

Collegio Carlo Alberto

The Origins of Ethnolinguistic Diversity

Stelios Michalopoulos

Working Paper No. 110

May 2009

www.carloalberto.org

The Origins of Ethnolinguistic Diversity¹

Stelios Michalopoulos²

First Draft: April, 2007

This Draft: February, 2009³

¹Part of this research circulated earlier under the title "Ethnolinguistic Diversity: Origins and Implications." I am indebted to Oded Galor for his constant advice and mentorship. Daron Acemoglu, Roland Benabou, Matteo Cervellati, James Fearon, Andrew Foster, Ioanna Grypari, Peter Howitt, Masayuki Kudamatsu, Nippe Lagerlof, David Laitin, Ashley Lester, Ross Levine, Glenn Loury, Ignacio Palacios-Huerta, Stephen Ross, Yona Rubinstein, Francesco Trebbi and David Weil provided valuable comments. I would like, also, to thank the participants at the 2007 NEUDC Conference, the 2007 LAMES Meetings in Bogotá, the 2007 NBER Summer Institute on Income Inequality and Growth and the 2008 Ethnicity Conference in Budapest, as well as the seminar participants at Brown University, Chicago GSB, Collegio Carlo Alberto, Dartmouth College, EIEF, IIES, Princeton University, Stanford GSB, Tufts University, UCL, University of Bologna, University of Copenhagen, University of Connecticut, University of Cyprus, University of Gothenburg, University of Houston, Warwick University, Yale University for the useful discussions. Lynn Carlsson's ArcGis expertise proved of invaluable assistance. Financial support from the Watson Institute's research project "Income Distribution across and within Countries" at Brown University is gratefully acknowledged.

²Tufts University.

³ c 2009 by Stelios Michalopoulos. Any opinions expressed here are those of the authors and not those of the Collegio Carlo Alberto.

Abstract

This research examines the economic origins of ethnolinguistic diversity. The empirical analysis constructs detailed data on the distribution of land quality and elevation across contiguous regions, virtual and real countries, and shows that variation in elevation and land quality has contributed significantly to the emergence and persistence of ethnic fractionalization. The empirical and historical evidence is consistent with the proposed hypothesis, 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.

JEL Classification: O11, O12, O15, O33, O40, J20, J24.

Keywords: Ethnic Diversity, Geography, Technological Progress, Human Capital, Colonization.

1 Introduction

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

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

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

This research proposes a mechanism through which heterogeneous land endowments shaped ethnic diversity in a stage of development when land was the single most important

¹See Hale (2004).

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

³Michalopoulos (2009) employs the proposed framework to uncover the causal impact of ethnolinguistic diversity on economic performance across regions and countries.

factor of production. Particularly, it suggests that differences in land characteristics across regions gave rise to location specific human capital,⁴ diminishing population mobility and leading to the formation of localized ethnicities. On the other hand, homogeneous land endowments facilitated population mixing, resulting eventually in the formation of a common ethnolinguistic identity.

The link between variable land endowments and ethnic diversity has a striking parallel to the relationship between biodiversity and variation within species. Darwin's observations that ecologically diverse places would bring about and sustain variation within finches is of particular relevance.⁵ Along the same lines, this study argues that variation in elevation and land qualities across regions is the ultimate cause of the emergence and persistence of ethnic diversity.

In the empirical section I employ new data on land's agricultural suitability at a resolution of 0.5 degrees latitude by 0.5 degrees longitude to construct the distribution of land quality at a regional and country level. Such disaggregated level data, never before used in an economic application, allow for the econometric analysis to be conducted at various levels of aggregation. Specifically, to mitigate the problem of endogenous borders, inherent to the literature on cross-country regressions, I arbitrarily divide the world into geographical entities of a fixed size, called virtual countries. Consistent with the hypothesis, I find that ethnic diversity, measured by the number of languages spoken in each virtual country, is systematically related to the underlying heterogeneity in land quality for agriculture. At the same time, the empirical analysis reveals that regions with more variable terrain sustain more ethnically fragmented societies. Overall, geographically diverse territories, that is places characterized by a wide spectrum of land qualities and variable altitudes, give rise and support more ethnic groups. The findings are robust to the inclusion of continental and country fixed effects which effectively capture any systematic elements related to the state and continental histories of these geographical units.

Taking further advantage of the information on where ethnic groups are located, I am able to sharply focus on the determinants of ethnic diversity at the local level. In particular, focusing on pairs of adjacent regions, each having a size of 0.5 degrees latitude by 0.5 degrees longitude, I find that the differences in land quality and elevation between any two adjacent regions 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

⁴Region specific human capital should be thought of as encompassing both the technical knowledge necessary to be productive in a given region and the capacity of the immune system to adapt to the local disease vectors.

⁵Darwin (Originally 1839, Reprinted in 2006) observed that a certain ecological niche was giving rise to an optimal shape of the finches' beaks.

and elevation between adjacent regions is a significant determinant of local ethnic diversity and (ii) the arrangement of a given heterogeneous land endowment in space matters in determining the degree of the overall cultural heterogeneity.

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

Historical accidents have also shaped contemporary fractionalization outcomes. The European colonization after the 15th century, for example, is an obvious candidate. Europeans substantially affected the ethnolinguistic spectrum of the places they colonized. In particular, their active manipulation of the original ethnolinguistic endowment, including the introduction of their own ethnicities and the replacement of the indigenous populations, introduced a man-made component of contemporary ethnic fractionalization, tipping the balance in favor of an ethnic spectrum whose identity and size is not a natural consequence of the primitive land characteristics. In particular, the empirics suggest that contemporary ethnic diversity is no longer systematically related to the underlying geographical heterogeneity in countries whose native populations as of 1500 *AD* represent less than 50% of the current population mix. This decomposition of contemporary ethnic fractionalization into a natural component, driven by the geographic variability, and a man-made one, offers new insights regarding the origins and implications of ethnic diversity.

The results of this study are directly related to the literature on state formation, see Alesina and Spolaore (1997). In this literature, preference heterogeneity is a key determinant of the optimal size of a state. Taking into account that heterogeneous land endowments may be associated with distinct needs for public goods,⁶ and establishing that these differences in land endowments are behind ethnic fragmentation, generate new insights about the relationship between state formation and ethnic diversity.

Another line of research, to which the findings are relevant, is a recent study by Spolaore and Wacziarg (2009). The authors document empirically the effect of genetic distance, a measure associated with the time elapsed since two populations' last common ancestors, on the pairwise income differences between countries. Larger genetic distance is associated with larger income differences. In the context of the proposed mechanism, population mixing, which affects

⁶Irrigation projects, for example, would be much more complementary to farmers' needs than herders.

genetic distance between two countries, is endogenous to the transferability of country specific human capital within the pair. The more similar the geographic endowments between two countries, the smaller should their genetic distance be, *ceteris paribus*. Therefore, the uneven diffusion of technology across countries may be an outcome of the differences in society's specific human capital. By introducing the pair-wise country differences in the distributions of land quality and elevation, one can decisively improve upon the interpretation of the existing results.

Despite the salience of ethnic diversity in shaping economic development there is only one recent working paper within economics that investigates the roots of ethnic diversity, see Ahlerup and Olsson (2008). The authors provide a theoretical setting where ethnic groups endogenously emerge among peripheral populations in response to an insufficient supply of public goods over time. Using novel data on the duration of human settlements since prehistorical times they show that countries where modern humans settled earlier sustain higher ethnic diversity today. Ahlerup and Olsson (2008) approach is complementary to the current study by linking ethnic diversity to public goods-provision and cultural drift, as well as providing evidence on the man-made component of ethnic diversity showing that longer modern state experience is negatively related to contemporary ethnic diversity.

The proposed hypothesis also bridges the divide in the literature regarding the formation of ethnicities, by identifying the economic mechanism at work. There are two main strands of thought. The primordial one qualifies ethnic groups as deeply rooted clearly drawn entities, see Geertz (1967), whereas the constructivists or instrumentalists, see Barth (1969), highlight the contingent and situational character of ethnicity with modern states being an important determinant of the latter. In the current framework, it is the heterogeneity in regional land endowments that initially gives rise to relatively stable ethnic diversity, an element of primordialism. However, as the process of development renders land increasingly unimportant ethnic identity is ultimately bound to become less attached to a certain set of region specific skills and, thus, more situational and ambiguous in character. For example, Miguel and Posner (2006) provide evidence that ethnic identification in Africa becomes more pronounced as political and economic competition increases. Similarly, Rao and Ban (2007) provide evidence on the man-made component of ethnic diversity in India by showing how state policies and local politics have had an important impact on shaping caste structures over the last fifty years. In another recent study Caselli and Coleman (2006) provide a theory where ethnic traits provide a dimension along which voluntary coalitions may be formed and Esteban and Ray (2007) investigate the salience of ethnic identity on the eruption of civil conflict. Along similar lines, Laitin (2007) views ethnic identity as providing a mechanism along which individuals coordinate to demand national recognition.

According to the proposed hypothesis, to the extent that ethnic 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 characteristics. This prediction generates new insights for understanding the pattern of population movements like the spread of the first agriculturalists and herders following the Neolithic Revolution, the settlement intensity of colonizers across the colonized world as well as the contemporary spatial distribution of ethnic groups in general. (Pre)historic evidence consistent with the proposed mechanism documenting the formation of homogeneous linguistic areas across regions of common geographic endowments is presented in Section 2.

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

The rest of the paper is organized as follows. Section 2 describes the hypothesis and alternative mechanisms and presents evidence on language spreads. Section 3 discusses the data and shows empirically how geographic heterogeneity shapes production decisions. Section 4 presents the main part of the empirical analysis. This is conducted in a (i) cross-virtual country (ii) cross-pair of adjacent regions and (iii) cross-country framework. It includes the various robustness checks and concludes by focusing on the impact of the European colonizers on the ethnolinguistic endowment of the colonized world. Finally, section 5 summarizes the key findings and concludes.

2 The Hypothesis and Evidence on Language Spreads

The intuition regarding the role of geographical heterogeneity in producing ethnic fragmentation may be illustrated by the following simple example. Imagine a world composed of two isolated islands A and B (see Figure 1). Island B is perfectly homogeneous regarding its land endowment, suitable for cultivating a single common crop whereas island A is geographically diverse with the northern part only suitable for herding and the southern part suitable for farming a single crop. In a stage of development when land dominates production decisions, groups residing in these regions would develop and accumulate skills specific to their locations. Groups in the northern part of island A would become herders whereas those residing in the southern part would be farmers. This intrinsic difference in land endowments, manifested in the imperfect transferability of location specific knowledge, would limit population mixing between these two areas leading eventually to the formation of two ethnically distinct regions. On the contrary, the common geographic endowment in island B would allow for skills locally acquired

to be perfectly applicable across regions facilitating population mixing and leading eventually to the emergence of a common ethnic or linguistic identity. Assuming in the beginning of time that regions are either ethnically fragmented or homogeneous does not affect the pattern of ethnolinguistic assimilation. What is important is, that in absence of population mixing, the process of cultural drift would lead to the formation of distinct cultural traits over time, see Boyd and Richerson (1985).

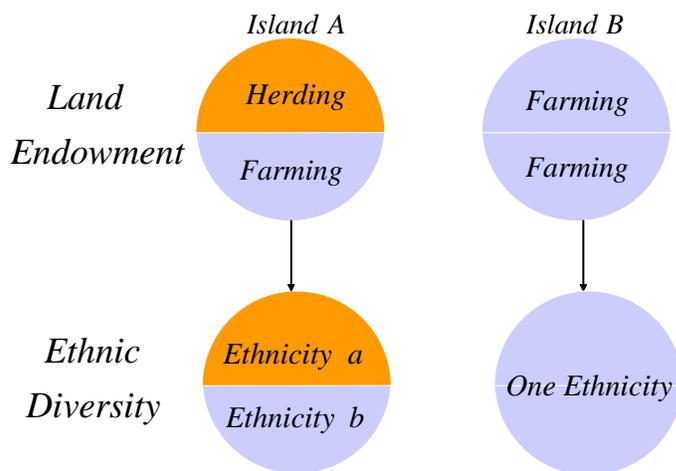


Figure 1: An Example of Geographic and Ethnic Diversity

One could argue that the intensity of trade between regions may be an independent force leading to a convergence in regional cultural traits. The absence of comprehensive historical data on trade intensity within and across countries makes difficult the estimation of the precise effect of trade on ethnic diversity. However, one would expect that trade would be more intense between regions with distinct factor endowments, i.e. with different land characteristics. Such a prediction, nevertheless, is at odds with the empirical findings suggesting that any trade induced force towards ethnic homogenization is not quantitatively dominant.⁷

The proposed mechanism rests upon two fundamental building blocks: (i) population movements influence the ethnolinguistic identity of the places involved and (ii) ethnic groups and languages tend to disperse along places with similar productive endowments.

Linguists have long recognized the role of population mixing in producing common lin-

⁷An additional reason why the quantitative importance of trade appears to be limited, may stem from the fact that whenever there are gains from trade to be made, customarily this is accompanied by the emergence of the merchant class within a society rather than a uniform participation in goods exchange across individuals. Similarly, the pursuit of economic diversification through marrying across regions of different productive endowments would also operate against finding a systematic positive relationship between ethnic and geographic diversity.

guistic elements between places. As Nichols (1997a) points out “almost all literature on language spreads focuses on either demographic expansion or migration as the basic mechanism.”⁸ Both instances are a result of population movements towards territories previously unoccupied by their ancestors. As an outcome of population mixing, the regional populations experience a language shift either to or from the immigrants’ language. Similarly, languages long in contact come to resemble one another in several dimensions like sound structure, lexicon, and grammar. This resultant structural approximation is called convergence.

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

Examples of groups that migrated along areas that were similar to their region of origin include Austronesians and speakers of Eskimoan languages, who are coastally adapted peoples, and have accordingly spread along coasts rather than inland. Along similar lines, Bellwood (2001) argues that the spread zones of agriculturalists and their languages following the Neolithic Revolution trace closely land qualities that were amenable to agricultural activities. Considering languages of the Indo-European family, their expansion after the Neolithic revolution is embedded to the notion of “spread” and “friction” or “mosaic” zones.⁹ Spread regions are characterized by similar land qualities where the early agriculturalists could easily apply their farming skills. Friction zones on the other hand, are areas less conducive to such activities. In these places the populations maintained their distinct ethnolinguistic behavior. Examples of the latter include regions like Melanesia, Northern Europe and Northern India, see Renfrew (2000) for a comprehensive review. Overall, early agriculturalists and pastoralists, perhaps not surprisingly, targeted and expanded into areas where their skills would best apply, homogenizing them linguistically.¹⁰

⁸Nichols (1997a) defines a spread zone as “an area of low density where a single language or family of languages occupies a large range.”

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).

¹⁰Other relatively more recent examples of ethnic groups that consistently migrated to places where they could utilize their ethnic human capital, include the Greeks and the Jews, among others, who belong to the historic trade diasporas (Curtin, 1984). In this case, it is the knowledge of how to conduct commerce that allowed these groups to spread into areas where merchandising was both possible and profitable. Botticini and Eckstein (2005), for example, document the religiously driven transformation of the Jewish ethnic human capital towards literacy and the resulting urban expansion.

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

It is to be noted that in absence of data directly measuring the human capital of an ethnic group¹¹ the empirical relationship between geographic and ethnic diversity is also consistent with other mechanisms. For example, one might argue that groups of people form an ethnic identity along a homogeneous land endowment in order to defend it against intruders and enforce property rights over it. If this is the case, places characterized by more diverse land endowments would automatically sustain more ethnic groups. Alternatively, one could think of the following scenario: originally ethnicities existed independent of the underlying geography, but every time a place was invaded the invading group was forcing the preexisting ethnicities to move in regions with more heterogeneous land endowments. This could be the case if ruling over homogeneous territories was easier than over heterogeneous ones. Such a scenario would deliver empirically a similar relationship, i.e. geographic diversity induces ethnic diversity.

Both mechanisms, however, are mute regarding how languages and ethnic groups spread which is a fundamental component for understanding linguistic and ethnic diversity as the evidence on language spreads suggests.

Keeping these points in mind I turn on the empirical part of this study.

3 Empirical section

3.1 The Data Sources

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

The notion of location specific skills applies also to understanding ethnic differentiation during the hunting and gathering regime. However, measures of cultural diversity before the advent of farming are not available. More importantly, the very spread of agriculture was

¹¹Appendix *B1* focuses on Kenyan ethnic groups and provides evidence on the direct link between the type of land endowments and the specific skills of the ethnic groups residing in different regions.

in many instances equivalent to the spread of languages spoken by the early pastoralists and agriculturalists, as the linguistic evidence reviewed in section 2 suggests. Such population movements significantly altered any preexisting spectrum of ethnic diversity. This justifies using data on land suitability for agriculture to capture differences in productive endowments for the period of human history following the Neolithic Revolution.

The intuition for using differences in land quality and elevation as the ultimate determinants of the differences in productive activities across regions is the following. Farming would be the dominant form of production in places characterized by high land quality, with the regions possibly differing in the optimal mix of plants and crops under cultivation. That is, even within agriculture, the specificity of human capital derives from the different crops produced regionally. However, pastoralism is bound to be more widespread at intermediate and low levels of land quality, exactly because agriculture is less suitable in such areas.¹² At very low levels of land quality being a middleman has been perhaps the most widespread activity as the case for cultures residing along trade routes suggests.¹³ Along similar lines, different altitudes are known to impose limits on the extent of agriculture as well as on the very choice of cultivated crops, see Grigg (1995). The next section provides empirical evidence which shows that geographic variability, as captured by the heterogeneity in land suitability for agriculture and elevation, is a significant determinant of actual crop diversity. It is to be noted that differences in elevation are likely also to be associated with higher transportation costs, further deterring population mobility.

The global data on agricultural suitability were assembled by Ramankutty et al. (2002) to investigate the effect of the future climate change on contemporary agricultural suitability.¹⁴ This dataset provides information on land quality characteristics at a resolution of 0.5 degrees latitude by 0.5 degrees longitude, representing an average region of about 55 km. by 35 km. In total there are 64004 observations.

Each observation takes a value between 0 and 1 and represents the probability that a particular grid cell may be cultivated. In order to construct this index, the authors (i) derive empirically the probability density function of the percentage of croplands around 1990 with

¹²Results available upon request show that land use changes as agricultural suitability changes. Constructing an index of land use at a country level which represents the ratio of land allocated to pasture versus croplands in the 1990's reveals that the relative pastoral intensity diminishes monotonically as the land suitability for agriculture increases. Not surprisingly in more fertile places people will mostly be farmers whereas in less fertile ones pastoralism becomes the preferred activity.

¹³A famous example includes the trading routes of West Africa from the 5th - 15th century AD. These routes ran north and south through the Sahara and traded commodities like gold from the African rivers, salt, ivory, ostrich feathers and the cola nut. In absence of these trading routes, such places would hardly maintain any other activity, and this is a prime example where the regional knowledge, of how to transfer goods safely through a certain passage, is entirely location specific and thus almost impossible to transfer to other places.

¹⁴Appendix G provides details on the data sources used in this study.

respect to climate and soil characteristics and (ii) combine this empirical probability density function with data on climate and soil quality at the resolution of 0.5 degrees latitude by 0.5 degrees longitude to predict the regional suitability for agriculture across the globe.

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

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

With respect to the cross-virtual country and cross-pair of adjacent regions analysis, ethnic diversity is constructed using information on the location of linguistic groups. In the case of virtual country regressions the number of languages within each geographical unit provides a measure of the overall ethnolinguistic diversity. In the adjacent region analysis, an index of ethnic similarity is constructed by calculating the percentage of common languages i.e. the number of common languages over the total number of languages spoken within a pair of adjacent regions. Data on the location of linguistic groups’ homelands are obtained from the Global Mapping International’s World Language Mapping System. This dataset is covering most of the world and is accurate for the years between 1990 and 1995. Languages are based on the 15th edition of the Ethnologue database on languages around the world.¹⁵

In the cross-real country analysis a wealth of alternative measures of ethnic diversity is available. The measure of fractionalization widely used is the probability that two *individuals* randomly chosen from the overall population will differ in the characteristic under consideration, i.e. ethnicity, language, religion. The results presented below use the index most widely

¹⁵The data are available at www.gmi.org. To identify which languages are spoken within the unit of analysis I use the information on the location of language polygons. Each of these polygons delineate a traditional linguistic homeland; populations away from their homelands (e.g. in cities, refugee populations, etc.) are not mapped. Also, the World Language Mapping System does not attempt to map immigrant languages. Finally, linguistic groups of unknown location, widespread languages i.e. languages whose boundaries coincide with a country’s boundaries and extinct languages are not mapped and, thus, not considered in the empirical analysis. The only exception for not mapping widespread languages is the case of English language which is mapped for the United States.

employed in the literature which is the ethnolinguistic fractionalization index, ELF , based on data from a Soviet ethnographic source, *Atlas Narodov Mira (Atlas of the People of the World) (1964)*, and augmented by Fearon and Laitin (2003).¹⁶ This index represents for each country the probability that two individuals randomly drawn from the overall population will belong to different ethnolinguistic groups. Using the linguistic, ethnic and religious fractionalization indexes constructed by Alesina et al. (2003), the absolute number of ethnic or linguistic groups derived by Fearon (2003) or the ethnic fractionalization measure proposed by Montalvo and Reynal-Querol (2005), the qualitative results are similar.¹⁷

3.2 The Properties of Geographic Variability and Productive Decisions

The distribution of land quality varies considerably across regions and across countries. For example, the following graph plots the distribution of regional land quality for Swaziland and Bhutan. In Swaziland the quality of land is concentrated around high values with average quality, $avg = 0.69$, and a *range* (this is the difference between the region with the highest land quality from that with the lowest) of 0.29.¹⁸ On the other hand, land quality in Bhutan averages 0.30 and it spans a much larger spectrum. In fact, $range_{Bhutan} = 0.69$. The difference in elevation between these two countries is similar with Bhutan exhibiting a much larger diversity in altitudes.

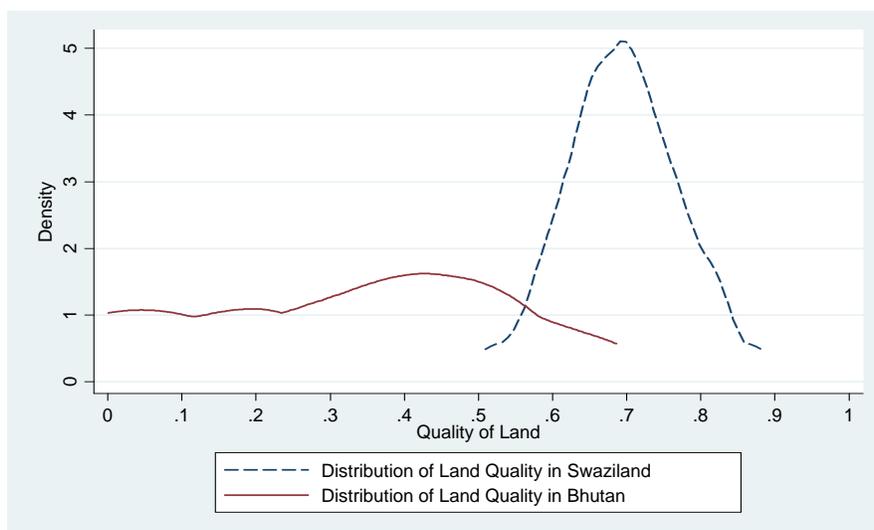


Figure 2

¹⁶According to Fearon and Laitin (2003) the sources used for augmenting *Atlas Narodov Mira (Atlas of the People of the World) (1964)* missing country observations were the CIA Factbook, Encyclopedia Britannica, and the Library of Congress Country Studies.

¹⁷Modifying the current framework to uncover the determinants of ethnic *polarization* is a topic for future research.

¹⁸The figure shows the kernel density estimate (weighted by the Epanechnikov kernel) of regional land qualities for each country.

The range of land quality, i.e. the support of the distribution within the respective unit of analysis, and the standard deviation of elevation, $elev_sd$, are the statistics used to capture the degree of geographical heterogeneity.¹⁹ These capture, albeit imperfectly, how readily location specific knowledge may be transferred across places. Intuitively, a larger *range* and a more variable topography imply that the geographical unit is composed of territories with increasingly different underlying productive characteristics, effectively enlarging the set of activities along which groups may specialize. The larger the spectrum of land qualities and the variation in elevation, the less transferable is the regional know-how. Thus, one would expect higher geographic diversity to increase the probability of ethnically distinct regions, ceteris paribus.²⁰ Indeed, going back to the example of Swaziland and Bhutan, ethnolinguistic fractionalization in Swaziland is only 0.38 compared to the highly ethnolinguistically fragmented society of Bhutan with $ELF_{Bhutan} = 0.69$.

The narrative so far suggests that geographic variability should manifest itself into different productive choices. Appendices *B1* and *B2* provide evidence on this direction. Appendix *B1*, in particular, demonstrates how different land qualities dictate the choice between pastoralism versus agriculture shaping diverse ethnic identities across regions in Kenya.

Appendix *B2* shows how geographic diversity shapes farming decisions. Specifically, using data on the global distribution of major crops cultivated around 1990, I derive the number of crops across countries, nmb_crops . The regression results in Table 1 show that countries endowed with larger variation in elevation, $elev_sd$, and more diverse land qualities, $range$, systematically cultivate a larger number of major crops. Figures 5*a* and 5*b* present the partial scatter plots as generated by the regression in Table 1 of the number of crops cultivated against the variation in elevation and diversity in land quality respectively. Regarding the rest of the controls included in the regression in Table 1 the statistically significant coefficient on $areakm2$, and average land quality, avg , imply that larger and more fertile countries grow more crops. Also, conditional on geography countries in Western Europe, denoted by reg_we , cultivate systematically fewer crops whereas those in Sub-Saharan Africa, denoted by reg_ssa , do not exhibit a systematic relationship with crop diversity. These results strengthen the claim that variation in elevation and land quality dispersion are the primitive elements behind productive choices.²¹

¹⁹In the robustness section for the cross-virtual country and cross-country analysis I show that using alternative statistics to capture heterogeneity in land quality and elevation the results are similar.

²⁰The average quality of land, avg , should not directly affect ethnic diversity, because if places are productively homogeneous then region specific skills are perfectly transferable across all pockets of land irrespective of the type of land quality. Nevertheless, a higher land quality by sustaining denser populations may affect the path of a country's economic development, indirectly influencing ethnic diversity. I return to this point in the regression analysis.

²¹It should be noted that using the actual crop diversity to explain ethnic diversity is not an appealing approach

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

Another concern is whether the observed geographic variability is an outcome of human interventions. Variation in elevation is plausibly exogenous and not subject to human interference at the regional scales the study investigates. However, diversity in land quality may be endogenous to human activities. In particular, the part of the index that depends on soil characteristics. This makes land quality possibly endogenous to the duration of agriculture and herding. Reassuringly, controlling for the timing of the rise of agriculture does not affect the results. Also, it is important to note that soil quality is itself endogenous to the regional climate. Comparing the global distribution of annual precipitation with the distribution of soil pH, it is evident that regions receiving a lot of precipitation are characterized by highly acidic soils, whereas in places with low precipitation the soil becomes alkaline.²²

Although one cannot rule out entirely the possibility of reverse causality running from exogenous group specific subsistence practises to soil diversity, this would only be operative at small changes in soil quality. It would seem unlikely to posit that herders in Kenya, for example, transformed their lands into semi-deserts because of their herding cattle and camels and that agriculturalists transformed their own territories into fertile lands by systematically planting certain crops. If anything it would be the agricultural practises leading to a deterioration of the land's soil properties. More generally, historical groups that severely damaged their environment either through overgrazing or overfarming did not survive over time. Reinforcing this point Diamond (2005) provides several examples of cultures whose subsistence practises

for several reasons. Crop choice is endogenous to a host of things like the level of economic development, among others, so if ethnic diversity affects economic development and development affects crops cultivated then in that case causality would run from ethnic diversity to crop diversity. Also, the number of crops grown around 1990 is a limited measure of productive diversity since it captures heterogeneity only *within* farming. These considerations advise against using the crop diversity as a predictor of ethnic diversity.

²²These 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.

were not sustainable given the underlying geographical capabilities triggering eventually an environmental collapse which was tantamount to these cultures' own demise.

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

4 Empirical Results

4.1 Cross-Virtual Country Analysis

Before going into the cross-country analysis, it is important to investigate whether the relationship between geography and ethnic diversity obtains at an arbitrary level of aggregation. Finding that geographical diversity leads to higher ethnic diversity, irrespective of a country's political boundaries, will greatly enhance the validity of the proposed hypothesis and alleviate any concerns related to border and country formation inherent to any cross-country analysis.

The way that the artificial countries are constructed is the following. First, I generate a global grid where each cell is 2.5 degrees longitude by 2.5 degrees latitude and then I intersect it with the global data on land quality and elevation (see map 1*b* in Appendix A with the resulting artificial countries which constitute the unit of analysis). Using alternative dimensions like 4 by 4 or 5 by 5 degrees does not change the results. Note that since the dimensions of the virtual countries are in decimal degrees the actual area of each cell is declining the further away from the equator, so the size of each virtual country and its distance from the equator are always controlled for.

For each virtual country, I construct the distribution of land quality and elevation and calculate the number of 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 *nmb_r_lang*. Including all languages irrespective of their spatial extent or only focusing on those languages with at least 2% of area coverage within a virtual country, the results remain qualitatively intact.

In the regression analysis virtual countries of at least 10000 square kilometers are included yielding an average virtual country of *areakm2* \approx 44000 square kilometers. The kernel density estimate of the distribution of the number of languages spoken across virtual countries is shown

in Figure 3:²³

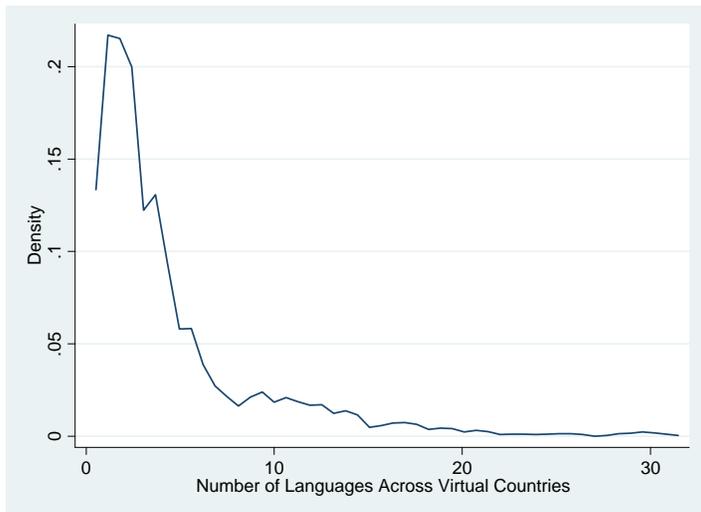


Figure 3

The resulting sample size is 1951 observations with a median of 25 regional land quality observations per virtual country. Descriptive statistics and the raw correlation between the variables used in the regressions are presented in Tables 2a and 2b. In each virtual country, there are on average 4.14 languages spoken and the pairwise correlations of both the dispersion of land quality, *range*, and the variation in elevation, *elev_sd*, with the number of languages are positive and large. Also, as one might expect, diversity in land quality is higher in larger virtual countries as well as in virtual countries characterized by more variable elevation, *elev_sd*.

Map 2 in Appendix C shows one example of a virtual country. The circles, which are the centroids of the original cells of 0.5 by 0.5 degrees, represent the regional land quality for agriculture. The differently colored polygons represent the locations of the linguistic groups. The virtual country in map 2 falls between two real countries with the squiggly line delineating the current borders between Iran on the east and Iraq on the west. There are in total 8 languages spoken in this area²⁴ and the spectrum of land qualities is 0.89, ranging from places that are totally inhospitable to agriculture to areas where the climate and the soil are highly conducive to cultivation.

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

²³Note that the distribution of the number of languages is skewed so instead of the levels the log of languages, *lnmbr_lang*, is used in the regressions below. Excluding the extremely linguistically fragmented artificial countries, i.e. those with more than 20 languages spoken, the qualitative results are similar.

²⁴Namely these are: Central Kurdish, Gurani, Koy Sanjaq Surat, North Mesopotamian Spoken Arabic, Sangisari, South Azerbaijani and Southern Kurdish. Languages' traditional homelands may overlap. For example, in this particular grid places where Gurani is spoken also speak Northern Kurdish. Three of linguistic these groups are split by the borders between Iran and Iraq. See (Alesina et al., 2006) for the role of border drawing in the creation of split ethnic groups.

$$\ln nmb_lang_i = \beta_0 + \beta_1 range_i + \beta_2 elev_sd_i + \beta_3 X_i + \xi_i \quad (1)$$

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

This main prediction is corroborated across all alternative specifications of Table 3.²⁵ In the first regression of Table 3 both $elev_sd$ and the $range$ have a large and significant positive impact on linguistic diversity. A two-standard deviation increase in $range$ increases linguistic diversity by 21% adding 0.86 languages to an average virtual country whereas a two-standard deviation increase in $elev_sd$ increases linguistic diversity by 22%, adding on average 0.90 languages. These are novel and economically important findings that reveal the geographic origins of contemporary ethnolinguistic diversity.

In the same specification, an array of additional geographical features are simultaneously accounted for. In particular, the size of each artificial country, $areakm2$, the average land quality, avg , the average elevation, $elev$, the latitudinal distance from the equator, abs_lat , the number of real countries a virtual country falls into, nmb_cntry , a dummy for the units that belong as a whole to an existing country, in_cntry , the area under water, $waterarea$, as well as the distance from the coastline, sea_dist , are controlled for. More countries a virtual country falls into, the more languages it sustains. This evidence has a dual interpretation. It may be suggestive of the effect of state formation on ethnic diversity and/or an artifact of modern states having drawn political boundaries along ethnic boundaries. The distance from the equator itself enters negatively and significantly, that is, even conditional on the area of a virtual country, fewer languages are spoken further away from the equator. Climatic volatility, which increases further from the equator, to the extent that it leads to persistent population mixing, would lower ethnic diversity.²⁶ Note also that biodiversity generally decreases further away from the equator, (Rosenzweig, 1995), effectively allowing for fewer productive niches along which groups of people may specialize.

Average land quality does not seem to affect linguistic diversity significantly, whereas

²⁵The 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 3000 km. Results are similar using 1000 km., 2000 km., and 6000 km.

²⁶Using the CRU TS 2.0 dataset on monthly temperatures from 1900 to 2000 one can show that distance from the equator is highly positively correlated with the standard deviation of monthly temperature looking both within and across seasons. Nettle (1996) provides further evidence that countries facing higher ecological risk sustain lower linguistic diversity.

places in higher altitudes sustain lower linguistic fragmentation. The variable capturing under water areas, *waterarea*, like rivers and lakes, enters negatively and is statistically insignificant. This raises the issue of whether water bodies are a barrier or a facilitator of population mobility. Finally, the distance from the shoreline of an artificial country, *sea_dist*, does not systematically affect linguistic diversity. Overall, these geographical characteristics capture 43% of the variation in linguistic diversity across virtual countries.

The statistical and quantitative importance of geographic diversity is robust to alternative specifications. In particular, taking advantage of the arbitrarily drawn borders of these geographical units one may explicitly control for real country and continental fixed effects.²⁷ This is done in all subsequent specifications. Such inclusion of powerful controls, not possible in a cross-country framework, allows to explicitly take into account any systematic elements related to the state histories of existing real countries and, thus, produce reliable estimates of the effect of geographic diversity on ethnic diversity. Since virtual countries within a country are geographically close, country fixed effects implicitly control for the common migratory distance of these cells from known original migration areas, like Addis Ababa in Ethiopia or the Fertile Crescent. The inclusion of country and continental fixed effects in the second column of Table 3 slightly changes the coefficients on *range* and *elev_sd*.

Columns 3 and 4 of Table 3 investigate whether the identified effect of geographic variability is driven by the inherent differences between regions in the tropics and the rest of the climatic zones. In column 3, the sample is restricted to virtual countries out of the tropics.²⁸ The estimated coefficient on *range* decreases slightly whereas the coefficient on the variation in elevation increases by almost 50%. This implies that out of the tropics variation in elevation is quantitatively a relatively more important determinant of linguistic diversity. This pattern reverses, however, when one examines the impact of geographic variability on ethnic diversity within the tropics, see column 4 of table 3. Across virtual countries in the tropics the coefficient of variation in elevation becomes less precisely estimated whereas diversity in land quality remains qualitatively and quantitatively significant. Tropical virtual countries with higher average land quality are characterized by larger linguistic diversity whereas the opposite is true for virtual countries out the tropics.

In column 5 of Table 3 the main specification is estimated focusing on cells that entirely belong to a single existing country. This robustness check allows to investigate whether the estimated strong positive relationship between geographic variability and ethnic diversity

²⁷For virtual countries falling into more than one country they are assigned the value of zero across the country dummies. Alternatively, for these virtual countries one could assign as country dummies instead of zeros the fraction of the virtual country's area that falls into each country. Doing so does not change the results.

²⁸The tropics extent from 23.5 latitude degrees south to 23.5 latitude degrees north.

obtains across regions within existing countries. Reassuringly, the dispersion of land quality across virtual countries systematically shapes ethnolinguistic diversity. Namely, territories within countries that display more heterogeneous land endowments give rise and sustain more ethnic and linguistic groups. A one standard deviation increase in both land quality diversity and variation in elevation increases by 30% the number of languages within a virtual country contributing significantly to the formation of ethnically diverse societies.

Geographical diversity within a virtual country so far has been captured by the range of land quality and the standard deviation of elevation. Tables 4a and 4b present the summary statistics and the correlation using alternative geographical diversity indexes. For land quality diversity three additional indexes are employed. Since *range* may be affected by extreme land quality values I compute land quality dispersion after excluding observations below the 5th and above the 95th percentile of the land quality distribution, denoted by *range5_95*. Another measure of heterogeneity is the standard deviation which is denoted by *lqsd*. Finally, given that *range*, *range5_95* and *lqsd* may still be affected by outliers I construct a measure of land quality fractionalization, *lqfrac*. This is the probability that two regions of 0.5 by 0.5 degrees randomly selected from a virtual country belong to different land quality groups. Land quality is grouped in three categories. The first group includes regions with $avg \leq 0.333$, the second group featuring regional land qualities between 0.333 and 0.66 and the third group having regions with land quality larger than 0.66. Given that these are broad quality classes, this index of land quality heterogeneity is less affected by small variation in land quality induced by human intervention. Two additional measures of elevation heterogeneity are also used. These are the dispersion of elevation, denoted *erange*, and the dispersion of elevation excluding values above the 95th percentile and below the 5th percentile of the elevation distribution. Table 4b shows that these new statistics are highly correlated with the measures already used.

Table 5 performs a series of robustness checks on the cross virtual country analysis using these alternative heterogeneity measures. Note that the additional regressors, not shown for brevity, are identical to those used in column 2 of Table 3. In specification 1 of table 5 the dispersion of land quality, *range*, and the dispersion of elevation, *erange*, are used whereas in specification 2 the *range5_95* and the *erange5_95* are introduced. In column 3 the standard deviation of both elevation, *elev_sd*, and land quality, *lqsd*, are used. The last regression employs the land quality fractionalization, *lqfrac*, and variation in elevation, *elev_sd*. Across all specifications the variables of interest remain both quantitatively and qualitatively significant at 1% level demonstrating the robustness of the findings to alternative indexes of geographical diversity.

This section establishes that heterogeneity in land quality and elevation across virtual

countries are both fundamental determinants of contemporary ethnic diversity. The fact that these results obtain at an arbitrary level of aggregation, in and out of the tropics and after controlling for country and continental fixed effects brings into light the, so far neglected, geographical origins of ethnic diversity.

4.2 Pairwise Analysis of Adjacent Regions

The example used in Section 2 focuses on how differences in the productive structure between *two* regions, due to differences in land endowments, deter the formation of common ethnic traits. Hence, a natural setting for testing the proposed hypothesis dictates pairs of adjacent regions as the unit of analysis. In this case, the empirically relevant question becomes how differences in geography, i.e. land quality and elevation, within a regional pair affects the degree of ethnic similarity between the two places. The information provided in the language dataset on the location of linguistic groups allows for such detailed investigation. To implement such a test I identify the neighboring cells of each 0.5 by 0.5 degrees cell. The neighbors of each cell are those who are adjacent at a distance of 0.5 degrees, i.e. directly to the: north, south, 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. In total, a single region may belong to at most eight pairs (see map 3 in Appendix *D* where the dots of regional land qualities are the centroids of the individual regions and the arrows pointing towards the within country neighbors). Out of the 64004 cells in the land quality dataset 18941 contain no information on languages and are dropped from the analysis. This is mostly due to the incomplete mapping of North and particularly Latin America. I also exclude pairs whose individual regions belong to different countries focusing on pairs of adjacent regions that fall entirely within a single country. There are 131772 unique regional pairs within countries.

For the pairwise regressions of adjacent regions the following specification is adopted:²⁹

$$pct_comlang_{ij} = \beta_0 + \beta_1 lqdiff_{ij} + \beta_2 eldiff_{ij} + \beta_3 X_{ij} + \xi_{ij} \quad (2)$$

where $pct_comlang_{ij}$ is the number of common languages divided by the total number of languages spoken within the regional pair and captures the degree of ethnic similarity between any two adjacent regions.³⁰ The variables $lqdiff_{ij}$ and $eldiff_{ij}$ stand for the absolute difference in land quality and elevation respectively between regions i and j . These indexes provide a measure of how dissimilar are the geographic characteristics of any two adjacent regions. Tables 6a and 6b present the summary statistics and the raw correlation of the variables used in the

²⁹Standard errors are clustered at the country level.

³⁰Using as a measure of local ethnic similarity, the number of languages spoken within each pair of regions, the results are unchanged.

analysis. Note that the mean of *pct_comlang* has an interesting economic interpretation: adjacent regions within countries, by virtue of proximity, have on average 80% of the total number of languages in common.

According to the hypothesis, 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 7 supports this focal prediction. The difference in land quality and elevation within a regional pair have a strong negative effect on the formation of common ethnic traits. In particular, a two standard deviation increase in the difference in land quality, *lqdiff_{ij}*, decreases the percentage of common languages by 4.5 points and a similar increase in the difference in elevation, *eldiff_{ij}*, decreases the percentage of common languages by 4.1 points, contributing significantly to the formation of ethnically distinct neighbors. In the same specification I take advantage of the relatively small size of the regional pairs to control for country and continental fixed effects. Regarding the country fixed effects each pair is assigned the dummy of the country it belongs to. This specification explicitly takes into account any systematic elements related to the state histories of each regional pair which might have independently affected the formation of common ethnic traits.

In column 2 controlling for several geographical characteristics the coefficients of interest remain fairly robust. Distance from the equator, *abs_lat*, systematically produces more linguistically homogeneous neighbors, whereas average elevation, *elev*, and the average land quality of the regional pair, *avg*, are not significantly affecting local ethnic similarity. The distance from the shoreline of a regional pair, *sea_dist*, the area under water within a pair, *waterarea*, and the difference in population density within the pair, *popdiff*, do not systematically affect local ethnic diversity. Finally, the overall area of the regional neighbors, denoted *areakm2*, enters positively and is statistically significant.

In column 3 of Table 7, I allow for the effect of the pairwise difference in regional land quality and elevation to vary across continents. The marginal effects of both *lqdiff_{ij}* and *eldiff_{ij}* differ significantly across continents.³¹ Within Africa and Oceania which includes Australia, New Zealand and parts of Papua New Guinea, changes in regional land quality have the greatest impact on local ethnic diversity, whereas changes in regional elevation are quantitatively and qualitatively less important. On the other hand, elevation differences are relatively more important than land quality differences in shaping ethnic diversity across regional pairs in Europe and North America. Within Asia changes in land quality and elevation have a roughly similar contribution to local ethnic diversity. For regional pairs in South America the extremely poor language coverage may be responsible for the insignificant findings.

³¹See Table 7 for a complete description of the marginal effects by continent.

Focusing on specific countries to investigate the impact of local geographic variability on ethnic diversity is possible thanks to the high resolution of the data. Column 4 in Table 7, for example, includes regional pairs that fall entirely within China. In regions across China those located at higher altitudes, *elev*, have more languages in common and those having larger bodies of water, *waterarea*, also share more common languages. A two-standard deviation increase in *lqdiff* decreases local ethnic similarity by 10% and a similar magnitude change in local elevation decreases ethnic similarity by 2.5%.

These findings demonstrate that (i) the difference in land quality and elevation between adjacent regions is a significant determinant of local ethnic diversity and (ii) the arrangement of a given heterogeneous land endowment in space matters in determining the degree of the overall cultural heterogeneity, i.e. the more spatially concentrated is a given land endowment the lower is the resulting ethnic diversity.

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

4.3 Cross-Real Country Analysis

Having established that the differences in land quality and elevation, between adjacent regions and within virtual countries affect systematically the local ethnic endowment, I now proceed into investigating the relationship between geographic variability and ethnolinguistic fractionalization across countries. Using the global data on suitability of land for agriculture and elevation I construct the desired measures of geographic variability for each country. The number of regional observations within country extend from a single observation for Monaco to 12279 for Russia. The median number per country is 82.

Existing countries vary widely in the distribution of land qualities. Figures 6a and 6b in Appendix E, map the regional land qualities for Lesotho and Malawi respectively. A visual inspection of these maps reveals the homogeneity of land quality in Lesotho, $range_{Lesotho} = 0.40$ compared to the apparent heterogeneity inherent to the land quality of Malawi, $range_{Malawi} = 0.61$. Note that these two countries have nonetheless comparable overall levels of land quality, i.e. $avg_{Lesotho} = 0.67$ and $avg_{Malawi} = 0.74$. Mapping the languages spoken in Lesotho and

Malawi a striking parallel emerges. The ethnically fragmented society of Malawi, $ELF_{Malawi} = 0.62$, reflects the large underlying spectrum of land qualities compared to the ethnically homogeneous Lesotho, $ELF_{Lesotho} = 0.22$.³²

As mentioned earlier the index of ethnolinguistic fractionalization, ELF , represents the probability that two individuals randomly drawn from a country's overall population will belong to different ethnolinguistic groups. This implies that the way people are distributed across ethnically distinct places affects measured fractionalization.³³ For example, consider a two-region country. It is easy to show that if these two regions are ethnically distinct then how population is distributed between the two locations will shape overall fractionalization. In particular, the more unequally is population distributed between the two regions the lower will be measured fractionalization. Land quality shapes population density with regions better suited for agriculture sustaining higher densities. Consequently, the gini coefficient of land quality for each country, denoted by $lqqini$, is constructed. The gini of land quality is highly correlated (0.55) with how unequally population density is distributed across regions within a country in 1990.^{34, 35}

Given the preceding discussion the following main specification is adopted:

$$ELF_i = a_0 + a_1range_i + a_2elev_sd_i + a_3avg_i + a_4lqqini_i + a_5X_i + \eta_i \quad (3)$$

where ELF_i is the level of ethnolinguistic fractionalization in country i , $range_i$ is the support of the distribution of land quality within a country, $elev_sd_i$ is the variation in elevation, avg_i stands for the average land quality in country i , and $lqqini_i$ is the gini coefficient measuring how unequally land quality is distributed among regions of country i .

In the regression analysis the sample is restricted in the following way. To make sure there are enough regional observations per individual country only those with at least 10 cells of

³²Lesotho is smaller than Malawi which may partially account for the observed ethnic and geographic heterogeneity. It is worth mentioning that the historical forces behind the formation of modern countries, by shaping the political boundaries have also determined the observed distribution of geographic characteristics.

³³This 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 population weighed transformation of these languages.

³⁴To measure the latter, I construct at the country level a gini index of population density. 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 0.5 by 0.5 degrees in order to make the inequality indexes comparable.

³⁵Results not shown also suggest that the gini coefficient of land quality is strongly correlated (the correlation is 0.55) with how clustered is land quality within a country, computed by the Moran's I index, a commonly used measure of spatial autocorrelation. That is, in countries with more unequal distribution of land quality, contiguous regions are on average of similar land characteristics. Consequently, the adjacency of productively similar regions would facilitate cross migration, due to low relocation costs, leading to lower fractionalization. Indeed, directly including in the regressions the level of clustering enters negatively, though insignificant, and decreases the coefficient of $lqqini$.

0.5 by 0.5 degrees with information on land quality and elevation are included. This limits the sample size to on average 149 countries. Descriptive statistics and the raw correlation between the variables of interest are presented in Tables 8a and 8b.

The results of the main specification (3) are presented in column 1 of Table 9. A two standard deviation increase in the dispersion of land quality, *range*, increases ethnolinguistic fractionalization by 22%. To better understand the magnitude of the effect note that the average difference in ethnolinguistic fractionalization between a Sub-Saharan and a non Sub-Saharan country is 0.33. The non-significant effect of variation in elevation on fractionalization in column 1, is driven by the fact that although Sub-Saharan Africa is the most ethnically diverse region, it has an average standard deviation of elevation of 0.24 km., whereas for a non Sub-Saharan country the average is 0.39 km. Indeed, controlling for continental fixed effects, see column 2, a more variable topography systematically increases ethnic diversity. The gini of land quality, *lqqini*, as expected, enters with a negative sign. Average land quality enters also negatively and statistically significant, it turns insignificant, though, once I control for population density in 1500 AD. This shows that average land quality by sustaining denser population densities historically may have indirectly influenced contemporary ethnic diversity. These purely geographical features account for 15% of the variation in contemporary ethnolinguistic fractionalization across countries.

In the second column of Table 9, dummies for Sub-Saharan Africa, *reg_ssa*, Latin America and Caribbean, *reg_lac*, and Western Europe, *reg_we*, and East Asia and Pacific, *reg_eap*, are introduced, in order to make sure that the results are not driven by a particular region. The coefficients of interest (except for *elev_sd*) generally decrease remain, though, both economically and statistically significant. Repeating the analysis excluding all the countries of Sub-Saharan Africa or focusing only within the latter produces qualitatively similar results.

In the third column of table 9 geographic and historical controls that could potentially affect fractionalization are accounted for. The pure size of a country, denoted by *areakm2*, enters positively but it is insignificant. The mean distance to the nearest coastline or sea-navigable river, denoted by *distcr*, though insignificant, weakly increases fractionalization. This is conforming with the view that places which are increasingly isolated from water passages have been experiencing limited population mixing and thus should on average display higher ethnolinguistic fractionalization. It should be noted, however, that mean distance from the sea, also captures the vulnerability of places to both the incidence and the intensity of colonization. Thus, the coefficient should be cautiously interpreted. The distance from the equator, denoted by *abs_lat*, has a strong negative effect on ethnolinguistic fractionalization.

To capture variation in historical contingencies across countries the population density

in 1500 *AD* and the country's year of independence are added. The log of the population density in 1500 *AD*, *lpd1500*,³⁶ enters negatively and significantly. This finding is evidence that conditional on geographic characteristics contemporary ethnic diversity may have been influenced by a country's historical levels of development as represented by the population density in 1500 *AD*. Also, the year when each country gained independence, *yrentry*, is negatively correlated with fractionalization. Specifically, the later the year of independence, the higher the level of fractionalization. This is consistent with the historical evidence suggesting that since their inception modern states systematically attempted to homogenize their populations along ethnolinguistic dimensions. The expansion of public schooling, for example, had exactly such an impact on linguistic diversity.³⁷ However, the causality may run in both directions. More fractionalized regions may cause a later emergence of modern states either because of being colonized or because of having a slower statehood formation. Figures 7*a* and 7*b* provide the partial scatter plots of the dispersion in land quality and the variation in elevation against *ELF*, as generated by the specification 3 in Table 9. Finally, in the last column of Table 9 the timing of the transition to agriculture, *agritran*, is introduced to account for differences in the timing of the emergence of cultivation across countries which might affect both the suitability of land for agriculture and ethnic diversity. Reassuringly, it enters insignificantly and the point estimates of *range*, and *elev_sd* are barely affected by its inclusion. However, the coefficient on population density in 1500 *AD* loses significance since it is highly correlated with the timing of the transition to agriculture.

Geographical diversity within a country has been captured by the dispersion of land quality, *range*, and the standard deviation of elevation, *elev_sd*. Table 10*a* presents the summary statistics using alternative geographical diversity indexes. The same alternative measures as in the case of virtual country analysis are used. These are the following for the case of land quality heterogeneity (i) land quality dispersion computed after excluding observations below the 5th and above the 95th percentile of the land quality distribution, denoted *range5_95*, (ii) the standard deviation which is denoted by *lqsd*, and a measure of land quality fractionalization, *lqfrac*, similarly constructed as in the case of the virtual country analysis. The two additional measures of elevation heterogeneity are (i) the dispersion of elevation, denoted *erange*, and the dispersion of elevation excluding values above the 95th percentile and below the 5th percentile of the elevation distribution, denoted *erange5_95*. Table 10*b* shows that these new statistics

³⁶This 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 *AD*. It may be useful to note that the term "state history" used throughout this study is distinct from the state antiquity index.

³⁷Laitin (1992) vividly describes the role of governments in promoting or demoting linguistic diversity along the process of state formation in African states.

are strongly correlated with the measures already used.

Table 11 performs a series of robustness checks on the cross country analysis using these alternative heterogeneity measures. Controls for continental fixed effects, the area of a country, its distance from the sea and the equator are also included. In specification 1 the dispersion of land quality, *range*, and the dispersion of elevation, *erange*, are used whereas in specification 2 the *range5_95* and the *erange5_95* are introduced. In column 3 the standard deviation of both elevation, *elev_sd*, and land quality, *lqsd*, are used. The last regression employs the land quality fractionalization, *lqfrac*, and variation in elevation, *elev_sd*. Across all specifications the variables of interest remain both quantitatively and qualitatively significant demonstrating the robustness of the findings to alternative indexes of geographical diversity.

The cross-country analysis, so far, highlights the fundamental role of the distribution of land quality and elevation in the formation of ethnically diverse societies and hints towards the endogeneity of the contemporary ethnolinguistic endowment to the divergent state histories across countries. In the next section the latter is explored in more detail.

4.4 Colonization and Ethnic Diversity

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

Consequently, ethnic diversity across countries colonized by Europeans is itself endogenous to their colonial experience, the identity of the colonizers and how intensely the colonizers settled, among other things. Column 1 in Table 12 presents several correlations between ethnic diversity and the identity of the colonizers. Conditional on geographical characteristics, countries colonized by Germans, French, Dutch, British and Portuguese display consistently higher levels of contemporary ethnic fractionalization compared to places where the Italians, Belgians and Spaniards landed.³⁸

The role of geography in shaping the endowment of ethnicities across space is predicated

³⁸An alternative reading of these correlations is that colonizers differed in the way they chose which places to colonize depending on the level of preexisting ethnic diversity. In absence of time series data on ethnic diversity before and after colonization one cannot disentangle between these two hypotheses.

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

In column 2 of Table 12, the sample is restricted into countries where the percentage of the current population's composition which was indigenous in these countries as of 1500 *AD* is *less* than 50%. The coefficients of the variables of interest decrease substantially becoming insignificant and even change sign in the case of *elev_sd*. Overall and as expected, within this subset of countries whose indigenous populations have been dramatically reduced, geographic variability cannot account for the observed ethnic diversity emphasizing the power of historical events in dramatically altering the spectrum of contemporary ethnic diversity.

The last column of Table 12, investigates the impact of the European colonizers' identity on the percentage of indigenous people living in the colonized countries today. Countries colonized by the Spaniards lost on average 51.2% of their indigenous population, 34.2% was lost in countries colonized by the Dutch, 25.7% was lost across places colonized by the British and 16.4% was lost across French colonies.

The findings of table 12 suggest that European colonizers substantially affected the ethnolinguistic spectrum of the places they colonized. The introduction of their own ethnicities and the replacement of the indigenous populations, in particular, introduced a man-made component of contemporary ethnic fractionalization tipping the balance in favor of an ethnic spectrum whose identity and size is not a natural consequence of the primitive land characteristics.

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

5 Concluding Remarks

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

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

A mechanism, albeit not exclusive, which rationalizes the relationship between geographic and ethnic diversity is also proposed. It argues that differences in geographical characteristics shaped the intensity of population mixing. Places exhibiting homogeneous land endowments were characterized by high transferability of region specific human capital. This facilitated population mobility leading to the formation of a common ethnolinguistic identity. On the contrary, among regions characterized by distinct land attributes, population mixing would be limited leading to the formation of local ethnicities and languages giving rise to a wider cultural spectrum.

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

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

Additionally, the proposed way of thinking about ethnicities as bearers of specific human capital may be used to understand how and why inequality emerges across ethnic groups. Along the process of development the advent of new technologies, being differentially complementary to the specific human capital of each ethnicity, would lead to differential rates of technology

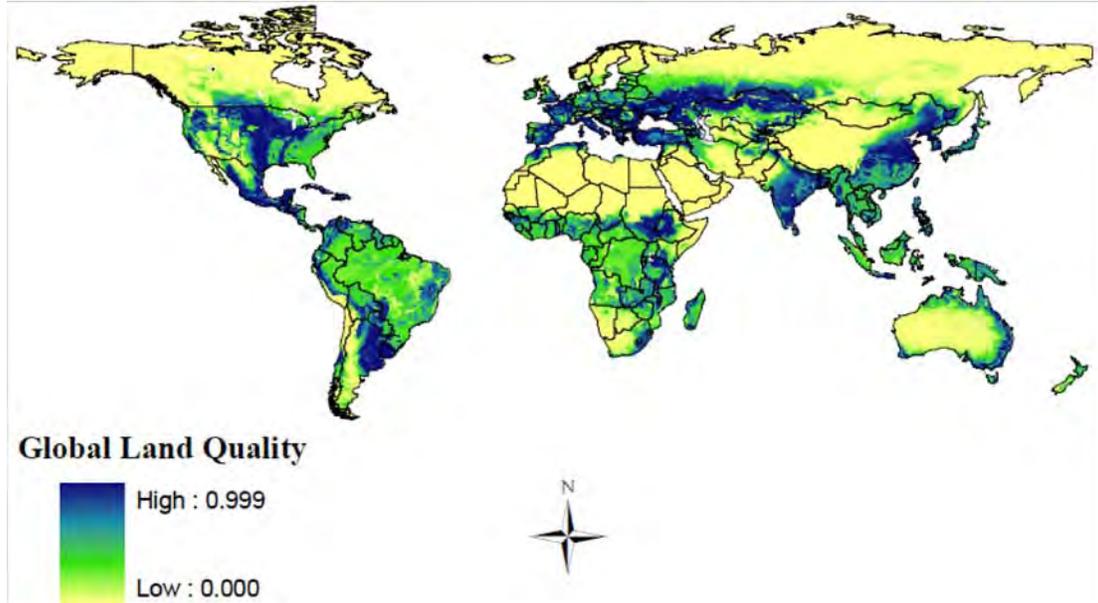
adoption and thus inequality across groups. This notion of location specific skills, driven by the underlying distribution of land endowments, could also be applied at a societal level generating new insights about the diffusion of development both within and across countries.

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

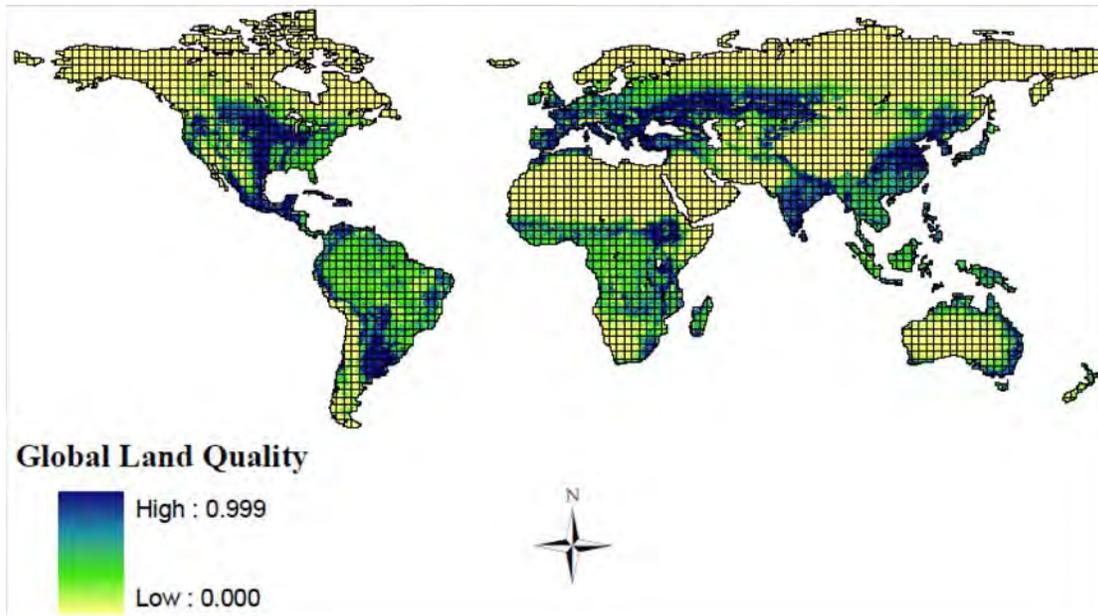
6 Appendix

Appendix A - Global Maps

Map 1a: Land Quality Across Countries



Map 1b: Land Quality Across Virtual Countries



Appendix B1 - Ethnic Groups and Land Quality in Kenya

The hypothesis put forward by this study is that ethnic groups are bearers of specific human capital and this specificity derives from the attributes of the land where 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 33 regions of 0.5 degrees latitude by 0.5 degrees longitude). Land suitability for agriculture (described in the empirical section) is in the horizontal axis, whereas the vertical axis displays the name of each group. The boxes map the interquartile range of land quality with the dots representing regions with land quality more than three standard deviations further from the mean.

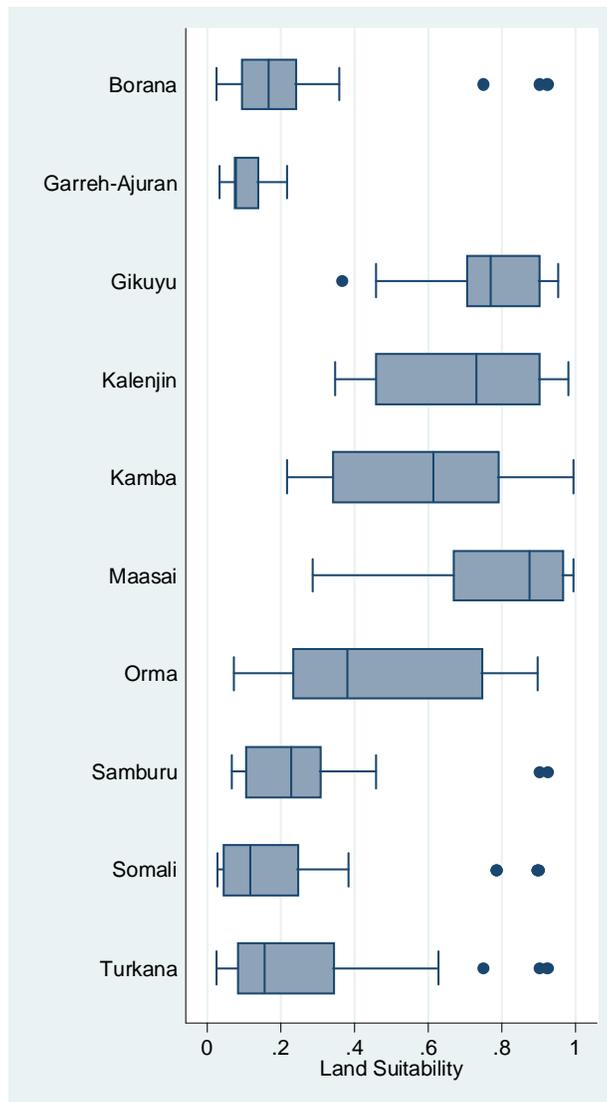


Figure 4: Land Quality within ethnic groups in Kenya

A cursory inspection of the box plots reveals that ethnic groups are not randomly dispersed across regional land qualities within Kenya. In fact, they seem to cluster in territories of distinct and homogenous land endowments. The Borana, the Garreh-Ajuran, the Samburu, the Somali and the Turkana are all located at relatively low levels of land quality where agriculture is almost impossible to maintain.³⁹ The Samburu, the Borana, the Turkana are semi-nomadic pastoralists who herd mainly cattle but also keep sheep, goats and camels, see Pavitt (2001). The Garreh-Ajuran and the Somali are semi-nomadic shepherds. These groups have the human capital to undertake the productive activities which are optimal for the places in which they are located and furthermore the linguistic distance among them is small. On the other hand, the Gikuyu and the Kalenjin are concentrated in territories of high land quality and they are mainly engaged in agriculture, producing: sorghum, millet, beans, sweet potatoes, maize, potatoes, cassava, bananas, sugarcane, yams, fruit, tobacco and coffee. The Kamba people are often found in different professions; some are agriculturalists others hunters, and a large number are pastoralists. This is an outcome of Kamba residing at intermediate levels of land quality which may sustain different activities.⁴⁰ The Orma people are mainly pastoralists who herd cattle, sheep and goats however, people within the Orma group who speak the dialect of Munyo are agriculturalists. This would explain the spread out distribution of the Orma people.

An interesting example is the case of the Maasai people. As it is evident from the map they are located at regions endowed with climatic and soil characteristics very favorable to farming. Nevertheless, the Maasai are semi-nomadic pastoralists with the herding of cattle being the dominant activity. At first, this observation may seem at odds with the proposed hypothesis which maintains that groups should develop skills best suited to their region. The history of Maasai, however, sheds important light on this issue, see Olson (1990). Upon the arrival of the British colonizers in territories that today constitute modern day Kenya, two treaties, one in 1904 and another in 1911, reduced the Maasai lands 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 lands 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.

³⁹The 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/>).

⁴⁰Reinforcing this point, anthropologists believe that the Kamba are a mixture of several East African people, and bear traits of the Bantu farmers as well as those of the Nilotic pastoralists.

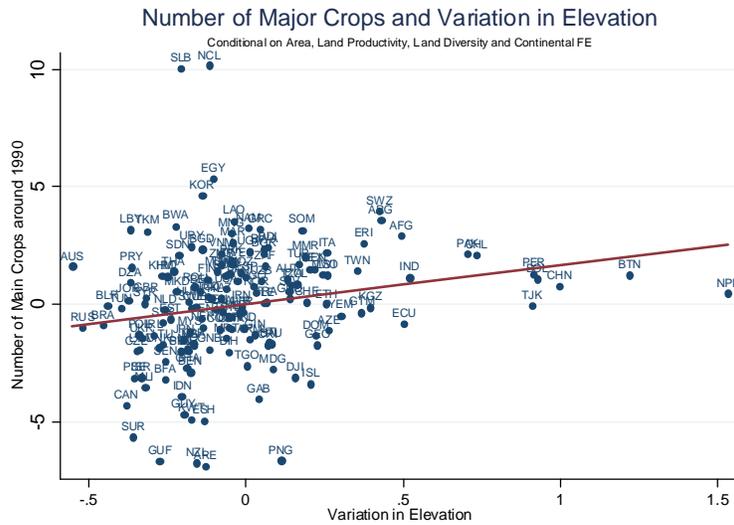
Appendix B2 - Crops and Geographic Variability

Table 1 : Geographic Diversity and Number of Cultivated Crops

| Dep. Var. | range | elev_sd | avg | areakm2 | reg_ssa | reg_we |
|-------------------|------------------------|------------------------|------------------------|------------------------|---------|-----------------------|
| nmbr_crops | 3.677 | 1.685 | 3.018 | 0.031 | 0.399 | -1.189 |
| | (1.085) ^{***} | (0.471) ^{***} | (1.011) ^{***} | (0.008) ^{***} | (0.502) | (0.531) ^{**} |

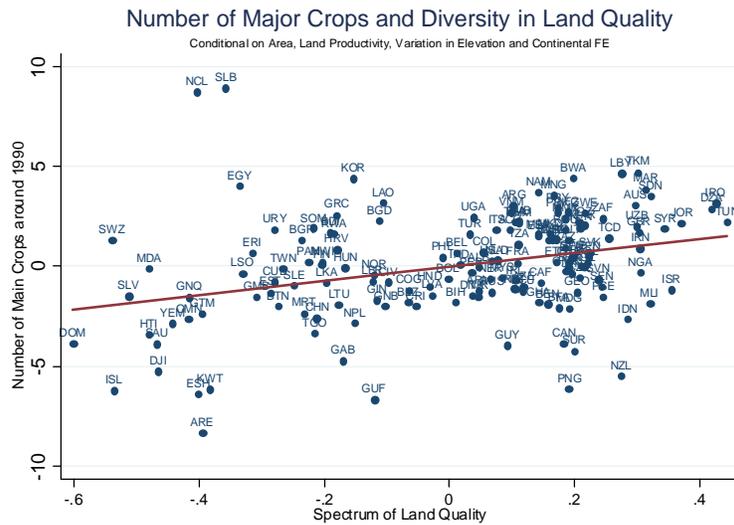
Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1; See Appendix G for variable's definitions

Figure 5a



See Appendix G for variables' definitions

Figure 5b



See Appendix G for variables' definitions

Appendix C - Virtual Country Analysis

Map 2: Example of a Virtual Country

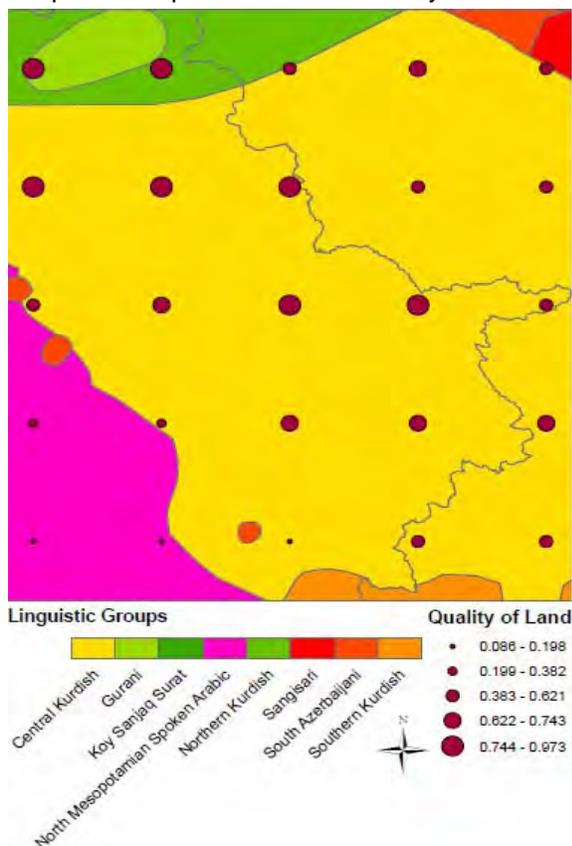


Table 2a: Summary Statistics for the Virtual Country Analysis

| <i>statistics</i> | ln#_lang | range | elev_sd | avg | areakm2 | sea_dist | waterarea | in_cntry | #cntry |
|-------------------|-----------------|--------------|----------------|------------|----------------|-----------------|------------------|-----------------|---------------|
| <i>mean</i> | 1.03 | 0.32 | 0.20 | 0.33 | 44.16 | 0.61 | 0.81 | 0.64 | 1.50 |
| <i>sd</i> | 0.84 | 0.28 | 0.23 | 0.30 | 20.01 | 0.56 | 1.25 | 0.48 | 0.78 |
| <i>max</i> | 3.53 | 1.00 | 2.20 | 0.98 | 76.89 | 2.66 | 15.89 | 1.00 | 6.00 |
| <i>min</i> | 0.00 | 0.00 | 0.00 | 0.00 | 10.01 | 0.00 | 0.00 | 0.00 | 1.00 |

See Appendix G for variables' definitions

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

| | ln#_lang | range | elev_sd | avg | areakm2 | abs_lat | sea_dist | waterarea | in_cntry | #cntry |
|------------------|-----------------|--------------|----------------|------------|----------------|----------------|-----------------|------------------|-----------------|---------------|
| ln#_lang | 1.00 | | | | | | | | | |
| range | 0.29 | 1.00 | | | | | | | | |
| elev_sd | 0.23 | 0.30 | 1.00 | | | | | | | |
| avg | 0.21 | 0.60 | 0.05 | 1.00 | | | | | | |
| areakm2 | 0.27 | 0.24 | 0.11 | 0.14 | 1.00 | | | | | |
| abs_lat | -0.58 | -0.19 | -0.13 | -0.23 | -0.40 | 1.00 | | | | |
| sea_dist | -0.07 | -0.04 | 0.09 | -0.18 | 0.20 | 0.13 | 1.00 | | | |
| waterarea | -0.06 | -0.08 | -0.08 | -0.14 | 0.18 | 0.05 | 0.03 | 1.00 | | |
| in_cntry | -0.35 | -0.19 | -0.12 | -0.10 | -0.22 | 0.23 | 0.00 | -0.06 | 1.00 | |
| #cntry | 0.37 | 0.22 | 0.16 | 0.13 | 0.22 | -0.19 | -0.01 | 0.06 | -0.85 | 1.00 |

See Appendix G for variables' definitions

Appendix C - Virtual Country Analysis

Table 3: Main Specification for the Virtual Country Analysis

| VARIABLES | (1) | (2) | (3) | (4) | (5) |
|--------------|----------------------|----------------------|---------------------|----------------------|---------------------|
| | ln#lang | ln#lang | ln#lang | ln#lang | ln#lang |
| range | 0.382*** (0.146) | 0.409*** (0.112) | 0.350*** (0.012) | 0.448*** (0.164) | 0.305** (0.133) |
| elev_sd | 0.482*** (0.149) | 0.357*** (0.123) | 0.530*** (0.091) | 0.314 (0.213) | 0.484*** (0.146) |
| avg | -0.082 (0.148) | -0.057 (0.143) | -0.153* (0.091) | 0.750** (0.313) | -0.105 (0.162) |
| elev | -0.090** (0.035) | -0.037 (0.036) | -0.009 (0.022) | -0.115 (0.074) | -0.034 (0.036) |
| areakm2 | 0.0001 (0.002) | 0.002 (0.002) | 0.001 (0.002) | 0.006*** (0.002) | 0.002 (0.002) |
| abs_lat | -0.022*** (0.003) | -0.024*** (0.004) | -0.007 (0.005) | -0.041*** (0.008) | -0.013** (0.007) |
| sea_dist | 0.020 (0.063) | -0.030 (0.057) | -0.063 (0.054) | 0.197 (0.128) | -0.098** (0.048) |
| waterarea | -0.021 (0.015) | -0.014 (0.014) | -0.001 (0.012) | -0.001 (0.023) | -0.012 (0.016) |
| in_centry | -0.025 (0.068) | 0.293 (0.195) | -0.018 (0.272) | 0.226 (0.223) | |
| #centry | 0.239*** (0.042) | 0.186*** (0.014) | 0.144*** (0.035) | 0.294*** (0.049) | |
| Observations | 1951 | 1951 | 1312 | 639 | 1241 |
| R^2 | 0.43 | 0.59 | 0.43 | 0.53 | 0.59 |

Standard errors in parentheses corrected for spatial autocorrelation, Conley (1999)

*** p< 0.01; ** p<0.05; * p<0.1

Specifications (2), (3) (4) and (5) include country and continental fixed effects .
 (3) focuses on virtual countries **out of the tropics**, (4) on virtual countries **in the tropics** and (5) on virtual countries belonging entirely to an existing real country
 See Appendix G for variables' definitions

Appendix C - Virtual Country Analysis

Table 4a: Summary Statistics for the Virtual Country Analysis - Geographical Diversity Indexes

| <i>_stats</i> | ln#_lang | range | range5_95 | lqsd | lqfrac | elev_sd | erange | erange5_95 |
|---------------|-----------------|--------------|------------------|-------------|---------------|----------------|---------------|-------------------|
| <i>mean</i> | 1.03 | 0.32 | 0.29 | 0.10 | 0.21 | 0.20 | 0.68 | 0.60 |
| <i>sd</i> | 0.84 | 0.28 | 0.26 | 0.09 | 0.23 | 0.23 | 0.73 | 0.66 |
| <i>max</i> | 3.53 | 1.00 | 0.99 | 0.47 | 0.67 | 2.20 | 5.89 | 5.83 |
| <i>min</i> | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

See Appendix G for variables' definitions

Table 4b: The Correlation Matrix for the Virtual Country Analysis - Geographical Diversity Indexes

| | ln#_lang | range | range5_95 | lqsd | lqfrac | elev_sd | erange | erange5_95 |
|-------------------|-----------------|--------------|------------------|-------------|---------------|----------------|---------------|-------------------|
| ln#_lang | 1.00 | | | | | | | |
| range | 0.29 | 1.00 | | | | | | |
| range5_95 | 0.27 | 0.98 | 1.00 | | | | | |
| lqsd | 0.27 | 0.97 | 0.98 | 1.00 | | | | |
| lqfrac | 0.31 | 0.85 | 0.86 | 0.85 | 1.00 | | | |
| elev_sd | 0.23 | 0.30 | 0.30 | 0.33 | 0.21 | 1.00 | | |
| erange | 0.23 | 0.31 | 0.30 | 0.33 | 0.21 | 0.98 | 1.00 | |
| erange5_95 | 0.22 | 0.30 | 0.30 | 0.32 | 0.20 | 0.99 | 0.98 | 1.00 |

See Appendix G for variables' definitions

Appendix C - Virtual Country Analysis

Table 5: Robustness Checks for the Virtual Country Analysis

| VARIABLES | (1) ln#lang | (2) ln#lang | (3) ln#lang | (4) ln#lang |
|--------------|---------------------|---------------------|---------------------|---------------------|
| range | 0.391*** (0.011) | | | |
| erange | 0.134*** (0.041) | | | |
| range5_95 | | 0.359*** (0.012) | | |
| erange5_95 | | 0.134*** (0.046) | | |
| lqsd | | | 1.048*** (0.341) | |
| elev_sd | | | 0.364*** (0.123) | |
| frac_suit | | | | 0.404*** (0.123) |
| elev_sd | | | | 0.412*** (0.122) |
| Observations | 1951 | 1951 | 1951 | 1951 |
| R^2 | 0.59 | 0.58 | 0.58 | 0.58 |

Standard errors are corrected for spatial autocorrelation, Conley (1999)

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

All specifications include country and continental fixed effects as well as standard controls used in all specifications of Table 3

See Appendix G for variables' definitions

Appendix D - Pairwise Analysis of Adjacent Regions

Map 3: Examples of Pairs of Adjacent Regions

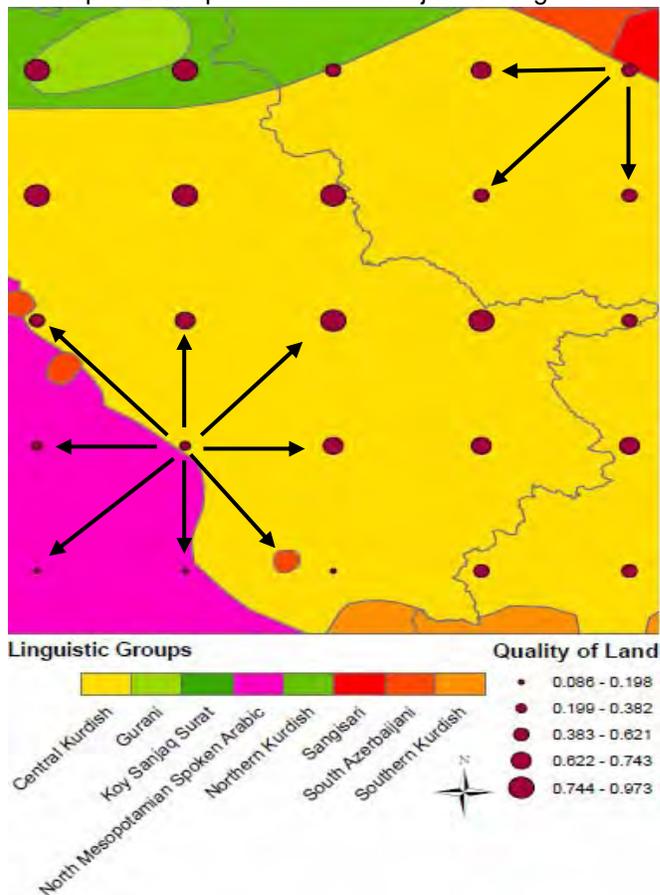


Table 6a: Summary Statistics for the Pairwise Analysis of Adjacent Regions

| <i>_stats</i> | pct_comlang | lqdiff | eldiff | elev | avg | sea_dist | waterarea | popdiff | areakm2 |
|---------------|--------------------|---------------|---------------|-------------|------------|-----------------|------------------|----------------|----------------|
| <i>mean</i> | 0.80 | 0.07 | 0.14 | 0.67 | 0.32 | 0.67 | 0.08 | 0.04 | 4.23 |
| <i>sd</i> | 0.28 | 0.12 | 0.22 | 0.81 | 0.32 | 0.57 | 0.18 | 0.23 | 1.37 |
| <i>max</i> | 1.00 | 1.00 | 3.54 | 5.80 | 1.00 | 2.75 | 4.95 | 14.68 | 6.16 |
| <i>min</i> | 0.00 | 0.00 | 0.00 | -0.66 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |

See Appendix G for variables' definitions

Table 6b: The Correlation Matrix for the Pairwise Analysis of Adjacent Regions

| | pct_comlang | lqdiff | eldiff | elev | avg | sea_dist | waterarea | popdiff | abs_lat | areakm2 |
|--------------------|--------------------|---------------|---------------|-------------|------------|-----------------|------------------|----------------|----------------|----------------|
| pct_comlang | 1.00 | | | | | | | | | |
| lqdiff | -0.14 | 1.00 | | | | | | | | |
| eldiff | -0.15 | 0.23 | 1.00 | | | | | | | |
| elev | -0.05 | 0.06 | 0.42 | 1.00 | | | | | | |
| avg | -0.13 | 0.33 | 0.02 | -0.13 | 1.00 | | | | | |
| sea_dist | 0.06 | -0.05 | 0.06 | 0.32 | -0.15 | 1.00 | | | | |
| waterarea | 0.04 | -0.04 | -0.06 | -0.04 | -0.08 | -0.01 | 1.00 | | | |
| popdiff | -0.02 | 0.05 | 0.01 | -0.05 | 0.14 | -0.07 | 0.00 | 1.00 | | |
| abs_lat | 0.40 | -0.13 | -0.12 | -0.13 | -0.21 | 0.11 | 0.03 | -0.03 | 1.00 | |
| areakm2 | -0.18 | 0.13 | 0.08 | 0.20 | 0.22 | 0.02 | 0.04 | 0.03 | -0.70 | 1.00 |

See Appendix G for variables' definitions

Appendix D - Pairwise Analysis of Adjacent Regions

Table 7: Main Specification for the Pairwise Analysis of Adjacent Regions

| VARIABLES | (1) pct_comlang | (2) pct_comlang | (3) pct_comlang | (4) pct_comlang |
|--------------|----------------------|----------------------|----------------------|----------------------|
| lqdiff | -0.188*** (0.049) | -0.154*** (0.037) | -0.303*** (0.084) | -0.405*** (0.087) |
| eldiff | -0.092*** (0.019) | -0.104*** (0.014) | -0.090* (0.051) | -0.058*** (0.022) |
| abs_lat | | 0.005*** (0.002) | 0.005*** (0.001) | 0.008** (0.004) |
| elev | | 0.004 (0.009) | 0.003 (0.009) | 0.030** (0.012) |
| avg | | -0.040 (0.030) | -0.040 (0.031) | 0.075 (0.110) |
| sea_dist | | 0.008 (0.012) | 0.008 (0.012) | -0.019 (0.041) |
| waterarea | | 0.006 (0.014) | 0.005 (0.014) | 0.048* (0.024) |
| popdiff | | 0.002 (0.006) | 0.002 (0.006) | 0.040 (0.032) |
| areakm2 | | 0.031*** (0.005) | 0.030*** (0.005) | 0.024 (0.019) |
| Constant | 0.583*** (0.002) | 0.376*** (0.042) | 0.376*** (0.036) | 0.351*** (0.039) |
| Observations | 131772 | 131772 | 131772 | 12659 |
| R^2 | 0.29 | 0.30 | 0.30 | 0.07 |

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

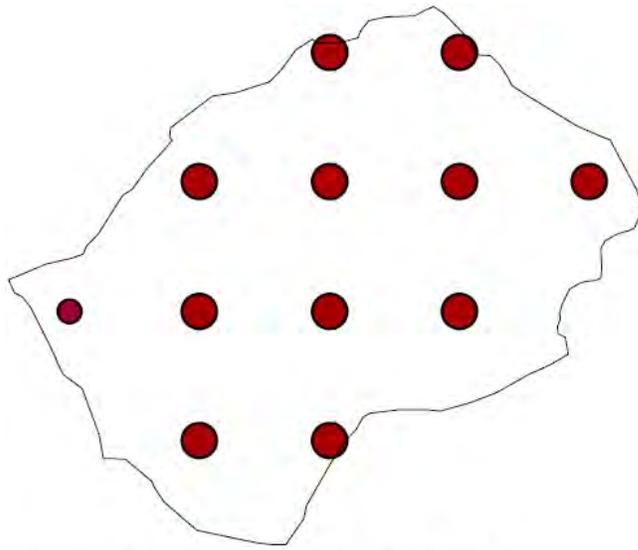
Specifications (1), (2) and (3) include country and continental fixed effects. (3) allows for the marginal effect of pair differences in elevation, eldiff, and lqdiff to vary across continents.

The reported coefficients on eldiff, lqdiff are for Africa. The marginal effects by continent are:

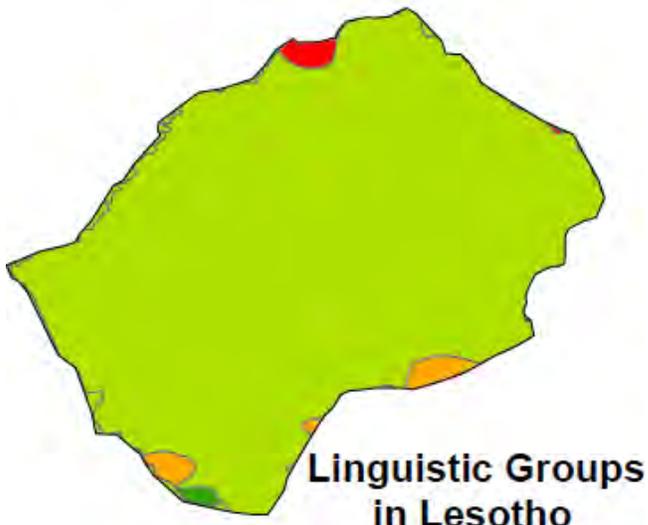
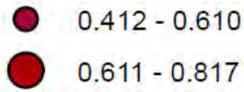
$$\begin{aligned}
 \frac{\partial \text{pct_comlang}}{\partial \text{eldiff}_{ij}} \Big|_{\text{Europe}} &= -.109^{**}; \quad \frac{\partial \text{pct_comlang}}{\partial \text{eldiff}_{ij}} \Big|_{\text{Asia}} = -.111^{***}; \\
 \frac{\partial \text{pct_comlang}}{\partial \text{eldiff}_{ij}} \Big|_{\text{Oceania}} &= -.040; \quad \frac{\partial \text{pct_comlang}}{\partial \text{eldiff}_{ij}} \Big|_{\text{N_America}} = -.205^{***} \\
 \frac{\partial \text{pct_comlang}}{\partial \text{eldiff}_{ij}} \Big|_{\text{S_America}} &= -.033; \quad \frac{\partial \text{pct_comlang}}{\partial \text{lqdiff}_{ij}} \Big|_{\text{Europe}} = -.101^{**}; \\
 \frac{\partial \text{pct_comlang}}{\partial \text{lqdiff}_{ij}} \Big|_{\text{Asia}} &= -.162^{**}; \quad \frac{\partial \text{pct_comlang}}{\partial \text{lqdiff}_{ij}} \Big|_{\text{Oceania}} = -.526^{***} \\
 \frac{\partial \text{pct_comlang}}{\partial \text{lqdiff}_{ij}} \Big|_{\text{N_America}} &= -.051; \quad \frac{\partial \text{pct_comlang}}{\partial \text{lqdiff}_{ij}} \Big|_{\text{S_America}} = 0.131;
 \end{aligned}$$

Specification (4) focuses on pairs of regions within China. In this case the standard errors are corrected for spatial autocorrelation, Conley (1999), using a cutoff of 1000 km's. See Appendix G for variables' definitions

Appendix E - Country Maps



Quality of Land in Lesotho



Linguistic Groups in Lesotho

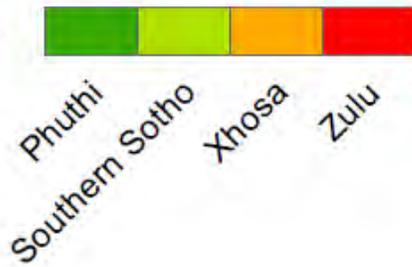
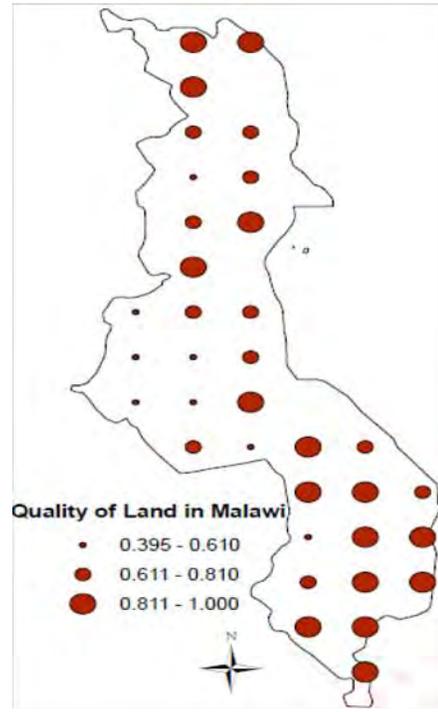
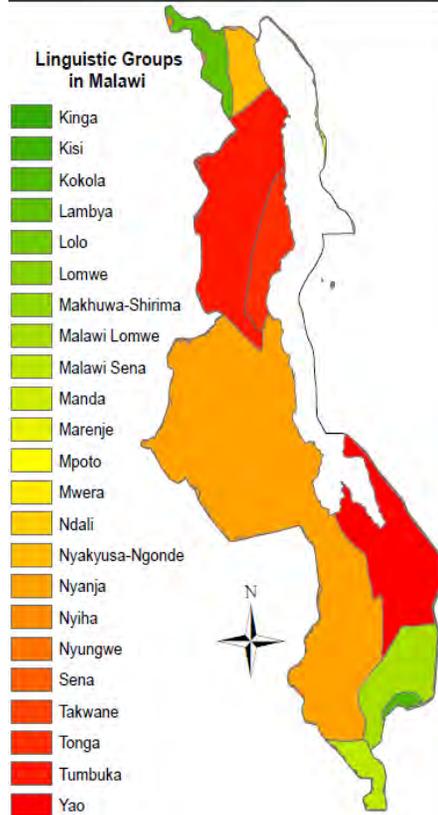
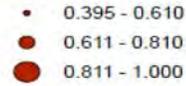


Figure 6a: Land Quality and Languages in Lesotho



Quality of Land in Malawi



Linguistic Groups in Malawi

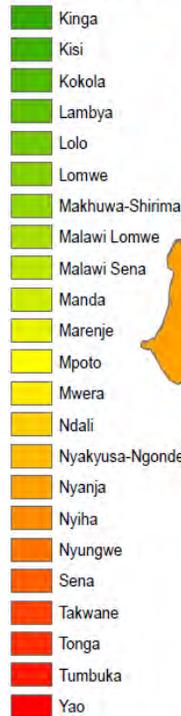


Figure 6b: Land Quality and Languages in Malawi

Appendix F - Real Country Analysis

Table 8a: Summary Statistics for the Real Country Analysis

| <i>_stats</i> | ELF | range | avg | lqqini | elev_sd | lpd1500 | yrentry | agritran | abs_lat | distcr |
|---------------|------------|--------------|------------|---------------|----------------|----------------|----------------|-----------------|----------------|---------------|
| <i>mean</i> | 0.41 | 0.74 | 0.44 | 0.33 | 0.39 | 0.90 | 1926.77 | 4835.56 | 27.52 | 0.36 |
| <i>sd</i> | 0.28 | 0.26 | 0.25 | 0.23 | 0.36 | 1.50 | 56.91 | 2386.57 | 17.41 | 0.46 |
| <i>max</i> | 0.93 | 1.00 | 0.96 | 0.88 | 2.10 | 3.84 | 1993.00 | 10500.00 | 64.92 | 2.39 |
| <i>min</i> | 0.00 | 0.01 | 0.00 | 0.03 | 0.01 | -3.82 | 1,816.00 | 400.00 | 1.36 | 0.02 |

See Appendix G for variables' definitions

Table 8b: The Correlation Matrix for the Real Country Analysis

| | ELF | range | avg | lqqini | elev_sd | lpd1500 | yrentry | agritran | abs_lat | distcr | reg_ssa |
|-----------------|------------|--------------|------------|---------------|----------------|----------------|----------------|-----------------|----------------|---------------|----------------|
| ELF | 1.00 | | | | | | | | | | |
| range | 0.21 | 1.00 | | | | | | | | | |
| avg | -0.15 | 0.26 | 1.00 | | | | | | | | |
| gini | 0.06 | 0.14 | -0.75 | 1.00 | | | | | | | |
| elev_sd | 0.11 | 0.33 | -0.07 | 0.26 | 1.00 | | | | | | |
| lpd1500 | -0.18 | 0.12 | 0.37 | -0.31 | 0.01 | 1.00 | | | | | |
| yrentry | 0.36 | -0.23 | -0.14 | -0.04 | -0.19 | -0.09 | 1.00 | | | | |
| agritran | -0.27 | 0.11 | -0.03 | 0.16 | 0.23 | 0.48 | -0.10 | 1.00 | | | |
| abs_lat | -0.37 | 0.00 | -0.07 | 0.22 | 0.07 | 0.16 | -0.21 | 0.44 | 1.00 | | |
| distcr | 0.36 | 0.19 | -0.37 | 0.50 | 0.21 | -0.30 | 0.17 | -0.02 | 0.09 | 1.00 | |
| reg_ssa | 0.53 | -0.05 | -0.13 | -0.03 | -0.24 | -0.12 | 0.36 | -0.52 | -0.53 | 0.17 | 1.00 |

See Appendix G for variables' definitions

Appendix F - Real Country Analysis

Table 9: Specifications for the Cross-Country Analysis

| VARIABLES | (1) ELF | (2) ELF | (3) ELF | (4) ELF |
|--------------|----------------------|----------------------|----------------------|----------------------|
| range | 0.417*** (0.102) | 0.339*** (0.091) | 0.316*** (0.099) | 0.333*** (0.101) |
| elev_sd | 0.052 (0.059) | 0.145*** (0.051) | 0.153*** (0.057) | 0.149*** (0.055) |
| avg | -0.715*** (0.159) | -0.387** (0.149) | -0.244 (0.158) | -0.222 (0.159) |
| lqgini | -0.700*** (0.190) | -0.444*** (0.165) | -0.478*** (0.170) | -0.449*** (0.168) |
| reg_ssa | | 0.282*** (0.050) | 0.160** (0.070) | 0.107 (0.089) |
| reg_we | | -0.170*** (0.055) | 0.059 (0.078) | 0.031 (0.084) |
| reg_lac | | -0.126** (0.055) | -0.192** (0.078) | -0.233*** (0.088) |
| reg_eap | | -0.058 (0.069) | -0.093 (0.064) | -0.101 (0.073) |
| abs_lat | | | -0.004** (0.002) | -0.004** (0.002) |
| areakm2 | | | 0.001 (0.001) | 0.000 (0.001) |
| distr | | | 0.088 (0.055) | 0.099* (0.059) |
| lpd1500 | | | -0.036* (0.021) | -0.029 (0.023) |
| yrentry | | | 0.001* (0.000) | 0.001 (0.000) |
| agritran | | | | -0.000 (0.000) |
| Observations | 149 | 149 | 145 | 142 |
| R^2 | 0.15 | 0.45 | 0.51 | 0.51 |

Robust standard errors in parentheses

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

See Appendix G for variables' definitions

Appendix F - Real Country Analysis

Figure 7a

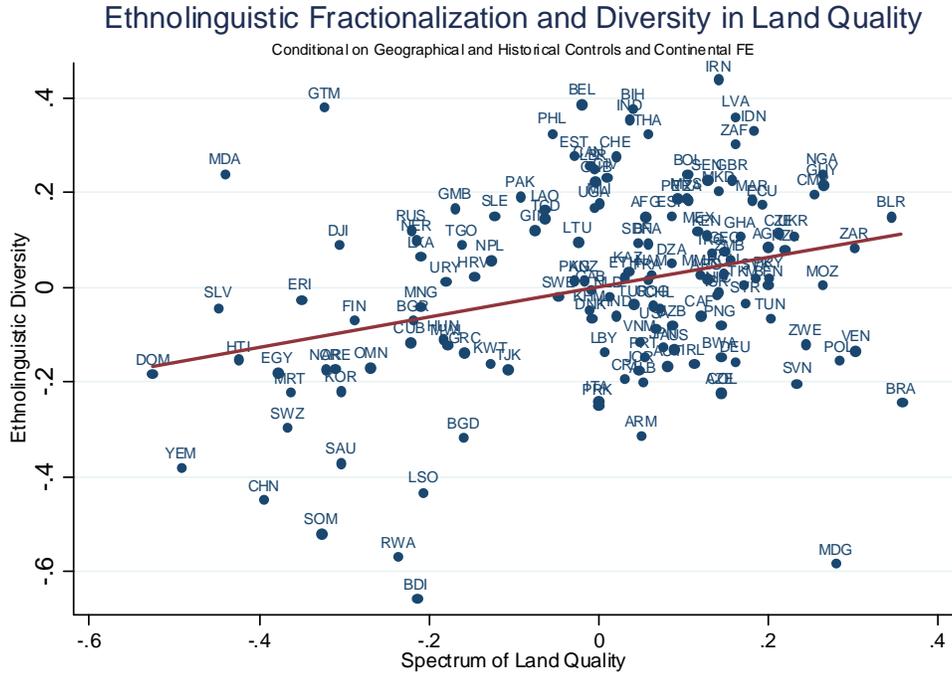
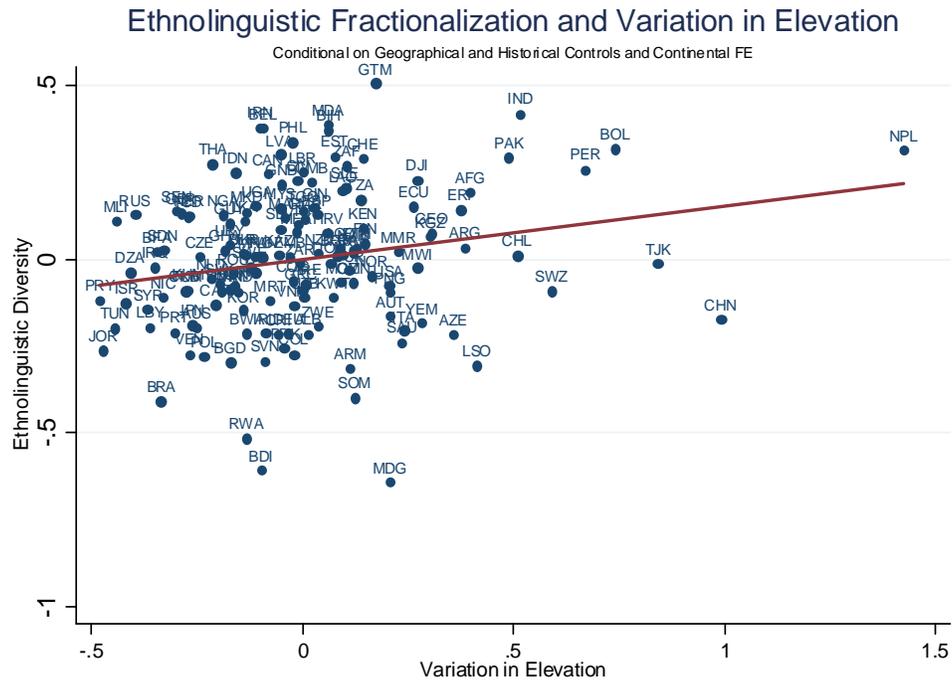


Figure 7b



Appendix F - Real Country Analysis

Table 10a: Summary Statistics for the Real Country Analysis - Geographical Variables

| <i>_stats</i> | ELF | range | range5_95 | lqsd | lqfrac | avg | lqqini | elev_sd | erange | erange5_95 |
|---------------|------------|--------------|------------------|-------------|---------------|------------|---------------|----------------|---------------|-------------------|
| <i>mean</i> | 0.41 | 0.73 | 0.57 | 0.19 | 0.39 | 0.44 | 0.33 | 0.40 | 1.71 | 1.23 |
| <i>sd</i> | 0.28 | 0.27 | 0.27 | 0.10 | 0.21 | 0.25 | 0.23 | 0.37 | 1.33 | 1.09 |
| <i>max</i> | 0.93 | 1.00 | 0.99 | 0.41 | 0.67 | 0.96 | 0.88 | 2.10 | 6.09 | 5.51 |
| <i>min</i> | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.03 | 0.01 | 0.04 | 0.04 |

See Appendix G for variables' definitions

Table 10b: The Correlation Matrix for the Real Country Analysis - Geographical Variables

| | ELF | range | range5_95 | lqsd | lqfrac | avg | lqqini | elev_sd | erange | erange5_95 |
|-------------------|------------|--------------|------------------|-------------|---------------|------------|---------------|----------------|---------------|-------------------|
| ELF | 1.00 | | | | | | | | | |
| range | 0.21 | 1.00 | | | | | | | | |
| range5_95 | 0.10 | 0.89 | 1.00 | | | | | | | |
| lqsd | 0.08 | 0.86 | 0.97 | 1.00 | | | | | | |
| lqfrac | 0.19 | 0.72 | 0.82 | 0.78 | 1.00 | | | | | |
| avg | -0.15 | 0.26 | 0.36 | 0.30 | 0.38 | 1.00 | | | | |
| lqqini | 0.06 | 0.14 | 0.05 | 0.14 | -0.21 | -0.75 | 1.00 | | | |
| elev_sd | 0.11 | 0.33 | 0.38 | 0.41 | 0.21 | -0.07 | 0.26 | 1.00 | | |
| erange | 0.17 | 0.42 | 0.41 | 0.44 | 0.21 | -0.13 | 0.34 | 0.93 | 1.00 | |
| erange5_95 | 0.12 | 0.33 | 0.37 | 0.41 | 0.2 | -0.09 | 0.28 | 0.99 | 0.94 | 1.00 |

See Appendix G for variables' definitions

Appendix F - Real Country Analysis

Table 11: Robustness Checks for the Cross-Country Analysis
Alternative measures of Geographical Diversity

| VARIABLES | (1) ELF | (2) ELF | (3) ELF | (4) ELF |
|--------------|----------------------|---------------------|---------------------|----------------------|
| range | 0.268*** (0.098) | | | |
| erange | 0.042** (0.016) | | | |
| range5_95 | | 0.177** (0.085) | | |
| erange5_95 | | 0.046** (0.020) | | |
| sd_suit | | | 0.463** (0.233) | |
| elev_sd | | | 0.136** (0.058) | 0.133** (0.056) |
| lqfrac | | | | 0.212** (0.095) |
| avg | -0.336** (0.158) | -0.251* (0.150) | -0.246 (0.153) | -0.162 (0.124) |
| lqgini | -0.473*** (0.169) | -0.356** (0.168) | -0.367** (0.172) | -0.218 (0.140) |
| reg_ssa | 0.215*** (0.068) | 0.212*** (0.071) | 0.215*** (0.072) | 0.197*** (0.073) |
| reg_we | -0.084 (0.071) | -0.072 (0.070) | -0.068 (0.069) | -0.077 (0.070) |
| reg_lac | -0.151** (0.062) | -0.160** (0.063) | -0.161** (0.063) | -0.165*** (0.063) |
| reg_eap | -0.072 (0.065) | -0.035 (0.070) | -0.033 (0.071) | -0.034 (0.071) |
| abs_lat | -0.002 (0.002) | -0.003 (0.002) | -0.003 (0.002) | -0.003 (0.002) |
| areakm2 | -0.000 (0.001) | 0.000 (0.001) | 0.000 (0.001) | 0.000 (0.001) |
| distr | 0.101* (0.051) | 0.137** (0.052) | 0.138*** (0.052) | 0.143*** (0.052) |
| Observations | 148 | 148 | 148 | 148 |
| R^2 | 0.48 | 0.45 | 0.45 | 0.46 |

Robust standard errors in parentheses,

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

See Appendix G for variables' definitions

Appendix F - Real Country Analysis

Table 12: Colonization and Man-Made Ethnic Diversity

| VARIABLES | (1) ELF | (2) ELF | (3) indig |
|--------------|----------------------|-------------------|----------------------|
| range | 0.204** (0.102) | 0.053 (0.196) | |
| elev_sd | 0.132** (0.057) | -0.113 (0.129) | |
| avg | -0.462*** (0.168) | -0.300 (0.291) | |
| gini | -0.751*** (0.175) | -0.011 (0.441) | |
| areakm2 | -0.002* (0.001) | 0.001 (0.002) | |
| distr | 0.286*** (0.043) | | |
| spanish_col | -0.001 (0.059) | | -0.512*** (0.067) |
| german_col | 0.439*** (0.040) | | -0.240 (0.176) |
| french_col | 0.237*** (0.059) | | -0.164** (0.078) |
| dutch_col | 0.380*** (0.090) | | -0.342* (0.200) |
| belgian_col | -0.246 (0.200) | | 0.063*** (0.020) |
| portu_col | 0.248** (0.117) | | -0.181 (0.170) |
| british_col | 0.205*** (0.053) | | -0.257*** (0.079) |
| italian_col | 0.153 (0.128) | | 0.040 (0.040) |
| Observations | 145 | 29 | 148 |
| R^2 | 0.45 | 0.19 | 0.30 |

Robust standard errors in parentheses

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

See Appendix G for variables' definitions

Appendix G - Data Sources

Geographical Variables

abs_lat: Absolute latitudinal distance from the equator.

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

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

Source: Center for International Development, CID.⁴¹ For the cross-virtual country and pair of adjacent regions analysis the area is constructed using ArcGIS. In the calculation are considered only areas over which both language and land quality data are available.

avg: average land quality within the respective unit of analysis

Source: Constructed by the author. The dataset is available at the Atlas of the Biosphere accessible at http://www.sage.wisc.edu/iamdata/grid_data_sel.php.

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

Source: Center for International Development, CID.

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

Source: Constructed by the author using information on elevation above sea level at a grid level. The data is aggregated at the same level as the land quality data i.e. at 0.5 degrees latitude by 0.5 degrees longitude.

Source: National Oceanic and Atmospheric Administration (NOAA) and U.S. National Geophysical Data Center, TerrainBase, release 1.0 (CD-ROM), Boulder, Colo.

Available at: <http://www.sage.wisc.edu/atlas/data.php?incdataset=Topography>

elev: average elevation within the unit of analysis in km.

Source: see **el_diff**

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

Source: see **el_diff**

erange: dispersion of elevation within the respective unit of analysis; i.e. the difference in elevation between the region with the highest elevation from that with the lowest.

Source: See **eldiff**

erange5_95: dispersion of elevation focusing between the 5th and the 95th percentile of the elevation distribution within the respective unit of analysis; i.e. the difference in elevation

⁴¹All geographical data from CID are available at: <http://www.ksg.harvard.edu/CID>

between the region with the highest elevation from that with the lowest excluding observations below the 5th and above the 95th percentile.

Source: See **eldiff**

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

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

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

lqdiff: absolute difference in land quality between adjacent regions.

Source: See **avg**

lqgini: the gini coefficient of land quality within country.

Source: See **avg**

lqfrac: The probability that two regions randomly selected from the unit of analysis will belong to different land quality groups. Land quality is grouped in three categories with the first group including regions with $avg \leq .333$, the second group featuring regional land qualities between 0.333 and 0.66 and the third group having regions with land quality larger than 0.66

Source: See **avg**

lqsd: the standard deviation of land quality within the unit of analysis.

Source: See **avg**

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

Source: This global data set, constructed by Leff et al. (2004), is intended to provide very rough indications of the probability of finding 18 major crops across the world in the early 1990s. Available at: http://www.sage.wisc.edu/iamdata/grid_data_sel.php

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

pct_comlang: number of common languages divided by the total number of languages spoken within a regional pair.

Source: see **lnnbr_lang**

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

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

http://www.sage.wisc.edu/iamdata/grid_data_sel.php

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

Source: See **avg**

range5_95: spectrum of land qualities focusing between the 5th and the 95th percentile of the land quality distribution within the respective unit of analysis; i.e. the difference in land quality between the region with the highest land quality from that with the lowest excluding regions below the 5th and above the 95th percentile.

Source: See **eldiff**

reg_eap: dummy variable equals 1 for countries in East Asia and Pacific.

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

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

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

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

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

waterarea: total area within the respective unit of analysis covered by river or lake.

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

Historical Variables

ELF: level of ethnolinguistic fractionalization within a country.

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

lpd1500: log population density in 1500.

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

yrentry: year a country achieved independence.

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

agritran: Year when the first significant region within a present-day country underwent a transition from reliance mainly on hunting and gathering to reliance mainly on cultivated crops (and livestock).

Source: Putterman, L., Agricultural Transition Data Set, Brown University, http://www.econ.brown.edu/fac/Louis_Putterman/agricultural%20data%20page.htm

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

http://www.econ.brown.edu/fac/Louis_Putterman/world%20migration%20matrix.htm

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

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

Source: "Determinants and Economic Consequences of Colonization: A Global Analysis"

Ertan, A., Putterman, L.,

Supplemented by entries from Encyclopedia Britannica where necessary.

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

Source: see **belgian_col**

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

Source: see **belgian_col**

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

Source: see **belgian_col**

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

Source: see **belgian_col**

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

Source: see **belgian_col**

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

Source: see **belgian_col**

References

- Ahlerup, Pelle and Ola Olsson**, “The Roots of Ethnic Diversity,” 2008. Working Papers, University of Gothenburg.
- Alesina, Alberto and Enrico Spolaore**, “On the Number and Size of Nations,” *Quarterly Journal of Economics*, 1997, *112*, 1027–1056.
- , **Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg**, “Fractionalization,” *Journal of Economic Growth*, 2003, *8*, 155–194.
- , **William Easterly, and Janina Matuszeski**, “Artificial States,” 2006. NBER Working Paper 12328.
- Atlas Narodov Mira (Atlas of the People of the World)*, Moscow: Glavnoe Upravlenie Geodezii i Kartograi, Bruck, S.I., and V.S. Apenchenko, 1964.
- Banerjee, Abhijit and Rohini Somanathan**, “The Political Economy of Public Goods: Some Evidence from India,” *Journal of Development Economics*, 2006, *82*, 287–314.
- Barth, Frederik**, *Ethnic Groups and Boundaries: The Social Organization of Cultural Difference*, Boston: Little, Brown, 1969.
- Bellwood, Peter**, “Early Agriculturalist Population Diasporas? Farming, Languages, and Genes,” *Annual Review of Anthropology*, 2001, *30*, 181–207.
- Bockstette, Valerie, Areendam Chanda, and Louis Putterman**, “States and Markets: The Advantage of an Early Start,” *Journal of Economic Growth*, 2002, *7* (4), 347–369.
- Botticini, Maristella and Zvi Eckstein**, “From Farmers to Merchants, Voluntary Conversions and Diaspora: A Human Capital Interpretation of Jewish History,” *Journal of Economic History*, 2005, *65*, 922–948.
- Boyd, Robert and Peter J. Richerson**, *Culture and the Evolutionary Process*, Chicago, IL: University of Chicago Press, 1985.
- Caselli, Francesco and Wilbur J. Coleman**, “On the Theory of Ethnic Conflict,” 2006. Working Paper, mimeo London School of Economics.
- Conley, Timothy G.**, “GMM Estimation with Cross Sectional Dependence,” *Journal of Econometrics*, 1999, *92*, 1–45.

- Curtin, Philip D.**, *Cross-Cultural Trade in World History*, Cambridge: Cambridge University Press, 1984.
- Darwin, Charles**, *The Voyage of the Beagle*, Reprinted by Black Dog and Leventhal Publishers, Originally 1839, Reprinted in 2006.
- Diamond, Jared**, *Collapse: How Societies Choose to Fail or Succeed*, New York, NY: Viking Press, 2005.
- Easterly, William and Ross Levine**, “Africa’s Growth Tragedy: Policies and Ethnic Divisions,” *Quarterly Journal of Economics*, 1997, 112 (4), 1203–1250.
- Esteban, Joan and Debraj Ray**, “On the Saliency of Ethnic Conflict,” 2007. Working Paper, mimeo New York University.
- Fearon, James**, “Ethnic Structure and Cultural Diversity by Country,” *Journal of Economic Growth*, 2003, 8, 195–222.
- and **David Laitin**, “Ethnicity, Insurgency and Civil War,” *American Political Science Review*, 2003, 97, 75–90.
- Geertz, Clifford**, *Old Societies and New States: The Quest for Modernity in Asia and Africa*, New York: Free Press, 1967.
- Gray, Russell D. and Quentin D. Atkinson**, “Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin,” *Nature*, 2003, 426, 435–439.
- Grigg, David B.**, *An Introduction to Agricultural Geography*, Routledge, London and New York, 1995.
- Hale, Henry E.**, “Explaining Ethnicity,” *Comparative Political Studies*, 2004, 37 (4), 458–485.
- Herbst, Jeffrey**, *State and Power in Africa*, Princeton, NJ: Princeton University Press, 2002.
- Laitin, David D.**, *Language Repertoires and State Construction in Africa*, Cambridge: Cambridge University Press, 1992.
- , *Nations, States, and Violence*, Oxford: Oxford University Press, 2007.
- Leff, B, N Ramankutty, and J.A. Foley**, “Geographic Distribution of Major Crops Across the World,” *Global Biogeochemical Cycles*, 2004, 18 (1).

- Michalopoulos, Stelios**, “Natural versus Man-Made Ethnolinguistic Diversity: Implications for Comparative Economic Development,” 2009. Mimeo, Department of Economics, Tufts University.
- Miguel, Edward and Daniel Posner**, “Sources of Ethnic Identification in Africa,” 2006. Working Paper, mimeo University of California, Berkeley.
- Montalvo, José G. and Marta Reynal-Querol**, “Ethnic Polarization, Potential Conflict and Civil War,” *American Economic Review*, 2005, *95*, 796–816.
- Nettle, Daniel**, *Linguistic Diversity*, Oxford: Oxford University Press, 1996.
- Nichols, Johanna**, “Modeling Ancient Population Structures and Movement in Linguistics,” *Annual Review of Anthropology*, 1997a, *26* (6), 359–384.
- Olson, Paul A.**, *Struggle for the Land: Indigenous Insight and Industrial Empire in the Semiarid World*, Nebraska: University of Nebraska Press, 1990.
- Pavitt, Nigel**, *Samburu*, Kyle Kathie Limited, 2001.
- Ramankutty, Navin, Jonathan A. Foley, John Norman, and Kevin McSweeney**, “The Global Distribution of Cultivable Lands: Current Patterns and Sensitivity to Possible Climate Change,” *Global Ecology and Biogeography*, 2002, *11*, 377–392.
- Rao, Vijayendra and Radu Ban**, “The Political Construction of Caste in South India,” 2007. mimeo, The World Bank.
- Renfrew, Colin**, “At the Edge of Knowability: Towards a Prehistory of Languages,” *Cambridge Archaeological Journal*, 2000, *10*, 7–34.
- Rosenzweig, Michael L.**, *Species Diversity in Space and Time*, New York, NY: Cambridge University Press, 1995.
- Spolaore, Enrico and Romain Wacziarg**, “The Diffusion of Development,” *Quarterly Journal of Economics*, 2009, *124* (2).