

Ethnolinguistic Diversity: Origins and Implications

Stelios Michalopoulos*
Brown University

January 20, 2008

Abstract

This research examines theoretically and empirically the economic origins of ethnolinguistic fractionalization. The empirical analysis constructs detailed data on the distribution of land quality across regions and countries, and shows that variation in land quality has contributed significantly to the emergence and persistence of ethnic diversity. The evidence supports the theoretical analysis, according to which heterogeneous land endowments generated region specific human capital, limiting population mobility and leading to the formation of localized ethnicities and languages. The research contributes to the understanding of the emergence of ethnicities and their spatial distribution and offers a distinction between the natural, geographically driven, versus the artificial, man-made, components of contemporary ethnic diversity. An application of the proposed approach casts some doubt on the influential findings that ethnic diversity has significant adverse effects on economic outcomes. Instrumenting ethnic diversity using measures of variation in land quality suggests that its effect on contemporary development is not significantly different from zero.

Keywords: Ethnic Diversity, Geography, Technological Progress, Population Mixing, Colonization, Economic Growth

JEL classification Numbers: O11, O12, O15, O33, O40, J20, J24.

*I am indebted to Oded Galor for his constant advice and mentorship. Daron Acemoglu, Roland Benabou, Andrew Foster, Ioanna Grypari, Peter Howitt, Nippe Lagerlof, Ashley Lester, Ross Levine, Glenn Loury, Ignacio Palacios-Huerta, Yona Rubinstein and David Weil provided valuable comments. I would like, also, to thank the participants at the NEUDC Conference at Harvard University, October 2007, the Latin American Econometric Society Meetings 2007 in Bogotá, and the NBER Summer Institute 2007 on Income Inequality and Growth, as well as the seminar participants at Brown University, Chicago GSB, University of Copenhagen, University of Gothenburg, Princeton University, and Warwick University for the useful discussions. Lynn Carlsson's ArcGis expertise proved of invaluable assistance. Financial support from the Watson Institute's research project "Income Distribution across and within Countries" is gratefully acknowledged.

1 Introduction

Ethnicity has been widely viewed in the realm of social sciences as instrumental for the understanding of socioeconomic processes. A rich literature in the fields of economics, political science, psychology, sociology, anthropology and history attests to this¹. Nevertheless, the economic origins of ethnic diversity have not been identified, limiting our understanding of the phenomenon and its implications for comparative economic development.

This research examines the economic origins of ethnic diversity. It establishes empirically that variation in land quality has contributed significantly to the emergence and persistence of ethnic diversity. The evidence supports the proposed theory, according to which heterogeneous land endowments generated region specific human capital, limiting population mobility and leading to the formation of localized ethnicities. In contrast to the influential finding about the adverse effect of ethnolinguistic fractionalization on economic development, the analysis demonstrates that ethnic diversity has no effect on comparative development.

The empirical investigation, conducted at various levels of aggregation, establishes that variation in regional land quality is a fundamental determinant of ethnic diversity. In particular, the analysis shows that contemporary ethnic diversity displays a natural component and a man-made one. The natural component is driven by the diversity in land quality across regions, whereas the man-made part captures the idiosyncratic state histories of existing countries, reflecting primarily their colonial experience and the timing of modern statehood.

The proposed distinction between the natural versus the man-made components of contemporary ethnic diversity raises the question whether the well documented negative relationship between ethnolinguistic fractionalization and countries' economic performance (Easterly and Levine (1997), Fearon and Latin (2003), Alesina et. al. (2003), Banerjee and Somanathan (2006), among others) reflects the direct effect of divergent state histories across countries, rather than a true effect of ethnic diversity on economic outcomes. Preliminary results challenge the influential finding that ethnic diversity has significant adverse effects on economic outcomes. Specifically, instrumenting ethnic diversity using measures of variation in land quality, suggests that its effect on contemporary development is not significantly different from zero.

The identification of the geographical origins of ethnic formation generates a wide range of applications. For example, the basic results may be used to explain the pattern of technology diffusion within and across countries as well as across ethnic groups. Technology would diffuse

¹See Hale (2004).

more quickly over places characterized by homogeneous land endowments, whereas in relatively heterogeneous ones, and according to the evidence more ethnically diverse, the diffusion would be less rapid leading to the emergence of inequality across ethnic groups.

This research argues that ethnicities and languages were formed in a stage of development when land was the single most important factor of production. Particularly, the theory suggests that heterogeneous land endowments across regions gave rise to region specific human capital, diminishing population mobility and leading to the formation of localized ethnicities. On the other hand, homogeneous land endowments facilitated population mixing, resulting eventually in the formation of a common ethnolinguistic identity.

The link between variation in land quality and ethnic diversity has a striking parallel to the relationship between biodiversity and variation within species. Darwin's observations that ecologically diverse places would bring about and sustain variation within finches is of particular relevance.² Along the same lines, this study argues that variation in land qualities across regions is the ultimate cause of the emergence and persistence of ethnic diversity.

The model uses a two-region overlapping generations framework. Human capital specific to each area accumulates over time through learning by doing, and is available to the region's 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 the human capital of those who relocate may not be perfectly applicable to the production process of the receiving place. According to the theory it is the interaction of these two elements, the ease of transferring region specific human capital and the incentive to change locality, induced by variation in the regional productivity shocks, that gave rise to regional variation in population mobility and ultimately to distinct ethnolinguistic traits.³

In the empirical section I employ new detailed information on land's agricultural suitability at a resolution of 0.5 degrees latitude by 0.5 degrees longitude to construct the distribution of

²Darwin observed that a certain ecological niche was giving rise to an optimal shape of the finches' beaks.

³From a theoretical point of view the intensity of trade between regions could be an independent force leading to a convergence of the regional cultural traits. However, one would expect that trading would be more intense between regions with distinct comparative advantages, i.e. having sufficiently different types of land quality, for example. Such prediction, nevertheless, is at odds with the empirical findings implying that any effect towards ethnic homogenization operating through trade is dominated by the forces identified in the theory. Similarly, the pursuit of economic diversification through marrying across regions of different productive traits would operate against finding a systematic positive relationship between ethnic diversity and heterogeneity in regional land qualities.

land quality at a regional and country level. Such disaggregated level data, never before used in an economic application, allow for the econometric analysis to be conducted in a cross-artificial country framework. Specifically, to mitigate the problem of endogenous borders inherent to the cross-country regressions, I arbitrarily divide the world into geographical entities of a fixed size. As predicted by the theory I find that ethnic diversity, measured by the number of languages⁴ spoken in each artificial country, is systematically related to the underlying variation in land quality. Those characterized by a wider spectrum of land qualities give rise and support more ethnic groups. The findings are robust to the inclusion of continental and country fixed effects which effectively capture any systematic elements related to the state and continental histories of these geographical units.

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

Moving into a cross-country framework, the relationship between variation in land quality and ethnic diversity is further validated. Existing countries characterized by more heterogeneous land qualities, exhibit higher levels of 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 and additional geographical characteristics, the qualitative predictions remain intact. Interestingly, the finding that distance from the equator has a negative impact on ethnic diversity is consistent with the prediction that places experiencing persistent productivity shocks would be characterized by lower ethnic diversity.⁵

Historical accidents have influenced contemporary fractionalization outcomes. The European colonization after the 15th century, for example, is an obvious candidate. Analyzing

⁴There are no worldwide data on the distribution of ethnicities. Reassuringly, measures of ethnic and linguistic diversity available for existing countries are very highly correlated.

⁵This interpretation derives from the observation that distance from the equator correlates with seasonality. Note also that biodiversity generally decreases further away from the equator (Rosenzweig (1995)) allowing for fewer productive niches along which people may specialize.

the role of the colonizers in affecting the ethnolinguistic diversity of the colonized world reveals important patterns. The evidence is supportive of the historically documented arbitrariness of border drawing, (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 not only affected the geographically determined level of fractionalization. As a consequence of the introduction of their own ethnicity and the active interference 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 contemporary ethnic fractionalization into a natural component, driven by the distribution of land qualities, and a man-made one, offers new insights regarding the origins of cultural diversity, highlighting the role of variation in land quality and colonial history in particular.

The results of this study are directly related to the literature on state formation (Alesina and Spolaore (1997)). In this literature preference heterogeneity is a key determinant of the optimal state's size. The facts that public goods may not be equally complementary across different land endowments, and that these very differences in land endowments are behind ethnic fragmentation, have important implications about the relationship between current state sustainability and ethnic diversity.

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. Naturally, population mixing between two regions, may directly reduce genetic distance. According to the proposed theory the latter is endogenous to the transferability of country specific human capital within the pair. As a result, countries that are relatively dissimilar in the distribution of land endowments, will be populated by people displaying larger genetic distance, *ceteris paribus*. Consequently, the uneven diffusion of technology across countries may be an outcome of the differences in society's specific human capital. Introducing the pair-wise country differences in the distribution of land qualities, one can decisively improve upon the interpretation of the existing results.

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. 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 regional land quality that gives rise initially to relatively stable ethnic diversity, an element of primordialism. However, as the process of development renders land increasingly unimportant ethnic identity is ultimately bound to become less attached to a certain set of region specific skills and, thus, more situational and ambiguous in character.⁶

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

This study is a stepping stone for further research. Equipped with a more substantive understanding of the origins and determinants of ethnolinguistic diversity, long standing questions among development and growth economists, in which ethnic diversity plays a significant role, may be readdressed. Such topics include the origins of inequality across ethnic groups, the factors that affect the formation of states and the determinants of the diffusion of development within and across countries.

The rest of the paper is organized as follows. In section 2 historical evidence on the causes and spatial pattern of linguistic spreads is presented and Appendix E offers anecdotal evidence associating the distribution of land quality with the human capital endowments of ethnic groups in Kenya. Section 3 advances the theory and its predictions. Section 4 discusses the data and lays out the empirical analysis conducted in a i) cross-artificial country ii) cross-pair of adjacent regions and iii) cross-country framework, including the various robustness checks. This section also quantifies the impact of the European colonizers on the ethnolinguistic endowment of the colonized world. The econometric analysis concludes in section 5 by investigating the causal impact of ethnic diversity on a variety of economic outcomes. The last section summarizes the

⁶In other words, as the importance of region specific knowledge diminishes, ethnicity gradually transforms into a deliberate choice/consumption good and/or becomes predominantly an outcome of modern states pursuing discrete ethnic policies. For example, Miguel and Posner (2006) provide evidence that ethnic identification in Africa becomes more pronounced as political and economic competition increases, similarly Rao and Ban (2007) provide evidence of the man-made component of ethnic diversity in India by showing how state policies and local politics have had an important impact on shaping caste structures over the last fifty years. In another recent study Caselli and Coleman (2006) have a theory where ethnic traits provide a dimension along which voluntary coalitions may be formed and Esteban and Ray (2007) investigate the salience of ethnic identity on the eruption of civil conflict.

key findings and concludes.

2 Evidence on Migrations and Language Spreads

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

Linguists have long recognized the role of population mixing in producing common linguistic elements between places. As Nichols (1997) points out "almost all literature on language spreads⁷ focuses on either demographic expansion or migration as the basic mechanism". Both instances are a result of population movements towards territories previously unoccupied by their ancestors. As a result of population mixing the resulting regional populations experience a language shift (either to or from the immigrants' language). Similarly, languages long in contact come to resemble each other in several dimensions like sound structure, lexicon, and grammar. This resultant structural approximation is called convergence. To the 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 in the long run as an increasing function of population mixing between places is justified.

Regarding the effect of differential climatic shocks in generating movements of people, evidence suggests that this was indeed an important factor.⁸ For example, Nichols (1997) suggests that at least since the advent of the Little Ice Age in the late middle ages highland economies have been precarious, whereas the lowlands, with their longer growing seasons, were relatively prosperous offering winter employment for the essentially transhumant male population of the highlands. This caused lowland dialects to spread uphill. Prior to the global cooling, however, lowlands were dry and uplands moist and warm. Under these conditions, with highlands being relatively more economically secure, upland dialects spread downhill, through a similar process. The linguistic patterns found in regions like central Caucasus and the highland spread of Quechua fall in this category, (Nichols 1997b).

⁷Nichols (1997) defines a spread zone as "an area of low density where a single language or family of languages occupies a large range"

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

There are several examples showing that migrations have been occurring between places 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 is the case for the grasslands of central Eurasia.

Examples of groups that migrated along areas that were similar to their region of origin include Austronesians and speakers of Eskimoan languages who are coastally adapted peoples, and, have accordingly spread along coasts rather than inland. Along similar lines, Bellwood (2001) argues that the spread zones of agriculturalists and their languages following the Neolithic Revolution trace closely the distribution of land qualities that were amenable to 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 languages⁹, 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. "Friction zones" on the other hand, 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 into areas where their specific human capital would best apply, homogenizing them linguistically.¹⁰

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

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

¹⁰Whether this process of language shift occurred through replacement of the local populations or by extensive intermarrying is yet an open question.

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

The (pre)historical evidence on the spread of peoples and languages provides 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. In every period the economy produces a single homogeneous good using land, labor and region specific technology as inputs to the production process. The supply of land is exogenous and fixed over time. In fact, there are two regions i and j . The regional labor supply is governed by the evolution of the region specific know-how, its transferability between the places and the state of the relative temporary idiosyncratic productivity shock.

Each individual lives two periods and population is fixed. 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 lives¹¹, c_{t+1} , and are represented by a strongly monotone and strictly quasi-concave utility function, $U = u(c_{t+1})$.

3.1 Production of Final Output

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

Suppose that there are no property rights over land.¹² The return to land in every period

¹¹Allowing both for endogenous fertility and intergenerational altruism the predictions would not be reversed.

¹²The 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

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

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

3.2 Accumulation of region specific technology

The level of regional technology available to the indigenous population at time t in region r advances as a result of learning by doing $h_{t+1}^r = \psi(h_t^r)$, $r \in \{i, j\}$ with $h_0^r = 1$, $\psi_{h_t^r} > 0$ and $\psi_{h_t^r h_t^r} < 0$. Since both region specific technologies start from the same initial level and follow the same law of motion, the technology available to the indigenous in each region is identical in every period. 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).

As adults, individuals may move freely from one region to another.¹³ However, this comes at a cost arising from differences in the 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 the individual may apply as a worker in the receiving place. This erosion increases as places become increasingly different in the set of productive activities.

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

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

where k_t^r are the units of knowledge that a migrant may apply should she move to region r and ε captures the degree of erosion within a regional pair. Those characterized by more

affect the qualitative results. Specifically, if property rights were preassigned to the indigenous then the rental price of land would adjust as a result of the demand from migrants. Alternatively, property rights could be endogenized in a conflict model sharing the same basic properties as the current set up leading to qualitatively similar predictions.

¹³Including additional costs associated to moving, either as a result of time expended on relocating or in the form of a transfer to the indigenous in the receiving area would not change the results. It would, however, add an additional dimension along which places might differ.

heterogeneous endowments score higher along this dimension. Note that within a regional pair erosion of region-specific knowledge is symmetric. The properties of transferring region-specific technology across places, follow directly by differentiating (2). In particular, the migrant's know-how relevant to the receiving place decreases in the level of erosion between the regions, $\frac{\partial k_t^r}{\partial \varepsilon} < 0 \forall r \in \{i, j\}$. Second, the migrant's know-how relevant to the receiving place increases in the human capital of the place of origin, $\frac{\partial k_t^r}{\partial h_t^q} > 0, \forall r, q \in \{i, j\}, r \neq q$. Third, there exist diminishing returns to the transferability of the know-how of the place of origin, $\frac{\partial^2 k_t^r}{\partial^2 h_t^q} < 0, \forall r, q \in \{i, j\}, r \neq q$. This captures that the accumulation of technology becomes increasingly region specific and, as a result, less useful in case of relocation.¹⁴ Lastly, the transferability of region-specific knowledge decreases with the level of erosion, $\frac{\partial^2 k_t^r}{\partial h_t^q \partial \varepsilon} < 0, \forall r, q \in \{i, j\}, r \neq q$. In other words, an additional unit of domestic know-how is less applicable to the receiving region in pairs characterized by higher erosion.

Taking into account the common evolution of region specific human capital and the preceding discussion, it follows that the indigenous population of region r , that is individuals who work in the same region they are born to, have higher level of know-how compared to that of the migrants during the period the migrants arrive, that is the output per worker is higher for the indigenous population.¹⁵ Specifically, using (1)

$$y_t^r = (z_t^r h_t^r) (m^r / L_t^r)^{1-\alpha} \quad \& \quad y_t^{q \rightarrow r} = (z_t^r k_t^r) (m^r / L_t^r)^{1-\alpha} \quad (3)$$

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

3.3 Defining Common Ethnicity

A probabilistic framework regarding the formation of shared ethnolinguistic elements is adopted. Particularly, it is conjectured that the probability that individuals from *regions* i and j will share common traits increases in the intensity of population mixing between the two regions over time.¹⁶ As individuals cross-migrate, they add their cultural traits from the place of origin

¹⁴Such diminishing returns could be conceived as an outcome of increasing specialization in the set of activities relevant for each region. At any given level of heterogeneity within a regional pair, further specialization in the respective activities diminishes the transferability of the additional know-how.

¹⁵It is useful to note that migrants' offspring have the same level of region specific human capital as the offspring of non-migrants. Gradual accumulation of the region specific technology for the offspring of immigrants would not alter the results. It could, however, create selection into reverse migration of the people whose ancestors were immigrants.

¹⁶Assuming 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

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 in probability distinct ethnolinguistic characteristics.

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

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

where I_t is an indicator function that takes the value of 1 if migration occurs in period t between regions i and j , irrespective of the direction, and 0 otherwise. Such formulation could alternatively be interpreted as an inverse measure of ethnic distance between the two regions. Note that this relationship applies in the long-run, so T should be thought as relatively large.¹⁷ According to this definition pairs of places whose populations never mixed until period T would have zero probability of sharing common ethnic traits, or alternatively put, maximal ethnolinguistic distance. Alternative specifications of (4) could accommodate a potential “founder” effect in 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 is shaped by the forces of the environment.

3.4 Labor Allocation Across Regions

Given preferences individuals in each period t maximize earnings. In the beginning of every period t regional productivity shocks, z_t^r , which last for one period, are realized. Adults observe the realization of the shock and decide whether or not to migrate by comparing the respective

regionally as time evolves due to cultural drift, Boyd and Richardson (1985).

¹⁷Indeed, in the short run population mixing may increase diversity in the receiving place, (Williamson, 2006).

incomes in (3).¹⁸ 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. Let $\{\lambda_t\}_{t=0}^T$ denote the sequence of the ratios of productivity shocks of region i relative to region j , that is $\lambda_t = \frac{z_t^i}{z_t^j}$. It follows that $\lambda_t > 0$ and $\lambda_t \geq 1$ iff $z_t^i \geq z_t^j$. Using (3) and substituting L_t^i, L_t^j with the respective values of the preceding period, individuals from region i have an incentive to move to region j in the beginning of period t iff:

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

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

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

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

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

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

¹⁸Migration in this framework lasts for at least one generation. It would be straightforward to incorporate short term migration by allowing for several productivity shocks per generation per region. Accounting for seasonality in the climatic fluctuations, would strengthen the theoretical predictions. Conditional on the similarity of productive endowments, places characterized by higher seasonality would exhibit larger and more frequent short-term migration movements.

the size of the migration $M_t^{i \rightarrow j}$ in the population of the receiving region, j , subtracting it from the region of origin, i , and manipulating (5) the level of migration may be explicitly derived

$$M_t^{i \rightarrow j} = \frac{L_{t-1}^i - (\lambda_t (h_t^i)^\epsilon)^{\frac{1}{1-\alpha}} \frac{m^i}{m^j} L_{t-1}^j}{1 + (\lambda_t (h_t^i)^\epsilon)^{\frac{1}{1-\alpha}} \frac{m^i}{m^j}} \quad (7)$$

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

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

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

3.4.1 Past Migrations

As it is evident from (7) and (8) the size of the migration movement in period t depends on the level of regional population densities in period $t - 1$. The latter is a function of past migration movements. In particular, in the beginning of any period t , and before any labor movement occurs (if any) the ratio of the regional population densities equals $\frac{L_{t-1}^i}{L_{t-1}^j}$.¹⁹ Depending on the direction of the last migration either (5) or (6) should hold with equality when evaluated at the regional population densities after the occurrence of migration in period, s . Consequently, solving for the ratio of regional population in period s , $\frac{L_s^i}{L_s^j}$, the following two cases obtain:

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

$$\frac{L_{t-1}^i}{L_{t-1}^j} = \frac{L_s^i}{L_s^j} = \left(\lambda_s (h_s^i)^\epsilon \right)^{\frac{1}{1-\alpha}} \frac{m_i}{m_j} \quad \text{if } M_s^{i \rightarrow j} > 0 \quad (9)$$

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

$$\frac{L_{t-1}^i}{L_{t-1}^j} = \frac{L_s^i}{L_s^j} = \left(\lambda_s (h_s^j)^{-\epsilon} \right)^{\frac{1}{1-\alpha}} \frac{m_i}{m_j} \quad \text{if } M_s^{j \rightarrow i} > 0 \quad (10)$$

In Appendix A the properties of the migration size between places given by (7) and (8) are established.

¹⁹The latter is identical to the ratio of population densities realized in the last occurrence of migration.

3.5 The $M^i M^j$ and $M^j M^i$ loci

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

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

$$y_t^{i \rightarrow j} = y_t^i \Rightarrow \lambda_t|_{M^i M^j} = \left(\frac{L_{t-1}^i m_j}{L_{t-1}^j m_i} \right)^{1-\alpha} (h_t^i)^{-\varepsilon} \quad (11)$$

Similarly, $M^j M^i$ 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, $y_t^{j \rightarrow i} = y_t^j$. In particular, $M^j M^i \equiv \left\{ \left(h_t^j, \lambda_t, \frac{L_{t-1}^i}{L_{t-1}^j}, \varepsilon\right) : y_t^{j \rightarrow i} = y_t^j \right\}$. Thus, the level of the relative productivity shock in period t , $\lambda_t|_{M^j M^i}$, that makes people from region j indifferent to moving is:

$$y_t^{j \rightarrow i} = y_t^j \Rightarrow \lambda_t|_{M^j M^i} = \left(\frac{L_{t-1}^i m_j}{L_{t-1}^j m_i} \right)^{1-\alpha} (h_t^j)^{\varepsilon} \quad (12)$$

As it is evident in (11) and (12) the ratio of the regional population densities from the last period is important in determining the no-migration loci. The ratio of regional population densities in period $t - 1$ may be expressed by either (9) or (10) 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 1 *The properties of the non-migration loci:*

The $M^i M^j$ locus

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

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

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

The $M^j M^i$ locus

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

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

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

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 1, conditional on the past that is on λ_s , h_s^j , and h_s^i , the distance between the no-migration loci, $M^j M^i$ and $M^i M^j$, increases at the level of erosion. This implies that given the contemporary relative 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 .

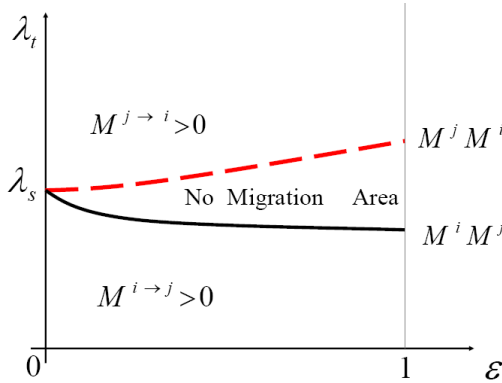


Figure 1a

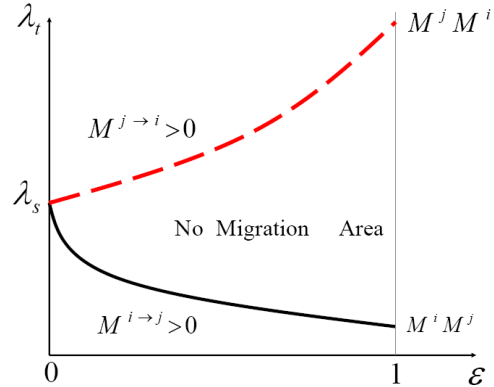


Figure 1b

In the set of figures above 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 of migration. Lemma 2 in Appendix A summarizes the cases of migration occurrences.

3.6 The Formation of Common Traits Over Time

Having established how the environment shapes population mixing, the formation of common ethnolinguistic elements may be traced over time. In period $t = 0$ the region specific technology is at its minimum, $h_0^i = h_0^j = 1$, since no accumulation has occurred yet, and individuals distribute themselves in places i and j such that the output per capita at time $t = 0$ is the same across regions. It is assumed that the relative productivity shock, λ_t , is a discrete random variable independently and identically distributed over time. In particular,

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

with $\lambda_{\min} < \lambda_{\max}$.²⁰ The following Proposition shows how erosion, ε , the ratio of the relative productivity shocks, λ_t/λ_s , and the level of region specific technology determine the probability that two regions will share common cultural elements.

Proposition 1 *Under (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 ,*

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

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 ,*

$$\frac{\partial f_T(\lambda_t; \varepsilon, \lambda_s, h_T)}{\partial \text{var}(\lambda_t)} \geq 0$$

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 ,*

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

²⁰This 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.

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.²¹ 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 productivity shocks are bound to form homogeneous ethnolinguistic traits.²²

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.

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. The ideal index could be derived looking into how similar was the

²¹To the extent that the duration of human settlements is a proxy of the level of region specific human capital, the empirical finding of Ahlerup and Olsson (2007) that the former positively affects ethnic diversity is consistent with the third prediction of Proposition 1.

²²This prediction is in line with Nettle's (1999) finding that countries with higher ecological risk sustain lower linguistic diversity.

distribution of productive activities across regions in a period of human history when the formation of cultural traits was taking place. Such quest for detailed data, 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 and in absence of historical data, I use contemporary disaggregated data on the suitability of land for agriculture as a proxy for the regional productive characteristics.

The intuition for using differences in land quality as a proxy for differences in the distribution of productive activities is the following. Farming would be the dominant form of production in places characterized by high land quality, with the regions possibly differing in the optimal mix of plants and crops under cultivation. That is, even within agriculture, the specificity of human capital derives from the different crops produced regionally. However, herding/pastoralism is bound to be more widespread at intermediate and low levels of land quality, exactly because agriculture is less suitable in such areas. At very low levels of land quality, also, being a middleman has been perhaps the most widespread activity as the case for cultures residing 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 future climatic change on agricultural suitability.²³ This dataset provides information on land quality characteristics at a disaggregated level. Each observation takes a value between 0 and 1 and represents the probability that a particular grid cell may be cultivated. The authors construct this index by (i) empirically fitting a relationship between existing croplands and both climate indices and soil characteristics and then (ii) combine the derived relationship with the available regional climatic and soil characteristics to predict the regional suitability of agriculture worldwide.

The climatic characteristics are based on mean-monthly climate conditions for the 1961–1990 period and capture i) monthly temperature ii) precipitation and iii) potential sunshine hours. All these measures monotonically increase the suitability of land for agriculture. Re-

²³The dataset is available at the Atlas of the Biosphere accessible at <http://www.sage.wisc.edu/atlas/data.php?incdataset=Suitability%20for%20Agriculture>

garding 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.

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.²⁴

This detailed dataset, never used in an economic application, provides an accurate description of the global distribution of land quality. The map in Appendix B shows the worldwide distribution of land quality. Using these raw global data I construct the distribution of land quality at the desired level of aggregation.

Regarding the cross-artificial country and cross-pair of adjacent regions analysis, ethnic diversity is captured using information on the location of linguistic groups. In the case of artificial country regressions the number of linguistics groups within each geographical unit provides a measure of overall ethnic diversity. Regarding the adjacent region analysis an index of ethnic similarity is constructed by calculating the percentage of common languages within any pair of adjacent regions. Data on the location of linguistic groups' homelands are 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 database of languages around the world.²⁵

Regarding the cross-real country analysis a wealth of alternative measures of ethnic diversity is available. The measure of fractionalization widely used is the probability that two *individuals* randomly chosen from the overall population will differ in the characteristic under consideration, 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)) and augmented by Fearon and Laitin (2003). This index represents for each country the

²⁴There 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.

²⁵The data are available at www.gmi.org. To identify which languages are spoken within the unit of analysis I use the information on the location of language polygons. Each of these polygons delineate a traditional linguistic homeland; populations away from their homelands (e.g. in cities, refugee populations, etc.) are not mapped. Also, the World Language Mapping System does not attempt to map immigrant languages. Finally, linguistic groups of unknown location, widespread languages and extinct languages are not mapped and, thus, not considered in the empirical analysis.

probability that two individuals randomly drawn from the overall population will belong to different ethnolinguistic groups. Using the linguistic, ethnic and religious fractionalization indexes constructed by Alesina et al. (2003), the absolute number of ethnic or linguistic groups derived by Fearon (2003) or the ethnic fractionalization measure proposed by Montalvo and Reynal-Querol, (2005), the qualitative results are similar.²⁶

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 figure²⁷ 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, $range_{Nepal} = 0.84$.

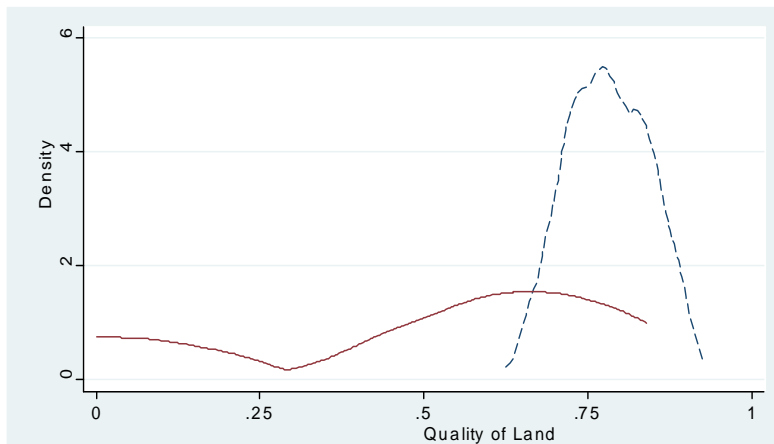


Figure 2: Dashed line - Greece, Solid Line - Nepal

The range of land quality, i.e. the support of the distribution within the respective unit of analysis, is the statistic used to capture the degree of heterogeneity in land quality.²⁸ It is an

²⁶Modifying the current framework to uncover the determinants of ethnic *polarization* is a topic for future research.

²⁷The figure shows the kernel density estimate (weighted by the Epanechnikov kernel) of regional land qualities for each country.

²⁸The standard deviation of regional land quality is an alternative measure of a country's productive heterogeneity. Such proxy inherently captures variation both in the extensive, that is, in the extremes of the distribution of the land endowment, and the intensive margin. Conditional on the range, however, increases in the standard deviation of the endowment increase the weight towards the fixed extremes of the land quality distribution. This effectively results in fewer distinct land qualities along which groups may specialize. A further

index of how readily location specific knowledge may be transferred across places. Intuitively, a larger range implies that the geographical unit considered is composed of territories which are increasingly different in the underlying land qualities, effectively enlarging the set of regionally distinct activities along which groups may specialize. Consequently, the larger is the spectrum of land qualities, i.e. *range*, within the unit of analysis the less transferable is the regional know-how. Thus, according to the theory²⁹ a larger range would increase the probability that the underlying areas are ethnically distinct, *ceteris paribus*. Indeed, going back to the example of Greece and Nepal, ethnolinguistic fractionalization in Greece is only 0.10 compared to the highly ethnolinguistically fragmented society of Nepal with $ELF_{Nepal} = 0.70$.

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 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.³⁰

4.2.1 Cross-Artificial Country Analysis

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

The way that the artificial countries are constructed is the following. First, I generate

consequence of such an increase is that it causes a more unequal distribution of population across regions and since by construction the fractionalization indexes at the real country level are affected by the distribution of the population across ethnic groups (see below) an increase in the intensive margin may decrease fractionalization. Results not shown, indeed suggest that controlling for the range of land quality and the standard deviation in the cross-real country regressions both enter significantly, the range with a positive sign and the standard deviation with a negative one. Same pattern obtains at the cross-artificial country regressions. 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.

²⁹The implications of the theory have been derived for pairs of regions. Extending the model to allow for multiregional population mixing I conjecture that it would not affect the qualitative predictions. It would, however, deliver a cumbersome analysis.

³⁰Nevertheless, 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-real country regressions but not in the cross-artificial country ones.

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 deliver artificial countries with geographical characteristics comparable to an average real country.³¹

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

In the regression analysis the sample of artificial countries is restricted in the following way. Territories for which there are at least 3 regions with information on land quality and languages are included. Also, to ensure that the findings are not driven by including in the regressions regions with negligible population density, only artificial countries whose individual regions have at least 1 person per sq. km. are considered.³² Given these considerations the distribution of the number of languages spoken across artificial countries is shown in Figure 3³³:

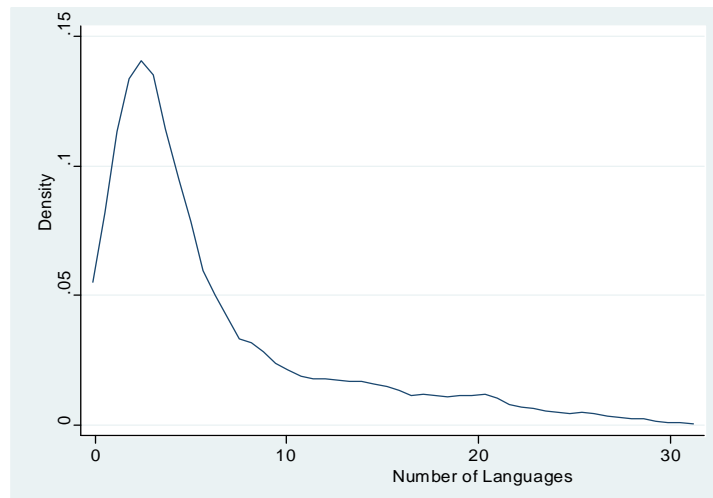


Figure 3: Distribution of languages across artificial countries

The resulting sample size is 548 artificial countries with a median of 53 regional land

³¹Using alternative dimensions like 2.5 by 2.5 or 5 by 5 degrees does not change the results.

³²The 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.

³³Note that the distribution of the number of languages is skewed so instead of the levels the log of languages is used in the regressions below. Excluding the extremely linguistically fragmented artificial countries, i.e. those with more than 20 languages spoken, the qualitative results are similar.

quality observations per artificial country. Descriptive statistics and the raw correlation between the variables used in the regressions are presented in Tables 1a, 1b. Among other geographical characteristics the standard deviation of elevation, *elev_sd*, is constructed in order to accurately quantify the variation in topography within an artificial country. Note that this variable except for capturing the pure transportation costs associated with relocation it may also proxy, like the spectrum of land qualities, *range*, for differences in the productive activities across regions. So, one may jointly interpret the effect of *elev_sd* and *range* on languages spoken as the impact of geographic variability on ethnolinguistic diversity.

In each artificial country there are on average 6.13 languages spoken and the raw correlation between the spectrum of land qualities, *range*, and the number of languages is positive and large, 0.29. Figure 4 in Appendix B shows one example of what now constitutes the unit of analysis. The circles represent the regional land qualities (they are the centroids of the original grids) and the different colors represent the locations of the different linguistic groups. This artificial country in Figure 4 falls between two existing countries with the squiggly line delineating the current borders between Iran on the east and Iraq on the west. There are in total 12 languages³⁴ spoken in this area and the spectrum of land qualities is 0.72 ranging from places that are totally inhospitable to agriculture to areas where the climate and the soil are highly conducive to cultivation.

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

$$\ln \#_lang_i = \beta_0 + \beta_1 range_i + \beta_2 X_i + \xi_i \tag{13}$$

where $\ln \#_lang_i$ is the log number of languages spoken in artificial country i , $range_i$ is the support of the distribution of land quality, and X_i is a vector of geographical and political 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 2.³⁵ Specifically, in the first regression of Table 2 only the *range* is included. It has a large and

³⁴Namely these are: Assyrian Neo-Aramaic, Central Kurdish, Chaldean Neo-Aramaic, Gurani, Koy Sanjaq Surat, Mesopotamian Spoken Arabic, Najdi Spoken Arabic, North Mesopotamian Spoken Arabic, Northern Kurdish, Sangisari, South Azerbaijani and Southern Kurdish. Languages' traditional homelands may overlap. In this particular grid, for example, places that speak Assyrian Neo-Aramaic also speak Chaldean Neo-Aramaic.

³⁵The results presented here are OLS estimates with the standard errors adjusted for spatial correlation following Conley (1999). This correction requires the choice of a cutoff distance, beyond which artificial countries do not influence each other. After projecting the world into the euclidean space using the Plate Carrée projection I use a cutoff distance of 6000 km. Results are similar using 1000 km, 2000 km, 3000 km or 8000 km. Using Tobit or Poisson estimators the predictions remain qualitatively intact.

significant positive impact on linguistic diversity. The variation in land qualities alone explains 9% of the variation in the number of languages. This is a novel and economically important finding that reveals the geographical origins of contemporary ethnolinguistic diversity.

In the second column of Table 2 I introduce an array of geographical features of these areas.³⁶ In particular, the size of each artificial country, *areakm2*, the average land quality, *avg*, the latitudinal distance from the equator, *abs_lat*, the standard deviation of elevation, *elev_sd*, the number of real countries that an artificial country falls into, *#_cntry*, a dummy for the units that belong as a whole to an existing country, *in_country*, the area under water, *water_area*, as well as the distance from the coastline, *sea_dist*, are controlled for. As expected artificial units with more variable topography sustain larger linguistic diversity. Areas that entirely belong to a single real country display systematically lower ethnic fragmentation whereas more languages are spoken in artificial units falling into more real countries. The distance from the equator itself enters negatively and significantly consistent with the prediction that more seasonal environments lead to lower ethnic diversity. Also, larger artificial units display more languages although the point estimate is insignificant. Average land quality is not significantly related to linguistic diversity conforming with the theoretical prediction. With respect to the variable capturing water barriers, *water_area*, it enters positively and it is statistically insignificant. This insignificant result raises the question whether water bodies should be considered as a barrier or a facilitator of population mobility. Finally, the distance from the shoreline of an artificial country does not systematically affect linguistic diversity. These important controls make the coefficient of *range* reduce by half, remains, however, both economically and statistically significant. One standard deviation increase in *range* increases linguistic diversity by 14% adding on average 0.89 languages to an average artificial country.

This finding is robust to alternative specifications. In particular, taking advantage of the arbitrarily drawn borders of these geographical units one may explicitly control for real country and continental fixed effects.³⁷ This is done in all subsequent specifications. Such inclusion of powerful controls, naturally not possible in a cross-real country framework, allows to explicitly take into account any systematic elements related to the state histories of existing real countries

³⁶The following measures are constructed by the author. For the details on the construction of these variables see Data Appendix D.

³⁷For an artificial country that falls into more than one real countries the respective dummies assigned represent the fraction of the artificial country's area that falls into each real country. For example, the artificial country in Figure 4, which falls into two real countries, gets the value of 0.333 for the country dummy of Iran because 33.3% of the artificial country belongs to Iran and 0.677 for the country dummy of Iraq. Similar reasoning applies to obtaining the continental dummies.

and, thus, produce reliable estimates of the effect of diversity in land quality on ethnic diversity. The inclusion of country and continental fixed effects in the third column of Table 2 slightly increases the coefficient on *range*. One standard deviation increase in the spectrum of land qualities increases by 15% the number of languages within an artificial country contributing significantly to the formation of ethnically diverse societies.

In the fourth column of Table 2 the interaction of the average land quality with the range, *avgxrange* is included. The coefficient of *range* slightly increases and remains precisely estimated. The direct effect of the average land quality is insignificant, as the theory predicts, and the interaction term enters highly insignificant. Thus, it is dropped from the subsequent analysis.

In column 5 of Table 2 the main specification (13) is estimated focusing on artificial units that entirely belong to a single existing country. This robustness check allows to investigate whether the estimated strong positive relationship between the spectrum of land qualities and ethnic diversity obtains across regions within existing countries. Reassuringly, the variation of land quality across regions within countries systematically shapes ethnolinguistic diversity. Namely, territories within countries that display more heterogeneous land endowments give rise and sustain more ethnic and linguistic groups.

In the last column of table 2 specification (13) is estimated allowing for a differential effect of diversity in land quality depending on whether an artificial country falls in or out of the tropics.³⁸ This specification allows 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 climatic zones. The effect of *range* on linguistic plurality is positive and significant at 1% level for the artificial countries out of the tropics. Also, the significantly positive effect of the interaction of range with tropics, denoted by *tropicsxrange*, implies that a certain level of heterogeneous land qualities is more conducive to ethnic differentiation in the climatically less variable and rich in species tropical environments.

This section establishes that the variation in land quality and elevation across artificial countries are a significant causal determinant of contemporary ethnic diversity. The fact that these results obtain at an arbitrary level of aggregation, and after controlling for country and continental fixed effects brings into light the, so far neglected, geographical origins of ethnic diversity.

³⁸The tropics extent from 23.5 latitude degrees south to 23.5 latitude degrees north.

4.2.2 Pairwise Analysis of Adjacent Regions

The theoretical framework has focused on how differences in the productive structure between *two* regions contribute or deter the formation of common ethnic traits. Hence, a direct test of the theory naturally dictates pairs of regions as the unit of analysis. In this setting, the empirically relevant question becomes how differences in land quality within a regional pair affects the ethnic similarity between the two places. The information provided in the language dataset on the location of linguistic groups allows for such detailed investigation. In particular, to implement such a test I identify the neighboring regions of each grid. Neighbors of each area are considered those who are immediately adjacent at a distance of 0.5 degrees, i.e. directly to the: south, north, east and west as well as those that are immediately and diagonally contiguous at a distance of 0.71 degrees i.e. to the northwest, southwest, northeast and southeast.³⁹ In total a single region may belong to at most eight pairs (see Figure 4). Out of the 58920 regions contained in the land quality dataset in 15982 grids there is no information on the languages spoken and consequently, are dropped from the analysis. In total, there are 159358 unique pairs of adjacent areas spanning the whole world averaging, 3.7 neighbors per region.

For the pairwise regressions of adjacent regions the following specification is adopted:⁴⁰

$$pct_comlang_{ij} = \beta_0 + \beta_1 lqdiff_{ij} + \beta_2 X_{ij} + \xi_{ij} \quad (14)$$

where $pct_comlang_{ij}$ is the percentage of common languages, i.e. the number of common languages divided by the total number of unique languages spoken in pair i, j , and captures the degree of ethnic similarity between any two adjacent regions.⁴¹ The variable $lqdiff_{ij}$ is the absolute difference in land quality between regions i and j and is an inverse measure of how similar are the primitive productive characteristics of any two adjacent areas. Tables 3a and 3b present the summary statistics and the raw correlation of the variables used in the analysis. Note that the mean of $pct_comlang$ has an interesting economic interpretation: adjacent regions, by virtue of proximity, have on average 77% of the total number of languages

³⁹Ioanna Grypari graciously provided the code for identifying the pairs of adjacent regions.

⁴⁰In specifications (1) and (3) standard errors are corrected for spatial dependence following Conley (1999). This correction requires the choice of a cutoff distance, beyond which regional pairs do not influence each other. After projecting the world into the euclidean space using the Plate Carrée projection I use a cutoff distance of 500 km. This is a highly computationally intensive method and currently I am working on implementing it for different cutoff distances and applying it to the remaining specifications. Specifications (2) and (4) are heteroskedastically robust OLS regressions clustered at the level of each individual region. This allows for the residuals of pairs having a common individual region to be correlated.

⁴¹Using as an inverse measure of local ethnic similarity the number of languages spoken within each pair of regions, the results are unchanged.

in common.

According to the theory regions characterized by large differences in their productive characteristics, would hinder regional population mixing eventually giving rise to ethnically distinct populations. The first column in Table 4 corroborates this focal prediction. The difference in land quality and elevation within a regional pair has a strong negative effect on the formation of common ethnic traits. In particular, a two standard deviation increase in the difference in land quality, $lqdiff_{ij}$, decreases the percentage of common languages by 3.9 points and a similar increase in the difference in elevation, $eldiff_{ij}$, decreases the percentage of common languages by 6 percentage points contributing significantly to the formation of ethnolinguistically distinct neighbors. In the same specification several geographical characteristics are taken into account. In particular, distance from the equator, abs_lat , systematically produces more linguistically homogeneous neighbors, the average elevation, $elev$, is not significantly affecting local ethnic diversity whereas the average land quality of the regional pair, $land_quality$, decreases ethnic similarity. Distance from the shoreline of a regional pair, sea_dist , as well as both regions being in the same country, $same_country$, significantly increase the likelihood of sharing common languages. As a proxy of water barriers in the pairwise regressions I construct and use the number of distinct water bodies that are found in each regional pair, $\#_body_waters$. This measure does not systematically affect local ethnic diversity. Finally, a control for the difference in the area of language coverage between the regional neighbors is included. As expected pairs whose individual regions differ in the spatial extent of their languages' coverage show lower linguistic similarity.⁴² Overall, these geographical characteristics capture 18% of the variation in local ethnic diversity.

In column 2 of Table 4 I take advantage of the high resolution data to control for country and continental fixed effects. Regarding the country fixed effects each region within a pair is assigned the dummy of the country it belongs to. This specification explicitly takes into account any systematic elements related to the state histories of each individual region, and might have independently affected the formation of common ethnic traits. The point estimates of $lqdiff$ and $eldiff$ slightly decrease remain nevertheless economically and statistically highly significant. One may naturally wonder whether the estimates on geographic variability derived in the first two specifications are applicable to an individual country as well. To shed some light on this point in column (3) the baseline specification focuses on regional pairs that belong entirely to India. Reassuringly, the point estimates on the difference in elevation and land

⁴²Introducing the pairwise difference in population density neither changes the results nor affects ethnic similarity independently.

quality are very similar to the estimates in specification (1). Interestingly, the number of water bodies within regional pairs in India significantly increases linguistic similarity implying that within India water bodies have been facilitating rather impeding the formation of common ethnic traits. The significant negative effect of latitudinal distance from the equator within India is harder to interpret.

Finally, in column 5 I allow for differential effect of the difference in regional land quality, $lqdiff_{ij}$, by continent. The marginal effect of $lqdiff_{ij}$ differs significantly across continents. As one might expect the effect is large and significant within Africa, Asia and Pacific whereas it quantitatively less so within Europe and the Americas.⁴³

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

4.2.3 Cross-Real Country Analysis

Having established that the differences in land quality and elevation between adjacent regions and within artificial countries affects systematically the respective ethnic endowment I now proceed into investigating the relationship between the spectrum of land qualities and ethnolinguistic fractionalization across existing countries. Using this global data on suitability of land for agriculture the distribution of land quality for each country is constructed. The number of regional observations per country range from a single observation for Luxemburg to 11515 for Russia. The median number of data points is 80.

Existing countries vary widely in the variety of land qualities covered by their territories. 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

⁴³In particular, $\left. \frac{\partial pct_comlang}{\partial lqdiff_{ij}} \right|_{Europe} = -.063$ significant at 1% level, $\left. \frac{\partial pct_comlang}{\partial lqdiff_{ij}} \right|_{Asia} = -.206$ significant at 1% level, $\left. \frac{\partial pct_comlang}{\partial lqdiff_{ij}} \right|_{Pacific} = -.542$ significant at 1% level, $\left. \frac{\partial pct_comlang}{\partial lqdiff_{ij}} \right|_{Americas} = .021$, insignificant. Note that the Pacific includes Australia, New Zealand and Papua New Guinea. The language coverage within Americas is the poorest across continents, which may partially explain the insignificant finding.

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, a striking parallel emerges. The ethnically fragmented society of Malawi, $ELF_{Malawi} = 0.62$, reflects the large underlying spectrum of land qualities compared to the ethnically homogeneous Lesotho, $ELF_{Lesotho} = 0.22$.

As 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.⁴⁴

It is straightforward to manipulate (4) to elucidate how population density across places affects measured fractionalization. The expected fractionalization, $E(ELF)$, for a pair of places in particular reads:

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

where $(1 - f_T)$ is the probability that the two regions i and j will have different ethnic traits and $\left(1 - \left(\frac{L_T^i}{L_T^j + L_T^i} \right)^2 - \left(\frac{L_T^j}{L_T^j + L_T^i} \right)^2 \right)$ is the probability that two randomly chosen individuals will belong to *different regions*. It is evident from (15) 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 $lqgini$, is constructed. As expected the gini of land quality is highly correlated (0.59) with how unequally

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

population density is distributed across regions within country in 1990.^{45, 46}

Given the preceding discussion the following main specification is adopted:

$$ELF_i = a_0 + a_1 range_i + a_2 avg_i + a_3 avg_i \times range_i + a_4 lqqini_i + \eta_i \quad (16)$$

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 $lqqini_i$ is the gini coefficient measuring how unequally is land quality distributed among regions of country i . The interaction term, $avg_i \times range_i$, 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 negligible population density the relevant statistics are derived after taking out from each country the 10% of the observations with the lowest population density.⁴⁷ 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 5a, 5b.

⁴⁵To 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>.

⁴⁶Results 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 $lqqini$, however, it is significant only in regressions using as dependent variable the ethnic fractionalization index derived by Alesina et. al. (2003).

⁴⁷Using 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.

In Table 6 the regressors of the main specification are added sequentially.⁴⁸ 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 (16) are presented in column 3 of Table 6. The inclusion of *lqqini* as expected enters 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 explain 21% of the variation of contemporary ethnolinguistic fractionalization across countries. As predicted, an increase in the spectrum of land qualities within a 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 existing countries become increasingly suitable for agricultural production it becomes easier to transfer region specific technology. This lowers the barriers to population mixing reducing ethnic diversity.

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 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*, *lqqini* and *avg_range* are significant at 1% level. The average land quality, *avg*, is not statistically different from zero which is consistent with the theoretical prediction.⁴⁹

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,⁵⁰ in Table 7 different

⁴⁸The 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.

⁴⁹Nevertheless, the overall effect of the average land quality on ethnic fractionalization is negative and statistically significant. Note that in the cross-artificial country regressions a similar effect was not obtained once country fixed effects were included in the analysis. This raises the possibility that this overall negative impact of *avg* on ethnic diversity is partly driven by state histories which depend on the level of land quality.

⁵⁰The 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 diversity

specifications are employed. I explore alternative hypotheses for the emergence of ethnicities, namely, other geographical characteristics and historical contingencies.

In the first column of Table 7 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. 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 negative but insignificant.⁵¹ Repeating the analysis excluding all the countries of Sub-Saharan Africa produces qualitatively similar results.

Other Geographical Characteristics

In the third column of table 7 geographic controls that could potentially affect fractionalization are accounted for. The distance from the equator, denoted by *abs_lat*,⁵² has a strong negative effect on ethnolinguistic fractionalization. To the extent that distance from the equator increases seasonality, this is consistent with the theory's prediction that places subject to more variable productivity shocks should display lower levels of fractionalization, *ceteris paribus*. 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, conditional on any regional fluctuation in productivity, and thus should on average display higher ethnolinguistic fractionalization. It should be noted, however, that mean distance from the sea, also captures the vulnerability of places to both the incidence and the intensity of colonization. Thus, the coefficient should be cautiously interpreted.

in land quality on ethnic diversity. More importantly, soil quality is mostly affected by the regional climate. In particular, comparing the global distribution of annual precipitation with the distribution of soil pH (these maps are available at http://www.sage.wisc.edu/atlas/maps/anntotprecip/atl_anntotprecip.jpg and http://www.sage.wisc.edu/atlas/maps/soilph/atl_soilph.jpg respectively) it is evident that regions receiving lots of precipitation are characterized by highly acidic soils whereas in places with low precipitation the soil is alkaline. This demonstrates that regional soil quality is overwhelmingly an outcome of the local climatic conditions.

⁵¹Throughout all specifications the marginal effect of *range* on ethnic diversity turns negative after a certain high threshold of land quality. Reassuringly, once negative it never obtains statistical significance across any specification.

⁵²See Appendix D for a detailed description of the data used.

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 7, is driven by the fact that although Sub-Saharan Africa, is the most ethnically diverse continent, has an average standard deviation of 0.28 km. whereas for a non Sub-Saharan country the average is 0.48 km. Indeed, controlling for continental fixed effects, see column 5, a more variable topography affects ethnic diversity 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 7 controls accounting for the variation in historical contingencies across countries, are added. The log of the population density in 1500 *AD*, *lpd1500*,⁵³ enters negatively but not significantly and the year when each country gained independence, *yrentry*, has a significant impact on 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.⁵⁴

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

⁵³This measure is highly correlated, around 0.56, with the index of state antiquity constructed by Bockstette et al. (2002). Including both makes them insignificant. Consequently, I only include in the regressions the log of the population density in 1500. It may be useful to note that the term "state history" used throughout this study is distinct from the state antiquity index.

⁵⁴Of 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.

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 7 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 25% of the observed ethnolinguistic variation as opposed to 21% 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 (16), is the natural level of fractionalization, nat_ELF , that a region would exhibit if left largely undisturbed. On the contrary, artificial fractionalization is the part of the observed fractionalization that is not driven by the characteristics of land endowment. 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 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 also having 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 frac-

tionalization outcomes endogenous to their colonial experience, the identity of the colonizers and how intensely the colonizers settled, among other things. Table 8 presents the main specification (16) 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 explains 32% of the variation in the ethnolinguistic fractionalization for the non-colonized world compared to only 16% for the colonized one. 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 Englebort et al. (2002), essentially shaping the extent of land qualities 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 main specification (16). Since both country borders and the size of ethnic groups are endogenous to the incidence and nature of colonization, to obtain the natural level of fractionalization of the colonized world, the point estimates used are those from the non-colonized sample in column 2 of table 8. 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 9 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 0.35 for the non-colonized countries and 0.40 for the colonized ones and the difference is significant at 5%.⁵⁵ It is possible that colonization itself could have been induced in the first place by the relatively high ethnic diversity of the regions. Nevertheless, the borders themselves, that is the distribution of land quality, were an outcome of the colonial intervention.

Summarizing the impact of the European colonizers on the ethnolinguistic diversity the evidence suggests that they substantially altered the ethnolinguistic endowment of the places

⁵⁵Including 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.

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.

4.4 Ethnic Diversity and Economic Outcomes: Is There a Causal Relationship?

The negative relationship between ethnic diversity and economic outcomes is well established across several studies within the economic growth and development literature (Easterly and Levine (1997), Alesina et. al. (2003), Banerjee and Somanathan (2006) among others). However, the findings of the previous section regarding the endogeneity of contemporary ethnic diversity to a country’s state history casts doubt as to whether this stylized fact may be causally interpreted.

This section is a first attempt to empirically disentangle the causal effect of ethnic diversity on economic outcomes. Specifically, the distribution of land quality across countries is employed as the source of variation to instrument for the level of contemporary ethnic diversity. However, using all the determinants of the natural level of ethnic fractionalization captured in the main specification (16) is not recommended since it is plausible that some of the summary measures of the distribution of a country’s land quality may impact economic outcomes directly and not only through the formation of more or less ethnically diverse societies. For example, the average agricultural land quality does not satisfy a priori the exclusion restriction since it may impact a country’s economic performance either directly or indirectly, through the formation of different institutions (Acemoglu et al. (2002)). In the specification presented below the excludable instruments are the *range* and the *lqqini*.⁵⁶ Also, given the discussion

⁵⁶Using the interaction of average land quality and range, *avgxrange*, as an additional excludable instrument the point estimates in the 2SLS specifications do not change. The Sargan statistics, however, become tenuous.

in the preceding sections about the impact of the European colonizers on the ethnolinguistic endowment and the strong effect of latitude on ethnic diversity, in all specifications controls for *abs_latitude* and a dummy equals 1 for countries that have been colonized by Europeans, *eucolony*, are included.

In the spirit of Alesina et al. (2007) I focus on three sets of dependent variables capturing economic, political and quality of life dimensions. For each variable the first column presents the OLS estimates and the second the 2SLS counterparts. As measures of economic outcomes I use the log income per capita in 2002 and the growth rate of income per capita from 1960 to 2000. The OLS estimates in Table 10 reproduce a well established empirical regularity. Ethnic diversity is highly negatively related with a country's economic performance. However, this strong significant impact vanishes as soon as ethnic diversity is instrumented by measures of variation in land quality, the *range* and the *lqqini* in particular. This is not a pure artifact of standard errors increasing, it is also because the point estimates on ELF decrease dramatically. For example, in the income per capita regression the 2SLS coefficient on ELF is 85% smaller than the OLS one, whereas in the income growth regression the coefficient on ELF not only decreases in magnitude but also becomes positive, though insignificant.⁵⁷ A similar picture emerges comparing the OLS to the 2SLS coefficients focusing on indexes of government effectiveness and corruption which capture how well the political system performs. The strong negative OLS coefficients on ELF drop substantially and lose significance in the 2SLS regressions. A word of caution is in order, though. Due to the inflated standard errors in the 2SLS estimation across all specifications one cannot reject that the OLS estimate of ELF is consistent. For example, the p-value of the Hausman test on the exogeneity of ELF in the log income per capita 2002 regression is 0.18.

Table 11 focuses on variables that capture the quality of life, like the percentage of the population within a country that has access to clean water and infant mortality. These variables proxy for the efficacy of the government in providing public goods. The same pattern is detected here. A strong negative and significant OLS coefficient on *ELF* is consistently supplanted by less precisely and of smaller magnitude point estimate in the 2SLS specifications.

Across all specifications in tables 10 and 11 the large *F* statistics which jointly test that the coefficients on the excluded instruments *range* and *lqqini* in the first stage equal zero, suggest that the 2SLS regressions do not suffer from weak instruments problem. Also, for the instruments to be valid they must not affect the dependent variables through any channel other

⁵⁷In the income growth regressions the log income per capita of the beginning of the period considered, here 1960, is included.

than ethnic diversity, since otherwise the effects attributed to ethnic diversity might actually be effects of other omitted channels. This restriction is tested using the standard Sargan test, whose null hypothesis is that the instruments are uncorrelated with the 2SLS residuals. The large p-values reported show that the instruments pass the test in all cases.⁵⁸

These preliminary findings shed some doubt on recent work suggesting that country-level ethnic diversity has significant adverse effects on economic outcomes. In fact, I find that instrumenting ethnic diversity using measures of variation in land quality results in measured effects that are not significantly different from zero.

5 Concluding Remarks

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

Constructing detailed data on the distribution of land quality across regions and countries I find that a larger spectrum of land qualities increases ethnic diversity. Both cross-artificial country and cross-real country regressions are examined. The former is of particular significance since the proposed relationship between the variation in land quality and ethnic diversity obtains at an arbitrary level of aggregation, explicitly avoiding the endogeneity of current countries' borders and after controlling for continental and real country fixed effects. These results are further corroborated by looking into how differences in land quality shape the extent of ethnic similarity within pairs of adjacent regions. Regional neighbors characterized by common land qualities are ethnically more similar than pairs of adjacent regions with different land endowments. Overall, the importance of the distribution of land quality in determining the natural level of ethnic diversity is a recurrent finding which reassuringly obtains across different levels of aggregation.

⁵⁸Ramcharan (2006) suggests that a higher gini of land allocation across different biome classes deters financial development by increasing sectorial concentration of total output. To the extent that *lqqini* captures such a dimension of geography i.e. inequality of land distribution across different biomes, then this would bias the 2SLS estimates (note that such bias, if any, is not detected by the Sargan statistic). Nevertheless, using only the *range* as an excludable instrument all the results are very similar.

The empirical results also show 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 fact that the natural level of fractionalization is higher among former European colonies than non-colonized countries is evidence of artificial drawing of the borders in the colonized world. Additionally, the inflated man-made component of fractionalization across the colonized countries is consistent with the widespread interference of the colonizers with the indigenous ethnic endowment.

The findings provide a stepping stone for further research. Equipped with a more substantive understanding of the origins of ethnic diversity, long standing questions among development and growth economists in which ethnic diversity plays a significant role may be readdressed. Specifically, the distinction between the natural versus the man-made components of contemporary ethnic diversity calls for a careful reinterpretation of the documented negative relationship between ethnic diversity and economic outcomes. Indeed, preliminary results suggest that instrumenting ethnic diversity with the diversity in land quality, there is no evidence in favor of a negative impact of ethnic fractionalization on contemporary economic development.

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

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

6 Appendix

Appendix A - Proofs

Properties of the migration size

Conditional on positive migration in period t , that is if either (5) or (6) 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)

$$\frac{\partial M_t^{j \rightarrow i}}{\partial \lambda_t} > 0 \quad \& \quad \frac{\partial M_t^{i \rightarrow j}}{\partial \lambda_t} < 0$$

2. decreasing in the size of the erosion, ε

$$\frac{\partial M_t^{j \rightarrow i}}{\partial \varepsilon}, \frac{\partial M_t^{i \rightarrow j}}{\partial \varepsilon} < 0$$

3. decreasing in the region specific technology of the place of origin, h_t^i, h_t^j

$$\frac{\partial M_t^{j \rightarrow i}}{\partial h_t^j}, \frac{\partial M_t^{i \rightarrow j}}{\partial h_t^i} < 0$$

Proof. Substituting (10) into (8) or (9) into (8) and differentiating produces the results \square

Proof of Lemma 1.

First, substitute in (11) the two possible realizations of the past population densities, either (9) or (10), and differentiate accordingly. Repeat the same process for (12). This completes the proof. \square

The following Lemma summarizes the cases of migration occurrences.

Lemma 2 *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*

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

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

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

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

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

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

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

Proof. Substituting the relevant ratio of the past population densities, either (9) or (10) depending on the direction of the last migration, in both (7) and (8) and solving for the required inequalities completes the proof. \square

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 2) that $\lambda_t > \lambda_s$. This implies that $\lambda_t = \lambda_{\max}$ and $\lambda_s = \lambda_{\min}$. Consequently, it follows that since in period s migration also occurred, the direction of this last migration could have only taken place from region i towards region j , i.e. $\lambda_s = \lambda_{\min}$ and $\lambda_{s-b} = \lambda_{\max}$. Similar reasoning rules out successive migration towards region i . This simplifies the analysis considerably since one may focus only on the cases of Lemma 2 where a current migration, should it take place, is always in the opposite direction of the last one. If $\lambda_t/\lambda_s = \lambda_{\min}/\lambda_{\max} < \left(h_s^j h_t^i \right)^{-\varepsilon}$ migration occurs towards region j . So, conditional on $\lambda_{\min}/\lambda_{\max}$, any regional pair characterized by higher ε and higher region specific technology, h_t^i , will experience fewer migrations towards region j . Similarly, migration occurs towards region i in period t iff $\lambda_t/\lambda_s = \lambda_{\max}/\lambda_{\min} > \left(h_s^j h_t^i \right)^\varepsilon$. It is evident that the 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 2 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 \square

Interpreting Expected Fractionalization, (15), in terms of regional land qualities:

Manipulating (15) may be rewritten as:

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

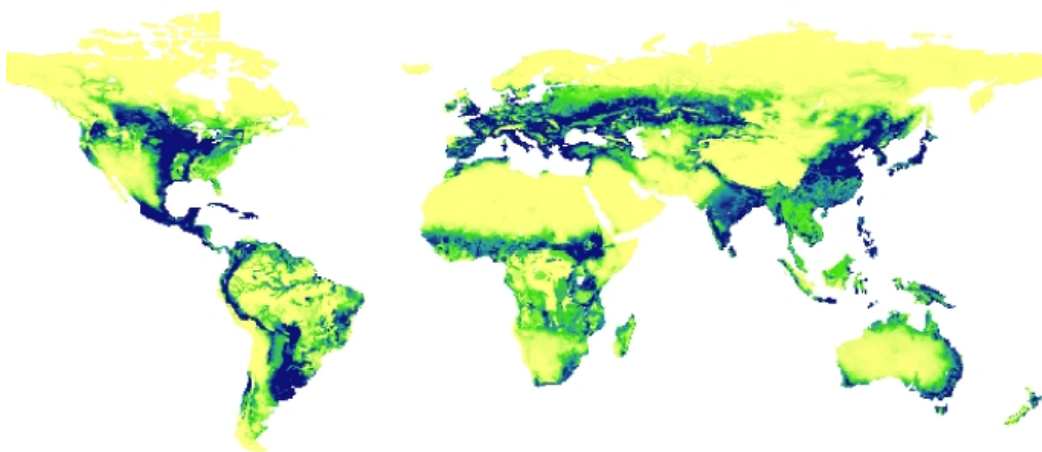
Noting (10) evaluated at $h_s^j = 1$ for example, the ratio of regional population densities is substituted accordingly and $E(ELF)$ may be rewritten as:

$$E(ELF) = (1 - f_T) \left(\frac{m^i}{2m^j} + \frac{m^j}{2m^i} + 1 \right)^{-1} \quad (17)$$

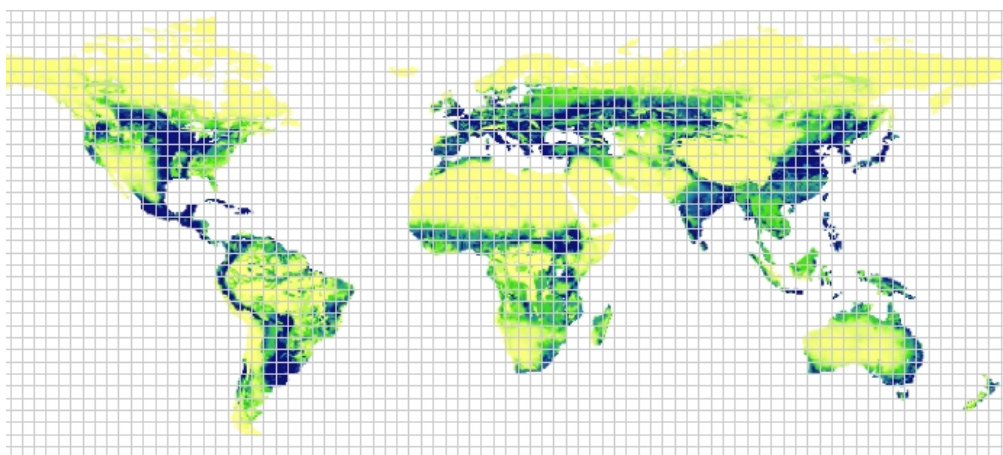
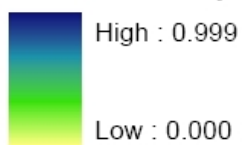
It is easy to show that conditional on the probability that two places will not share the same cultural traits, $(1 - f_T)$, a more unequal distribution of the quality of land will decrease measured fractionalization. For example, let $m^i > m^j$ then an increase in m^i and/or a decrease in m^j will decrease $E(ELF)$. This obtains by differentiating (17) 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 (17) shows, affects negatively fractionalization outcomes.

Appendix B - Maps



Global land quality



Global land quality

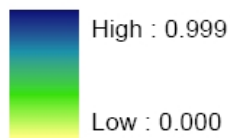


Figure 4: Example of an Artificial Country

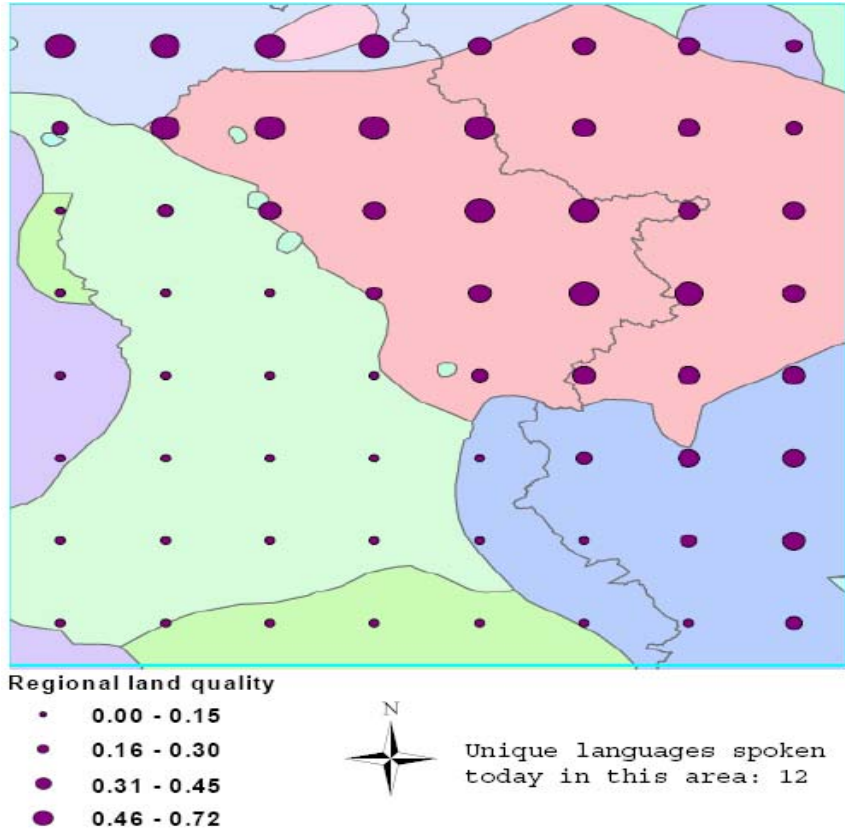


Table 1a: Summary Statistics for the Cross-Artificial Country Analysis

<i>statistics</i>	#_lang	range	avg	elev_sd	areakm2	in_country	#_cntry	water_area	sea_dist
<i>mean</i>	6.13	0.53	0.40	0.27	98.76	0.48	1.93	1.93	0.49
<i>sd</i>	6.08	0.29	0.27	0.28	60.16	0.50	1.17	2.54	0.57
<i>max</i>	30.00	1.00	0.98	2.17	196.98	1.00	8.00	19.45	2.40
<i>min</i>	1.00	0.00	0.00	0.00	0.61	0.00	1.00	0.00	0.00

#_lang: number of languages spoken with at least 1% area coverage within an artificial country;
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;
elev_sd: standard deviation of elevation measured in kilometers within artificial country; *in_country*: dummy equals 1 if artificial country's falls as a whole into a single real country; *areakm2*: size of each artificial country in thousands of sq. km.
#_cntry: number of real countries an artificial country falls into; *water_area*: area in thousand's of sq km of artificial country under water i.e. river or lake; *sea_dist*: distance of the centroid of an artificial country from the coastline;

Data Sources: See Appendix D

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

	#_lang	range	avg	elev_sd	areakm2	in_country	#_cntry	water_area	sea_dist
#_lang	1								
range	0.29	1							
avg	0.07	0.38	1						
elev_sd	0.27	0.20	-0.02	1					
areakm2	0.32	0.40	-0.06	0.11	1				
in_country	-0.27	-0.31	0.01	-0.11	-0.35	1			
#_cntry	0.34	0.32	0.01	0.19	0.36	-0.77	1		
water_area	0.04	0.15	-0.13	-0.04	0.37	-0.25	0.18	1	
sea_dist	0.02	0.25	-0.11	0.09	0.42	-0.20	0.13	0.26	1

See variables' description in Table 1a

Table 2: Main Specification and Robustness in Cross-Artificial Country Regressions

Dependent Variable: Log Number of Languages Spoken						
	OLS	OLS	OLS	OLS	OLS	OLS
	Baseline	Geographical Controls	Continental and Country fixed effects	Average Land Quality and Interaction	Regions Within Countries	Tropics and Non-Tropics
	(1)	(2)	(3)	(4)	(5)	(6)
range	0.952 (4.66) ^{***}	0.488 (2.27) ^{**}	0.524 (3.88) ^{***}	0.545 (3.03) ^{***}	0.460 (2.85) ^{***}	0.345 (2.84) ^{***}
avg		-0.104 (0.50)	-0.184 (1.13)	-0.160 (0.70)	-0.229 (1.15)	-0.183 (1.15)
avgxrange				-0.059 (0.17)		
areakm2		0.002 (1.48)	0.002 (1.83) [*]	0.002 (1.79) [*]	0.004 (3.10) ^{***}	0.002 (1.73) [*]
abs_lat		-0.027 (5.95) ^{***}	-0.026 (3.17) ^{***}	-0.026 (3.18) ^{***}	-0.026 (2.68) ^{***}	-0.022 (2.56) ^{**}
elev_sd		0.498 (2.18) ^{**}	0.684 (3.07) ^{***}	0.682 (3.02) ^{***}	1.320 (4.21) ^{***}	0.702 (3.26) ^{***}
in_country		-0.172 (1.98) ^{**}	-0.094 (0.95)	-0.094 (0.95)		-0.101 (1.03)
#_cntry		0.170 (5.33)	0.161 (2.47) ^{**}	0.161 (2.45) ^{**}		0.157 (2.34) ^{**}
sea_dist		0.068 (0.64)	-0.038 (0.42)	-0.038 (0.42)	-0.160 (1.37)	-0.029 (0.37)
water_area		0.017 (1.15)	-0.009 (0.87)	-0.009 (0.88)	-0.013 (0.57)	-0.006 (0.64)
tropics						-0.079 (0.54)
tropicsxrange						0.633 (2.79) ^{***}
Observations	548	548	548	548	263	548
R-squared	0.09	0.52	0.81	0.81	0.78	0.82

OLS regressions with absolute value of t statistics in parentheses;

Standard errors are corrected for spatial correlation following Conley (1999);

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

Specifications (3), (4), (5) and (6) include 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 x range*: the interaction between range and avg; *elev_sd*: standard deviation of elevation within artificial country measured in kilometers; *abs_lat*: artificial country's latitudinal distance from the equator; *areakm2*: size of each artificial country in thousands of sq. km. *in_country*: dummy variable equals 1 if an artificial country falls completely within a real country; *#_cntry*: number of real countries in which the artificial country belongs to; *tropics*: dummy equals 1 if the average absolute latitude of an artificial country is less than 23.5, zero otherwise; *tropics x range*: the interaction between range and tropics; *water_area*: area in thousand's of sq km of artificial country; under water i.e. river or lake; *sea_dist*: distance of the centroid of an artificial country from the coastline;

Table 3a: Summary Statistics for the Pairwise Analysis of Adjacent Regions

<i>statistics</i>	pct_comlang	lqdiff	eldiff	#_lang	sea_dist	same_cntry	#_body_waters	diff_langarea
<i>mean</i>	0.77	0.08	0.15	2.19	0.70	0.94	4.33	0.26
<i>sd</i>	0.30	0.12	0.25	2.22	0.59	0.24	7.70	0.49
<i>max</i>	1.00	0.99	4.00	67.00	2.68	1.00	176.50	3.11
<i>min</i>	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00

pct_comlang: number of common languages divided by the total number of unique languages spoken within a pair of adjacent regions; *lqdiff*: absolute difference in land quality within a pair of adjacent regions; *eldiff*: absolute difference in elevation in km's within a pair of adjacent regions; *sea_dist*: distance from the coastline in 1000s of km of a regional pair; *same_country*: dummy equals 1 if a regional pair belongs to the same country; *#_body_waters*: number of distinct body waters within a regional pair; *#_lang_pair*: total number of unique languages spoken within a pair; *diff_langarea*: difference in area of linguistic coverage between the regional neighbors in 1000 of sq kilometers;

Table 3b: The Correlation Matrix for the Pairwise Analysis of Adjacent Regions

	pct_comlang	lqdiff	eldiff	#_lang	sea_dist	same_cntry	#_body_waters	diff_langarea
pct_comlang	1.00							
lqdiff	-0.15	1.00						
eldiff	-0.15	0.25	1.00					
#_lang	-0.60	0.15	0.16	1.00				
dist_sea	0.08	0.00	0.02	-0.10	1.00			
same_cntry	0.15	-0.08	-0.06	-0.15	0.00	1.00		
#_body_waters	0.10	-0.07	-0.14	-0.10	-0.10	0.02	1.00	
diff_langarea	-0.09	0.00	0.08	0.03	-0.06	-0.07	-0.02	1.00

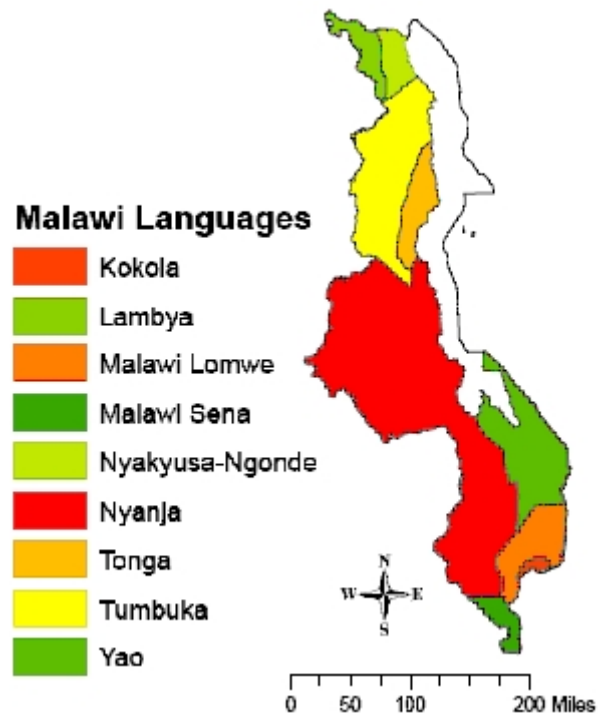
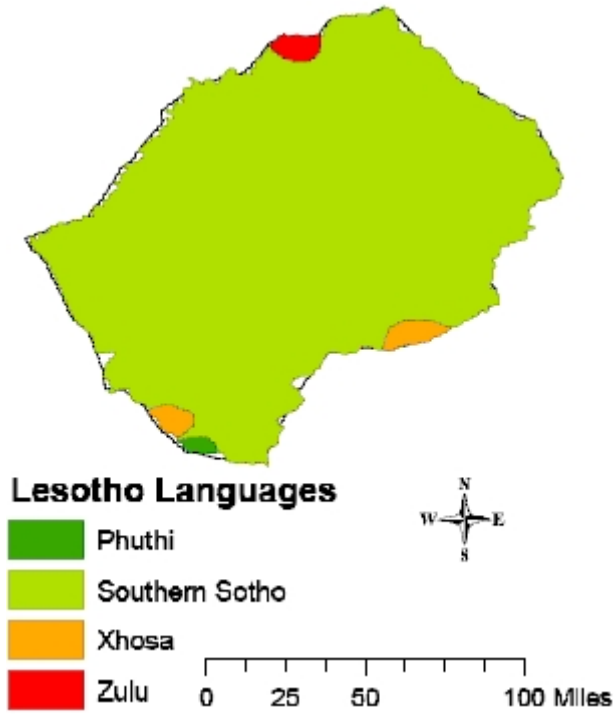
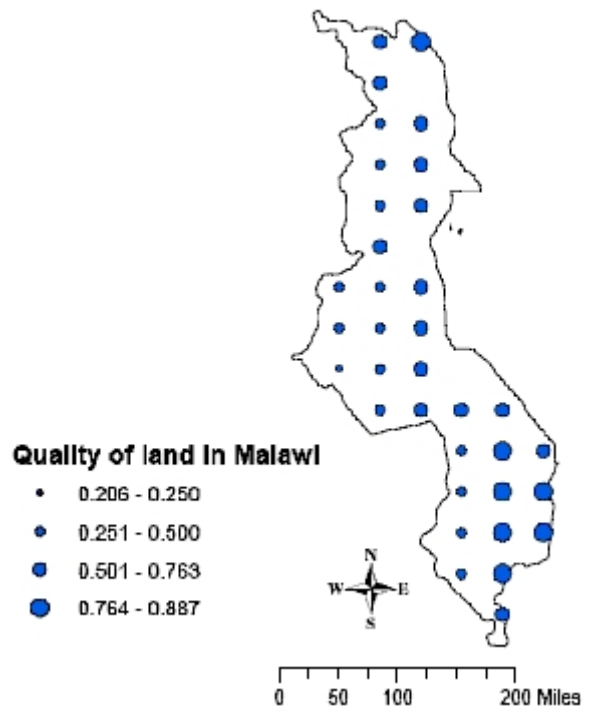
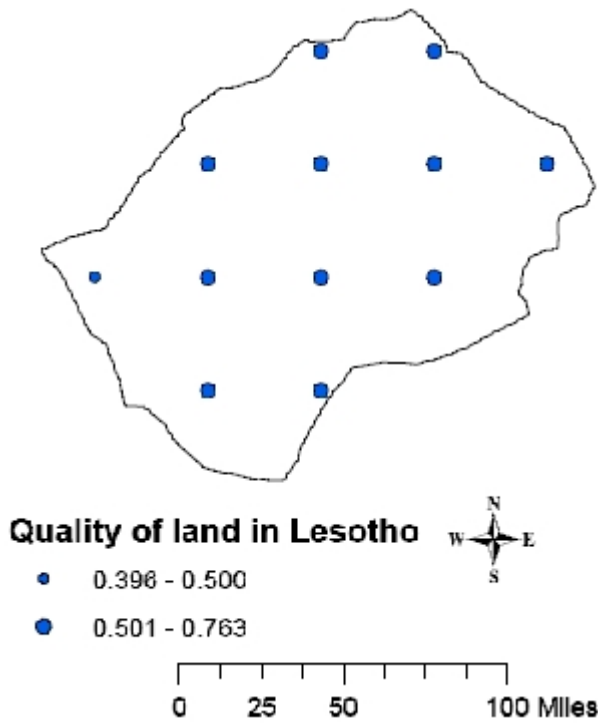
pct_comlang: number of common languages divided by the total number of unique languages spoken within a pair of adjacent regions; *lqdiff*: absolute difference in land quality within a pair of adjacent regions; *eldiff*: absolute difference in elevation in km's within a pair of adjacent regions; *sea_dist*: distance from the coastline in 1000s of km of a regional pair; *same_country*: dummy equals 1 if a regional pair belongs to the same country; *#_body_waters*: number of distinct body waters within a regional pair; *#_lang_pair*: total number of unique languages spoken within a pair; *diff_langarea*: difference in area of linguistic coverage between the regional neighbors in 1000 of sq kilometers;

Table 4: Main Specification and Robustness in the Pairwise Analysis of Adjacent Regions

Dependent Variable: Percentage of Common Languages				
	OLS	OLS	OLS	OLS
	Baseline	Country and Continental fixed effects	Baseline within India	Marginal Effects by Continent
	(1)	(2)	(3)	(4)
<i>lqdiff</i>	-0.163 (7.13) ^{***}	-0.122 (15.66) ^{***}	-0.172 (2.13) ^{**}	-0.266 (13.04) ^{**}
<i>eldiff</i>	-0.120 (9.48) ^{***}	0.095 (22.29) ^{***}	-0.127 (3.53) ^{***}	-0.090 (21.19) ^{***}
<i>abs_lat</i>	0.005 (20.20) ^{***}	0.003 (16.70) ^{***}	-0.010 (2.33) ^{**}	0.004 (17.30) ^{***}
<i>elev</i>	0.002 (0.37)	0.001 (0.50)	-0.042 (3.57) ^{***}	-0.001 (0.55)
<i>land_quality</i>	-0.029 (1.80) [*]	-0.027 (5.97) ^{***}	-0.311 (3.70) ^{***}	-0.025 (5.44) ^{***}
<i>sea_dist</i>	0.019 (2.51) ^{**}	0.011 (4.83) ^{***}	0.082 (1.15)	0.009 (3.97) ^{***}
<i>same_country</i>	0.132 (14.96) ^{***}	0.108 (30.37) ^{***}		0.108 (30.26) ^{***}
<i>#_body_waters</i>	-0.006 (0.15)	0.0002 (1.95) [*]	0.006 (5.78) ^{***}	0.0002 (1.70) [*]
<i>diff_langarea</i>	-0.013 (2.41) ^{**}	-0.009 (4.35) ^{***}	-0.042 (1.46)	-0.008 (4.25) ^{***}
<i>lqdiff_asia</i>				0.061 (2.54) ^{**}
<i>lqdiff_americas</i>				0.287 (11.50) ^{***}
<i>lqdiff_europe</i>				0.204 (8.17) ^{***}
<i>lqdiff_pacific</i>				-0.275 (3.72) ^{***}
constant	0.475 (27.05) ^{***}	0.368 (10.97) ^{***}	0.958 (10.41) ^{***}	0.795 (7.03) ^{***}
R-squared	0.18	0.28	0.21	0.29
Observations	159358	159358	4138	159358

OLS regressions with absolute value of t statistics in parentheses; In (1), (3) standard errors are corrected for spatial autocorrelation following Conley (1999); In (2) and (4) standard errors are clustered at the level of each individual region; * significant at 10%; ** significant at 5%; *** significant at 1%; (2), (4) include both continental and country and fixed effects for each region within a pair; (5) allows for differential marginal effect of the difference in land quality for pairs of different continents; Africa is the omitted continent; *pct_comlang*: number of common languages divided by the total number of unique languages spoken within a pair of adjacent regions; *lqdiff*: absolute difference in land quality within a pair of adjacent regions; *eldiff*: absolute difference in elevation in km's within a pair of adjacent regions; *sea_dist*: distance from the coastline in 1000s of km of a regional pair; *same_country*: dummy equals 1 if a regional pair belongs to the same country; *#_body_waters*: number of distinct body waters within a regional pair; *elev*: average elevation within a regional pair; *land_quality*: average land quality in a pair; *diff_langarea*: difference in area of linguistic coverage between the regional neighbors in 1000 of sq kilometers;

Appendix C - Maps



Upper map land quality; lower map languages

Upper map land quality; lower map languages

Table 5a: Summary statistics for Cross-Real Country Analysis

<i>statistics</i>	ELF	range	avg	avgxrange	lqqini	lpd1500	elev_sd	yreentry
<i>mean</i>	0.410	0.697	0.395	0.282	0.364	0.906	0.409	1927.120
<i>sd</i>	0.281	0.265	0.249	0.182	0.225	1.504	0.348	56.920
<i>max</i>	0.925	0.990	0.958	0.787	0.859	3.842	1.867	1993.000
<i>min</i>	0.001	0.002	0.003	0.000	0.028	-3.817	0.019	1816.000

Table 5b: Correlation Matrix for Cross-Real Country Analysis

	ELF	range	avg	avgxrange	lqqini	lpd1500	elev_sd	yreentry
ELF	1.00							
range	0.20	1.00						
avg	-0.21	0.15	1.00					
avgxrange	-0.12	0.59	0.78	1.00				
lqqini	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	
yreentry	0.36	-0.31	-0.20	-0.30	-0.06	-0.09	-0.21	1.00

ELF: ethnolinguistic fractionalization; *range*: 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 x range*: the interaction between range and avg; *lqqini*: 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 measured in kilometers, country; *yreentry*: year when each country gained independence as a modern state. Data Sources: Appendix D;

Table 6: Main Specification for Cross-Real Country Analysis

<u>Dependent Variable: Ethnolinguistic Fractionalization (ELF)</u>			
	(1)	(2)	(3)
<i>range</i>	0.203 (2.43)**	0.455 (3.55)***	0.889 (5.73)***
<i>avg</i>		0.076 (0.42)	-0.283 (1.49)
<i>avgxrange</i>		-0.659 (2.18)**	-1.21 (3.90)***
<i>lqqini</i>			-0.874 (4.45)***
Adj R-squared	0.03	0.11	0.21
Observations	147	147	147

OLS regression with absolute value of t statistics in parentheses

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

range: 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 x range*: the interaction between range

and avg, *lqqini*: the gini of coefficient of land quality within country; Data Sources: Appendix D;

Table 7: Robustness Checks for Cross-Real Country Analysis

Dependent Variable: Ethnolinguistic Fractionalization (ELF)						
	Baseline	Continental fixed effects	Additional geography	Historical controls	Full specification	Indigenous >75%
	(1)	(2)	(3)	(4)	(5)	(6)
range	0.889 (5.73)***	0.641 (4.62)***	0.692 (4.71)***	0.889 (5.92)***	0.668 (4.87)***	1.032 (5.65)***
avg	-0.283 (1.49)	-0.101 (0.58)	-0.261 (1.54)	-0.07 (0.37)	0.092 (0.52)	-0.241 (0.85)
avgxrange	-1.21 (3.90)***	-0.653 (2.28)**	-0.834 (2.83)***	-1.043 (3.39)***	-0.765 (2.75)***	-1.422 (3.73)***
lqqini	-0.874 (4.45)***	-0.416 (2.30)**	-0.742 (3.91)***	-0.626 (3.27)***	-0.508 (2.81)***	-0.921 (3.78)***
reg_ssa		0.255 (5.73)***			0.149 (2.48)**	
reg_lac		-0.102 (1.79)*			-0.18 (2.18)**	
reg_we		-0.172 (2.70)***			0.075 -0.87	
lpd1500				-0.02 (1.36)	-0.034 (2.03)**	
yrentry				0.002 (4.73)***	0.001 (1.86)*	
abs_lat			-0.005 (4.27)***		-0.004 (2.23)**	
areakm2			-0.002 (1.61)		-0.001 (0.61)	
distcr			0.202 (4.14)***		0.105 (2.18)**	
elev_sd			0.037 (0.65)		0.141 (2.37)**	
Adj R-squared	0.21	0.42	0.38	0.35	0.5	0.25
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%

range: 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 x range*: the interaction between range and avg, *lqqini*: 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 than 75% of the current population

Data Sources: See Appendix D

Table 8: Colonization and Artificial Fractionalization

Dependent Variable: Ethnolinguistic Fractionalization (ELF)

	Colonized by Europeans	Non-Colonized by Europeans
	(1)	(2)
range	0.792 (3.80) ^{***}	1.09 (4.02) ^{***}
avg	-0.324 (1.37)	-0.34 (0.97)
avgxrange	-0.892 (1.83) [*]	1.323 (2.79) ^{**}
lqqini	-0.813 (3.14) ^{***}	-0.952 (3.16) ^{**}
Adj R-squared	0.16	0.32
Observations	93	47

OLS regressions with absolute value of t statistics in parentheses

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

range: 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 x range*:

the interaction between range and avg; *lqqini*: the gini of coefficient of land quality

within country; Colonized: colonized by Europeans after 1500 AD; Non-Colonized:

not colonized by Europeans after 1500 AD excluding the colonizers;

Data Sources: See Appendix D

Table 9: Colonization and Natural Fractionalization

nat_ETF if colonized: **0.40**

nat_ETF if not colonized: **0.35**

Pr(T < t) = **0.04**

nat_ETF: natural level of fractionalization computed using the predicted values
of regression (2) in Table 8; *Colonized*: colonized by Europeans after 1500

Non-Colonized: not colonized by Europeans after 1500 excluding the colonizers;

Table 10: Instrumental Variable Regressions for Economic and Political Variables

	<u>Dep. Var: log income per capita, 2002</u>		<u>Dep. Var: log income per capita growth rate, 1960-2000</u>		<u>Dep. Var: Government Effectiveness 1996-2005</u>		<u>Dep. Var: Corruption 1996-2005</u>	
	<i>OLS</i>	<i>2SLS</i>	<i>OLS</i>	<i>2SLS</i>	<i>OLS</i>	<i>2SLS</i>	<i>OLS</i>	<i>2SLS</i>
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ELF	-1.045 (0.29)***	-0.147 (0.77)	-1.130 (0.58)*	0.319 (1.99)	-0.545 (0.27)**	0.051 (0.73)	-0.764 (0.27)***	-0.546 (0.70)
avg	-0.729 (0.42)*	-0.364 (0.53)	0.145 (0.91)	0.820 (1.26)	-0.610 (0.44)	-0.382 (0.50)	-0.675 (0.44)	-0.591 (0.50)
avgxrange	1.299 (0.59)**	1.113 (0.63)*	-0.095 (1.35)	-0.510 (1.38)	1.086 (0.60)*	0.953 (0.62)	0.719 (0.60)	0.670 (0.61)
abs_lat	0.043 (0.006)***	0.051 (0.009)***	0.040 (0.02)***	0.052 (0.02)**	0.041 (0.006)***	0.046 (0.009)***	0.041 (0.006)***	0.043 (0.009)***
eucolony	0.237 (0.22)	0.331 (0.24)	-0.902 (0.48)*	-0.845 (0.48)*	0.596 (0.22)***	0.674 (0.23)***	0.635 (0.22)***	0.664 (0.23)***
lngdppc1960			-0.446 (0.21)**	-0.379 (0.22)*				
F-test of excluded instruments		9.97***		4.00**		11.52***		11.52***
Sargan statistic (p-value)		0.55		0.40		0.56		0.44
Observations	137	137	95	95	147	147	147	147
R-squared	0.56		0.36		0.38		0.40	

In the OLS, 2SLS regressions absolute values of standard errors reported in parentheses;

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

In the 2SLS regressions the excluded instruments are the *range* and the *lqqini*; The Over-Identification Test is based on the Sargan statistic, distributed as Chi-squared with one degree of freedom; The p-value of the F statistic reported jointly tests that the coefficients on *range* and *lqqini* in the first stage equal zero;

lngdppc1960: log income per capita in 1960; *range*: spectrum of land qualities within a 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 x range*: the interaction between range and avg; abs_lat: average latitudinal distance from the equator of each country; *lqqini*: the gini coefficient of land quality within country; eucolony: dummy for countries colonized by Europeans after 1500 AD; **ELF**: level of ethnolinguistic fractionalization within a country;

Data Sources: See Appendix D

Table 11: Instrumental Variable Regressions for Quality of Life Variables

	<u>Dep. Var: Access to Clean water 2000</u>		<u>Dep. Var: Infant Mortality 2000</u>	
	<i>OLS</i>	<i>2SLS</i>	<i>OLS</i>	<i>2SLS</i>
	(1)	(2)	(3)	(4)
ELF	-12.204 (5.50)**	5.279 (16.00)	50.157 (9.99)***	30.133 (27.65)
avg	5.127 (8.99)	12.302 (10.98)	2.743 (16.14)	-4.618 (18.47)
avgxrange	10.88 (12.36)	7.538 (13.00)	-45.364 (21.91)**	-40.759 (22.64)*
abs_lat	0.645 (0.13)***	0.816 (0.20)***	-1.313 (0.24)***	-1.51 (0.35)***
eucolony	5.055 (4.64)	8.097 (5.40)	-8.085 (8.09)	-10.965 (8.84)
F-test of excluded instruments	8.64***		10.08***	
Sargan statistic (p-value)	0.76		0.89	
Observations	134	134	144	144
R-squared	0.38		0.54	

In the OLS, 2SLS regressions absolute values of standard errors reported in parentheses;

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

In the 2SLS regressions the excluded instruments are the *range* and the *lqqini*; The Over-Identification Test is based on the Sargan statistic, distributed as Chi-squared with one degree of freedom; The p-value of the F statistic reported jointly tests that the coefficients on *range* and *lqqini* in the first stage equal zero;

range: spectrum of land qualities within a 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; avgxrange: the interaction between range and avg; abs_lat: average latitudinal distance from the equator of each country; lqqini: the gini coefficient of land quality within country; euclony: dummy for countries colonized by Europeans after 1500 AD; **ELF**: level of ethnolinguistic fractionalization within a country;

Data Sources: See Appendix D

Appendix D - Data Sources

Geographical Variables

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

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

eldiff: difference in elevation between adjacent regions

Source: see **elev_sd**

areakm2: land area (km²)

Source: Center for International Development, CID.⁵⁹ For the cross-artificial country analysis the area within an artificial country is constructed by the author using ArcGIS. In the calculation are considered only areas over which both language and land quality data are available.

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

Source: Center for International Development, CID.

abs_lat: Absolute Latitudinal Distance from the Equator.

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

in_country: dummy variable equals 1 if an artificial country falls completely within a real country; constructed by the author using ArcGIS.

same_country: dummy variable equals 1 if a regional pair falls completely within a real country; constructed by the author using ArcGIS.

#_cntry: number of real countries in which an artificial country belongs to; constructed by the author using ArcGIS.

water_area: area in thousand's of sq km. of an artificial country; under water i.e. river or lake;

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

⁵⁹All geographical data from CID are available at: <http://www.ksg.harvard.edu/CID>

#_body_waters: number of distinct body waters within a regional pair.

Source: see **water_area**.

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

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

Historical Variables

ELF: level of ethnolinguistic fractionalization within a country.

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

lpd1500: log population density in 1500.

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

yrentry: year a country achieved independence.

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

indigenous: percentage of indigenous population as of 1500 still comprising more than 75% of the current population's composition.

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

eucolony: 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.

Economic Variables

log income per capita for 1960 and 2000; log income per capita growth rate 1960-2000; Source: Summers-Heston updated with World Bank per capita growth rates.

Corruption and Government Effectiveness indexes are 10 year averages, 1996-2005. Source: Kaufmann-Kraay indices of institutions for 2004 (increase means better institutions).

Percentage of the Population with Access to Clean water, 2004; Infant Mortality, 2000. Source: WDI.

Appendix E - Examples of Ethnic Groups in Kenya

Ethnic Groups in Kenya

The theoretical premise of this study is that ethnic groups are endowments of specific human capital and this specificity derives from the land quality in which an ethnic group resides. This section presents anecdotal evidence in support of the hypothesis. The graph below plots the distribution of land quality within ethnic groups in Kenya with similar spatial extent (a group of those examined here spans on average 25 regions of 0.5 degrees latitude by 0.5 degrees longitude). Land suitability for agriculture (described in the empirical section) is in the horizontal axis, whereas the vertical axis displays the name of each group. The boxes map the interquartile range of land quality with the dots representing regions with land quality more than three standard deviations further from the mean.



Distribution of Land Quality within ethnic groups in Kenya

A careful inspection of the box plots reveals that ethnic groups are not randomly dispersed across regional land qualities within Kenya. In fact, they seem to cluster in territories of distinct and homogenous land endowments. The Samburu people, the Orma and the Garreh-Ajuran are all exclusively located at low levels of land quality where agriculture is almost impossible to maintain.⁶⁰ The Samburu are semi-nomadic pastoralists who herd mainly cattle but also keep sheep, goats and camels, see Pavitt (2001). The Orma are semi-nomadic shepherds and the Garreh-Ajuran are semi-nomadic pastoralists. These groups have the human capital to undertake the productive activities which are optimal for the places in which they are located. On the other hand, the Gikuyu and the Kalenjin are concentrated in territories of high land quality and they are mainly engaged in agriculture producing: sorghum, millet, beans, sweet potatoes, maize, potatoes, cassava, bananas, sugarcane, yams, fruit, tobacco and

⁶⁰The description of the main productive activities of each ethnic group, unless otherwise noted, comes from the entries found in the Ethnologue website, (<http://www.ethnologue.com/>).

coffee. The Kamba are often found in different professions: some are agriculturalists others hunters, and a large number are pastoralists. This according to the theory is an outcome of the fact that Kamba reside in intermediate levels of land quality which may sustain different optimal activities.

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

References

- [1] Acemoglu, D., Johnson, S., and Robinson, J., (2002) “Reversal Of Fortune: Geography And Institutions In The Making Of The Modern World Income Distribution”, Quarterly Journal of Economics, v107.
- [2] Ahlerup, P. and Olsson, O., (2007) “The Roots of Ethnic Diversity”, University of Gothenburg, mimeo.
- [3] Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S. and R. Wacziarg (2003) “Fractionalization”, Journal of Economic Growth, v8, pp. 155-194.
- [4] Alesina, A., Easterly, W., Matuszeski, M., (2006) ”Artificial States”, Working Paper 12328.
- [5] Alesina, A., and Spolaore, E., (1997) “On the Number and Size of Nations”, Quarterly Journal of Economics. v112, pp. 1027-1056.
- [6] Ashraf, Q., and Michalopoulos, S., (2007) “The Climatic Origins of the Neolithic Revolution: A Theory of Long-Run Development via Climate-Induced Technological Progress”, SSRN: <http://ssrn.com/abstract=903847>
- [7] Atlas Narodov Mira (Atlas of the People of the World) (1964), Moscow: Glavnoe Upravlenie Geodezii i Kartograi, Bruck, S.I., and V.S. Apenchenko (eds.).
- [8] Banerjee, A., and Somanathan R., (2006) “The Political Economy of Public Goods: Some Evidence from India”, Journal of Development Economics, v82, Issue 2, pp. 287-314.
- [9] Barth, F., (1969) “Ethnic Groups and Boundaries: The Social Organization of Cultural Difference”, Boston: Little, Brown.
- [10] Bellwood, P., (2001) “Early Agriculturalist Population Diasporas? Farming, Languages, and Genes”, Annual Review of Anthropology, v30., pp. 181-207.
- [11] Bockstette, V., Chanda, A., Putterman, L., (2002) “States and Markets: the Advantage of an Early Start”, Journal of Economic Growth, v7, Issue 4, pp. 347-369.
- [12] Boyd, R., and P.J. Richardson., (1985) “Culture and the Evolutionary Process”, University of Chicago Press, Chicago.
- [13] Boserup, E., (1965). “The Conditions of Agricultural Progress”, Aldine Publishing Company, Chicago.
- [14] Botticini, M., Eckstein, Z., (2005) “From Farmers to Merchants, Voluntary Conversions and Diaspora: A Human Capital Interpretation of Jewish History”, Journal of Economic History, v65, Issue 4, pp. 922-48.
- [15] Caselli, F., and Coleman, II. W., (2006) “On the Theory of Ethnic Conflict”, London School of Economics, mimeo.
- [16] 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.

- [17] Conley, T.G., (1999) "GMM Estimation with Cross Sectional Dependence", *Journal of Econometrics*, v92, Issue 1, pp. 1–45.
- [18] Curtin, P., (1984) "Cross-Cultural Trade in World History", Cambridge: Cambridge University Press.
- [19] Darwin, C., (1839) "The Voyage of the Beagle", Available at: http://www.online-literature.com/darwin/voyage_beagle/
- [20] Easterly, W., and Levine, R., (1997) "Africa's Growth Tragedy: Policies and Ethnic Divisions", *Quarterly Journal of Economics*, v112, Issue 4, pp. 1203-50.
- [21] Englebert, P., Tarango, S., and Carter, M., (2002) "Dismemberment and Suffocation: A Contribution to the Debate on African Boundaries", *Comparative Political Studies*. v35, Issue 10, pp. 1093-1118.
- [22] Esteban, J., Ray., D., (2007) "On the Salience of Ethnic Conflict", New York University, mimeo.
- [23] Fearon, J., (2003) "Ethnic Structure and Cultural Diversity by Country", *Journal of Economic Growth*, v8, Issue 2, pp. 195-222.
- [24] Fearon, J., Laitin, D. (2003) "Ethnicity, Insurgency and Civil War", *American Political Science Review*, v97, pp. 75-90.
- [25] Galor, O. and Weil, D.N., (2000), "Population, Technology and Growth: From the Malthusian Regime to the Demographic Transition", *American Economic Review* v110, pp. 806-828.
- [26] Geertz, C., (1967) "Old Societies and New States: The Quest for Modernity in Asia and Africa", New York: Free Press.
- [27] Gray, R. Atkinson, Q. (2003), "Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin", *Nature*, v426, pp. 435-439.
- [28] Hale, H., (2004) "Explaining Ethnicity", *Comparative Political Studies* 2004, pp. 37- 458.
- [29] Herbst, J., (2002) "State and Power in Africa", Princeton, NJ: Princeton University Press.
- [30] La Porta, R., Lopez de Silanes, F., Shleifer, A., Vishny, R., (1999) "The Quality of Government", *Journal of Law Economics and Organization*, pp. 315-388.
- [31] Miguel, E., and Posner, D., N., (2006) "Sources of Ethnic Identification in Africa", mimeo.
- [32] Montalvo, and Reynal-Querol, (2005), "Ethnic Polarization, Potential Conflict and Civil War", *American Economic Review*, 2005.
- [33] Nettle, D., (1999) "Linguistic Diversity", Oxford University Press, Oxford.
- [34] Nichols, J., (1997) "Modeling Ancient Population Structures and Movement in Linguistics", *Annual Review of Anthropology*, v26., pp. 359-384.
- [35] Nichols, J., (1997b) "Chechen Phonology", In *Phonologies of Asia and Africa*, ed. AS Kaye, P Daniels, pp. 941-71. Bloomington, Ind:Eisenbrauns.

- [36] Olson, P., (1990) “Struggle for the Land: Indigenous Insight and Industrial Empire in the Semiarid World”, University of Nebraska Press.
- [37] Pavitt, N., (2001) “Samburu”, Kyle Cathie Limited.
- [38] Ramankutty, N., J.A., Foley, J., Norman, and K., McSweeney, (2002) “The Global Distribution of Cultivable Lands: Current Patterns and Sensitivity to Possible Climate Change”, *Global Ecology and Biogeography*, v11, pp. 377–392.
- [39] Ramcharan, R., (2006) “Does Economic Diversification Lead to Financial Development? Evidence From Topography”, IMF Working Paper 06/35.
- [40] Rao V., and Ban R. (2007) “The Political Construction of Caste in South India”, World Bank, mimeo.
- [41] Renfrew, C., (1992) “Archaeology, Genetics and Linguistic Diversity”, *Man, New Series*, v27, Issue 3, pp. 445-478.
- [42] Renfrew, C., (2000) “At the Edge of Knowability: Towards a Prehistory of Languages”, *Cambridge Archaeological Journal*, v10, Issue 1, 7-34.
- [43] Rosenzweig, M. L., (1995) “Species Diversity in Space and Time”, Cambridge University Press, New York, NY.
- [44] Spolaore, E., and Wacziarg R., (2006) “The Diffusion of Development”, NBER Working Paper #12153.
- [45] Williamson J., (2006) “Poverty Traps Distance and Diversity: The Migration Connection”, NBER Working Paper #12549.