd*	*avg*	*areakm2*	*reg_ssa*	*reg_we*
	4.693	1.457	0.524	0.016	2.077	−2.995
	(1.152)***	(0.463)***	(1.014)	(0.009)*	(0.502)***	(0.751)***

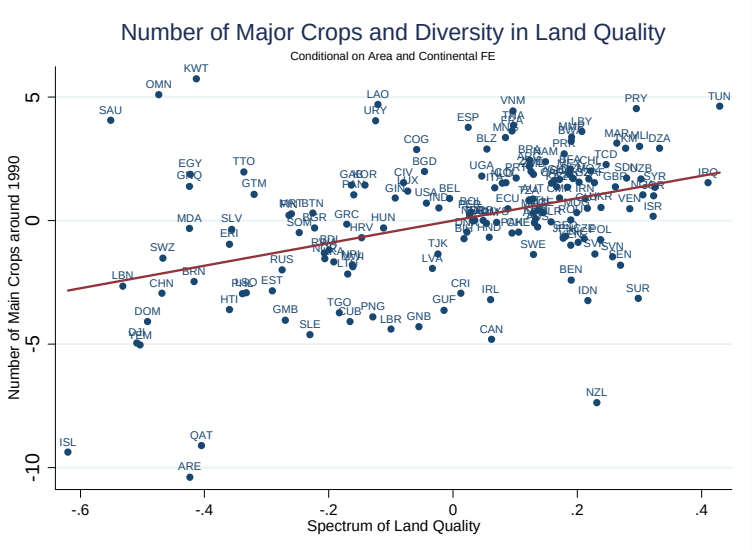
Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1; See Appendix H for variable's definitions

Figure 5*a*



See Appendix H for variables' definitions

Figure 5*b*



See Appendix H for variables' definitions

Appendix D - Virtual Country Analysis

Map 2: Example of a Virtual Country

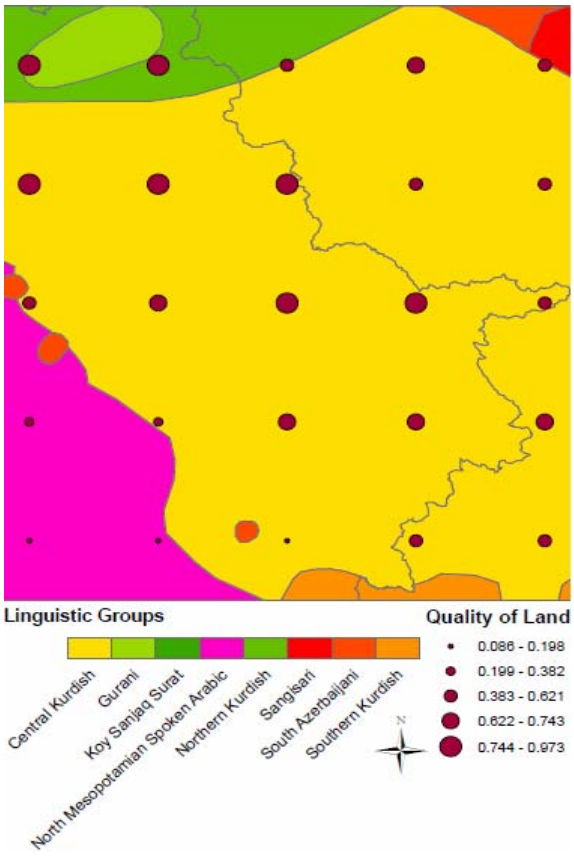


Table 2a: Summary Statistics for the Virtual Country Analysis

statistics	**lnnmbr_lang**	**range**	**avg**	**elev_sd**	**areakm2**	**in_cntry**	**nmbr_cntry**	**sea_dist**	**waterarea**
mean	1.11	0.39	0.44	0.23	43.97	0.59	1.58	0.57	0.68
sd	0.86	0.27	0.29	0.25	22.28	0.49	0.83	0.60	1.06
max	3.43	1.00	0.99	2.20	76.92	1.00	6.00	2.53	15.44
min	0.00	0.00	0.00	0.00	0.12	0.00	1.00	0.00	0.00

See Appendix H for variables' definitions

Table 2b: The Correlation Matrix for the Virtual Country Analysis

	lnnmbr_lang	**range**	**avg**	**elev_sd**	**areakm2**	**in_cntry**	**nmbr_cntry**	**sea_dist**	**waterarea**
lnnmbr_lang	1.00								
range	0.27	1.00							
avg	0.07	0.39	1.00						
elev_sd	0.27	0.32	0.01	1.00					
areakm2	0.31	0.21	-0.03	0.02	1.00				
in_cntry	-0.35	-0.18	0.03	-0.11	-0.18	1.00			
nmbr_cntry	0.36	0.22	0.01	0.14	0.20	-0.84	1.00		
sea_dist	-0.02	0.14	-0.08	0.01	0.25	-0.07	0.04	1.00	
waterarea	-0.04	-0.04	-0.17	-0.10	0.24	-0.08	0.08	0.09	1.00

See Appendix H for variables' definitions

Appendix D - Virtual Country Analysis

Table 3: Main Specification for the Virtual Country Analysis

	(1)	(2)	(3)	(4)	(5)
VARIABLES	lnnmbr_lang	lnnmbr_lang	lnnmbr_lang	lnnmbr_lang	lnnmbr_lang
range	0.440***	0.381***	0.377***	0.464***	0.302***
	(0.149)	(0.126)	(0.111)	(0.179)	(0.101)
elev_sd	0.402***	0.493***	0.683***	0.441	0.912***
	(0.154)	(0.118)	(0.103)	(0.285)	(0.184)
avg	-0.053	-0.063	-0.231***	0.709*	-0.083
	(0.166)	(0.143)	(0.088)	(0.361)	(0.146)
areakm2	0.005***	0.004**	0.001	0.008***	0.006***
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
abs_lat	-0.023***	-0.028***	-0.015***	-0.030***	-0.017***
	(0.003)	(0.004)	(0.005)	(0.008)	(0.067)
sea_dist	0.049	0.052	-0.001	0.340**	-0.006
	(0.076)	(0.067)	(0.053)	(0.156)	(0.062)
waterarea	-0.029*	-0.017	0.016	-0.021	-0.028
	(0.021)	(0.019)	(0.021)	(0.022)	(0.033)
in_cntry	-0.174***	-2.969**	-1.484	-0.588***	
	(0.068)	(1.304)	(1.164)	(0.197)	
nmbr_cntry	0.189***	0.141***	0.103***	0.220***	
	(0.037)	(0.033)	(0.027)	(0.057)	
Observations	1373	1373	869	504	816
R^2	0.45	0.63	0.51	0.54	0.64

Standard errors in parentheses are corrected for spatial autocorrelation following Conley (1999).

***p < 0:01; **p < 0:05; *p < 0:1

Specifications (2), (3) (4) and (5) include country and continental fixed effects. Specification (3) focuses on virtual countries **out of the tropics**. Specification (4) on virtual countries **in the tropics** and specification (5) on virtual countries belonging entirely to an existing real country.

Appendix E - Pairwise Analysis of Adjacent Regions

Table 4a: Summary Statistics for the Pairwise Analysis of Adjacent Regions

statistics	**pct_comlang**	**lqdiff**	**eldiff**	**elev**	**avg**	**sea_dist**	**waterarea**	**popdiff**
mean	0.80	0.07	0.14	0.67	0.31	0.71	0.07	0.03
sd	0.29	0.12	0.23	0.81	0.32	0.59	0.16	0.15
max	1.00	0.99	3.67	5.75	1.00	2.68	4.95	9.75
min	0.00	0.00	0.00	-0.90	0.00	0.00	0.00	0.00

See Appendix H for variables' definitions

Table 4b: The Correlation Matrix for the Pairwise Analysis of Adjacent Regions

	pct_comlang	**lqdiff**	**eldiff**	**elev**	**avg**	**sea_dist**	**waterarea**	**popdiff**
pct_comlang	1.00							
lqdiff	-0.14	1.00						
eldiff	-0.14	0.23	1.00					
elev	-0.06	0.06	0.40	1.00				
avg	-0.13	0.34	0.03	-0.12	1.00			
sea_dist	0.08	0.00	0.01	0.25	0.00	1.00		
waterarea	0.03	-0.04	-0.07	-0.03	-0.07	0.01	1.00	
popdiff	-0.03	0.08	0.03	-0.06	0.18	-0.10	0.00	1.00

See Appendix H for variables' definitions

Appendix E - Pairwise Analysis of Adjacent Regions

Table 5: Main Specification for the Pairwise Analysis of Adjacent Regions

VARIABLES	(1) pct_comlang	(2) pct_comlang	(3) pct_comlang	(4) pct_comlang
lqdiff	-0.142***	-0.148***	-0.319***	-0.270***
	(0.051)	(0.038)	(0.086)	(0.059)
eldiff	-0.115***	-0.090***	-0.059	-0.035**
	(0.019)	(0.014)	(0.049)	(0.015)
abs_lat	0.005***	0.003*	0.003**	-0.004
	(0.001)	(0.001)	(0.001)	(0.004)
elev	0.007	0.003	0.003	-0.023*
	(0.008)	(0.009)	(0.009)	(0.013)
avg	-0.021	-0.028	-0.028	-0.046
	(0.038)	(0.032)	(0.032)	(0.078)
sea_dist	0.016	0.015	0.014	-0.078*
	(0.013)	(0.010)	(0.010)	(0.039)
waterarea	0.032	0.017	0.015	0.019
	(0.023)	(0.017)	(0.017)	(0.022)
popdiff	0.004	0.003	0.002	-0.018
	(0.023)	(0.012)	(0.018)	(0.041)
area_diff	-0.018*	-0.007	-0.007	0.009
	(0.009)	(0.007)	(0.007)	(0.025)
Constant	0.617***	0.677***	0.712***	0.891***
	(0.034)	(0.038)	(0.036)	(0.154)
Observations	134657	134657	134657	12757
R^2	0.21	0.27	0.28	0.20

Standard errors are clustered at the country level, ***p < 0.01; **p < 0.05; *p < 0.1

Specifications (2) and (3) include country and continental fixed effects. (3) allows for the marginal effect of pair differences in elevation, eldiff, and lqdiff to vary across continents. The omitted continent is Africa. The marginal effects for each continent are the following:

$\frac{\partial pct\_comlang}{\partial eldiff_{ij}}\Big|_{Europe} = -.151^{**}$; $\frac{\partial pct\_comlang}{\partial eldiff_{ij}}\Big|_{Asia} = -.102^{***}$;

$\frac{\partial pct\_comlang}{\partial eldiff_{ij}}\Big|_{Pacific} = -.181^{***}$; $\frac{\partial pct\_comlang}{\partial eldiff_{ij}}\Big|_{N\_America} = -.183^{***}$

$\frac{\partial pct\_comlang}{\partial eldiff_{ij}}\Big|_{S\_America} = -.0004$; $\frac{\partial pct\_comlang}{\partial lqdiff_{ij}}\Big|_{Europe} = -.098^{***}$;

$\frac{\partial pct\_comlang}{\partial lqdiff_{ij}}\Big|_{Asia} = -.155^{*}$; $\frac{\partial pct\_comlang}{\partial lqdiff_{ij}}\Big|_{Pacific} = -.631^{***}$

$\frac{\partial pct\_comlang}{\partial lqdiff_{ij}}\Big|_{N\_America} = -.043$; $\frac{\partial pct\_comlang}{\partial lqdiff_{ij}}\Big|_{S\_America} = 0.142$;

Specification (4) focuses on pairs of regions within China. In this case the standard errors errors are clustered at the level of each administrative unit of China.

See Appendix H for variables' definitions

Appendix F - Country Maps

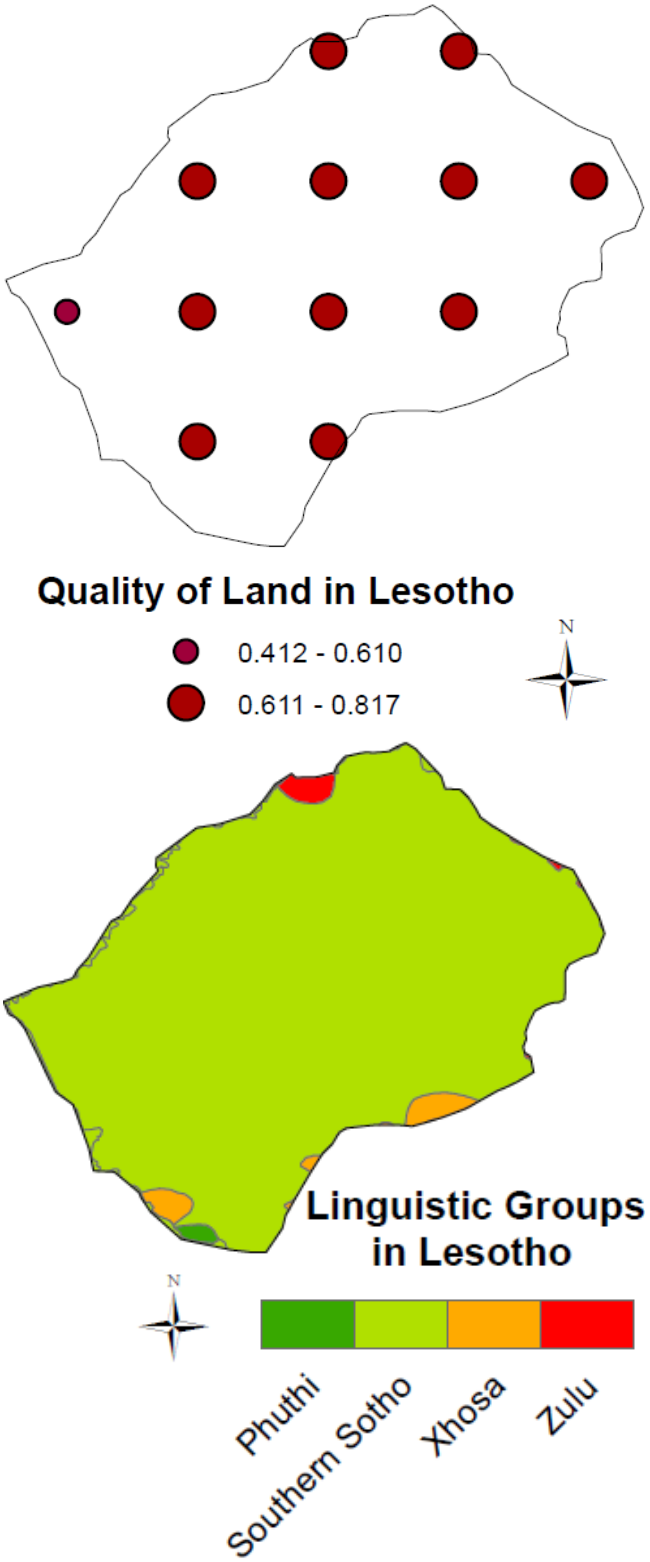


Figure 6*a*: Land Quality and Languages in Lesotho

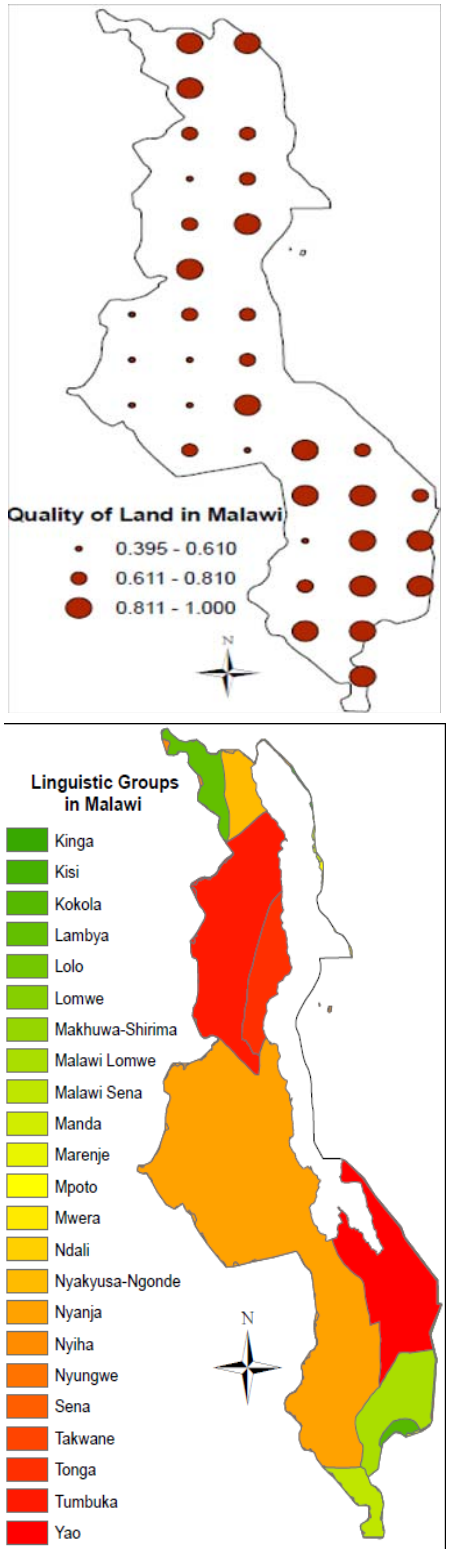


Figure 6*b*: Land Quality and Languages in Malawi

Appendix G - Real Country Analysis

Table 6a: Summary Statistics for the Real Country Analysis

statistics	ELF	range	avg	lqgini	elev_sd	lpd1500	yrentry
mean	0.41	0.72	0.45	0.33	0.41	0.91	1927.85
sd	0.28	0.27	0.25	0.23	0.35	1.49	56.58
max	0.93	1.00	0.95	0.88	1.87	3.84	1993.00
min	0.00	0.01	0.00	0.03	0.01	-3.82	1816.00

See Appendix H for variables' definitions

Table 6b: The Correlation Matrix for the Real Country Analysis

	ELF	range	avg	lqgini	elev_sd	lpd1500	yrentry
ELF	1						
range	0.18	1					
avg	-0.18	0.07	1				
lqgini	0.07	0.30	-0.79	1			
elev_sd	0.08	0.39	-0.02	0.25	1		
lpd1500	-0.18	0.09	0.38	-0.33	0	1	
yrentry	0.35	-0.23	-0.15	-0.03	-0.24	-0.08	1

See Appendix H for variables' definitions

Appendix G - Real Country Analysis

Table 7: Specifications for the Cross-Country Analysis

VARIABLES	(1) ELF	(2) ELF	(3) ELF
range	0.392***	0.331***	0.325***
	(0.103)	(0.096)	(0.099)
elev_sd	0.044	0.124*	0.142**
	(0.070)	(0.063)	(0.063)
avg	-0.714***	-0.378**	-0.224
	(0.165)	(0.147)	(0.152)
lqgini	-0.685***	-0.396**	-0.452***
	(0.193)	(0.168)	(0.169)
reg_ssa		0.294***	0.143*
		(0.049)	(0.077)
reg_we		-0.152***	0.083
		(0.058)	(0.077)
americas		-0.083	-0.179*
		(0.059)	(0.092)
reg_eap		-0.025	-0.092
		(0.076)	(0.069)
abs_lat			-0.004**
			(0.002)
areakm2			0.001
			(0.001)
distcr			0.085
			(0.056)
lpd1500			-0.038*
			(0.023)
yrentry			0.001**
			(0.000)
Observations	146	146	146
R^2	0.15	0.43	0.50

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

See Appendix H for variables' definitions

Appendix G - Real Country Analysis

Figure 7*a*

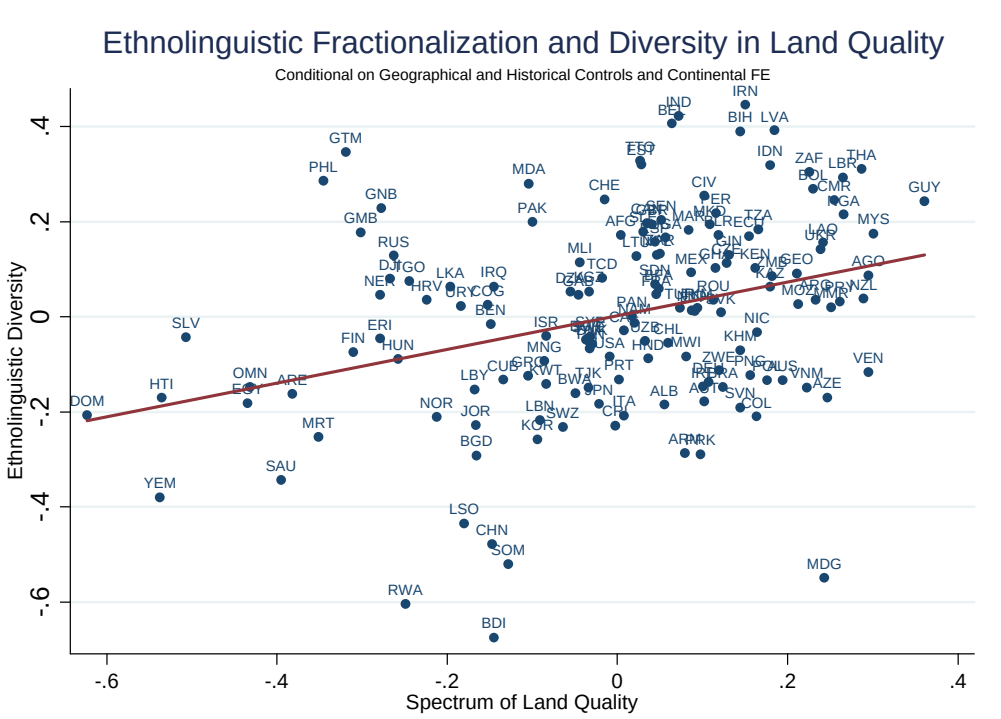
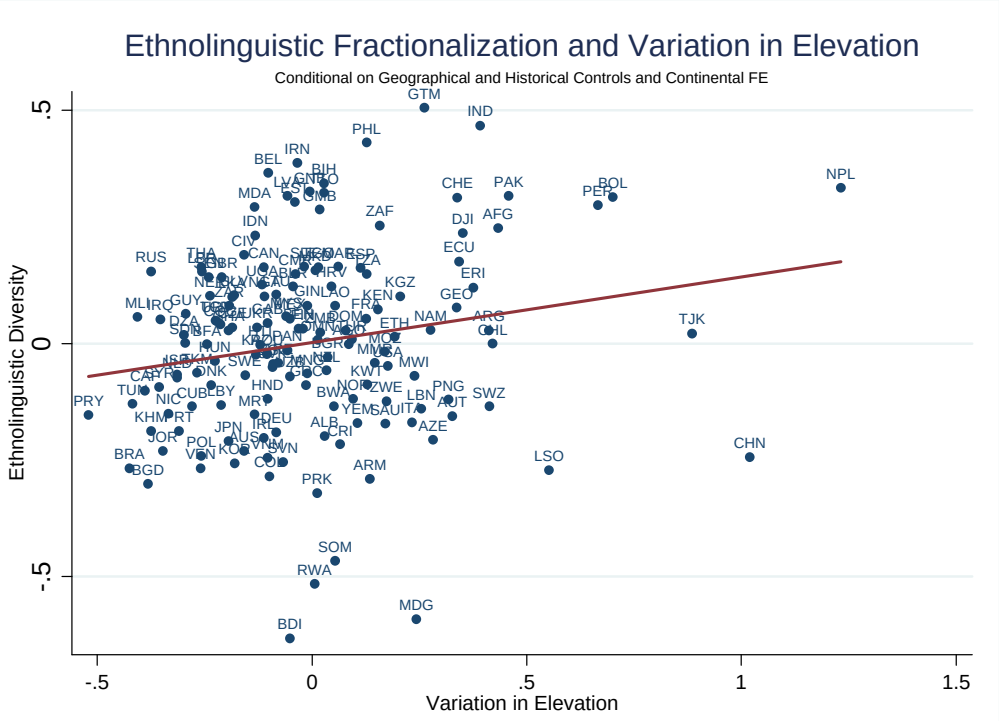


Figure 7*b*



Appendix G - Real Country Analysis

Table 8: Colonization and Man-Made Ethnic Diversity

VARIABLES	(1) ELF	(2) ELF	(3) indigenous
range	0.214**	0.110	
	(0.104)	(0.180)	
elev_sd	0.133**	-0.207	
	(0.060)	(0.131)	
avg	-0.469***	-0.269	
	(0.164)	(0.262)	
lqgini	-0.742***	0.056	
	(0.172)	(0.414)	
areakm2	-0.002*		
	(0.001)		
distcr	0.293***		
	(0.043)		
spanish_col	-0.000		-0.512***
	(0.060)		(0.067)
german_col	0.428***		-0.240
	(0.040)		(0.176)
french_col	0.221***		-0.166**
	(0.057)		(0.075)
dutch_col	0.379***		-0.208
	(0.089)		(0.196)
belgian_col	-0.261		0.063***
	(0.204)		(0.020)
portu_col	0.250**		-0.181
	(0.117)		(0.170)
british_col	0.202***		-0.261***
	(0.053)		(0.077)
italian_col	0.145		0.040
	(0.125)		(0.040)
Observations	146	28	148
R^2	0.44	0.22	0.30

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

See Appendix H for variables' definitions

Appendix H - Data Sources

Geographical Variables

abs_lat: Absolute latitudinal distance from the equator.

Source: Available from Development Research Institute, NYU. For the cross-virtual country analysis and the regional pairs analysis the distance from the equator is calculated from the centroid of the respective unit of analysis.

areakm2: land area in 1000's of sq. km.

Source: Center for International Development, CID.[52] For the cross-virtual country the area of the virtual countries are constructed using ArcGIS. In the calculation are considered only areas over which both language and land quality data are available.

area_diff: difference in 1000's of sq km. in the linguistic coverage between the regions in the pair.

Source: See **areakm2**

avg: average land quality within the respective unit of analysis

Source: Constructed by the author. The dataset is available at the Atlas of the Biosphere accessible at http://www.sage.wisc.edu/atlas/data.php?incdataset=Suitability%20for%20Agriculture. I wish to thank Navin Ramankutty, the author of this dataset for sharing an update of this land suitability index which is used in this study.

distcr: distance from centroid of a country to nearest coast or sea-navigable river (1000's of km).

Source: Center for International Development, CID.

eldiff: difference in elevation within pairs of adjacent regions in km.

Source: Constructed by the author using information on elevation above sea level at a grid level. The data is aggregated at the same level as the land quality data i.e. at 0.5 degrees latitude by 0.5 degrees longitude. Source: The Atlas of Biosphere: http://www.sage.wisc.edu:16080/atlas/.

elev: average elevation within pairs of adjacent regions in km.

Source: see **el_diff**

elev_sd: standard deviation of elevation within actual and virtual countries in km.

Source: see **el_diff**

in_cntry: dummy equals 1 if a virtual country falls completely within a real country; constructed using ArcGIS.

lnnmbr_lang: log number of languages spoken within a virtual country.

[52] All geographical data from CID are available at: http://www.ksg.harvard.edu/CID

Source: 15th edition of the Ethnologue database of languages obtained from Global Mapping International's World Language Mapping System.

lqdiff: absolute difference in land quality between adjacent regions.

Source: See **avg**

lqgini: the gini coefficient of land quality within country.

Source: See **avg**

nmbr_crops: Number of crops that are cultivated in a country during the year.

Source: This global data set, constructed by Leff et al. (2004), is intended to provide very rough indications of the probability of finding 18 major crops across the world in the early 1990s.

nmbr_cntry: number of real countries in which a virtual country belongs to; constructed using ArcGis.

pct_comlang: percentage of common languages spoken within a pair of adjacent regions.

Source: see **lnnmbr_lang**

popdiff: difference in the population density between adjacent regions in thousand's of people per sq km.

Source: Center for International Earth Science Information Network (CIESIN).

range: spectrum of land qualities within the respective unit of analysis; i.e. the difference in land quality between the region with the highest land quality from that with the lowest.

Source: See **avg**

reg_lac: dummy variable equals 1 for countries in Latin America and Caribbean.

reg_ssa: dummy variable equals 1 for countries in Sub-Saharan Africa.

reg_we: dummy variable equals 1 for countries in Western Europe.

sea_dist: distance from the nearest coastline in 1000s of km's of the centroid of the unit of analysis, i.e. regional pair or virtual country.

Source: Constructed using the Coastlines of seas, oceans, and extremely large lakes dataset after excluding the lakes. Publisher and place: Global Mapping International, Colorado Springs, Colorado, USA. Series name: Global Ministry Mapping System. Series issue: Version 3.0

waterarea: total area within the respective unit of analysis covered by water.

Source: Constructed using the "Inland water area features" dataset from Global Mapping International, Colorado Springs, Colorado, USA. Series name: Global Ministry Mapping System.

Historical Variables

ELF: level of ethnolinguistic fractionalization within a country.

Source: Fearon and Laitin (2003) available at http://www.stanford.edu/~jfearon/

lpd1500: log population density in 1500.

Source: McEvedy and Jones (1978), "Atlas of World Population History".

yrentry: year a country achieved independence.

Source: Fearon J., "Ethnic and Cultural Diversity by Country", originally from the Correlated of War database (COW).

indigenous: percentage of the current population's composition which was indigenous in these countries as of 1500 *AD*.

Source: Putterman, L., 2007, World Migration Matrix, 1500 – 2000, Brown University.

belgian_col: dummy equals 1 if a country was a Belgian colony after 1500 AD.

Source: "Determinants and Economic Consequences of Colonization: A Global Analysis"
Ertan, A., Putterman, L.,
Supplemented by entries from Encyclopedia Britannica where necessary.

british_col: dummy equals 1 if a country was a British colony after 1500 AD.

Source: see **belgian_col**

dutch_col: dummy equals 1 if a country was a Dutch colony after 1500 AD.

Source: see **belgian_col**

french_col: dummy equals 1 if a country was a French colony after 1500 AD.

Source: see **belgian_col**

italian_col: dummy equals 1 if a country was an Italian colony after 1500 AD.

Source: see **belgian_col**

portu_col: dummy equals 1 if a country was a Portuguese colony after 1500 AD.

Source: see **belgian_col**

spanish_col: dummy equals 1 if a country was a Spanish colony after 1500 AD.

Source: see **belgian_col**

References

Ahlerup, Pelle and Ola Olsson, "The Roots of Ethnic Diversity," 2007. Working Papers, University of Gothenburg.

Alesina, Alberto and Enrico Spolaore, "On the Number and Size of Nations," *Quarterly Journal of Economics*, 1997, *112*, 1027–1056.

____, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg, "Fractionalization," *Journal of Economic Growth*, 2003, *8*, 155–194.

Ashraf, Quamrul and Stelios Michalopoulos, "The Climatic Origins of the Neolithic Revolution: A Theory of Long-Run Development via Climate-Induced Technological Progress," 2007. Working Paper, mimeo Brown University.

Atlas Narodov Mira (Atlas of the People of the World), Moscow: Glavnoe Upravlenie Geodezii i Kartograi, Bruck, S.I., and V.S. Apenchenko, 1964.

Banerjee, Abhijit and Rohini Somanathan, "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.

Bellwood, Peter, "Early Agriculturalist Population Diasporas? Farming, Languages, and Genes," *Annual Review of Anthropology*, 2001, *30*, 181–207.

Bockstette, Valerie, Areendam Chanda, and Louis Putterman, "States and Markets: The Advantage of an Early Start," *Journal of Economic Growth*, 2002, *7* (4), 347–369.

Boserup, Ester, *The Conditions of Agricultural Growth*, Chicago, IL: Aldine Publishing Company, 1965.

Botticini, Maristella and Zvi Eckstein, "From Farmers to Merchants, Voluntary Conversions and Diaspora: A Human Capital Interpretation of Jewish History," *Journal of Economic History*, 2005, *65*, 922–948.

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

Caselli, Francesco and Wilbur J. Coleman, "On the Theory of Ethnic Conflict," 2006. Working Paper, mimeo London School of Economics.

Conley, Timothy G., "GMM Estimation with Cross Sectional Dependence," *Journal of Econometrics*, 1999, *92*, 1–45.

Curtin, Philip D., *Cross-Cultural Trade in World History*, Cambridge: Cambridge University Press, 1984.

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

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

Esteban, Joan and Debraj Ray, "On the Salience of Ethnic Conflict," 2007. Working Paper, mimeo New York University.

Fearon, James, "Ethnic Structure and Cultural Diversity by Country," *Journal of Economic Growth*, 2003, *8*, 195–222.

___ and David Laitin, "Ethnicity, Insurgency and Civil War," *American Political Science Review*, 2003, *97*, 75–90.

Galor, Oded and David N. Weil, "Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond," *American Economic Review*, 2000, *90* (4), 806–828.

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

Gray, Russell D. and Quentin D. Atkinson, "Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin," *Nature*, 2003, *426*, 435–439.

Grigg, David B., *An Introduction to Agricultural Geography*, Routledge, London and New York, 1995.

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

Herbst, Jeffrey, *State and Power in Africa*, Princeton, NJ: Princeton University Press, 2002.

Leff, B, N Ramankutty, and J.A. Foley, "Geographic Distribution of Major Crops Across the World," *Global Biogeochemical Cycles*, 2004, *18* (1).

Michalopoulos, Stelios, "Natural versus Man-Made Ethnolinguistic Diversity: Implications for Comparative Economic Development," 2008. Mimeo, Department of Economics, Tufts University.

Miguel, Edward and Daniel Posner, "Sources of Ethnic Identification in Africa," 2006. Working Paper, mimeo University of California, Berkeley.

Montalvo, José G. and Marta Reynal-Querol, "Ethnic Polarization, Potential Conflict and Civil War," *American Economic Review*, 2005, *95*, 796–816.

Nettle, Daniel, *Linguistic Diversity*, Oxford: Oxford University Press, 1996.

Nichols, Johanna, "Chechen Phonology," in P Daniels AS Kaye, ed., *Phonologies of Asia and Africa*, Bloomington: Eisenbrauns, 1997.

___ , "Modeling Ancient Population Structures and Movement in Linguistics," *Annual Review of Anthropology*, 1997a, *26* (6), 359–384.

Olson, Paul A., *Struggle for the Land: Indigenous Insight and Industrial Empire in the Semiarid World*, Nebraska: University of Nebraska Press, 1990.

Pavitt, Nigel, *Samburu*, Kyle Kathie Limited, 2001.

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

Rao, Vijayendra and Radu Ban, "The Political Construction of Caste in South India," 2007. mimeo, The World Bank.

Renfrew, Colin, "At the Edge of Knowability: Towards a Prehistory of Languages," *Cambridge Archaeological Journal*, 2000, *10*, 7–34.

Rosenzweig, Michael L., *Species Diversity in Space and Time*, New York, NY: Cambridge University Press, 1995.

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

Willliamson, Jeffrey G., "Poverty Traps Distance and Diversity: The Migration Connection," 2006. NBER Working Paper No. 12549.