The Origins of Ethnolinguistic Diversity: Theory and Evidence

Stelios Michalopoulos*

July 2, 2007

Abstract

This research examines theoretically and empirically the economic origins of cultural diversity and sheds new light on the emergence of ethnolinguistic fractionalization. The study argues that differences in the productive activities across regions led to the emergence of region specific human capital. Among regions characterized by dissimilar productive endowments, population mixing was limited leading to the formation of localized ethnicities and languages, producing a wider ethnolinguistic spectrum. Using new detailed data on the global distribution of land quality, the empirical analysis conducted in a cross country as well as cross-region framework, reveals that variation in land quality contributed significantly to the emergence and persistence of ethnolinguistic diversity. The empirical results also document the impact of European colonization on the ethnic diversity of the colonized world both through the drawing of the borders and the active manipulation of the underlying ethnicities. This research contributes to an understanding of the emergence and the distribution of languages and ethnicities and constitutes a first step towards comprehending the natural, i.e. geographically driven, and artificial, i.e. man-made, components of contemporary ethnolinguistic diversity.

Keywords: Ethnolinguistic Diversity, Geography, Technological Progress, Population Mixing, Colonization

JEL classification Numbers: O11, O15, O33, O40, J20, J24.

*I am indebted to Oded Galor for his constant advice and mentorship. Comments from Andrew Foster, Ioanna Grypari, Peter Howitt, Nippe Lagerlof, Ashley Lester, Ross Levine, Glenn Loury, Ignacio Palacios-Huerta and David Weil as well as seminar participants at Brown University were very helpful. Lynn Carlsson's ArcGis expertise proved of invaluable assistance.

1 Introduction

This study provides a theoretical and empirical framework for understanding the economic origins of ethnic diversity. The formation of ethnic diversity has been a long standing topic of research in the realm of social sciences. A rich literature in the fields of political science, psychology, sociology, anthropology and history attests to it, see Hale (2004). However, the economic origins of ethnic diversity are poorly understood both from an empirical and a theoretical point of view, limiting the conclusions that may be drawn from the existing intensive discourse across disciplines. A similar concern also applies to the large and growing literature within economics which has focused on the relationship between ethnolinguistic diversity and economic outcomes. Consequently, identifying the foundations of ethnic diversity will decisively improve upon the interpretation of the existing literature. In particular, uncovering the forces behind the emergence of differential ethnic traits will have important implications for understanding comparative economic development today.

Providing a theory of how ethnic identity, cultural practises, religion and language are constructed is beyond the scope of this research. However, given that such elements of human behavior have emerged universally in all societies the puzzle remains as to why some places exhibit higher or lower levels of cultural diversity. Exploring the rise of ethnic diversity and identifying its underlying components is the goal of this research.

The key finding of this study is that diversity in land qualities across regions contributed significantly to the emergence and persistence of ethnic diversity. The empirical results, in particular show that contemporary ethnic diversity displays a natural component and a man-made one. The natural component is driven by the diversity of land quality across regions, whereas the man-made part reflects the idiosyncratic state histories of each country, including the colonial experience and the emergence of modern states among other things.

There are three elements that form the basis of the theory. The first is that variation in the set of optimal productive activities across regions, generated by the underlying variation in land qualities, gave rise to region specific human capital. These differences in region specific human capital constituted a barrier to population mixing. Subsequently, the extent to which localities overlapped regarding their productive characteristics determined how easy was for the local populations to transfer their region specific human capital. Distribution of land qualities conducive to regionally distinct sets of productive activities effectively hindered population mobility between places. On the other hand, places exhibiting more homogeneous productive structures would facilitate mixing of the local populations resulting in the formation of common

ethnolinguistic behavior.

Over time site specific productivity shocks generated incentives to relocate. According to the theory it is the interaction of these two elements, the easiness to transfer regional human capital and the incentive to change locality, induced by variation in the regional productivity shocks, that gave rise to differences in ethnic diversity both within and across countries.

As already mentioned the formation of common cultural and ethnolinguistic traits for a pair of regions is positively related to the intensity of population mixing within this pair. Such formulation derives from the observation that a region experiencing infrequent population exchanges is bound to give rise to distinct ethnolinguistic traits as cultural drift may dominate the evolution of the characteristics of such places. By not imposing a binary relationship in the ethnic similarity between two regions, the analysis may also be applied to understanding ethnic or linguistic distance, with higher intensity of population mixing leading to lower ethnic distance.[1]

Ethnicities and languages were formed in a stage of development when land was the single most important factor of production. The theory, thus, predicts that ethnic diversity should be prominent as long as land is the major input in the production process. On the other hand, during an era when general human capital,[2] rather than region specific ethnic capital, is the individual input to the production process, ethnic and linguistic markers would gradually become less salient since population mixing would be more frequent.[3]

In this respect the proposed theory 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 within. The primordial one qualifies ethnic groups as deeply rooted clearly drawn entities, Geertz (1967), whereas the constructivists or instrumentalists, Barth (1969), highlight the contingent and situational character of ethnicity. In the current framework, it is the heterogeneity in land's productive traits that gives rise initially to relatively stable ethnic diversity, an element of primordialism. However, as the process of development renders land increasingly unimportant then ethnic identity is bound to become less attached to a certain set of region specific skills and, thus, more situational and ambiguous in character.[4]

[1] Note that this statement applies in the long run. In the short run migration movements may increase diversity in the receiving place, see Williamson (2006).

[2] This obtains under the following assumptions. First, there is no assortative mating according to ethnicity, second all ethnicities in the non land-intensive stage of development have the same opportunities to acquire human capital and the institutions in place do not generate a systemic bias against any of them.

[3] For a discussion of the salience of ethnic identity on the eruption of civil conflict see Esteban and Ray (2007).

[4] In other words, as the importance of region specific knowledge diminishes, ethnicity gradually transforms into a consumption good.

The model developed employs a stochastic, one sector, two-region overlapping generations framework. Land, labor and region specific technology are employed in each regional production function. The technology in every area develops over time through learning by doing, and is available to the indigenous population. People in the beginning of each period compare the potential income of their place of origin to that in case of moving and act accordingly. The incentive to move stems from the differential impact of temporary regional productivity shocks. Transferring region specific know-how across places, however, is costly in the sense that it may not be applicable to the receiving place. In fact, this cost increases in the heterogeneity of productive activities between places. Consequently, conditional on the productivity shocks, regions with larger overlap in their productive characteristics would experience more frequent population exchanges. Similarly, pairs of areas characterized by larger regional productivity fluctuations would display consistently more intense population mixing, ceteris paribus.

The proposed framework may also be used to understand both long-range migratory movements like the spread of the first agriculturalists and herders following the Neolithic Revolution as well as migratory patterns within shorter ranges. The historical evidence in section 2, centers on both the cause and effect of long range and short range population movements on linguistic spreads.

In the empirical section the regional heterogeneity in productive structures, which is the focus of the theory, is proxied using detailed regional data on the distribution of land quality for agriculture for the whole world. The econometric analysis is conducted in a cross-region as well as a cross country framework. For the cross-region regressions I arbitrarily divide the world into geographical entities of a given size and consistent with the theory I find that the number of languages spoken in these regions is systematically related to the underlying variation in land quality. Regions characterized by a wider spectrum of land qualities give rise and support larger linguistic fragmentation. Including continental and country fixed effects, effectively taking into account the idiosyncratic state and continental histories, the findings remain robust. Moving into a cross-country framework the proposed hypothesis is also validated. Countries characterized by more heterogeneous land qualities, exhibit higher ethnolinguistic fractionalization. This highlights the fundamental role that the spectrum of regional land qualities has played in the formation of more or less culturally diverse societies.

Testing alternative hypotheses regarding the formation of ethnolinguistic diversity, focusing on differential historical paths like the timing of the emergence of modern states, the

population in 1500 as a proxy for early economic development, and additional geographical characteristics like elevation and distance from the sea among other features, the qualitative predictions remain intact. Interestingly, the identified strong negative impact of the distance from the equator on ethnic diversity is consistent with the prediction that places experiencing persistent productivity shocks are conducive to low ethnic diversity. Note that distance from the equator correlates with seasonality. This emphasizes the economic basis of the origins of cultural diversity.

Historical accidents have influenced fractionalization outcomes. The European colonization after the 15th century, for example, is an obvious candidate. Analyzing the role of the colonizers in affecting the ethnolinguistic diversity of the colonized countries, reveals important patterns. The evidence is suggestive of the historically documented arbitrariness of border drawing, see Englebert et al. (2002). In particular, the results show that the way borders were drawn, generated a spectrum of land qualities which was conducive to higher ethnolinguistic diversity. However, colonizers did not only affect the geographically determined level of fractionalization. As a consequence of the introduction of their own ethnicity and the active interfering with the local populations, they generated artificial fractionalization, that is a component of ethnolinguistic diversity which was not an outcome of the underlying geography. This decomposition of observed fractionalization into natural, i.e. driven by the distribution of land qualities, and man-made components, offers new insights regarding the origins of cultural diversity, highlighting the role of variation in land quality and colonial history in particular.

By identifying the role of the European colonizers in affecting both the natural and artificial elements of ethnolinguistic diversity this research adds to the literature on the impact of European colonization on the indigenous economies (La Porta et al. (1999), Acemoglu, Johnson and Robinson (2001)). The findings are also closely related to a recent study by Alesina et al. (2006) in which new measures of state artificiality are proposed. Man-made fractionalization, measured by the fraction of ethnolinguistic diversity not explained by the underlying distribution of land quality, increases significantly the probability of being included in the top 13 most artificial states that the authors provide. Naturally, it remains to be seen whether such relationship is relevant to the whole dataset or is only a feature of the identified countries.

This study is also directly related to the strand of literature that concerns the relationship between ethnolinguistic fractionalization and countries' economic performance, see Easterly and Levine (1997), Fearon and Latin (2003) and Alesina et. al. (2003) among others.

The theoretical and empirical thesis shows that ethnic diversity is driven by the distribution of land quality within a country. At the same time the empirical analysis shows that the divergent state histories of existing countries, evident in the presence or absence of colonization as well as in the levels of early economic development, have influenced significantly the contemporary ethnolinguistic endowment. Consequently, the documented negative relationship between ethnolinguistic fractionalization and economic outcomes may reflect the direct effect of state history rather than a true effect of ethnic diversity. Thus, further research on the causal impact of ethnic diversity on comparative economic development today is warranted.[5]

Another line of research to which the findings are relevant is a recent study by Spolaore and Wacziarg (2006). 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 inversely affects the adoption of technology. In the proposed framework population mixing between two regions, may directly reduce genetic distance. Thus, the latter is endogenous to both the regional productivity shocks and the transferability of region specific technology within the pair. As a result, countries that are relatively dissimilar in the distribution of productive possibilities, will be populated by people displaying larger genetic distance, ceteris paribus. Consequently, the uneven diffusion of development across countries may be an outcome of the differences in country specific human capital rather than genetic distance itself. It would be interesting to replicate their empirical analysis introducing the pair-wise country distances of the distribution of land quality. An inclusion of such control is bound to partially account for the documented significant effect of genetic distance on pair-wise income comparisons.

The results could be also used to understand the diffusion of technology not only across but also within countries. Technology would diffuse more quickly in more homogeneous countries, land quality wise, whereas in relatively heterogeneous ones, and according to the theory and evidence more culturally diverse, the diffusion would be less rapid leading to the emergence of inequality among ethnic groups. This would obtain because of the differential complementarity between a new technology and the preexisting variation in ethnic specific human capital reflecting the variation in regional land qualities. Intuitively speaking, herders unlike farmers would be less likely to adopt a new technology specific to farming.

This research sheds new light on the emergence and the distribution of languages and ethnicities and constitutes a first step within economics towards the understanding of natural

[5] Michalopoulos (2007b) uses the proposed framework to uncover the causal impact of ethnolinguistic diversity on the economic performance across countries looking on a variety of economic indicators.

and man-made components of ethnic diversity. This study is a stepping stone for further research. Equipped with a more substantive understanding of the origins and determinants of ethnolinguistic diversity, new ways of addressing long standing important questions among development and growth economists may be offered. These range from the formation of states, to the inequality across ethnic groups, to the effect of ethnolinguistic diversity on the eruption of civil wars, public good provision and economic outcomes in general.

The paper is organized as follows. In Section 2 (pre)historical evidence regarding the occurrence of migrations and the spread of linguistic groups is briefly reviewed. Section 3 presents the theory and its predictions. Section 4 discusses the data and covers the empirical analysis conducted both in a cross-region and a cross-country framework, including the various robustness checks and finally focusing on the impact of the European colonizers on the observed fractionalization outcomes. The last section concludes.

2 Evidence on Migrations and Language Spreads

The theory rests upon three fundamental building blocks: (*i*) population movements influence the ethnolinguistic diversity of the places involved, leading eventually to a convergence in the underlying traits (*ii*) migration of ethnic groups and languages occurs more often among places with similar productive endowments (*iii*) regional productivity shocks generate the incentive to relocate from one place to another.

Among linguists it has been long recognized the role of population mixing in producing common linguistic elements between places. As Nichols (1997) points outs "almost all literature on language spreads[6] focuses on either demographic expansion or migration as the basic mechanism". Both instances are a result of the movement of populations towards territories previously unoccupied by their ancestors. In these new regions population mixing leads to language shift (either to or from the immigrants' language). Also, 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 extent that recurrent contact between regional populations may occur through repetitive cross migrations (short-term or long-term), the modeling of the emergence of common ethnolinguistic characteristics as an increasing function of population mixing between places is justified.

There are several examples showing that migrations have been occurring between places

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

of similar productive characteristics. Linguistic research has identified several regions of the world which are spread zones of languages, that is regions characterized by low linguistic diversity. A common characteristic of such regions is the underlying homogeneity in the endowment of land quality, as it is the case for the grasslands of central Eurasia. Generally, large spread zones are associated with high latitudes where seasonality is evident and arid interiors, whereas linguistic heterogeneity increases in the less seasonal climates, Nichols (1997). These distributional features highlight the role of variations in productivity shocks in shaping migration movements and, ultimately, linguistic patterns.

Examples of groups that migrated along areas that were similar to their region of origin are Austronesians and speakers of Eskimoan languages who are coastally adapted peoples, and accordingly they have 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 the distribution of land qualities that were relevant for agricultural activities. In fact, the pattern of the languages' expansion, belonging to the Indo European family, after the Neolithic revolution is embedded to the notion of "spread" and "friction" or "mosaic" zones. "Spread" regions were characterized by similar land qualities where the early agriculturalists in the case Indo-European languages[7], or nomad pastoralists in the case of the Turkic and Mongolian languages (these belong to the Altaic language family) could easily apply their own specific knowledge and friction zones were places less conducive to either activity. In such places the populations maintained their distinct ethnolinguistic behavior. Examples of the latter include regions like Melanesia, Western and Northern Europe and Northern India, see Renfrew (2000) for a comprehensive review. This implies that early agriculturalists and pastoralists, perhaps not surprisingly, targeted and expanded at areas where their specific knowledge would best apply, homogenizing them linguistically. If this process of language shift occurred through replacement of the local populations or by extensive intermarrying is yet an open question.

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, Cushin (1984). It is in this case the knowledge of how to conduct commerce that allowed these groups to spread in areas where merchandising was both possible and profitable. Botticini and Eckstein (2006), for example, document the religiously driven transformation of the Jewish ethnic human capital towards

[7] Gray and Atkinson (2003) produce evidence showing that IndoEuropean languages expanded with the spread of agriculture from Anatolia around 8,000–9,500 years BP.

literacy and the resulting expansion.

Generally, according to the theory migratory movements should be relatively more frequent among ethnic groups whose knowledge is less attached to specific land attributes, as in the case of trade diasporas. Should the ethnic knowledge be region specific, though, then such groups would disperse in places that are similar to the place of origin regarding the underlying productive activities, minimizing, thus, erosion of their specific human capital.

Regarding the effect of differential climatic shocks in generating movements of people evidence suggests that this was indeed an important factor.[8] For example, as Nichols (1997) suggests, 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 present in regions like central Caucasus (Nichols 1997b) and the highland spread of Quechua fall in this category.

The linguistic and (pre)historical evidence for the spread of peoples and languages provide ample support to the building blocks of the theory presented below.

3 The Basic Structure of the Model

Consider an overlapping-generations economy in which economic activity extends over infinite discrete time. Each individual lives two periods and bears exactly one child in the second period of her life, i.e. population is fixed. In every period the economy produces a single homogeneous good using land, labor and region specific technology as inputs in the production process. The supply of land is exogenous and fixed over time. In fact, there are two regions in the economy i and j. The supply of labor in each place is determined by the evolution of the region specific know-how, its transferability between the places and the state of the temporary idiosyncratic productivity shock relative to the other region.

[8] The independent role of regional climatic fluctuations in generating differential timing of the transition to agriculture has been proposed by Ashraf and Michalopoulos (2006).

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, Y_t^r, is

$$Y_t^r = (z_t^r h_t^r)\,(L_t^r)^{\alpha}\,(m^r X^r)^{1-\alpha}; \qquad \alpha \in (0,1),\ r \in \{i,\ j\} \tag{1}$$

where z_t^r is the productivity shock in period t in region r, h_t^r is the level of knowledge in period t relevant to region r which evolves over time through learning by doing - it may be interpreted as region specific human capital - L_t^r is the total labor employed in period t in region r, m^r represents the land quality of region r and X^r is the size of land used in production in every period in region r (which for simplicity is normalized to 1 for all r).

Suppose that there are no property rights over land.[9] The return to land in every period is therefore zero, and the wage rate in period t is equal to the output per worker produced at time t, y_t^r.

$$y_t^r = (z_t^r h_t^r)\,(m^r/L_t^r)^{1-\alpha} \tag{2}$$

3.2 Preferences

In every period a generation, which consists of a continuum of individuals of measure L, is born. Specifically, an agent born in period t, gives birth at the beginning of period $t+1$ in the region where she works at that period. People, within as well as across generations, are identical in their preferences and their ability in utilizing the technology of the region they are born to. Individuals live for two periods. In the first period, they 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 and consume the earnings.

Individuals' preferences are defined over consumption in the second period of their lives, c_{t+1}.[10] The preferences of an individual n born in period t are, thus, represented by the utility function,

$$u^{t,n} = u\left(c_{t+1}^{t,n}\right) \tag{3}$$

[9] The modeling of the production side is based upon two simplifying assumptions. First, capital is not an 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 primitive characteristics as the current set up leading to qualitative similar predictions.

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

where $c_{t+1}^{t,n}$ is the consumption of a member n of generation t in period $t+1$. The utility function is strongly monotone and strictly quasi-concave.

3.3 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.

$$h_{t+1}^r = \psi\left(h_t^r\right),\ r \in \{i,\ j\}$$

with $h_0^r = 1$, $\psi_{h_r^r} > 0$ and $\psi_{h_t^r h_t^r} < 0$, that is, the level of regional know-how in any period is a monotonically non-decreasing concave function of the level of know-how in the preceding period. 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. That is,

$$h_t^j = h_t^i \quad \forall \quad t \geq 0 \tag{4}$$

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 enjoying a higher technological growth rate less willing to move, ceteris paribus. Furthermore, it's 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).

3.4 Transferring region specific technology across places

As adults, individuals may move freely from one region to another.[11] However, this comes at a cost arising from differences in territory-specific human capital. In particular, since the level of technology, h_t^r, is region r specific, relocation renders obsolete part of the knowledge that the individual may apply as a worker in the receiving place. This erosion increases as places become increasingly different in the feasible and/or optimal set of productive activities.

[11] Including additional costs associated with migration, 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.

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:

$$k_t^r = (h_t^q)^{1-\varepsilon} \ \forall \ r, \ q \in \{i,j\}, r \neq q, \\ 0 \leq \varepsilon \leq 1, \quad h_t^q \geq 1 \tag{5}$$

where k_t^r are the units of knowledge that a migrant will be able to apply should she move to region r and ε captures the degree of erosion within regional pairs. Those characterized by more heterogeneous endowments score higher along this dimension. Note that within a regional pair erosion of region-specific knowledge is symmetric, thus it is quantitatively identical irrespective of the direction of the migration.

The properties of transferring region-specific technology across places, follow directly by differentiating (5):

1. The migrant's level of know-how relevant to the receiving place decreases in the level of erosion between the regions, $\frac{\partial k_t^r}{\partial \varepsilon} < 0 \ \forall \ r \in \{i,j\}$

2. The migrant's level of know-how relevant to the receiving place increases in the level of know-how of the place of origin, $\frac{\partial k_t^r}{\partial h_t^q} > 0, \forall \ r, \ q \in \{i,j\}, \ r \neq q$.

2. 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 the fact that the accumulation of technology becomes increasingly region specific and, as a result, less useful in case of migration.[12]

3. 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 (4) 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.[13] Specifically, using (2)

[12] 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 pair of regions, further specialization in the respective activities diminishes the transferability of the additional know-how.

[13] 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

$$y_t^r = (z_t^r h_t^r)\,(m^r/L_t^r)^{1-\alpha}$$
$$y_t^{q\to r} = (z_t^r k_t^r)\,(m^r/L_t^r)^{1-\alpha} \tag{6}$$

$\forall\ r,\ q \in \{i,j\}, r \neq q.$

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.

3.5 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.[14] As individuals cross-migrate, they add their cultural traits from the place of origin 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 incessantly 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 past cross–migrations would evolve to exhibit distinct ethnolinguistic characteristics.

Formally, let f_T denote the probability that places, i and j, observed in period T will exhibit common ethnolinguistic elements.

$$f_T = \frac{\sum_{t=1}^{T} I_t}{T} \tag{7}$$

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. As already mentioned since this relationship applies in the long-run, T should be thought as relatively large. According to this definition pairs of places whose populations never mixed

not alter the results. It could, however, create selection into reverse migration of the people whose ancestors were immigrants.

[14] Assuming either perfect initial ethnolinguistic heterogeneity or perfect homogeneity across regions does not affect the pattern of ethnolinguistic assimilation. Should the latter be the case, then cultural practices are formed regionally as time evolves due to cultural drift, Boyd and Richardson (1985).

until period T would have zero probability of sharing common ethnic traits, or alternatively put, maximal ethnolinguistic distance. Alternative specifications of (7) could accommodate a potential "founder" effect in case that earlier migrations have a larger impact than later ones in the formation of common ethnicity. Also, 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 cultural diversity across places. The analysis below establishes how this intensity varies according to the transferability of regional human capital.

3.6 Labor Allocation Across Regions

Since the utility of members of generation t depends only on their consumption as adults, utility maximization is equivalent to maximizing lifetime income. In the beginning of every period t productivity shocks, z_t^r, which last for one period, are realized in each region. Adult individuals observe the realization of the shock[15] and decide whether or not to migrate by comparing the respective incomes in (6).[16] Erosion of region-specific technology decreases potential income in case of relocation, whereas a relatively higher productivity shock in the host area acts as an incentive to the prospective migrant. This is the fundamental trade-off created by the forces in the environment.

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 given the ratio of the regional population densities bequeathed from period $t-1$. Let $\{\lambda_t\}_{t=0}^{T}$ denote the sequence of the ratio 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 (6) and substituting L_t^i, L_t^j with the respective values of the preceding period, individuals from region i have an incentive to migrate to region j in the beginning of period t iff;

$$y_t^{i\to j} > y_t^i \Rightarrow \lambda_t < \left(h_t^i\right)^{-\varepsilon} \left(\frac{m^j}{m^i}\frac{L_{t-1}^i}{L_{t-1}^j}\right)^{1-\alpha} \tag{8}$$

[15] The analysis abstracts from imperfect information regarding the size of the regional productivity shocks.

[16] Migration in this framework lasts for at least one generation. It would be straightforward to incorporate short term migration by allowing for more subsequent 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.

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

$$y_t^{j\to i} > y_t^j \Rightarrow \lambda_t > \left(h_t^j\right)^{\varepsilon} \left(\frac{m^j}{m^i}\frac{L_{t-1}^i}{L_{t-1}^j}\right)^{1-\alpha} \tag{9}$$

It is obvious from (8) and (9) 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.

Lemma 1 *When individuals in one region strictly prefer to migrate then individuals in the other region strictly prefer not to, i.e.,*

$$\begin{aligned} y_t^{i\to j} &> y_t^i \Rightarrow y_t^{j\to i} < y_t^j \\ y_t^{j\to i} &> y_t^j \Rightarrow y_t^{i\to j} < y_t^i \end{aligned}$$

Proof. It is straightforward to show that the right-hand side of (8) is always smaller than the right-hand side of (9). They coincide when $\varepsilon = 0$ or $h_t^j = h_t^j = 1$. □

Given the absence of mobility barriers, as long as either of inequalities in Lemma 1 obtain 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 exact size of the realized migration is the one that makes the marginal individual from the place of origin indifferent between moving and staying in the land 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, (8) should hold with equality. Adding the size of the migration $M_t^{i\to j}$ in the receiving region, j, subtracting it from the region of origin, i, and manipulating (8) the level of migration may be explicitly derived

$$M_t^{i\to j} = \frac{L_{t-1}^i - \left(\lambda_t \left(h_t^i\right)^{\epsilon}\right)^{\frac{1}{1-\alpha}} \frac{m^i}{m^j} L_{t-1}^j}{1 + \left(\lambda_t \left(h_t^i\right)^{\epsilon}\right)^{\frac{1}{1-\alpha}} \frac{m^i}{m^j}} \tag{10}$$

Note that the numerator of (10) is always positive as a long as (8) 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,

$$M_t^{j\to i}=\frac{\left(\lambda_t\left(h_t^j\right)^{-\epsilon}\right)^{\frac{1}{1-\alpha}}\frac{m^i}{m^j}L_{t-1}^j-L_{t-1}^i}{1+\left(\lambda_t\left(h_t^j\right)^{-\epsilon}\right)^{\frac{1}{1-\alpha}}\frac{m^i}{m^j}} \tag{11}$$

Again, note that the numerator in (11) is positive as long as (9) holds in the beginning of period t.

3.6.1 Past Migrations

As it is evident from (10) and (11) the size of the migration movement in period t depends on the level of regional population densities in the period $t-1$. The following Lemma derives the ratio of population densities in the beginning of every period t as a function of past migration movements (if any).

Lemma 2 *In the beginning of any period t, and before any labor movement occurs (if any) the ratio of the regional population densities equals $\frac{L_{t-1}^i}{L_{t-1}^j}$.*[17] *There are two cases:*

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

$$\frac{L_{t-1}^i}{L_{t-1}^j}=\frac{L_s^i}{L_s^j}=\left(\lambda_s\left(h_s^{i}\right)^{\epsilon}\right)^{\frac{1}{1-\alpha}}\frac{m_i}{m_j}\quad if\quad M_s^{i\to j}>0 \tag{12}$$

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

$$\frac{L_{t-1}^i}{L_{t-1}^j}=\frac{L_s^i}{L_s^j}=\left(\lambda_s\left(h_s^{j}\right)^{-\epsilon}\right)^{\frac{1}{1-\alpha}}\frac{m_i}{m_j}\quad if\quad M_s^{j\to i}>0 \tag{13}$$

Proof. Depending on the direction of the last migration either (8) or (9) should hold with equality when evaluated at the regional population densities after the occurrence of migration in period, s. Solving for the ratio of regional population in period s, $\frac{L_s^i}{L_s^j}$, completes the proof.□

Corollary 1 *If last migration occurred in period $s=0$, that is, it represented the initial settlement of people across regions i and j, then the ratio of regional population densities in period $t-1$ equals:*

$$\frac{L_{t-1}^i}{L_{t-1}^j}=\frac{L_0^i}{L_0^j}=\lambda_0\frac{m_i}{m_j} \tag{14}$$

[17] The latter is identical to the ratio of population densities realized in the last occurrence of migration.

Proof. Follows from Lemma 2 and noting that (12) and (13) are identical evaluated at $h_0^j = h_s^j = 1$. □

In Appendix A Lemma 4 establishes the properties of the size of the migration between places given by (10) and (9).

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

Given the definition of common ethnicity in (7) 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 migration 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 or not, that is, $y_t^{i\to j} = y_t^i$.

$$M^iM^j \equiv \left\{ \left(h_t^i, \lambda_t, \frac{L_{t-1}^i}{L_{t-1}^j}, \varepsilon\right) : y_t^{i\to j} = y_t^i \right\}$$

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:

$$y_t^{i\to j} = y_t^i \Rightarrow \lambda_t|_{M^iM^j} = \left(\frac{L_{t-1}^i}{L_{t-1}^j}\frac{m_j}{m_i}\right)^{1-\alpha} \left(h_t^i\right)^{-\varepsilon} \tag{15}$$

Similarly, M^jM^i is the geometric locus of all tuples $\left(h_t^j, \lambda_t, \frac{L_{t-1}^i}{L_{t-1}^j}, \varepsilon\right)$ such that the marginal individual in region j is indifferent between moving or not, that is, $y_t^{j\to i} = y_t^j$. In particular,

$$M^jM^i \equiv \left\{ \left(h_t^j, \lambda_t, \frac{L_{t-1}^i}{L_{t-1}^j}, \varepsilon\right) : y_t^{j\to i} = y_t^j \right\}$$

Thus, the level of the relative productivity shock in period t, $\lambda_t|_{M^jM^i}$, that makes people from region j indifferent to moving is:

$$y_t^{j\to i} = y_t^j \Rightarrow \lambda_t|_{M^jM^i} = \left(\frac{L_{t-1}^i}{L_{t-1}^j}\frac{m_j}{m_i}\right)^{1-\alpha} \left(h_t^j\right)^{\varepsilon} \tag{16}$$

As it is evident in (15) and (16) the ratio of the regional population densities from the last period is important in determining the no-migration loci. From Lemma 2 the ratio of regional population densities in period $t-1$ may be expressed by either (12) or (13) depending on the

direction of the last movement across places in period s. The following lemma summarizes the properties of the migration indifference curves.

Lemma 3 *Using Lemma 2 at any period t these are the properties of the non-migration loci for each region*

1. *The M^iM^j locus*

$$\frac{\partial \lambda_t}{\partial h_t^i}\Big|_{M^iM^j} < 0 \quad \& \quad \frac{\partial^2 \lambda_t}{\partial^2 h_t^i}\Big|_{M^iM^j} > 0$$

$$\frac{\partial \lambda_t}{\partial \varepsilon}\Big|_{M^iM^j} < 0 \quad \& \quad \frac{\partial^2 \lambda_t}{\partial^2 \varepsilon}\Big|_{M^iM^j} > 0$$

$$\frac{\partial \lambda_t}{\partial \lambda_s}\Big|_{M^iM^j} > 0 \quad \& \quad \frac{\partial^2 \lambda_t}{\partial^2 \lambda_s}\Big|_{M^iM^j} = 0$$

2. *The M^jM^i locus*

$$\frac{\partial \lambda_t}{\partial h_t^j}\Big|_{M^jM^i} > 0 \quad \& \quad \frac{\partial^2 \lambda_t}{\partial^2 h_t^j}\Big|_{M^jM^i} < 0$$

$$\frac{\partial \lambda_t}{\partial \varepsilon}\Big|_{M^jM^i} > 0 \quad \& \quad \frac{\partial^2 \lambda_t}{\partial^2 \varepsilon}\Big|_{M^jM^i} > 0$$

$$\frac{\partial \lambda_t}{\partial \lambda_s}\Big|_{M^jM^i} > 0 \quad \& \quad \frac{\partial^2 \lambda_t}{\partial^2 \lambda_s}\Big|_{M^jM^i} = 0$$

Proof. See Appendix A. □

The pair of Figures below $(1a, 1b)$ shows the effect of the erosion, ε, on the occurrence of migration. As it follows from Lemma 3, 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 at the level of erosion. This implies that given the contemporary productivity shock, λ_t, pairs of regions i and j which are more dissimilar with respect to their productive structures experience infrequent population mixing limiting the formation of common ethnolinguistic traits. Figure $1b$ is drawn with a higher level of region specific technology than in $1a$ to exemplify the adverse effect of the accumulation of region specific human capital on migration outcomes. This obtains because as people further specialize in their regions' specific productive activities the accumulating knowledge becomes increasingly less transferable, hindering cross-migration. 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.

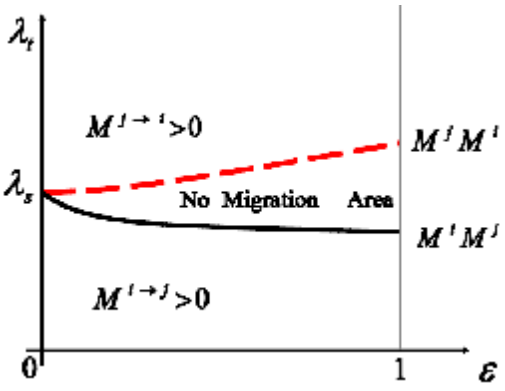


Figure 1a

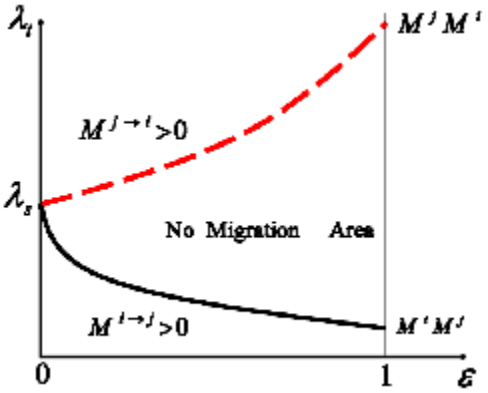


Figure 1b

In the set of figures above and from Lemma 3 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 (see figures $1a, 1b$), the larger the difference between the temporary shock λ_t and λ_s the more probable is the occurrence and the larger is the size of migration. The latter is established in Lemma 4.

Lemma 5 in Appendix A summarize the cases of migration occurrences.

3.8 The Formation of Common Traits Over Time

Being equipped with the relevant Lemmas about how population mixing is shaped by the environment, 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.[18] In particular,

$$\lambda_t = \begin{cases} \lambda_{\min} & \text{with probability} \quad p \\ \\ \lambda_{\max} & \text{with probability} \quad 1-p \end{cases} \tag{A1}$$

[18] 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.

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

Proposition 1 *Under (A1)*

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. 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 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. The latter, will be the focus of the empirical analysis.

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 and/or seasonality, 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 variable productivity shocks are bound to form homogeneous ethnolinguistic traits.

It is important to note that the theory is about individuals from different geographical entities sharing or not common cultural elements. Consequently, the distribution of population across regions needs to be taken into account in order to translate these predictions into statements about the overall level of ethnolinguistic fractionalization within a country.

The following section presents the data and the empirical strategy bringing the theoretical predictions into econometric analysis conducted both in a cross-region and cross-country framework.

4 Empirical section

4.1 The Data Sources

To test the predictions generated by the theory, an index of the transferability of region specific human capital is needed. An ideal index could be derived looking into how similar were the regional distributions of productive activities across places in a period of human history when the formation of cultural traits was taking place. Such quest for detailed data though 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 factor in the production process, I use contemporary regional detailed data on the suitability of land for agriculture.[19]

The intuition for using differences in land quality as a proxy for differences in the distribution of productive activities is the following. Farming is bound to 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 more common for intermediate and low levels of land quality, exactly because agriculture is less suitable in such areas. At very low levels of land quality, also, being a middleman has been perhaps the most widespread activity as the case for cultures residing

[19] Detailed disaggregated data on land quality going sufficiently back in time do not exist. Reassuringly, the measure of quality of land used is, as wouls be expected, highly correlated (magnitude of 0.40) with population density in 1500 AD.

along trade routes suggests. 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. Such places in absence of these trading routes 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 in other places.

The global data on agricultural suitability, originally in grid format, were assembled by Ramankutty et al. (2002) to investigate the effect of the expected climatic change on agricultural suitability.[20] This dataset provides regional detailed information on land quality characteristics (see below). The resolution is 0.5 degrees (latitude *by* longitude), thus the average land plot has a size of about 55 km by 35 km. In total there are 58920 observations. Using this global data I derive the distribution of land quality for each country. Number of regional observations per country range from a single observation for Luxemburg to 11515 for Russia. The median number of points per country is 80.[21]

Each observation is a value between 0 and 1 and represents the probability that a particular grid cell may be cultivated. The authors construct this index by examining the relationships between existing croplands and both climate indices and soil characteristics and predict the suitability of agriculture for the entire world using the observed relationship.

The climatic characteristics which are based on mean-monthly climate conditions for the 1961–1990 period capture i) monthly temperature ii) precipitation and iii) potential sunshine hours. All these measures 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 and the 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, (see Ramankutty et al. (2002) for a thorough discussion of the index).

This detailed dataset, never used in an economic application, provides an accurate description of the distribution of land quality both within and across countries. The map in

[20] The dataset is available at the Atlas of the Biosphere accessible at http://www.sage.wisc.edu/atlas/data.php?incdataset=Suitability%20for%20Agriculture

[21] There are some missing countries, mostly islands whose size is not large enough to make it in the dataset. Regarding a subset of the existing countries, there are few pockets of land for which there is no information.

Appendix B shows the worldwide distribution of land quality.

For the regional analysis ethnic diversity is captured by the number of unique languages spoken in each region. I calculate the number of languages for each region using data on the locations of language groups obtained from 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 linguistics database of languages around the world.

Regarding the cross-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 a population will differ in the characteristic under consideration, like 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 (1964)), as 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 Reynal-Querol (2002) the qualitative results are roughly similar.[22]

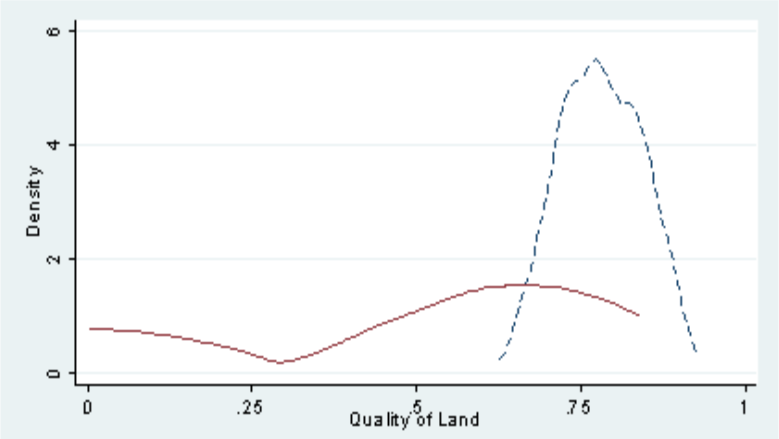
4.2 The empirical analysis

The distribution of land quality varies considerably across regions and across countries. For example, the following graph plots the distribution of regional land qualities for Greece and Nepal. These countries are of similar size. As it is evident in the figure[23] below, in Greece the quality of land is very concentrated around high values with average quality, $avg = 0.78$, and a range (this is the difference between the region with the highest land quality from that with the lowest) of 0.25. On the other hand, the land quality in Nepal averages 0.47 but it spans a much larger spectrum with a sizeable left tail. In fact $range_{Nepal} = 0.84$. The large difference in the spectrum of land qualities between these two countries is evident, as the theory would predict, in their respective degree of cultural diversity. Ethnolinguistic fractionalization

[22] Using the polarization index constructed by Reynal-Querol (2002) as a measure of ethnic diversity produces results that are qualitatively similar. Significance, though, varies according to the specification, becoming insignificant as more controls are added in the regression analysis.

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

in Greece is only 0.10 compared to the highly ethnolinguistically fragmented society of Nepal with $ELF_{Nepal} = 0.70$.



Dashed line: Greece, Solid Line: Nepal

The range of land quality, i.e. the support of the distribution within the respective unit of analysis, either at a regional or country level, is the statistic used to illustrate the degree of heterogeneity in the quality of land.[24] It captures how readily location specific knowledge may be transferred across places. Intuitively, a larger range implies that territories are increasingly different in their underlying qualities, which would lead to regionally distinct sets of activities. Consequently, the larger is the spectrum of land qualities, i.e. *range*, within the unit of analysis the higher is the erosion of the regional know-how in case of relocation. Thus, according to the theory[25] a larger range would increase the probability that the underlying areas are ethnically distinct, ceteris paribus.

The average quality of land, *avg*, according to the theory, should not have any direct effect on ethnic diversity, since it is only the difference in the productive structure across places that

[24] The standard deviation of the quality of land 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 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 change, nevertheless, essentially produces a more unequal distribution of population across regions and since by construction the fractionalization indexes are affected by the distribution of the population across ethnolinguistic groups (see below) an increase in the intensive margin may decrease fractionalization. Results not shown indeed suggest that when controlling for the range of land quality and the standard deviation simultaneously both enter significantly with the range positive and the standard deviation negative. 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 instead.

[25] The implications of the theory have been derived for pairs of regions. Extending the model to allow for multiregional migration i conjecture that would not affect the qualitative predictions. It would, however, deliver a cumbersome analysis.

matters. If places are perfectly 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.[26]

4.2.1 Cross-region analysis

Before turning into the cross-country analysis it is important to investigate whether the predictions of the theory are relevant to any arbitrary unit of analysis. Finding that at any level of regional aggregation a larger spectrum of land qualities leads to higher ethnic diversity will greatly enhance the validity of the proposed theory.

The way that the regions are constructed is the following. First, I generate a global grid where each regional unit is 4 degrees longitude by 4 degrees latitude and then I intersect it with the global data on land quality, see the map in Appendix B with the resulting "artificial countries" which constitute the unit of analysis. The dimensions are chosen to guarantee that there are sufficient observations of land quality per "artificial country". Using alternative dimensions for the grid does not change the results.

For each "artificial country" i derive the distribution of land quality and calculate the number of unique languages spoken. The latter is computed by counting the number of languages spoken at each observation of land quality. Specifically, i count the number of languages at a distance of 0.25 degrees from the centroid of each observation of land quality. This guarantees that all languages within an "artificial country" are considered. Then, I aggregate the number of unique languages spoken over all land qualities that fall into each "artificial country" generated by the grid. The variable representing the number of languages spoken within each "artificial country" is denoted $number\_lang$.

In the regression analysis the sample of "artificial countries" is restricted in the following way. Only those territories for which there are at least 10 regions with information on land quality are included. Additionally, to ensure that the findings are not driven by including in the regressions regions with very low, or even zero, population density, "artificial countries" with average population density less than 1 person per sq. km. are excluded. Finally, "artificial countries" in which the number of languages spoken exceeds the available observations of land

[26] Nevertheless, conditional on a positive qualitative distance across pockets of land, proxied by the *range*, increases in the average land quality may increase the easiness of transferring knowledge across places. The intuition is the following: as the average land quality increases, the distribution shifts to the right and agriculture becomes gradually the dominant activity. Within agriculture, though, the region-specific human capital is easier to transfer, since the production process is more homogeneous. Given the construction of the land quality index this implies that the **actual** heterogeneity in productive activities between places, that is the erosion in the transferability of region specific human capital, may decrease as the average level of land quality increases. As it will become evident such an effect is present in the cross-country regressions but not in the cross-region ones.

quality are not considered to avoid detecting any relationship driven by extremely linguistically fragmented units. The is the case for only 37 "artificial countries".[27]Such considerations produce a sample size of 887 "artificial countries" with a median of 64 observations on regional land qualities per "artificial country". Descriptive statistics and the raw correlation between the variables used in the regressions are presented in Tables $1a, 1b$. The additional variables included are: the absolute latitudinal distance from the equator, denoted abs_lat, the area of each "artificial country", denoted by $areakm2$ and measured in thousand of square kilometers, and the standard deviation of elevation measured in meters within each unit of analysis, denoted $elev\_sd$.[28] This measure, constructed by the author, is chosen because it captures accurately the variation in topography and, thus, the pure transportation cost associated with relocation. In each "artificial country" there are on average 7.45 languages spoken and the raw correlation between the spectrum of land qualities, $range$, and the number of languages is large, 0.27, and positive.

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

$$number\_lang_i = \beta_0 + \beta_1 range_i + \beta_2 X_i + \xi_i \tag{17}$$

where $number\_lang_i$ is the number of unique languages spoken in "artificial country" i, $range_i$ is the support of the distribution of land quality, and X_i is a vector of additional controls. The key prediction of the theory is that the larger is the spectrum of land qualities across places the higher is the probability that these places will develop distinct ethnic traits.

This main prediction is corroborated across all alternative specifications of Table 2.[29] Specifically, in the first regression of Table 2 only the $range$ is included. It has a large and significant positive impact on the number of languages spoken. The variation in land qualities itself explains 7% of the variation in languages spoken in the world today. This finding is robust to alternative specifications. In particular, in the second column of Table 2 the size of each "artificial country", the latitudinal distance from the equator and the standard deviation of elevation are included. As expected larger artificial units have more languages, those with more variable topography also sustain larger linguistic diversity and the distance from the equator itself has a strong negative impact on the number of languages spoken. These controls make

[27]Using alternative thresholds both for the minimum number of land quality observations per "artificial country" and for population density and/or including in the analysis those extremely linguistically fragmented "artificial countries" the qualitative results are similar.

[28]See Appendix C for a detailed description of the data used.

[29]The results presented here are OLS estimates with heteroskedastically robust standard errors. Using Tobit or Poisson estimators the predictions remain qualitative and quantitative intact.

the coefficient of *range* drop sufficiently, it remains however both economically and statistically highly significant. To the extent that the distance from the equator is associated with the presence of seasonality and more unpredictable climate in general, the strong negative impact of *abs_lat* on linguistic diversity is consistent with the prediction of the theory that areas characterized by variation in productivity shocks will give rise to more homogeneous ethnic entities. The introduction of the average quality of land, *avg*, and its interaction with the *range*, *avg_range*, is designed to capture a potential diminishing effect of the variation in land quality on the formation of ethnic diversity as average land quality increases. Such effect is not detected, the effect of the interaction is negative as expected but insignificant throughout. Consequently, the interaction is dropped from the rest of the cross-region analysis.

In columns 3 and 4 of Table 2 I take advantage of the cross-region framework to explicitly control for any idiosyncrasies of countries and continents. This is done by generating country and continental dummies for those "artificial units" that fall into a single country and/or a single continent respectively. Such inclusion of powerful controls, not possible in a cross-country framework, allows to fully take into account any idiosyncratic country histories and thus produce reliable estimates of the effect of variation in land qualities on ethnic diversity. The inclusion of country and continental fixed effects does not affect significantly the estimated coefficient on *range*. One standard deviation increase in the spectrum of land qualities increases by 1.56 the number of languages spoken contributing significantly to the formation of ethnically diverse societies. Both the latitudinal gradient, the standard deviation in elevation and the area of each "artificial unit" enter significantly and with the expected sign.

In the last two columns of table 2 specification (17) including continental and country fixed effects is estimated separately for "artificial units" that are outside the tropics[30] in column (5) and those that fall within the tropics in column (6). These regressions allow to investigate whether the identified impact of the variation in land quality is driven by the climatic differences between the tropics and the rest of the geographic zones. Reassuringly, in both regressions the effect of *range* on linguistic diversity is precisely estimated at 5% level. However, the impact of the variation in land quality is much larger in the tropics.

This section establishes that the variation in land quality across regions coupled with the distance from the equator are significant predictors of contemporary linguistic diversity. The fact that these results obtain at an arbitrary level of aggregation and after controlling for country and continental fixed effects, highlights the importance of the forces identified by the

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

theory.

4.2.2 Cross-country analysis

Having established that the variation of land quality across regions affects systematically the number of languages spoken today i now proceed into investigating the relationship between the spectrum of land qualities and ethnolinguistic fractionalization across countries.

Existing countries vary widely in the spectrum of land qualities that their territories cover. In Appendix C maps with the regional land qualities for Lesotho and Malawi are presented. A visual inspection of these maps reveals the homogeneity of land quality in Lesotho, $range_{Lesotho} = 0.37$ compared to the apparent heterogeneity inherent to the land quality of Malawi, $range_{Malawi} = 0.68$. Note that these two countries have nonetheless comparable overall levels of land quality, i.e. $avg_{Lesotho} = 0.66$ and $avg_{Malawi} = 0.56$

Superimposing the languages spoken in Lesotho and Malawi, see maps in Appendix C, the difference is clear. 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 already mentioned 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 how people are distributed across places affects measured fractionalization. For example, should one region have the largest fraction of the total population of the pair of places considered, this implies that even if these two regions have different ethnicities the measured fractionalization will be low compared to a case that these two places are equally densely populated.[31] In this respect it is important to consider that the theory provides a framework in which the probability that individuals from two *different regions* will share common cultural traits, is endogenous to how similar the productive structures of these regions are.

It is straightforward to manipulate (7) to elucidate how population density across places affects fractionalization. The expected fractionalization, $E(ELF)$, for a pair of places in particular reads:

$$E(ELF) = (1 - f_T)\left(1 - \left(\frac{L_T^i}{L_T^j + L_T^i}\right)^2 - \left(\frac{L_T^j}{L_T^j + L_T^j}\right)^2\right) \tag{18}$$

[31] This is not a concern in the cross-region regressions given that the dependent variable is the count of languages spoken rather than a transformation of the count of people speaking these languages.

where $(1 - f_T)$ is the probability that the two regions i and j will have different ethnic traits and $\left(1 - \left(\frac{L_T^i}{L_T^j + L_T^i}\right)^2 - \left(\frac{L_T^j}{L_T^j + L_T^j}\right)^2\right)$ is the probability that two randomly chosen individuals will belong to *different regions.* It is evident from (18) that the more unequally is population distributed across places the lower would be fractionalization, ceteris paribus. In Appendix A the regional population densities are expressed as a function of the regional land qualities and it is shown that in the two-region case, conditional on the probability that two places will have different ethnolinguistic elements, $(1 - f_T)$, a more unequal distribution of land quality decreases fractionalization.

Consequently, the gini coefficient of land quality for each country, denoted by $gini$, is constructed. As expected the gini of land quality is highly correlated (0.59) with how unequally population density is distributed across regions within country in 1990.[32], [33]

Given the preceding discussion the following main specification is adopted:

$$ELF_i = a_0 + a_1 range_i + a_2 avg_i + a_3 avg_i range_i + a_4 gini_i + \eta_i \tag{19}$$

where ELF_i is the level of ethnolinguistic fractionalization in country i, avg_i stands for the average land quality in country i, $range_i$ is the support of the distribution of land quality, and $gini_i$ is the gini coefficient measuring how unequally is land quality distributed across regions in country i. The interaction term, $avg_i range_i$, is intended to capture a diminishing effect of variation in land quality as the average quality increases and η_i is the error term. Given the theory and the preceding remarks the predictions are:

$$a_1 > 0, \quad a_2 = 0, \quad a_3 < 0, \quad a_4 < 0$$

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 are included. Additionally, to ensure that the findings are not driven by including in the regressions regions with very low,

[32] To measure the latter a gini index of population density is constructed by the author 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. The data is available at http://sedac.ciesin.columbia.edu/gpw.

[33] Results not shown also suggest that the gini coefficient of land quality is strongly related (the correlation is 0.55) to 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 $gini$, however, it is significant only in regressions using as dependent variable the ethnic fractionalization index derived by Alesina et. al. (2003).

or even zero, population density the relevant statistics are derived after taking out from each country the 10% of the observations with the lowest population density.[34] This amounts to taking out places with a median population density of 0.12 individuals per square km. Such considerations limit the sample size to on average 147 countries depending on the specification. Descriptive statistics and the raw correlation between the variables of interest are presented in Tables $3a, 3b$.

In Table 4 the regressors of the main specification are added sequentially. The standard errors presented all along are not corrected for heteroskedasticity since using White's (1980) general test the null hypothesis of homoskedasticity may not be rejected. Allowing for robust standard errors the results are the same. In column 1 only the *range* is introduced and the coefficient is positive and statistically significant. In column 2 the average land quality, avg, and its interaction with the support of the distribution of land quality, avg_range, are added and both the sign and the significance of a_1, a_2, a_3 are in accordance to the theoretical predictions.

The results of the main specification (19) are presented in column 3 of Table 4. The inclusion of *gini* as expected enters significantly with the predicted negative sign improving significantly the regression fit and increases the coefficient of *range* as would be expected given the positive correlation between these two. These dimensions of the distribution of land quality identified explain 23% of the variation of contemporary ethnolinguistic fractionalization across countries. As predicted, an increase in the spectrum of land qualities within country increases ethnolinguistic fractionalization significantly. The negative coefficient of the interaction term also implies that the effect of variation in land quality diminishes as average land quality improves. This is consistent with the view that as regions within country become increasingly suitable for agricultural production it becomes easier to transfer region specific technology. This lowers the barriers to mobility between populations residing in different areas leading to lower fractionalization outcomes.

The impact of land heterogeneity, measured by the *range*, is also economically significant. A two standard deviation increase in the spectrum of land quality, evaluated at the mean of land quality, increases fractionalization by 0.23. To better understand this magnitude note

[34] Using alternative thresholds both for the minimum number of observations per country and the regional population density the qualitative results are similar. Furthermore, we have also performed the regression analysis by weighting each region with the relevant population density as of 1990 and the results are largely unchanged. A concern with this approach has to do with the fact that it does not reflect the period during which the fractionalization measures were collected, around 1950. For the same reason using directly the gini coefficient of regional population density as of 1990 in the main regression, although it delivers similar results, is not pursued further.

that the average difference in ethnolinguistic fractionalization between a Sub-Saharan and a non Sub-Saharan country is 0.33. All coefficients for the *range*, *gini* and *avg_range* are fairly precisely estimated and are significant at 1% level. The average land quality, *avg*, is not statistically different from zero which is consistent with the prediction that the average land quality may impact the formation of ethnicities only through its interaction with the qualitative heterogeneity across regions.

To make sure that the results are not subject to omitted variables bias, reverse causality is less of a concern given the nature of the land quality characteristics,[35] in Table 5 different specifications are employed. I explore alternative hypotheses for the emergence of ethnicities, namely, other geographical characteristics and historical contingencies.

Continental Fixed effects

In the first column in Table 5 the main specification is repeated. In the second column continental dummies for Sub-Saharan Africa (*reg_ssa*), Latin America and Caribbean (*reg_lac*) and Western Europe (*reg_we*) are introduced, in order to make sure that the results are not driven by a particular continent. The coefficients of interest generally decrease remain, though, both economically and statistically significant. Note that the marginal overall effect of *range* does not change since both the direct and the interaction effect decrease by roughly the same magnitude. In fact the effect of land quality heterogeneity, *range*, is significantly positive for all countries with $avg \leq 0.69$. For countries larger than this threshold (21 out of 147) the effect of *range* is insignificant. Repeating the analysis after excluding all the countries of Sub-Saharan Africa produces qualitatively similar results.

Other Geographical Characteristics

In the third column of table 5 geographic controls that could potentially affect fractionalization are accounted for. As found in the cross-region regressions the distance from the equator, denoted by *abs_latitude*, has a strong negative effect on ethnolinguistic fractionalization. To the extent that distance from the equator increases seasonality, this is consistent

[35] The derivation of the land quality is partially based on the quality of the soil. This makes land quality possibly endogenous to the rise/duration of agriculture/herding. Controlling for the timing of the rise of agriculture is not significantly related to ethnic diversity and does not change the coefficients of the variables of interest (results available upon request). A priori there is no reason to expect that ethnic diversity per se would systematically impact the soil quality. Nevertheless, if for some reason ethnic diversity was reducing overall soil quality then the current results underestimate the true effect of variation in land quality on ethnic diversity.

with the theory's prediction that places subject to more variable productivity shocks should display lower levels of fractionalization, ceteris paribus. This is a robust finding, in column 5 of Table 3 even after controlling for a host of continental, historical and geographic characteristics, the coefficient on *abs_latitude* remains significant.[36] The pure size of a country, denoted by *areakm2*, perhaps surprisingly enters negatively although insignificant. The mean distance to the nearest coastline or sea-navigable river, denoted by *distcr*, increases fractionalization and this is conforming with the view that places which are increasingly isolated from water passages have been experiencing limited population mixing, given any regional fluctuation in productivity, and thus should on average display higher ethnolinguistic fractionalization. It should be noted 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.

An important geographic characteristic that might affect the formation of languages and ethnicities is the topography of each country. To account for elevation alternative measures are used. The one presented here uses a new index constructed by the author, namely, the standard deviation of elevation within a country, denoted *elev_sd*. This measure is chosen because it captures accurately the variation in topography within a country. The results are similar using average elevation, the % of mountainous land within country or the difference between the lowest and the highest point. The non-significant effect of the standard deviation of elevation on fractionalization in column 3 of table 5, is driven by the fact that although Sub-Saharan Africa, is the most fractionalized continent of the world, has an average standard deviation of 0.28 km whereas for a non Sub-Saharan country the respective number averages 0.48 km. Indeed, controlling for continental fixed effects, see column 5, a more variable topography affects fractionalization positively and significantly.

The inclusion of these additional geographical features reduces the magnitude of the coefficients of interest it does not alter, nevertheless, the qualitative predictions.

Historical Attributes

In column 4 of table 5 controls accounting for the variation in historical contingencies across countries, are added. The log of the population density in 1500, *lpd*1500,[37] enters

[36] Countries also vary in their latitudinal extent. Nevertheless, in all regressions explictly controlling for it came highly insignificant and did not affect the other estimates. Consequently, it was left out of the analysis.

[37] This measure is highly correlated, around 0.56, with the experience of each region with statehood as con-

negatively but not significantly and the year when each country gained independence, $yrentry$, has a significant impact of fractionalization. Specifically, the later is the year of independence the higher is the level of fractionalization. This is consistent with the historical evidence which suggests that modern states since their inception systematically attempted to homogenize the population along ethnolinguistic dimensions. The expansion of public schooling, for example, had exactly such an impact on linguistic diversity.[38]

Column 5 adds to the main specification all the additional controls regarding geographic characteristics, continental dummies and historical traits. The variables of interest remain both economically and statistically significant. These robustness checks underline the fundamental role of the distribution of land quality in shaping ethnolinguistic diversity. At the same time $lpd1500$ enters negatively and significantly. This finding is evidence that indeed contemporary ethnic diversity is endogenous to the developmental history of each country as captured by the population density in 1500.

So far, the empirical analysis includes countries whose ethnic mix is a relatively recent phenomenon. United States, Brazil, Australia, Canada etc. fall into this category. However, according to the theory the formation of ethnicities is an outcome of a long run process and a stage of development when land was the dominant factor of production. In column 6 of Table 5 the sample is restricted into countries whose percentage of indigenous population as of 1500 still comprises at least 75% of the current population mix. Under this specification, the results are even stronger and the distribution of land quality accounts for 27% of the observed ethnolinguistic variation as opposed to 23% in column 1 which included all countries.

4.3 The Effect of Colonialism on Fractionalization

The component of ethnic diversity driven by the distribution of land quality, captured in the main specification (19), is the natural level of fractionalization, nat_ELF, that a region would exhibit if left largely undisturbed. On the contrary, artificial fractionalization, denoted by art_ELF, is the part of the observed fractionalization that is not driven by the characteristics of land quality. According to the theory, in a world with common historical paths the natural component would in principle explain an equal share of the fractionalization outcomes across subsets of countries. However, it is certainly true that countries have experienced distinct

structed by Bockstette et al. (2002). Including both at the same time makes them insignificant. Consequently, i only include in the regressions the log of the population density in 1500.

[38] Of course, the causality may run both directions since more fractionalized regions may lead to a later emergence of modern states either because of being colonized or because of having a slower statehood formation.

historical events.

The previous section showed that the impact of heterogeneous land qualities on fractionalization is robust to alternative controls which accommodate for divergent historical paths, with the latter however having also an independent effect on contemporary ethnic diversity. This section investigates in detail an issue that has received particular attention within economics and this is 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, Brazil to settling at very low levels in other places, as in Congo for example. In several instances, they actively influenced preexisting groups by giving territories to those that were not the initial claimants, ignoring the fact that another group was already in the same territory or favoring some groups politically over others. Generally, the European colonization created an imbalance in the mix of the indigenous populations, directly affecting the preexisting ethnic spectrum.

The discussion above implies that countries colonized by Europeans should exhibit fractionalization outcomes endogenous to their colonial experience, the identity of the colonizers and how intensely the colonizers settled,[39] among other things. Table 6 presents the main specification (19) separately for countries that were colonized by European powers after the 15th century and for those that were not. As expected the R^2 coefficient is larger for the sample of countries that did not experience colonization. Specifically, the distribution of land quality explain 28% of the variation in the ethnolinguistic fractionalization for the non-colonized sample and 19% for the colonized one.[40] This finding is consistent with the view that colonizers extensively manipulated the underlying ethnicities augmenting significantly the artificial component of observed fractionalization outcomes. However, it is not only the man-made component through which colonizers affected the ethnolinguistic mix of the colonized world.

Historical accounts suggest that colonizers except for actively influencing the ethnic endowment of each region also drew borders in an arbitrary way, see Herbst (2002) and Englebert et al. (2002), essentially shaping the geographical spectrum whose ethnicities would compose each country's ethnic mix. The effect of border drawing may be uncovered by looking at the natural level of fractionalization, nat_ELF. This is derived using the predicted values of the

[39] The latter has been argued to depend on the very health environment of the colonized countries, Acemoglou et al. (2001).

[40] Excluding the western European countries from the non-colonized sample the results become even stronger. In particular, the distribution of land quality now explain 39% of the variation in the ethnolinguistic diversity in the non-colonized world.

main specification (19). Since both country borders and the size of ethnic groups are endogenous to the incidence and nature of colonization, to obtain the natural fractionalization of the colonized world, the point estimates used are those from the non-colonized sample in column 2 of table 6. Consequently, the estimate derived is effectively the level of fractionalization that would emerge in the colonized countries should the European colonization be limited to the arbitrary drawing of borders.

Table 7 presents the natural level of fractionalization, nat_ELF, for the colonized and the non-colonized sample. The results establish that the borders drawn by colonizers inflated significantly the natural component of ethnolinguistic diversity. Specifically, the geographically driven component of fractionalization is estimated to be 0.33 for the non-colonized countries and 0.38 for the colonized ones and the difference is significant at 5%.[41] It is possible that colonization itself could have been induced in the first place by the relatively high ethnic diversity of the regions, the borders themselves that is the distribution of land quality, nevertheless, were an outcome of the colonial intervention.

Given the decomposition of observed fractionalization into natural and man-made elements it is of interest to investigate whether any of these components is related to artificial statehood as defined by Alesina et. al. (2006). The authors propose two alternative indexes as to how artificial are current countries, one measures how straight the borders are and the other the percent of the population of a country that belongs to a group appearing in two or more adjacent countries. Although, the data are not publicly available yet, the authors cite the 13 most artificial countries according to both measures. To investigate whether our measure of artificial fractionalization increases the probability of belonging to the most artificial countries according to Alesina, a dummy that equals 1 if the country is one of these 13 countries is regressed on both the natural, nat_ELF, and the artificial component of fractionalization, art_ELF. Both the natural and the man-made components are the predicted values and the residuals respectively of main specification (19) estimated using only the non-colonized sample.

Table 8 presents the results.[42] The effect of artificial ethnolinguistic fractionalization is both statistically and economically significant. A one standard deviation increase in art_ELF, increases by 6% the probability of being one of the most artificial states. The negative insignificant coefficient on nat_ELF, implies that the natural component of fractionalization is not

[41] Including in the derivation of the natural component of fractionalization the variation in topography, i.e. the standard deviation of elevation, the difference in natural fractionalization between the colonized and the non-colonized world is similar.

[42] The coefficients reported are the marginal effects of a probit regression. Note that the residuals proxy for artificial fractionalization up to a constant.

correlated with the measure of artificial statehood of Alesina et al. Naturally, until the entire dataset for the state artificiality is released the results should be considered only as tentative.

Summarizing the impact of the European colonizers on the ethnolinguistic diversity the evidence suggests that they substantially altered the ethnolinguistic endowment of the places they colonized. Decomposing the existing fractionalization into a part driven by the distribution of land quality and another one which is unrelated to the underlying land endowment, i.e. man-made, the results suggest that colonizers increased both dimensions significantly. Namely, the European intervention imposed country borders that brought together regions whose land characteristics could in principle sustain a wider ethnic spectrum. This was an outcome of the intrinsic qualitative diversity of the land enclosed.

At the same time, their active manipulation of the original ethnolinguistic endowment, including the introduction of their own ethnicities, substantially altered the man-made component of the observed fractionalization tipping the balance in favor of an ethnic spectrum whose identity and size was not a natural consequence of the primitive land characteristics. These results suggest that contemporary fractionalization is endogenous to both the colonial experience and the historical levels of development captured by the population density in 1500. To the extent that state history has been shown to affect contemporary economic outcomes independently, the documented relationship between ethnolinguistic diversity and economic outcomes should be cautiously interpreted.

5 Concluding Remarks

This research provides and tests a theory on the emergence of cultural diversity. The study argues that the heterogeneity in the regional productive characteristics shaped the intensity of population mixing across places. The transferability of region specific human capital and the incentive to relocate, generated by the occurrence of regional productivity shocks, form the basis of the theory.

Regions similar along their productive characteristics would display persistent population mixing resulting in the formation of a common ethnolinguistic behavior. On the contrary, among places characterized by dissimilar productive endowments, population mobility would be limited leading to the formation of local ethnicities and languages giving rise to a wider cultural spectrum.

The theory's prediction about diversity in land quality and ethnic diversity has a striking parallel to the literature on biodiversity and variation within species. Darwin's observations

on finches (1839) is of particular relevance. He observed that a certain ecological niche was giving rise to a specific optimal shape of the finches' beaks. So, ecologically diverse places would bring about and sustain variation within finches. Along the same lines this study argues that a certain level of land quality generated specific human capital. Consequently, variation in land qualities across regions contributed significantly to the emergence and persistence of ethnic diversity.

Using new detailed data on the distribution of land quality within and across countries to proxy for the differences in region specific human capital I find that a larger spectrum of land qualities increases ethnic diversity. Both cross–region and a cross-country regressions are examined. The cross-region framework is of particular significance since the proposed relationship between the variation in land quality and ethnic diversity obtains at an arbitrary level of aggregation and after controlling for continental and country fixed effects.

These results are robust to alternative specifications. In particular, controlling for additional geographic characteristics and accounting for the divergent historical paths across countries, the main predictions of the theory remain largely unchanged.

The empirical results also uncover the impact of state history on contemporary ethnic diversity. In particular, exploring the role of European colonizers in shaping ethnolinguistic diversity within the colonized world, interesting regularities are revealed. The arbitrary border drawing becomes evident by looking at the level of natural fractionalization. This is systematically higher in the colonized world. The interfering of the colonizers, however, with the local ethnicities was also widespread. The augmented man-made component of ethnic diversity across the colonized countries attests to it. Preliminary results suggest that state artificiality as measured by Alesina et. al. (2006) correlates strongly with the derived measure of man-made fractionalization.

This research sheds new light on the emergence and the distribution of languages and ethnicities within countries and constitutes a first step towards conceptualizing the natural and man-made components of ethnic diversity.

The findings provide a stepping stone for further research. Equipped with a more substantive understanding of the economic origins and elements of cultural diversity, long standing questions within development and growth literature may be readdressed. Issues like the formation of states, the diffusion of development, the inequality across ethnic groups, and the causal effect of ethnolinguistic diversity on economic outcomes in general, analyzed through the proposed framework may offer new significant insights.

6 Appendix

Appendix A - Proofs

Using Lemma 2 is straightforward to establish the properties of the size of the migration between places, captured by (10) and (11).

Lemma 4 *Conditional on positive migration in period t, that is if either (8) or (9) obtain in the beginning of period t, the size of the population that migrates is*

1. *increasing (decreasing) in the relative regional productivity shock, λ_t, in case of migration from j to i (i to j)*
$$\frac{\partial M_t^{j\to i}}{\partial \lambda_t} > 0 \quad \& \quad \frac{\partial M_t^{i\to j}}{\partial \lambda_t} < 0$$
2. *decreasing in the size of the erosion, ε*
$$\frac{\partial M_t^{j\to i}}{\partial \varepsilon}, \frac{\partial M_t^{i\to j}}{\partial \varepsilon} < 0$$
3. *decreasing in the region specific technology of the place of origin, h_t^i, h_t^j*
$$\frac{\partial M_t^{j\to i}}{\partial h_t^j}, \frac{\partial M_t^{i\to j}}{\partial h_t^i} < 0$$

Proof of Lemma 4.

Let $M_t^{j\to i} > 0$ then there two distinct cases regarding the realization of the last migration movement in period s.

Case 1: $M_s^{j\to i} > 0$.

Substituting (13) into (11) and simplifying:

$$M_t^{j\to i} = \frac{L_{t-1}^j[1-(\frac{\lambda_s}{\lambda_t})^{\frac{1}{1-\alpha}}(\frac{h_t^j}{h_s^j})^{\frac{\varepsilon}{1-\alpha}}]}{\frac{m^j}{m^i}(\lambda_t)^{-\frac{1}{1-\alpha}}(h_t^j)^{\frac{\varepsilon}{1-\alpha}}+1} \tag{20}$$

Case 2: $M_s^{i\to j} > 0$.

Substituting (12) into (11) and simplifying:

$$M_t^{j\to i} = \frac{L_{t-1}^j[1-(\frac{\lambda_s}{\lambda_t})^{\frac{1}{1-\alpha}}(h_s^i h_t^j)^{\frac{\varepsilon}{1-\alpha}}]}{\frac{m^j}{m^i}(\lambda_t)^{-\frac{1}{1-\alpha}}(h_t^j)^{\frac{\varepsilon}{1-\alpha}}+1} \tag{21}$$

Direct differentiation of (20) and (21) with respect to the variables of interest, produces the results. A similar derivation applies to the case of migration from $i \to j$, $M_t^{i \to j}$. □

Proof of Lemma 3.

The steps are similar to those performed in the proof of Lemma 4. First, substitute in (15) the two possible realizations of the past population densities, either (12) or (13), and differentiate with respect to the variables of interest. Repeat the same process for (16). This completes the proof. □

The following Lemma summarizes the cases of migration occurrences.

Lemma 5 *In any period t there are the following cases as to the occurrence or not of migration.*

1. *If 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 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 (12) or (13) depending on the direction of the last migration, in both (10) and (11) and solving for the required inequalities completes the proof. □

Proof of Proposition 1.

Under Assumption (A1) the ratio λ_t/λ_s may take three unique values either $\lambda_{\min}/\lambda_{\max}$ or $\lambda_{\max}/\lambda_{\min}$ 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 5) 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 if 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 5 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 (and also smaller in magnitude, see Lemma 4) 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 left hand-side increases as erosion increases, precipitating the end of migratory movements towards region i.

Conditional on (A1) 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 5 that the larger is $\lambda_{\max}/\lambda_{\min}$ (equivalent the smaller is $\lambda_{\min}/\lambda_{\max}$) the more probable will be migration. 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, (18), in terms of regional land qualities:

Manipulating (18) may be rewritten as:

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

Noting (14) for example,[43] the ratio of regional population densities is substituted accordingly and $E(ELF)$ may be rewritten as:

$$E(ELF) = (1 - f_T)\left(\frac{m^i}{2m^j} + \frac{m^j}{2m^i} + 1\right)^{-1} \tag{22}$$

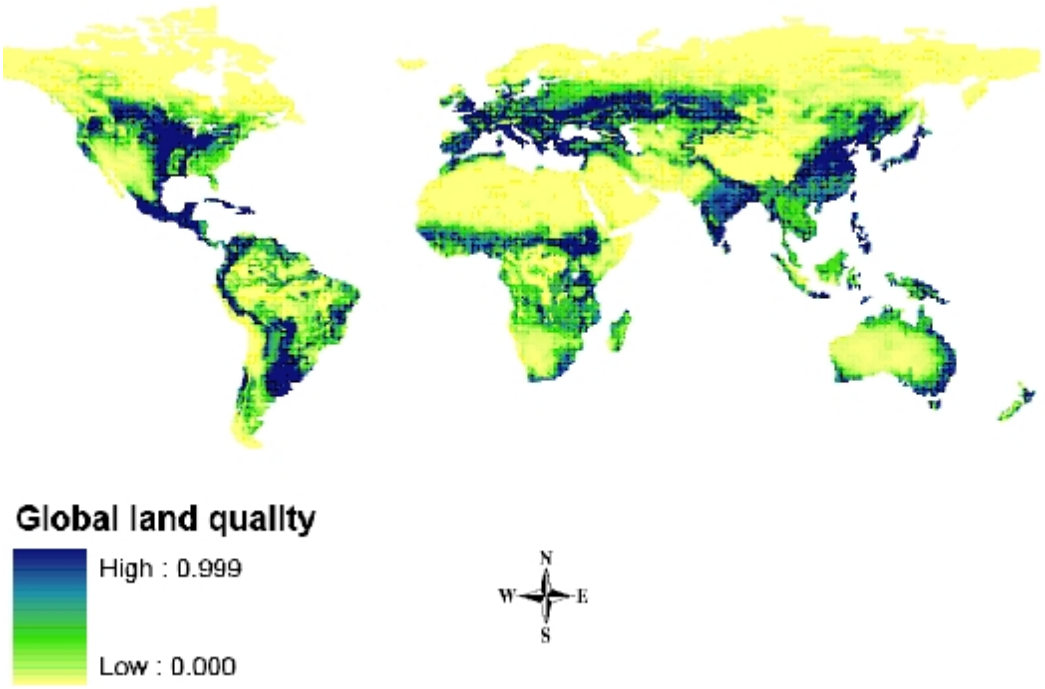
It is easy to show that conditional on the probability that two places will not share the same cultural traits, $(1 - f_T)$, a more unequal distribution of the quality of land will decrease measured fractionalization. For example, let $m^i > m^j$ then an increase in m^i and/or a decrease

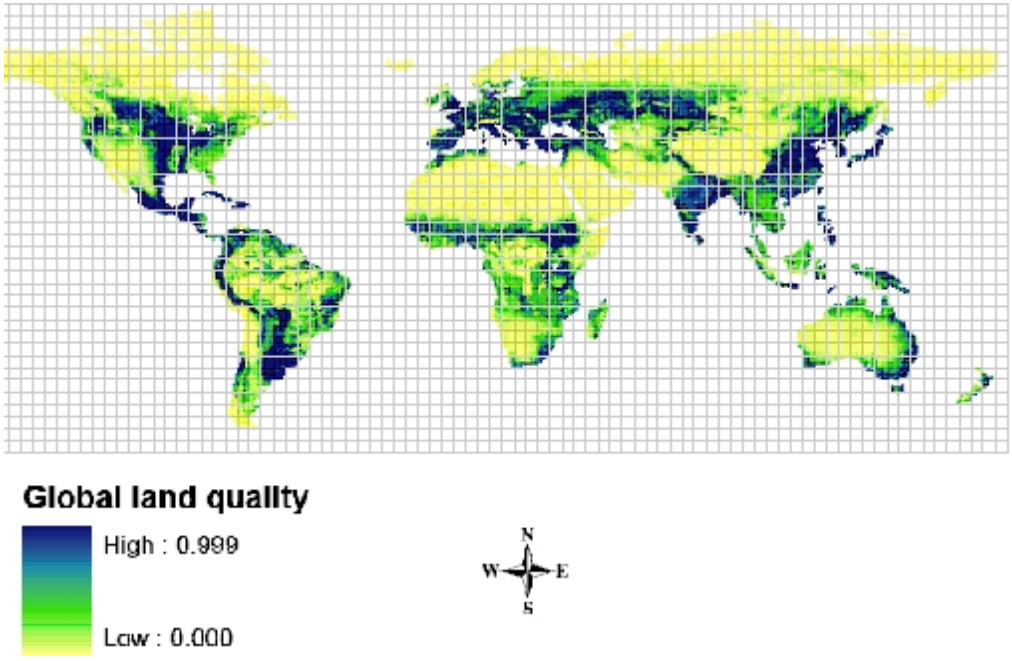
[43] Using either (12) or (13) the analysis remains qualitatively similar, however $E(elf)$ will be a more complicated function of regional land qualities.

in m^j will decrease $E(ELF)$. This obtains by differentiating (22) 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, as (22) shows, affects negatively fractionalization outcomes.

Appendix B - Maps





In the lower map the grids represent the "artificial countries" generated for the cross-region analysis. Each grid is 4 degrees latitude by 4 degrees longitude.

Table 1a: Summary Statistics for the Cross-Region Analysis

statistics	**number_lang**	**range**	**avg**	**avg_range**	**abs_lat**	**elev_sd***	**areakm2**
mean	7.45	0.45	0.27	0.18	35.24	254.32	107.86
sd	8.89	0.33	0.27	0.19	19.11	257.70	35.13
max	59.00	1.00	0.98	0.88	74.88	1733.14	151.24
min	1.00	0.00	0.00	0.00	0.38	0.83	6.71

**in meters*

number_lang : number of unique languages spoken within each "artificial country"; r*ange* : spectrum of land qualities within an "artificial country"; i.e. the difference in land quality between the region with the highest land quality from that with the lowest; *avg* : is the average land quality within "artificial country"; avg_range: the interaction between range and avg; elev_sd: standard deviation of elevation within "artificial country"; abs_lat: "artificial country's" latitudinal distance from the equator; areakm2: size of each "artificial country" in sq. km.
Data Sources: See Appendix D

Table 1b: The Correlation Matrix for the Cross-Region Analysis

	number_lang	**range**	**avg**	**avg_range**	**abs_lat**	**elev_sd**	**areakm2**
number_lang	1						
range	0.27	1					
avg	0.12	0.63	1				
avg_range	0.15	0.8	0.89	1			
abs_lat	-0.48	-0.3	-0.18	-0.16	1		
elev_sd	0.14	0.17	0.01	0.07	-0.13	1	
areakm2	0.43	0.29	0.06	0.13	-0.65	0.12	1

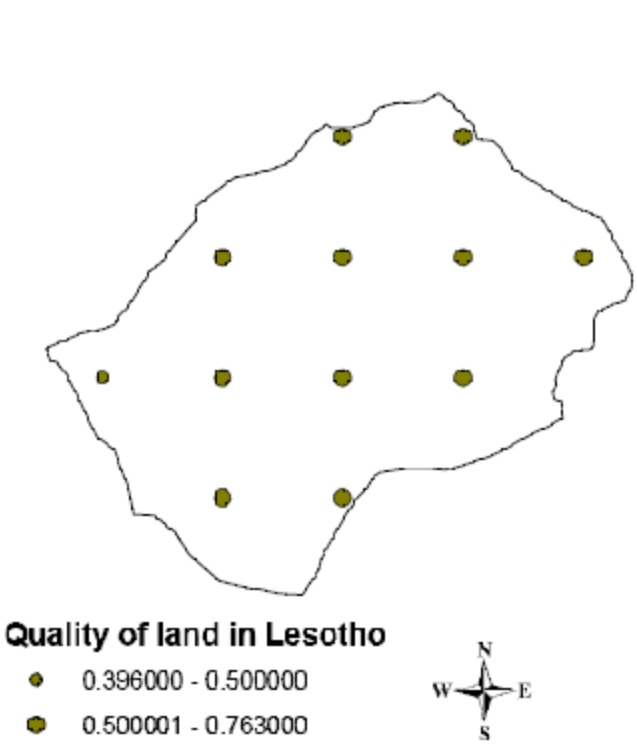
number_lang : number of unique languages spoken within each "artificial country"; r*ange* : spectrum of land qualities within an "artificial country"; i.e. the difference in land quality between the region with the highest land quality from that with the lowest; *avg* : is the average land quality within "artificial country"; avg_range: the interaction between range and avg; elev_sd: standard deviation of elevation within "artificial country"; abs_lat: "artificial country's" latitudinal distance from the equator; areakm2: size of each "artificial country" in sq. km.
Data Sources: See Appendix D

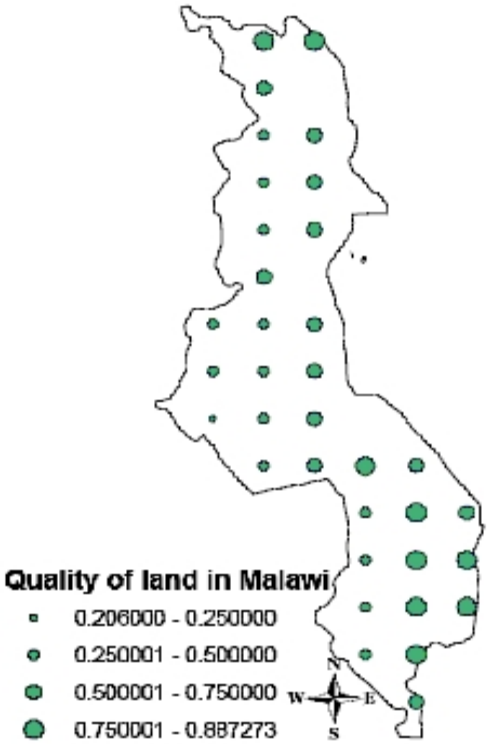
Table 2: Main Specification and Robustness Checks in Cross-Region Regressions

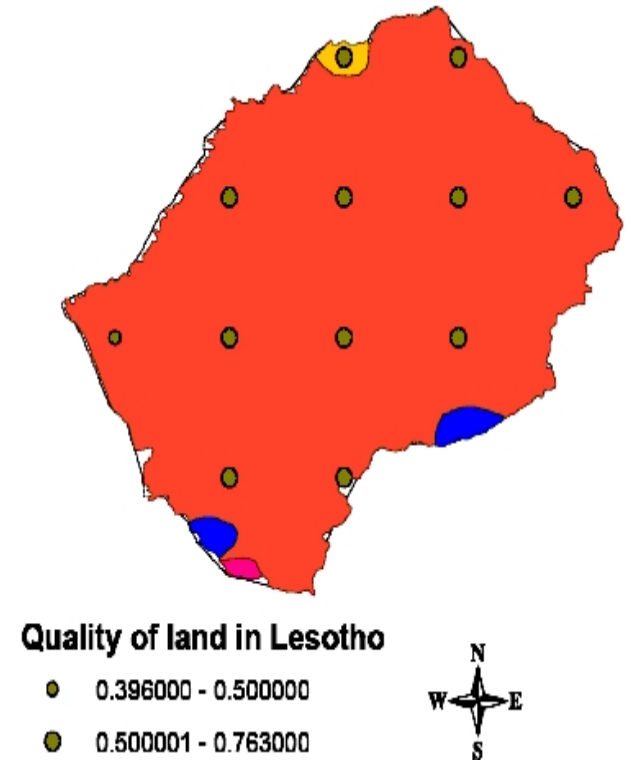
	Dependent Variable: Number of Languages Spoken					
	OLS	OLS	OLS	OLS	OLS	OLS
	Baseline	Elevation and Distance from the Equator	Continental fixed effects	Continental and Country fixed effects	Non-Tropics	Tropics
	(1)	(2)	(3)	(4)	(5)	(6)
range	**7.188** (9.28)***	**4.067** (3.04)***	**4.109** (4.32)***	**4.743** (4.22)***	**1.663** (2.37)**	**6.892** (2.10)**
avg		0.318 (0.17)	-1.011 (0.84)	-0.881 (0.65)	0.12 (0.15)	5.88 (1.39)
avg_range		-2.755 (0.76)				
areakm2		0.046 (6.26)***	0.04 (5.48)***	0.032 (3.35)***	0.01 (1.88)*	0.103 (5.83)***
abs_lat		-0.146 (8.84)***	-0.122 (6.75)***	-0.216 (7.41)***	-0.039 (1.75)*	-0.559 (4.94)***
elev_sd		0.002 (1.67)*	0.002 (1.80)*	0.003 (2.23)**	0.006 (3.94)***	-0.002 (0.84)
R-squared	0.07	0.27	0.31	0.45	0.36	0.49
Observations	887	887	887	887	601	269

OLS regressions with absolute value of robust t statistics in parentheses.
* significant at 10%; ** significant at 5%; *** significant at 1%
Regressions (4), (5) and (6) include both country and continental fixed effects.
range: spectrum of land qualities within an "artificial country"; i.e. the difference in land quality between the region with the highest land quality from that with the lowest; *avg:* is the average land quality within "artificial country"; *avg_range* : the interaction between range and avg; *elev_sd* : standard deviation of elevation within "artificial country"; *abs_lat* : "artificial country's" latitudinal distance from the equator; *areakm2* : size of each "artificial country" in square kilometers;
Data Sources: See Appendix D

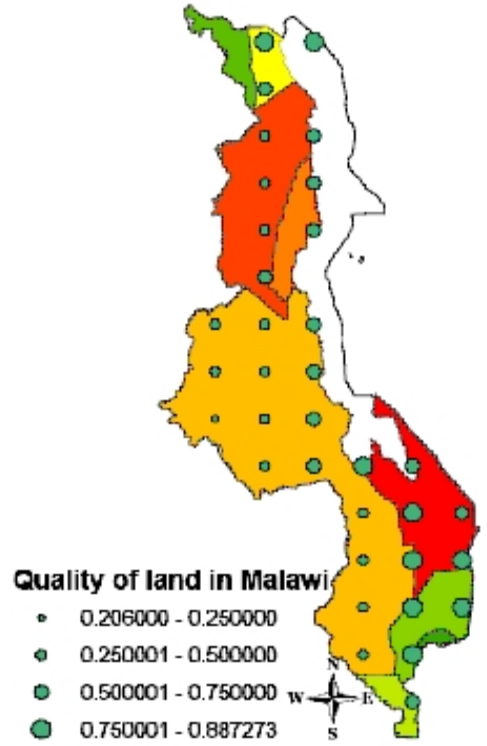
Appendix C - Maps







Upper map land quality; lower map languages and land quality



Upper map land quality; lower map languages and land quality

Table 3a: Summary statistics for Cross-Country Analysis

statistics	**ELF**	**range**	**avg**	**avg_range**	**gini**	**lpd1500**	**elev_sd***	**yrentry**
mean	0.410	0.697	0.395	0.282	0.364	0.906	0.409	1927.120
sd	0.281	0.265	0.249	0.182	0.225	1.504	0.348	56.920
max	0.925	0.990	0.958	0.787	0.859	3.842	1.867	1993.000
min	0.001	0.002	0.003	0.000	0.028	-3.817	0.019	1816.000

**in kilometers*

ELF : ethnolinguistic fractionalization; r*ange* : spectrum of land qualities within the unit of analysis, country, i.e. the difference in land quality between the region with the highest land quality from that with the lowest; *avg* : is the average land quality within the unit of analysis, country; *avg_range* : the interaction between range and avg; *gini* : the gini of coefficient of land quality within country; *lpd1500* : log of the population density in 1500; *elev_sd* : standard deviation of elevation within the unit of analysis, country; *yrentry* : year when modern state obtained independence.

Data Sources: See Appendix D

Table 3b: Correlation Matrix for Cross-Country Analysis

	ELF	**range**	**avg**	**avg_range**	**gini**	**lpd1500**	**elev_sd**	**yrentry**
ELF	1.00							
range	0.20	1.00						
avg	-0.21	0.15	1.00					
avg_range	-0.12	0.59	0.78	1.00				
gini	0.12	0.23	-0.78	-0.52	1.00			
lpd1500	-0.17	0.15	0.39	0.46	-0.34	1.00		
elev_sd	0.10	0.33	-0.02	0.15	0.23	0.01	1.00	
yrentry	0.36	-0.31	-0.20	-0.30	-0.06	-0.09	-0.21	1.00

ELF : ethnolinguistic fractionalization; r*ange* : spectrum of land qualities within the unit of analysis, country, i.e. the difference in land quality between the region with the highest land quality from that with the lowest; *avg* : is the average land quality within the unit of analysis, country; *avg_range* : the interaction between range and avg; *gini* : the gini of coefficient of land quality within country; *lpd1500* : log of the population density in 1500; *elev_sd* : standard deviation of elevation within the unit of analysis, country; *yrentry* : year when modern state obtained independence.

Data Sources: See Appendix D

Table 4: Main Specification for Cross-Country Analysis

	Dependent Variable: ethnolinguistic fractionalization (ELF)		
	(1)	(2)	(3)
range	0.203	0.455	0.889
	(2.43)**	(3.55)***	(5.73)***
avg		0.076	-0.283
		(0.42)	(1.49)
avg_range		-0.659	-1.21
		(2.18)**	(3.90)***
gini			-0.874
			(4.45)***
R-squared	0.04	0.13	0.23
Observatio	147	147	147

OLS regression with absolute value of t statistics in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

r*ange* : spectrum of land qualities within the unit of analysis, country, i.e. the difference in land quality between the region with the highest land quality from that with the lowest, avg: is the average land quality within country, avg_range: the interaction between range and avg, gini: the gini of coefficient of land quality within country.

Data Sources: See Appendix D

Table 5: Main Specification and Robustness Checks

	Dependent Variable: ethnolinguistic fractionalization (ELF)					
	Baseline	Continental fixed effects	Additional geography	Historical controls	Full specification	Indigenous >75%
	(1)	(2)	(3)	(4)	(5)	(6)
range	0.925	0.641	0.692	0.889	0.668	1.032
	(5.92)***	(4.62)***	(4.71)***	(5.92)***	(4.87)***	(5.65)***
avg	-0.198	-0.101	-0.261	-0.07	0.092	-0.241
	(0.79)	(0.58)	(1.54)	(0.37)	(0.52)	(0.85)
avg_range	-1.359	-0.653	-0.834	-1.043	-0.765	-1.422
	(3.85)***	(2.28)**	(2.83)***	(3.39)***	(2.75)***	(3.73)***
gini	-0.91	-0.416	-0.742	-0.626	-0.508	-0.921
	(4.30)***	(2.30)**	(3.91)***	(3.27)***	(2.81)***	(3.78)***
reg_ssa		0.255			0.149	
		(5.73)***			(2.48)**	
reg_lac		-0.102			-0.18	
		(1.79)*			(2.18)**	
reg_we		-0.172			0.075	
		(2.70)***			-0.87	
lpd1500				-0.02	-0.034	
				(1.36)	(2.03)**	
yrentry				0.002	0.001	
				(4.73)***	(1.86)*	
abs_latitude			-0.005		-0.004	
			(4.27)***		(2.23)**	
areakm2			-0.002		-0.001	
			(1.61)		(0.61)	
distcr			0.202		0.105	
			(4.14)***		(2.18)**	
elev_sd			0.037		0.141	
			(0.65)		(2.37)**	
R-squared	**0.23**	0.44	0.42	0.35	0.54	**0.27**
Observations	147	147	146	143	143	101

OLS regressions with absolute value of t statistics in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

r*ange* : spectrum of land qualities within country, i.e. the difference in land quality between the region with the highest land quality from that with the lowest, avg: is the average land quality within country, *avg_range* : the interaction between range and avg, *gini* : the gini of coefficient of land quality within country *reg_ssa:* dummy for Sub-Saharan countries, *reg_lac* : dummy for Latin-American and Caribbean countries *reg_we:* dummy for Western European countries*, abs_lat:* country's latitudinal distance from the equator *areakm2* : size of each country in square kilometers; *distcr* : distance from centroid of country to nearest coast or sea-navigable river (km); *elev_sd* : standard deviation of elevation within country; *lpd1500:* log of the population density in 1500*; yrentry:* year when modern state obtained independence.

indigenous: percentage of indigenous population as of 1500 comprising more that 75% of the current population

Data Sources: See Appendix D

Table 6: Colonization and Fractionalization

Dependent Variable: ethnolinguistic fractionalization (ELF)

	Colonized by Europeans	**Non-Colonized** by Europeans
	(1)	(2)
range	0.792	0.914
	(3.80)***	(3.31)***
avg	-0.324	-0.291
	(1.37)	(0.79)
avg_range	-0.892	-1.171
	(1.83)*	(2.41)**
gini	-0.813	-0.818
	(3.14)***	(2.63)**
R-squared	**0.19**	**0.28**
Observations	93	54

OLS regressions with absolute value of t statistics in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%
r*ange* : spectrum of land qualities within the unit of analysis, country, i.e. the difference in land quality between the region with the highest land quality from that with the lowest; avg: is the average land quality within country; avg_range: the interaction between range and avg; gini: the gini of coefficient of land quality within country; Data Sources: See Appendix D

Table 7: Colonization and Natural Fractionalization

nat_ELF if colonized: **0.38** nat_ELF if not colonized: **0.33**

Pr(T < t) = **0.03**

nat_ELF : natural level of fractionalization computed using the predicted values of regression (2) in Table 6; *Colonized* : colonized by Europeans after 1500
non-colonized : not colonized by Europeans after 1500

Table 8: Artificial Fractionalization and Artificial States

Dependent Variable: top 13 artificial states according to Alesina et al.

art_ELF	0.248
	(2.90)***
nat_ELF	-0.04
	(0.03)
Observations	147

Probit regression with absolute value of z statistics in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%
nat_ELF : natural level of fractionalization computed using the predicted values of regression (2) in Table 6; art_*ELF: artificial level of fractionalization computed* using the residuals of regression (2) in Table 6;

Appendix D - Data Sources

Geographical Variables

elev_sd: standard deviation of elevation for actual countries and "artificial countries".

Source: Constructed by the author using detailed grid level data on the average elevation above sea level for each country. Available at the Atlas of Biosphere: http://www.sage.wisc.edu:16080/atlas/

areakm2: land area (km2)

Source: Center for International Development, CID.[44] For the cross-region analysis the area is calculated by the author using the Haversine Formula, see http://www.movable-type.co.uk/scripts/GIS-FAQ-5.1.html

distcr: distance from centroid of country to nearest coast or sea-navigable river (km)

Source: Center for International Development, CID.

abs_latitude: Absolute Latitudinal Distance from the Equator.

Source: The World Bank. Available from Development Research Institute, NYU. For the cross-region analysis the distance from the equator is calculated by the author using the centroid of each constructed regional unit.

Historical Variables

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 indigenous population as of 1500 still comprising more that 75% of the current population mix.

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

colonized: is a dummy equals 1 if a country was colonized by a European power 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.

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

References

[1] Acemoglu, D., Johnson, S., and Robinson, J., "The Colonial Origins of Comparative Development: An Empirical Investigation," American Economic Review, December, XCI (2001a), 1369-1401.

[2] Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S. and R. Wacziarg (2003) "Fractionalization". Journal of Economic Growth, 8, 155-194.

[3] Alesina, A., Easterly, W., Matuszeski, M., (2006) "Artificial States", Working Paper 12328.

[4] Alesina, A., and Spolaore, E., (1997) "On the Number and Size of Nations." Quarterly Journal of Economics. v112, 1027-1056.

[5] Ashraf, Q., and Michalopoulos S., (2006), "The Climatic Origins of the Neolithic Revolution: A Theory of Long-Run Development via Climate-Induced Technological Progress", SSRN: http://ssrn.com/abstract=903847

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

[7] Barth, F., (1969). "Ethnic groups and boundaries: The social organization of cultural difference." Boston: Little, Brown.

[8] Bellwood, P., "Early Agriculturalist Population Diasporas? Farming, Languages, and Genes", Annual Review of Anthropology, Vol. 30. (2001), pp. 181-207.

[9] Bockstette, V., Chanda, A., Putterman, L.(2002) "States and Markets:the Advantage of an Early Start", Journal of Economic Growth Volume 7, Number 4, 347-369.

[10] Boyd, R., and P.J. Richardson., (1985), "*Culture and the Evolutionary Process*" (University of Chicago Press, Chicago).

[11] Boserup, E., (1965). "*The Conditions of Agricultural Progress*", (Aldine Publishing Company, Chicago).

[12] Botticini, M., Eckstein, Z., (2005) "From Farmers to Merchants, Voluntary Conversions and Diaspora: A Human Capital Interpretation of Jewish History", Journal of Economic History 65, no. 4, 922-48.

[13] Center for International Earth Science Information Network (CIESIN), Columbia University; and Centro Internacional de Agricultura Tropical (CIAT). 2005. Gridded Population of the World Version 3 (GPWv3) Palisades, NY: Socioeconomic Data and Applications Center (SEDAC), Columbia.

[14] Curtin, P., (1984) "Cross-Cultural Trade in World History", Cambridge: Cambridge University Press.

[15] Darwin, C. (1839) "The Voyage of the Beagle", Available at: http://www.online-literature.com/darwin/voyage_beagle/

[16] Easterly, W., and Levine, R., (1997) "Africa's growth tragedy: Policies and Ethnic divisions", Quarterly Journal of Economics, 112(4):1203-50.

[17] Englebert, P., Tarango, S., and Carter, M., (2002) "Dismemberment and Suffocation: A Contribution to the Debate on African Boundaries." Comparative Political Studies. v35:10, 1093-1118.

[18] Esteban, J., Ray., D., (2007) "On the salience of ethnic conflict", Working Paper.

[19] Fearon, J., (2003) "Ethnic Structure and Cultural Diversity by Country", Journal of Economic Growth, 8(2). 195-222.

[20] Fearon, J., Laitin, D. (2002) "Ethnicity, Insurgency and Civil War", American Political Science Review.

[21] Galor, O. and Weil, D.N., (2000), "Population, Technology and Growth: From the Malthusian Regime to the Demographic Transition." *American Economic Review* 110, 806-828.

[22] Geertz, C., (1967) "Old societies and new states: The quest for modernity in Asia and Africa." New York: Free Press.

[23] Gray, R. Atkinson, Q. (2003), "Language-tree divergence times support the Anatolian theory of Indo-European origin", Nature, 426, 435-439

[24] Hale, H., (2004) "Explaining Ethnicity", Comparative Political Studies 2004; 37- 458.

[25] Herbst, J., (2002) "State and Power in Africa". Princeton, NJ: Princeton University Press.

[26] La Porta, R., Lopez de Silanes, F., Shleifer, A., Vishny, R., (1999) "The Quality of Government" Journal of Law Economics and Organization, 315-388.

[27] Michalopoulos, S., (2007b),"The Origins of Ethnolinguistic Diversity: Natural and Artificial Components in the Process of Development", Brown University.

[28] Montalvo, and Reynal-Querol, (2005), "Ethnic polarization, potential conflict and civil war, American Economic Review", 2005.

[29] Nichols, J.,(1997) "Modeling Ancient Population Structures and Movement in Linguistics". Annual Review of Anthropology, Vol. 26., 359-384.

[30] Nichols, J., (1997b). "Chechen phonology", In Phonologies of Asia and Africa, ed. AS Kaye, P Daniels, 941-71. Bloomington, Ind:Eisenbrauns.

[31] Ramankutty, N., J.A. Foley , J. Norman, and K. McSweeney, "The global distribution of cultivable lands:current patterns and sensitivity to possible climate change", Global Ecology & Biogeography (2002) 11, 377–392.

[32] Renfrew, C., (1992) "Archaeology, Genetics and Linguistic Diversity", Man, New Series, Vol. 27, No. 3, 445-478.

[33] Renfrew, C., (2000) "At the edge of knowability: Towards a Prehistory of Languages", Cambridge Archaelogical Journal, Vol. 10, No. 1, 7-34.

[34] Spolaore, E., and Wacziarg R., (2006), "The Diffusion of Development," NBER Working Paper #12153.

[35] Williamson J., (2006), "Poverty Traps Distance and Diversity: The Migration Connection", NBER Working Paper #12549.