共同利用・共同研究拠点 Joint Usage/Research Center
東京大学 The University of Tokyo
空間情報科学研究センター
Center for Spatial Information Science

Discussion Paper Series

No. 169
CSIS Discussion Paper

September 2020

Linguistic Distance and Economic Prosperity: A Cross-Country Analysis

Mariko Nakagawa
(Center for Spatial Information Science, University of Tokyo & Tohoku University)
Shonosuke Sugasawa
(Center for Spatial Information Science, University of Tokyo)

Linguistic Distance and Economic Prosperity: A Cross-Country Analysis*

Mariko Nakagawa† and Shonosuke Sugasawa‡

September 23, 2020

Abstract

We investigate the impacts of access to domestic and international communication on economic development by constructing two indices of linguistic distance—domestic and international—, capturing language acquisition costs, which are higher when acquiring linguistically more distant languages. While the domestic linguistic distance index captures the constraints of communication among speakers of different mother tongues within a country, the international linguistic distance index captures the constraints of global communication via English. We find that domestic linguistic distance has a negative impact on GDP per capita, while international linguistic distance has no significant impact. In addition, we conduct quantile regressions to see if the negative effect of the domestic linguistic distance differs across quantiles of GDP per capita. The analysis reveals that the domestic linguistic distance index robustly has a negative impact on GDP per capita for countries over various quantiles, but that there is no clear tendency of a heterogeneous impact on economic prosperity across quantiles given by the domestic linguistic distance.

JEL code: O1, O5, F5

Key words: domestic linguistic distance, international linguistic distance, mismatch in language use, economic success

*We would like to thank Takatoshi Tabuchi for his thoughtful comments and suggestions. We also thank Ryo Ito, Marcus Berliant, Takaaki Takahashi, Dan Sasaki, Hikaru Ogawa, and Shota Fujishima for their comments which have improved this paper. We are grateful to the seminar participants at the Urban Economics Workshop at the University of Tokyo, the ARSC annual meeting at Ryukyu University, Asian Seminar in Regional Science at Tohoku University, and Policy Modeling Workshop at National Graduate Institute for Policy Studies for their valuable comments. All remaining errors are the authors' responsibility. This study is supported by the Grants-in-Aid for Scientific Research (Research project number: 13J10130) for the Japan Society for the Promotion of Science (JSPS) Fellows and Research Activity Start-up (Research project number: 16H06703).

†Center for Spatial Information Science, the University of Tokyo and the Graduate School of Information Sciences, Tohoku University. Email: mnakagawa@csis.u-tokyo.ac.jp

‡Center for Spatial Information Science, the University of Tokyo. Email: sugasawa@csis.u-tokyo.ac.jp

1 Introduction

In recent years, there has been increasing research interest on the impacts of ethno-linguistic heterogeneity on economic and political activities. With the realization that ethno-linguistic heterogeneity is important when considering various social phenomena, a vast body of empirical and theoretical literature on ethno-linguistic diversity has developed. Easterly and Levine (1997), Alesina et al. (2003), and Alesina and La Ferrara (2005) investigate the impacts of ethno-linguistic diversity on economic development. Easterly and Levine (1997) show that ethnic diversity negatively affects a country's income level and explains the low incomes in African nations in particular. According to Alesina et al.'s (2003) cross-country analysis, ethnically and linguistically diverse structures decrease growth levels, which implies that heterogeneous composition of ethno-linguistic characteristics negatively affects economic development. Alesina and La Ferrara (2005) argue that the costs of heterogeneity come from difficulty reaching agreement on productive public goods and policies common to all ethno-linguistic groups. The difficulty of harmonization between groups could negatively affect economically lagging countries, such as some African countries, which are kept from catching up with developed countries.

What factors hinder harmonized conformity and communication among different ethno-linguistic groups? Admittedly, lack of trust toward or perceived reliability of other groups, or cultural difference affects the difficulty in arriving at agreement among various groups. In addition, one of the essential barriers to sufficient harmonization between various types of ethno-linguistic groups, especially linguistic differences, stems from difficulty in smooth communication caused by the lack of effectively common languages, or at least common languages that can be shared with little acquisition effort. It is typical of countries with colonized histories to have the colonizer's language(s) as the official language(s). On the other hand, mother tongues of native residents in such post-colonial countries is likely to be a local ethnic language that is linguistically distant from colonizers' languages.

To capture the cost required to attain the ability to use domestic official languages different from one's mother tongue when a society consists of different linguistic groups, we use a measure of within-country linguistic distance, which expresses how distant residents' mother tongues is from the official language in terms of linguistic characteristics. This linguistic distance in the domain of domestic communication is reflective of language acquisition costs (i.e., effort made) when acquiring domestic central languages. As mentioned in Chiswick and Miller (2005) and Isphording and Otten (2011, 2013), it is more difficult to acquire a language if the native language is linguistically distant from the language to be learned (in this case, the official language(s) of a country) and, hence, the measure of

domestic linguistic distance to the official language(s) expresses the language acquisition costs incurred by individuals whose mother tongues are different from the official language(s). We conduct analysis of the impact of domestic linguistic distance on cross-country differences in economic success.

Similar language acquisition costs necessary to accomplish smooth communication might be found in the international context. Recent empirical studies suggest great importance of languages spoken worldwide, such as English, for economic success in the modern era, when things and individuals are densely connected globally. For example, in the realm of international migration, Adsera and Pytlikova (2015) argue that English as a global language is one of the crucial factors determining the configuration of migration decisions, by investigating whether potential immigrants prefer destinations where English is the local language. In addition, Ku and Zussman (2010) report that trade partners sharing no common native languages mitigate the language barriers through communication in a nonnative but highly influential language, English. They show that proficiency in English can promote international trade and, thus, that acquired English capability can overcome difficulty in communication. Similarly, Hutchinson (2005) argues that trade between the US and another country will be less intense if languages used in that country are different from English. Moreover, as pointed out in the literature, such as Jones (2001, Chapter 6) and Ku and Zussman (2010), which focus on the communicative benefits of English as a lingua franca, acquired proficiency in English could promote economic development by maintaining access to the world's stock of technological knowledge that has recently been accumulated in English. Given the substantive importance of English, acquisition cost of English, which can be captured by linguistic distance between residents' mother tongues in a country and English, may also impact economic outcomes of that country.

The impacts of distance to English on economic development are considered with a slightly different concept from language acquisition in Spolaore and Wacziarg (2009), who provide an alternative explanation for why distance to the US might matter. Their concept of distance is genetic distance, a measure of the time passed since two populations shared ancestors. The authors show that variation in genetic distance to the US, the most developed country, explains dispersion in economic development, based on an idea that individuals genetically close to the US residents are more likely to follow or to more easily catch up with the latest technology due to similarity of cultures or of methods of problem solving.

Below, we briefly summarize the paper. First, we construct linguistic indices. As for the difficulty in accessing effective domestic communication—that is, acquisition costs of the official language(s)

necessary to communicate with residents in a whole country—, we construct a domestic linguistic distance, DLD, for each country. DLD is calculated as a population-weighted average of linguistic distance between the official language and residents' mother tongues. Similarly, we construct an international linguistic distance, ILD, to capture the aspect of a cost to access the global communication via English. As for ILD, we construct two types of indices, ILD_{PC} and ILD_{CC}, the former calculated as the population-weighted average of distances between English and residents' mother tongues in a country, and the latter of which is calculated as the linguistic distance between English and the country's official language.

The main results are as follows. We find a significantly negative impact of DLD on GDP per capita, while no significant impact is found in the case of ILDs, which implies that difficulty in communication within a country from a viewpoint of mismatch in language use hinders economic success. By contrast, difficulty in English communication may not have a significant impact on a country's economic success, at least at the country-aggregated level. We also conduct an analysis to see if the DLD's impact differs between rich and poor countries by using quantile regression techniques. We find that DLD robustly has a negative impact on GDP per capita for countries over various quantiles of GDP per capita, but no clear tendency that DLD has a heterogeneous impact on economic prosperity across quantiles.

The remainder of this paper is organized as follows. Section 2 introduces notions of domestic and international linguistic centers and distances and provides indices for the linguistic distance. Section 3 presents an empirical framework of the baseline regression and a couple of comments on statistical concerns. We also address possible remedies for such concerns, although they are not a perfect way to resolve them. Section 4 conducts baseline regressions, showing that DLD has a negative impact on GDP per capita. Section 5 conducts quantile regressions to investigate possibility of heterogeneous impacts of DLD on GDP per capita. In Section 6, robustness checks of the negative impacts of linguistic distance indices are conducted, and Section 7 concludes the paper.

2 Indices of linguistic distance

2.1 Definition of linguistic distance

Our linguistic distance index data cover a wide cross-section of countries, where indices for each country are based on weighted averages of linguistic distances. For the first step, we calculate the linguistic distance for each pair of living languages listed in the 16th edition of Ethnologue (Lewis,

2009). In Ethnologue, the entire world's known living languages are listed and are categorized based on the similarities in their linguistic characteristics. If two arbitrarily chosen languages belong to the same linguistic family, these are thought to be similar and, thus to exhibit shorter linguistic distance. Before introducing the concept of linguistic distance, the notion of linguistic similarity is considered.

In order to provide a quantitative value for the abstract notion of linguistic similarity, linguistic dendrograms, constructed from the language categorization provided in Ethnologue, are utilized. The number of shared edges between languages i and j on a dendrogram are denoted by $e(i,j)$. If $e(i,j)$ is large, it implies that languages i and j are categorized into a meta-group, such that they are closer linguistically. When quantifying linguistic similarity, we employ the approach of Fearon (2003) and Desmet et al. (2009). First, we define $g_{\max}$ as the maximum number of $g(i)$ for all existing languages i in the world, where $g(i)$ is the generation to which language i belongs, which is used to convert $e(i,j)$ into proportions of cognates between i and j (normalization into the interval $[0,1]$). Put differently, $g_{\max}$ is the maximum of the number of edges that can be shared by two languages.[1] Then, linguistic similarity is formulated as follows:

$$\text{similarity}(i,j) = \frac{e(i,j)}{g_{\max}}.$$

Now, we define linguistic distance between languages i and j, $\tau(i,j)$. Higher similarities show shorter linguistic distances; hence, if similarity(i,j) increases, $\tau(i,j)$ decreases. Furthermore, we assume that $\tau(i,j) = \tau(j,i)$ for all languages i and j. $\tau(i,j)$ is a standardized metric (i.e., $\tau(i,j) \in [0,1]$), and $\tau(i,i) = 0$ for all i. Under these assumptions on metrics, $\tau(i,j)$, is defined as

$$\tau(i,j) = 1 - [\text{similarity}(i,j)]^{\delta} = 1 - \left[\frac{e(i,j)}{g_{\max}}\right]^{\delta} \qquad \text{for all} \;\; i,j (i \neq j),$$

where $\delta \in (0,1)$ is a parameter determining how fast the linguistic distance declines as the number of shared edges increases. More intuitively, as mentioned in Desmet et al. (2009), δ captures how much more distant two languages from different linguistic families are, compared with languages that belong to the same family.

[1] The linguistic dendrogram constructed from Ethnologue has 15 nested classifications, so that $g_{\max} = 15$.

2.2 Indices of linguistic distance

Our linguistic distance indices are based on the population-weighted averages of linguistic distances. Consider country i with a population of $N(i)$ individuals, who are partitioned into $K(i)$ distinct language groups according to their language use. $N_j(i)$ is a population of language group j in country i. We assume that each individual belongs to only one language group, and no individual is assumed completely multilingual, which leads to

$$N(i) = \sum_{j=1}^{K(i)} N_j(i).$$

The population share of group j in country i, $s_j(i)$, is defined as

$$s_j(i) = \frac{N_j(i)}{N(i)}$$

so that

$$\sum_{j=1}^{K(i)} s_j(i) = 1$$

for all countries.

First, we define the domestic linguistic distance index, $DLD(i)$, for each country i, which is interpreted as the cost incurred when accessing the "linguistic center" of country i or an official language in country i, and by accessing it, the residents can communicate with each other. In other words, without acquiring the nationally widely spoken language(s), communication among the residents of a nation is impossible. Language $c(i)$ is the central language of country i, and $\tau_{j,c(i)}$ is the linguistic distance between languages j and $c(i)$. $DLD(i)$ is defined as the population-weighted average of linguistic distances to the domestic central language:[2]

$$DLD(i) = \sum_{j=1}^{K(i)} s_j(i)\tau_{j,c(i)}. \tag{1}$$

Next, we define the international linguistic distance index, $ILD(i)$. Here, English is adopted as the

[2]Ginsburgh et al. (2005) consider which language(s) should be considered appropriate as the official language(s) of the European Union by calculating the population-weighted average of linguistic distances from residents' mother tongues to major European languages. In so doing, the authors search for languages whose population-weighted averages of linguistic distances are the smallest, because linguistic distance is considered the cost of acquiring other languages. Constructed linguistic indices in Ginsburgh et al. (2005) are similar to our $DLD(i)$, but in their analysis, linguistic distance indices are utilized to determine the central language(s), while our $DLD(i)$ is constructed given the linguistic center.

central language of the world, symbolized as C (by acquiring English, smooth global communication is possible). As for the choice of ILD, there are two possibilities, ILD_{PC} and ILD_{CC}:

$$ILD_{PC}(i) = \sum_{j=1}^{K(i)} s_j(i)\tau_{j,C}, \tag{2}$$

$$ILD_{CC}(i) = \sum_{j=1}^{K(i)} s_j(i)\tau_{c(i),C} = \tau_{c(i),C}. \tag{3}$$

$ILD_{PC}(i)$ refers to the linguistic distance cost that each person in group j in country i incurs to access the international center, besides the cost to access the domestic center in country i. $ILD_{CC}(i)$ is the linguistic distance between the domestic and international central languages (the domestic official language and English).

2.3 Linguistic center

To define linguistic centers, we introduce the notion of "language status" proposed in Ethnologue 17th edition (Lewis et al., 2014) on the criteria of the Expand Graded Intergenerational Description Scale (EGIDS), which ranges from status 0 to 10 according to importance or usage of the languages. For example, languages labeled status 0 (international) are widely used between nations in trade, knowledge exchange, and international policy. Status-1 languages (official languages) are used in education, work, mass media, and government at the national level.[3]

As for the international linguistic center, English is considered the most appropriate, because among languages with status 0, it has the biggest second-language user population.[4] For the domestic linguistic center(s), official language(s) (status-1 languages) are chosen. Given the definition of domestic linguistic center, some modifications to linguistic distance indices are necessary. First, we consider revising DLD. If country i has several status-1 languages, DLD in (1) is modified as

$$DLD(i) = \sum_{c(i)\in C(i)} \gamma_{c(i)} \sum_{j=1}^{K(i)} s_j(i)\tau_{j,c(i)}, \tag{4}$$

[3]For more details of language status definition, see www.ethnologue.com/about/language-status.
[4]Arabic, Chinese, English, French, Russian, and Spanish are labeled as status-0 languages.

where $C(i)$ is a set of status-1 languages in country i, and $\gamma_{c(i)}$ is defined as

$$\gamma_{c(i)} = \frac{N_{c(i)}}{\sum_{l \in C(i)} N_l}.$$

In short, DLD for a country with multiple official languages is the weighted average of the weighted averages of the domestic linguistic distances to status-1 languages. Similarly, ILD_{CC} in (3) index is modified as

$$ILD_{CC}(i) = \sum_{c(i) \in C(i)} \gamma_{c(i)} \sum_{j=1}^{K(i)} s_j(i) \tau_{c(i),C}. \tag{5}$$

Because it is not necessary to apply any modification for ILD_{PC}, the revised DLD and ILD_{CC} given in (4) and (5) respectively, as well as the original ILD_{PC} given in (2), are exploited in our empirical analysis.[5]

3 Model specification

Our main interest lies in empirically investigating the relationship between domestic and international distances and economic success; to measure the economic prosperity of a country, we adopt the country's income (GDP per capita) at real PPP from the Penn World Tables 8.0 (Feenstra et al., 2013b), and our empirical model is specified as follows:

$$\ln \text{GDP/capita}_i = \beta_0 + \beta_D DLD_i + \beta_I ILD_i + X_{\text{control},i} \beta_{\text{control}} + \epsilon_i, \tag{6}$$

where ILD is either ILD_{PC} or ILD_{CC}. In the PWT 8.0, expenditure- or output-side GDP is available, and by following a user guide of PWT 8.0 (Feenstra et al., 2013a), the output-side measure is adopted. Hereafter, we notate country indicator i as a subscript. In addition, because DLD and ILD depend on a given value of linguistic distance parameter δ_D and δ_I, respectively, we denote $DLD(\delta_D)$ and $ILD(\delta_I)$, respectively.

A vector of control variables (X_{control}) contains variables in terms of market size, education, trade openness, institutional qualities, and geographic characteristics. The choice of controls essentially follows the perspectives of those in Alesina et al. (2016). Since the use of a large number of control

[5] In some countries with multiple official languages, one (or more than one) official language(s) has no population of speakers of that language as their mother tongues. For example, Ethnologue reports that Cameroon, which has two national languages, English and French, has no population of speakers of these two official languages. For such countries, special treatments are needed to calculate the linguistic distance indices. Notes for constructing linguistic distances for such cases are listed in Table A3 in Appendix B.

variables under a moderate sample size (111 in this study) may lead to unstable fitting of quantile regressions in low or high quantiles that we will conduct in Section 5, it would be better to omit controls that are not necessarily recognized as having effects on GDP per capita. Then, we choose the following control variables for our baseline model. We adopt years of schooling (Barro and Lee, 2013) for education covariates, and population sizes from the Penn World Tables 8.0 (Feenstra et al., 2013b) for the market size control. Furthermore, we control for the standard trade volume measure, trade openness[6] (i.e., export + import share of GDP in real PPP prices) from PWT 8.0. As the vector of geographical determinants,[7] we adopt absolute latitude of the capital city and land area from CEPII (Head et al., 2010), and a ratio of population within 100 km of ice-free coast to total population from Gallup et al. (1999). For institutional quality, we choose the revised combined polity2 score from the Polity IV database (Marshall and Jaggers, 2012), which measures the extent to which political participation is unrestricted, open, and fully competitive, executive recruitment is elective, and to which constraints on the chief executive are substantial.[8] Further, following La Porta et al. (1999), legal origin dummy variables are included.[9]

Before running regressions, a couple of comments should be made. Occasionally, regional fixed effects are adopted as covariates to explain the extent of economic development. However, DLD, which is one of the primary points of interest in our context, has high correlation with the Sub-Saharan African continental dummy variable in the full range of the values of parameter δ_D (correlation between $DLD(\delta_D = 0.5)$ and the Sub-Saharan African dummy variable is 0.71 and they show high correlation throughout $\delta_D \in \{0.1, \ldots, 0.9\}$); and thus, we are forced to drop regional dummy variable from our regressions. Despite excluding regional fixed effects from the baseline model, in Section 6.6 we consider this aspect from a viewpoint of spatial econometrics.

Another noteworthy issue concerns endogeneity caused by reverse causality. As considered in Section 6.3, the effects of linguistic distance, especially in richer economies, might be affected by immigrants attracted to such countries. Although the following treatments are not perfect remedies, they reduce the reverse causality concerns related to immigrant attraction of the rich countries. First,

[6]The impacts of trade on a country's income or growth are discussed in Frankel and Romer (1999), Rodriguez and Rodrik (2001), and Yanikkaya (2003).

[7]As for the impacts of geographical determinants of economic development, see Sachs (2003) and Putterman and Weil (2010).

[8]Hall and Jones (1999), Acemoglu et al. (2001), Glaeser et al. (2004), and Rodrik et al. (2004) consider how institutional qualities affect economic growth or income difference.

[9]We only include socialist law and do not include British law, French law, and German law, in order not to have excessive the number of control variables. However, we also conducted a regression with all these legal origin variables, which revealed to have an effectively the same result as we obtain in the baseline estimation. A result is available upon request.

when constructing linguistic distance indices, we excluded the "immigrant languages" reported in Ethnologue. Ethnologue separately reports immigrant languages, which are categorized as such if they are spoken by relatively recently arrived or transient populations. As our linguistic distance indices do not include such immigrant populations/languages, the reverse causality concern of immigrant attraction may be mitigated. In addition, in order to tackle the possibility of rich economies' attraction of migrants, Section 6.3 directly includes either a stock migrant population size or a stock migrant population share in a set of explanatory variables. Moreover, in all specifications, independent variables precede the dependent variable, GDP per capita. Treating the linguistic distance data, as well as other control variables, as predetermined by having them preceding the dependent variable might mitigate the problem of reverse causality.[10][11]

4 Empirical Results

4.1 Results of the baseline model

This study concerns the effect of domestic and international linguistic costs, and our first prediction is that both intra- and international linguistic distances have negative impacts on GDP per capita. If it is difficult to access the domestic linguistic center (official language), it is difficult to communicate within a country. Furthermore, acquiring languages other than one's mother tongue is costly, which might reduce economic activity performance. The same can be said for international linguistic distance. If acquiring English is costly and individuals have difficulty with fluent English communication, then they are likely to lose opportunities to create global connections, which may prevent improvement to economic status of a country.

Table 1 displays the baseline results.

[Table 1 around here]

Column (1) shows the result based on DLD and ILD_{PC} and column (3) is based on DLD and ILD_{CC} without control variables, and columns (2) and (4) are those with baseline control variables. Throughout the manuscript, DLD, ILD_{PC}, and ILD_{CC} are mainly calculated under $\delta_D = 0.5$ and

[10] The GDP per capita is in 2010, while linguistic distance indices are calculated by using Ethnologue data published in 2009. Due to the updating scheme of the Ethnologue, not a small number of linguistic compositions are based on the data before 2009. To coordinate the Ethnologue data, other independent variables are mainly from year 2005.

[11] Tables A1 and A2 in Appendix A exhibit the data sources and the summary statistics, respectively.

$\delta_I = 0.5$. Below in this section, we check the robustness of the results by changing the values of δ_D and δ_I within the range of 0 to 1.

As for the results without controlling for other covariates, both DLD and ILD are significantly negative, as shown in column (1) for the specification with ILD_{PC} and column (3) for that with ILD_{CC}. However, after controlling for the baseline control variables, ILD loses significance and is indistinguishable from zero for both specifications. This may be partly because, especially in developing countries, only a small number of elites are required to acquire English as well as the domestic official language, but the rest of the citizens may not. Thus, when investigating the effect of access to English communication at the country level, the importance of English skills does not appear clear. By contrast, DLD keeps significantly negative at the 1% significance level, which is an indication of the negative effect on the output-side GDP per capita given by the difficulty in domestic communication in terms of acquisition costs of the official language other than one's mother tongue, if the interpretation of the correlation as a causal inference is permitted.

Previous literature focuses on the negative effect of ethno-linguistic diversity on economic performance, as reviewed in Section 1. Previous studies interpret that the negative effect of ethno-linguistic diversity is caused by the inability to agree on common public goods and policies. Individuals with different preferences have to share common policies, which would decrease the average utility level, leading to negative effects on some economic aspects. Instead, we tackle this negative effect from the viewpoint of a linguistic distance cost, which captures difficulty acquiring an official language other than one's mother tongue. We will turn to this point in relation with linguistic diversity in Section 6.4.

Since linguistic distance parameters can take various values in the range in which they are defined, we vary the parameter values and rerun regressions. Table 2 reports the coefficients of domestic and international linguistic distance indices with different linguistic parameter values (δ_D, δ_I), whose specifications are based on the regression including DLD and ILD_{PC} with the baseline control variables.[12] Values of δ_D changes in the vertical direction (δ_D increases from the top to the bottom in the range of 0 to 1) and those of δ_I varies in the horizontal direction (δ_I increases from left to right in the range of 0 to 1).

[Table 2 around here]

[12]Table A4 in Appendix C shows an analogous result based on the specification with DLD and ILD_{CC}. The result under this specification is effectively the same as that with DLD and ILD_{PC} shown in Table 2.

The behavior of coefficients on DLD listed in Table 2 shows strong robustness of negative significance to a change in different values of δ_D and δ_I. This stability of DLD negativity confirms the claim that difficulty in domestic communication from the viewpoint of language acquisition hinders economic prosperity as assessed by the output-side GDP per capita.

5 Quantile regression

In this section, we investigate whether the negative effect of DLD found in Section 4.1 differs across quantiles of the country's per-capita GDP. As discussed in Alesina et al. (2016), in which the authors show that the effect of birthplace diversity on GDP per capita differs between relatively poor countries and rich countries, we consider the possibility of heterogeneous impacts of DLD on GDP per capita. Unlike in Alesina et al. (2016), in which the whole sample is divided into two subsamples, poor countries are defined as countries with GDP per capita lower than the median GDP per capita and rich countries as those with GDP per capita higher than the median, and regressions run separately for two subsamples, we conduct quantile regression.[13] The advantage of quantile regression is that by directly modeling low or high quantiles of GDP per capita, it can avoid the dichotomized groupings such as "poor" and "rich", which are often sensitive to the method of grouping. Furthermore, it can provide a wide range of information regarding the potentially heterogeneous effects of DLD on different quantiles of GDP per capita.

The quantile regression estimator is obtained by solving the following optimization problem:

$$\min_{\beta} \left[\sum_{i=1}^{n} \rho_\theta \left(\ln \text{GDP/capita}_i - (\beta_0 + \beta_D DLD_i + \beta_I ILD_i + X_{\text{control},i}\beta_{\text{control}}) \right) \right],$$

where $\rho_\theta(u) = u(\theta - I(u < 0))$ for the θ-th quantile $(0 < \theta < 1)$.

Table 3 displays the result of quantile regressions for different quantiles, and Figure 1 is the corresponding graphical illustrations, based on the ILD_{PC} specification.[14]

[Table 3 and Figure 1 around here]

[13] Quantile regression methods are applied to the cross-country growth analysis. Foster (2008) examines the relationship between trade liberalization and per-capita GDP growth. The author utilizes the quantile regression techniques to consider the parameter heterogeneity of the impact of trade liberalization. Also, Dufrenot et al. (2010) employ quantile regression methods to see how the impact on the growth rate of per-capita income given by trade openness differs across different quantiles.

[14] Table A5 and Figure A1 in Appendix C exhibit the results for the specification with ILD_{CC}. The results do not change effectively.

The values of the goodness-of-fit (GOF) measure (Koenker and Machado, 1999) in Table 3 suggest that all the models fit the data reasonably well. It also shows that ILD does not have significant effects on all the quantiles of GDP per capita, whereas DLD has always negative effects on it. Although the magnitude of the negative effects of DLD can be heterogeneous over different quantiles, no clear tendency of the behavior of the coefficient of DLD across quantiles is observed. For example, if the negative effect of DLD were more severe in poor countries with lower GDP per capita than rich countries with higher GDP per capita, the absolute value of the coefficient of DLD might be larger for lower quantiles, but such a clear and easily interpretable tendency is not corroborated in our quantile regression results. In Figure 1, we report 81 paths of estimated regression coefficients in 9 quantiles, where each path corresponds to a quantile regression model using $\delta_D, \delta_I \in \{0.1, \ldots, 0.9\}$. This shows that the results in Table 3 are rather robust against alternative specifications of δ_D and δ_I.

6 Robustness checks

6.1 Natural characteristics

In this class of robustness checks, we consider the effects of natural characteristics on GDP per capita, in terms of climate and geographic aspects. First, we include the average annual temperature and precipitation in the baseline model, as mild climate is considered to facilitate human activities, which can lead to improved economic status of a country. Table 4 shows the result for this robustness check.

[Table 4 around here]

For the specification of ILD_{PC} and ILD_{CC}, column (1) and column (4), respectively, include the average annual temperature, column (2) and column (5), respectively, include the average annual precipitation, and column (3) and column (6), respectively, include both. In all columns, the significantly negative coefficient of DLD remains unchanged.

Next, we consider the effect of elevation. The rough and ragged land or mountainous geographic features may decrease efficiency of transportation within a country, which can affect economic activity in a negative way. To account for this geographical aspect, we add the average and standard deviation of elevation in the baseline model.[15] Table 5 displays the results.

[Table 5 around here]

[15]The elevation data are extracted from Michalopoulos (2012).

For the specification of ILD_{PC} and ILD_{CC}, column (1) and column (4), respectively, include the average elevation within a country, column (2) and column (5), respectively, include the standard deviation of the elevation, and column (3) and column (6), respectively, simultaneously include both variables. In this estimation, we can certify the significantly negative impact of DLD on the output-side GDP per capita, as the coefficients of DLD are negative at the 1% level in all columns.

Finally in this section, we take into account agricultural suitability, as countries with fertile land, which is captured by better agricultural suitability, may be in an advantageous position than those with barren land. Table 6 adds the average and variation of agricultural suitability within a country.[16]

[Table 6 around here]

For the specification of ILD_{PC} and ILD_{CC}, column (1) and column (4), respectively, include the average agricultural suitability within a country, column (2) and column (5), respectively, include the standard deviation of the agricultural suitability, and column (3) and column (6), respectively, include both agriculture variables. Although none of the agriculture variables are significant, possibly due to the industrial shift from agriculture to manufacturing or service industries, the significance of the negative coefficient of DLD is mostly unaffected in all columns.

6.2 Infectious disease

As the extent of medical diffusion, such as sufficient vaccination, differs across countries, and such improvement of medical facilities may enhance economic activities. To consider this point, we include variables expressing incidence of various infectious diseases. From the infectious diseases listed in the WHO database, we choose diseases with fewer missing observations to keep a sufficient number of observations. Then, the incidence or the number of reported cases per million people of the following six infectious diseases are taken into account: tuberculosis, malaria, measles, neonatal tetanus, pertussis, and total rubella. Instead of including these six disease variables, we conduct a principal component analysis (PCA), which is known as a variable-reduction method when there are multiple measures of similar concepts. Following a conventional procedure of running a PCA, we choose the first and second principal components with eigenvalues exceeding unity, as suggested by Kaiser's rule. Table 7 reports the result of the PCA for the first two principal components.[17] Table 7 implies that considering only

[16]The agricultural suitability data are extracted from Michalopoulos (2012).

[17]Figure A2 in Appendix C displays a scree plot which shows the eigenvalues of each principal component. The first and second principal components exceed one, but the others do not. In addition, by checking the diagnostic measures for sampling adequacy, the the Kaiser-Meyer-Olkin (KMO) index, the overall KMO measure exceeds 0.5, indicating that the data set is suitable for the PCA (see Table A6 in Appendix C).

the first and second principal components will explain about 50% (48.3%) of the information provided by all six infectious disease variables.[18]

By adding the predicted values of the first and second principal components of the infectious disease variables to the baseline model, we run the regressions (results shown in Table 8).

[Tables 7 and 8 around here]

As the number of observations drops due to inclusion of infectious disease variables (from 111 in the baseline estimation to 70 in the infectious disease estimation), we run the baseline estimation with the same 70 countries in column (1) for the ILD_{PC} specification and column (3) for the ILD_{CC} specification, simply for reference. In columns (2) and (4), the first principal component for the infectious disease is significant at the 10% level, so that some aspect of infectious disease is negatively associated with GDP per capita. More importantly in terms of our interest, even after controlling for infectious diseases, DLD maintains a significantly negative coefficient in both ILD specifications. Thus, difficulty in domestic communication from a viewpoint of language acquisition costs decreases a country's GDP per capita when the infectious disease effect is accounted for.

6.3 Immigrant

While one of the primary interests of this study is the impact of linguistic distance indices on GDP per capita, reverse causality could be an issue, because of rich countries' attraction of migrants, which can affect linguistic distance indices. Controlling for a migrant stock population size or its share to the total population, although not a perfect remedy to this problem, might mitigate this issue. Table 9 reports the result for inclusion of the immigrant variables.

[Table 9 around here]

Under ILD_{PC} specification and ILD_{CC} specification, column (1) and column (2), respectively, show the results with a migrant stock population size variable, and column (3) and column (4), respectively, show results with the migrant stock population share. In all specifications, the coefficient of DLD is significantly negative after immigrant effects are controlled. Although this is not an ideal solution to this endogeneity issue, the strong negative effect of the linguistic distance to the official language is robustly corroborated.

[18] We also conducted including the third principal component as well as the first and second components, which explains 63% of the infectious disease information. Including the predicted values of the first three principal components in the regression does not change the negative significance of DLD. Results are provided upon request.

6.4 Linguistic aspects

This section deals with robustness checks in relation with linguistic aspects. Specifically, we conduct two types of robustness checks, the first being about multiple official languages, and the second about linguistic fractionalization/polarization.

First, we consider the aspect of multiple official languages. Countries consisting of heterogeneous linguistic groups tend to have more than one official language (status-1 languages defined in Ethnologue) and, hence, some countries have a single official language, but others have several official languages. To shed a light on this part, we construct "multiple official language variable" which takes one if there are several status-1 languages in a country and zero otherwise. Table 10 displays the result including multiple official language dummy variable.

[Table 10 around here]

Both specifications (column (1) under ILD_{PC} and column (2) under ILD_{CC}) show the significantly negative coefficient on DLD. In addition, as for the multiple official language variable, the coefficient is negative at the 10% level. This result coordinates with the finding that DLD has a negative impact on GDP per capita, as the multiple official language may also capture linguistic heterogeneity. Even after this effect of multiple official languages is separated out, the significantly negative effect of the domestic linguistic distance remains.

Next, we consider linguistic fractionalization and polarization. As both the linguistic distance and linguistic fractionalization/polarization stem from linguistic heterogeneity within a society, it is worth trying to separate out these effects. To do so, we add linguistic fractionalization/polalization indices, such as GI, ELF, ER, RQ, and PH,[19] all of which come from Desmet et al. (2009). GI and ELF are categorized in a family of fractionalization indices, while ER and RQ are classified in a family of polarization indices. Peripheral heterogeneity, PH, is an intermediate index between fractionalization and polarization. Among these variables, GI, ER, and PH account for linguistic distance between languages, while ELF and RQ do not. As these linguistic diversity variables are based on Ethnologue, utilizing the variables is adequate and fits well to our linguistic distance data. Table 11 shows the estimation result both including linguistic distance indices and a linguistic fractionalization/polarization index.

[19] With our notation, these indices are formally defined as follows: $GI(i) = \sum_{j=1}^{K(i)} \sum_{k=1}^{K(i)} s_k s_j \tau_{jk}$, $ELF(i) = 1 - \sum_{j=1}^{K(i)} s_j^2$, $ER(i) = \sum_{j=1}^{K(i)} \sum_{k=1}^{K(i)} s_k s_j^2 \tau_{jk}$, $RQ(i) = \sum_{j=1}^{K(i)} s_j^2 (1 - s_j)$, and $PH(i) = 2 \sum_{j=1}^{K(i)} s_j s_{c(i)} \tau_{j,c(i)}$.

[Table 11 around here]

The first five columns are under the specification with ILD_{PC} , and the latter five are under ILD_{CC}. Columns (1) and (6) add GI, columns (2) and (7) add ELF, columns (3) and (8) add ER, columns (4) and (9) add RQ, and columns (5) and (10) add PH to the baseline control variables. All specifications other than those with ELF show a significantly negative coefficient of DLD when the linguistic fractionalization/polarization effects are controlled for. By contrast, for the inclusion of ELF as in columns (2) and (7), DLD loses its significance. In column (2), both DLD and ELF are insignificant, and in column (7), DLD is insignificant and ELF is severely significant at the 10% level. Such insignificance (or weak significance) of DLD and ELF may be due to high correlation between these two variables (correlation between them is 0.76). However, it can be asserted that DLD tends to give a negative impact on the output-side GDP per capita after linguistic fractionalization/polarization aspects are considered.

6.5 Genetic distance to the US

This section features ILD instead of DLD rather than checking the robustness of the negative effect of DLD. International linguistic distance in this paper is calculated based on the linguistic distance between English and each language. The idea behind the choice of English as the international linguistic center is simply that English is the most widely used language in the world. A similar measure to ILD in the context of genetic distance is the genetic distance to the US, proposed in Spolaore and Wacziarg (2009). Their concept of distance, genetic distance, measures the time passed since two populations shared ancestors. They calculate two genetic distance indices to the US from each country for cross-country regression, the first of which is a genetic distance between dominant groups and the second of which is a population weighted genetic distance based on the genetic composition in a country. By using these genetic distance indices, they find that shorter genetic distance to the US leads to higher income per capita. Their interpretation of this result is that individuals genetically close to the US residents are more likely to follow or more easily catch up with the latest complex technological and institutional innovations invented in the US. In addition, when the genetic distance to the US and the linguistic distance to the US are put together in the estimation model, the former is significantly negative (the closer to the US in terms of genetic composition, the more significantly higher the per capita income) while the latter is insignificant. They view the genetic distance indices as such, capturing a broad set of characteristics including the language aspect.

To see whether their results can be consistently observed in our framework, we add indices of the genetic distance to the US employed in Spolaore and Wacziarg (2009) to our baseline model. Table 12 shows the result.

[Table 12 around here]

Columns (1) and (2) include the index of genetic distance between genetically dominant groups in a country in the baseline ILD_{PC} and ILD_{CC} specifications, respectively, and columns (3) and (4), respectively, include the index of population-weighted genetic distance based on genetic compositions. Unlike in Spolaore and Wacziarg (2009), indices of the genetic distance to the US in our estimation are only significant at the 10% level (in columns (1)–(3)) or insignificant (in column (4)). However, in comparison with our ILD variables that are insignificant in all specifications, the finding that the genetic distance to the US more explains a country's GDP per capita than ILD is consistent with that in Spolaore and Wacziarg (2009). As for the coefficient of DLD, it is significantly negative in all specification, so the robustness of the negative impact of DLD is confirmed.

6.6 Spatial dependence

As mentioned in Section 3, due to high correlation between DLD and the Sub-Saharan African dummy variable, we excluded regional fixed effects from the baseline model in order to circumvent the multicollinearity problem. To deal with the possibility of spatial dependence among error terms due to omitting regional fixed effects, we conducted an empirical analysis in a spatial econometrics framework, which resolves spatial dependences among observations. This resettlement to the empirical field of spatial econometrics is natural, since regional dummy variables are considered the simplest way to account for the factors related to spatial characteristics of observations. For instance, Attfield et al. (2000) shed light on the relationship between the continent dummy variable and spatial econometrics model (based on distance among countries) in a cross-country analysis of economic growth. Moreno and Trehan (1997) and Maurseth (2003) also conduct empirical analysis of economic growth at the cross-country level using spatial econometrics techniques, and both conclude the effectiveness of employing those techniques to find the clustered economic growth. In addition, Romero and Burkey (2011) analyze the impact of debt/GDP ratio on GDP levels with spatial empirics, the scope of which is restricted to the Eurozone. Furthermore, cross-country differences and spillover effects of the quality of governance and institution are inspected using spatial econometrics models (Seldadyo et al., 2010;

Kelejian et al., 2013).

To check the possible spatial correlation in the data, we first computed Moran's I statistic of the residuals in the regression model (6), calculated under two types of spatial weight matrices: the first-order contiguity matrix and an inverse distance matrix based on the distance between capital cities. The obtained value is not statistically significant under the 5% nominal level, as shown in the last row in Table 13, implying that there is no significant spatial dependence among error terms even when the model does not include regional dummy variables.

[Table 13 around here]

Despite the insignificance of the Moran's I, we check if the significantly negative coefficient of DLD is unaffected even when the spatial correlation is considered and run a spatial regression specified as follows:

$$\ln \text{GDP/capita}_i = \rho \sum_{j=1}^{n} w_{ij} \ln \text{GDP/capita}_j + \beta_0 + \beta_D DLD_i + \beta_I ILD_i + X_{\text{control},i}\beta_{\text{control}} + \epsilon_i, \quad (7)$$

where ρ is an unknown correlation parameter, and w_{ij} is the (i,j)-th entry of a spatial weight matrix with the principal diagonal elements as zeros. The spatial weight matrix is either the contiguity matrix or inverse distance matrix, both of which are row-standardized (namely, $\sum_{j=1}^{n} w_{ij} = 1$ for all i). Results of the estimation based on (7) are given in Table 13. The first two columns are the results under the contiguity spatial weight matrix, and the last two columns are those under the inverse distance matrix. Columns (1) and (3) show the results with the ILD_{PC} specification, and columns (2) and (4) are those with ILD_{CC}. From Table 13, we can confirm that the estimated results of the parameters are not changed from those obtained in Section 4.1. That is, DLD is significantly negative in all specifications, which is natural because of the insignificance of ρ.[20]

7 Conclusion

In this study, we investigated the impacts of domestic and international linguistic distances on output-side GDP per capita, based on the idea that linguistic distance expresses language acquisition costs, which may hinder effective economic activities. First, we constructed pair-wise linguistic distances for all living languages in the world using a linguistic dendrogram. Then, we constructed two types

[20]We fit the spatial model with all the possible choices of δ_D and δ_I, but none of them detects significance of ρ even under the 5% nominal level.

of linguistic distance indices—domestic and international. A domestic linguistic distance index is calculated as a population-weighted average of linguistic distances between mother tongues of residents in a country and the official language. International linguistic distance indices are calculated in two ways: (i) population-weighted average of linguistic distances between a mother tongue of residents in a country and English, and (ii) linguistic distance between official languages and English.

This study found that the effects of domestic linguistic distance on output-side GDP per capita are significantly negative. As many African countries are likely to have longer domestic linguistic distance, Africa's poor economic status can be explained partly by its lesser access to the domestic linguistic center, causing more difficult nationwide communication due to mismatch of official languages and mother tongues. As for international linguistic distance, we did not find significant impact on GDP per capita. This result aligns with what was found in Spolaore and Wacziarg (2009), showing that the genetic distance to the US rather than the linguistic distance to the US explains a country's lower income per capita. We also conducted the quantile regression analysis to determine whether the negative impact of the domestic linguistic distance to GDP per capita is heterogeneous over different quantiles of GDP per capita. Unfortunately, we did not find a clear tendency of heterogeneous impacts of the domestic linguistic distance on GDP per capita across quantiles.

Despite this, our results are distinct from those in the previous literature, mostly featuring ethnolinguistic diversity, in that our linguistic indices capture language acquisition costs that are indispensable in communicating, even within a country, if the society is heterogeneous in terms of language use.

References

Acemoglu, D., S. Johnson, and J. A. Robinson (2001): "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 91, 1369–1401.

Adsera, A. and M. Pytlikova (2015): "The Role of Language in Shaping International Migration," *The Economic Journal*, 125, F49–F81.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003): "Fractionalization," *Journal of Economic Growth*, 8, 155–194.

Alesina, A., J. Harnoss, and H. Rapoport (2016): "Birthplace Diversity and Economic Prosperity," *Journal of Economic Growth*, 21, 101–138.

Alesina, A. and E. La Ferrara (2005): "Ethnic Diversity and Economic Performance," *Journal of Economic Literature*, 43, 762–800.

Attfield, C., E. S. Cannon, D. Demery, and N. W. Duck (2000): "Economic Growth and Geographic Proximity," *Economics Letters*, 68, 109–112.

Barro, R. J. and J. W. Lee (2013): "A New Data Set of Educational Attainment in the World, 1950–2010," *Journal of Development Economics*, 104, 184–198.

Chiswick, B. R. and P. W. Miller (2005): "Linguistic Distance: A Quantitative Measure of the Distance between English and Other Languages," *Journal of Multilingual and Multicultural Development*, 26, 1–11.

Desmet, K., S. Weber, and I. Ortuño-Ortín (2009): "Linguistic Diversity and Redistribution," *Journal of the European Economic Association*, 7, 1291–1318.

Dufrenot, G., V. Mignon, and C. Tsangarides (2010): "The Trade-growth Nexus in the Developing Countries: A Quantile Regression Approach," *Review of World Economics*, 146, 731–761.

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

Fearon, J. D. (2003): "Ethnic and Cultural Diversity by Country," *Journal of Economic Growth*, 8, 195–222.

Feenstra, R. C., R. Inklaar, and M. Timmer (2013a): *PWT 8.0–A User Guide*, Available for download at www.rug.nl/research/ggdc/data/penn-world-table.

Feenstra, R. C., R. Inklaar, and M. P. Timmer (2013b): *The Next Generation of the Penn World Table*, Available for download at www.ggdc.net/pwt.

Foster, N. (2008): "The Impact of Trade Liberalisation on Economic Growth: Evidence from a Quantile Regression Analysis," *Kyklos*, 61, 543–567.

Frankel, J. A. and D. Romer (1999): "Does Trade Cause Growth?" *American Economic Review*, 89, 379–399.

Gallup, J. L., J. D. Sachs, and A. D. Mellinger (1999): "Geography and Economic Development," *International Regional Science Review*, 22, 179–232.

Ginsburgh, V., I. Ortuño-Ortín, and S. Weber (2005): "Disenfranchisement in Linguistically Diverse Societies: The Case of the European Union," *Journal of the European Economic Association*, 3, 946–965.

Glaeser, E. L., R. La Porta, F. Lopez-de Silanes, and A. Shleifer (2004): "Do Institutions Cause Growth?" *Journal of Economic Growth*, 9, 271–303.

Hall, R. E. and C. Jones (1999): "Why Do Some Countries Produce So Much More Output Per Worker Than Others?" *Quarterly Journal of Economics*, 114, 83–116.

Head, K., T. Mayer, and J. Ries (2010): "The Erosion of Colonial Trade Linkages after Independence," *Journal of International Economics*, 81, 1–14.

Hutchinson, W. K. (2005): "'Linguistic Distance' as a Determinant of Bilateral Trade," *Southern Economic Journal*, 72, 1–15.

Isphording, I. E. and S. Otten (2011): "Linguistic Distance and the Language Fluency of Immigrants," *Ruhr Economic Paper No. 274*.

——— (2013): "The Costs of Babylon–Linguistic Distance in Applied Economics," *Review of International Economics*, 21, 354–369.

Jones, E. L. (2001): *The Record of Global Economic Development*, Cheltenham, UK; Northampton, MA, USA: Edward Elgar Publishing.

Kelejian, H. H., P. Murrell, and O. Shepotylo (2013): "Spatial Spillovers in the Development of Institutions," *Journal of Development Economics*, 101, 297–315.

Koenker, R. and J. A. Machado (1999): "Goodness of Fit and Related Inference Processes for Quantile Regression," *Journal of the American Statistical Association*, 94, 1296–1310.

Ku, H. and A. Zussman (2010): "Lingua Franca: The Role of English in International Trade," *Journal of Economic Behavior & Organization*, 75, 250–260.

La Porta, R., F. Lopez-de Silanes, A. Shleifer, and R. Vishny (1999): "The Quality of Government," *Journal of Law, Economics, and Organization*, 15, 222–279.

Lewis, M. P., ed. (2009): *Ethnologue: Languages of the World, 16th edition*, Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Lewis, M. P., G. F. Simons, and C. D. Fennig, eds. (2014): *Ethnologue: Languages of the World, 17th edition*, Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Marshall, M. G. and K. Jaggers (2012): "Polity IV Project: Political Regime Characteristics and Transitions, 1800–2012," .

Maurseth, P. B. (2003): "Geography and Growth: Some Empirical Evidence," *Nordic Journal of Political Economy*, 29, 25–46.

Michalopoulos, S. (2012): "The Origins of Ethnolinguistic Diversity," *American Economic Review*, 102, 1508–1539.

Moreno, R. and B. Trehan (1997): "Location and the Growth of Nations," *Journal of Economic Growth*, 2, 399–418.

Putterman, L. and D. N. Weil (2010): "Post-1500 Population Flows and the Long-Run Determinants of Economic Growth and Inequality," *Quarterly Journal of Economics*, 125, 1627–1682.

Rodriguez, F. and D. Rodrik (2001): "Trade Policy and Economic Growth: A Skeptic's Guide to the Cross-National Evidence," *NBER Macroeconomics Annual 2000, Volume 15*, 261–338.

Rodrik, D., A. Subramanian, and F. Trebbi (2004): "Institutions Rule: The Primacy of Institutions over Geography and Integration in Economic Development," *Journal of Economic Growth*, 9, 131–165.

Romero, A. A. and M. L. Burkey (2011): "Debt Overhang in the Eurozone: A Spatial Panel Analysis," *Review of Regional Studies*, 41, 49–63.

Sachs, J. D. (2003): "Institutions Don't Rule: Direct Effects of Geography on per Capita Income," *NBER Working Paper No. 9490*.

Seldadyo, H., J. P. Elhorst, and J. De Haan (2010): "Geography and Governance: Does Space Matter?" *Papers in Regional Science*, 89, 625–640.

Spolaore, E. and R. Wacziarg (2009): "The Diffusion of Development," *The Quarterly Journal of Economics*, 124, 469–529.

Yanikkaya, H. (2003): "Trade Openness and Economic Growth: A Cross-country Empirical Investigation," *Journal of Development Economics*, 72, 57–89.

Table 1: Baseline result

	(1)	(2)	(3)	(4)
Dependent variable (log)		GDP per capita		
$DLD(\delta_D = 0.5)$	-1.591***	-0.599***	-2.404***	-0.576***
	(0.29)	(0.21)	(0.24)	(0.22)
$ILD_{PC}(\delta_I = 0.5)$	-2.049***	0.042		
	(0.40)	(0.22)		
$ILD_{CC}(\delta_I = 0.5)$			-0.900***	0.043
			(0.31)	(0.18)
Years of schooling		0.231***		0.232***
		(0.03)		(0.03)
Population size		-0.002		-0.004
		(0.04)		(0.04)
Area size		0.092**		0.092**
		(0.04)		(0.04)
Trade openness		0.396**		0.396**
		(0.16)		(0.17)
Coastal population		0.640***		0.642***
		(0.23)		(0.23)
Absolute latitude		0.018***		0.018***
		(0.00)		(0.00)
Socialist law		-0.728***		-0.728***
		(0.17)		(0.17)
Polity2		-0.024*		-0.024*
		(0.01)		(0.01)
Observations	111	111	111	111
Adjusted R-squared	0.49	0.82	0.43	0.82

Robust standard errors are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 2: Linguistic distance indices and GDP per capita (full range of (δ_D, δ_I)): Baseline model with ILD_{PC}

$\delta_D \backslash \delta_I$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	-0.530**	-0.545**	-0.561***	-0.572***	-0.578***	-0.580***	-0.580***	-0.579***	-0.577***
	(0.21)	(0.21)	(0.21)	(0.20)	(0.20)	(0.20)	(0.20)	(0.20)	(0.20)
0.2	-0.545**	-0.562**	-0.578***	-0.590***	-0.596***	-0.598***	-0.598***	-0.597***	-0.595***
	(0.22)	(0.22)	(0.21)	(0.21)	(0.21)	(0.21)	(0.21)	(0.20)	(0.20)
0.3	-0.551**	-0.567**	-0.584***	-0.597***	-0.604***	-0.606***	-0.606***	-0.604***	-0.603***
	(0.22)	(0.22)	(0.22)	(0.21)	(0.21)	(0.21)	(0.21)	(0.21)	(0.21)
0.4	-0.549**	-0.566**	-0.583***	-0.596***	-0.604***	-0.606***	-0.606***	-0.605***	-0.603***
	(0.22)	(0.22)	(0.22)	(0.22)	(0.21)	(0.21)	(0.21)	(0.21)	(0.21)
0.5	-0.542**	-0.560**	-0.578**	-0.591***	-0.599***	-0.602***	-0.601***	-0.600***	-0.599***
	(0.22)	(0.22)	(0.22)	(0.22)	(0.21)	(0.21)	(0.21)	(0.21)	(0.21)
0.6	-0.533**	-0.551**	-0.570**	-0.584***	-0.591***	-0.594***	-0.594***	-0.593***	-0.591***
	(0.23)	(0.22)	(0.22)	(0.22)	(0.22)	(0.21)	(0.21)	(0.21)	(0.21)
0.7	-0.523**	-0.541**	-0.560**	-0.574***	-0.582***	-0.585***	-0.586***	-0.584***	-0.583***
	(0.23)	(0.22)	(0.22)	(0.22)	(0.22)	(0.21)	(0.21)	(0.21)	(0.21)
0.8	-0.512**	-0.530**	-0.550**	-0.564**	-0.573***	-0.576***	-0.576***	-0.575***	-0.574***
	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)	(0.21)	(0.21)	(0.21)	(0.21)
0.9	-0.501**	-0.520**	-0.539**	-0.554**	-0.563**	-0.566***	-0.567***	-0.566***	-0.564***
	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)	(0.21)	(0.21)	(0.21)	(0.21)

This table shows coefficients of DLD on a full range of linguistic distance index parameters, δ (δ_D and δ_I). ILD_{PC} as the international linguistic distance index. GDP/capita is the dependent variable. All results include the vector of baseline control variables. Robust standard errors are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 3: Quantile regression result (baseline control with ILD_{PC})

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent variable (log)	GDP per capita								
Quantile	10%	20%	30%	40%	50%	60%	70%	80%	90%
$DLD(\delta_D = 0.5)$	-1.077***	-0.735**	-0.642*	-0.718**	-0.935***	-0.919***	-0.817**	-0.833**	-0.748**
	(0.30)	(0.33)	(0.33)	(0.30)	(0.29)	(0.29)	(0.33)	(0.35)	(0.34)
$ILD_{PC}(\delta_I = 0.5)$	-0.163	0.074	0.187	0.186	0.076	0.135	0.008	0.037	-0.023
	(0.31)	(0.34)	(0.35)	(0.33)	(0.35)	(0.34)	(0.37)	(0.40)	(0.42)
Years of schooling	0.195***	0.205***	0.242***	0.219***	0.214***	0.195***	0.218***	0.210***	0.224***
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)
Population size	0.048	0.047	0.029	0.024	0.038	-0.003	0.025	0.023	0.063
	(0.05)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Area size	0.189***	0.166**	0.131**	0.114*	0.082	0.108	0.038	0.009	-0.080
	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)
Trade openness	0.672***	0.573**	0.459*	0.445*	0.445*	0.399*	0.540*	0.514*	0.300
	(0.23)	(0.25)	(0.25)	(0.23)	(0.23)	(0.23)	(0.28)	(0.30)	(0.29)
Coastal population	0.460*	0.609**	0.556**	0.668**	0.530*	0.475	0.310	0.282	0.197
	(0.24)	(0.27)	(0.28)	(0.26)	(0.30)	(0.30)	(0.32)	(0.32)	(0.31)
Absolute latitude	0.017***	0.020***	0.018**	0.016**	0.011	0.013**	0.011*	0.015**	0.020***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Socialist law	-0.671***	-0.753***	-0.804***	-0.489**	-0.429*	-0.496**	-0.698***	-0.823***	-1.078***
	(0.23)	(0.27)	(0.28)	(0.23)	(0.23)	(0.23)	(0.24)	(0.24)	(0.24)
Polity2	-0.016	-0.003	-0.012	-0.001	-0.008	0.002	-0.025	-0.028	-0.050**
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Observations	111	111	111	111	111	111	111	111	111
GOF	0.63	0.66	0.67	0.67	0.66	0.65	0.64	0.61	0.56

Standard errors based on 10,000 bootstrap replications are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

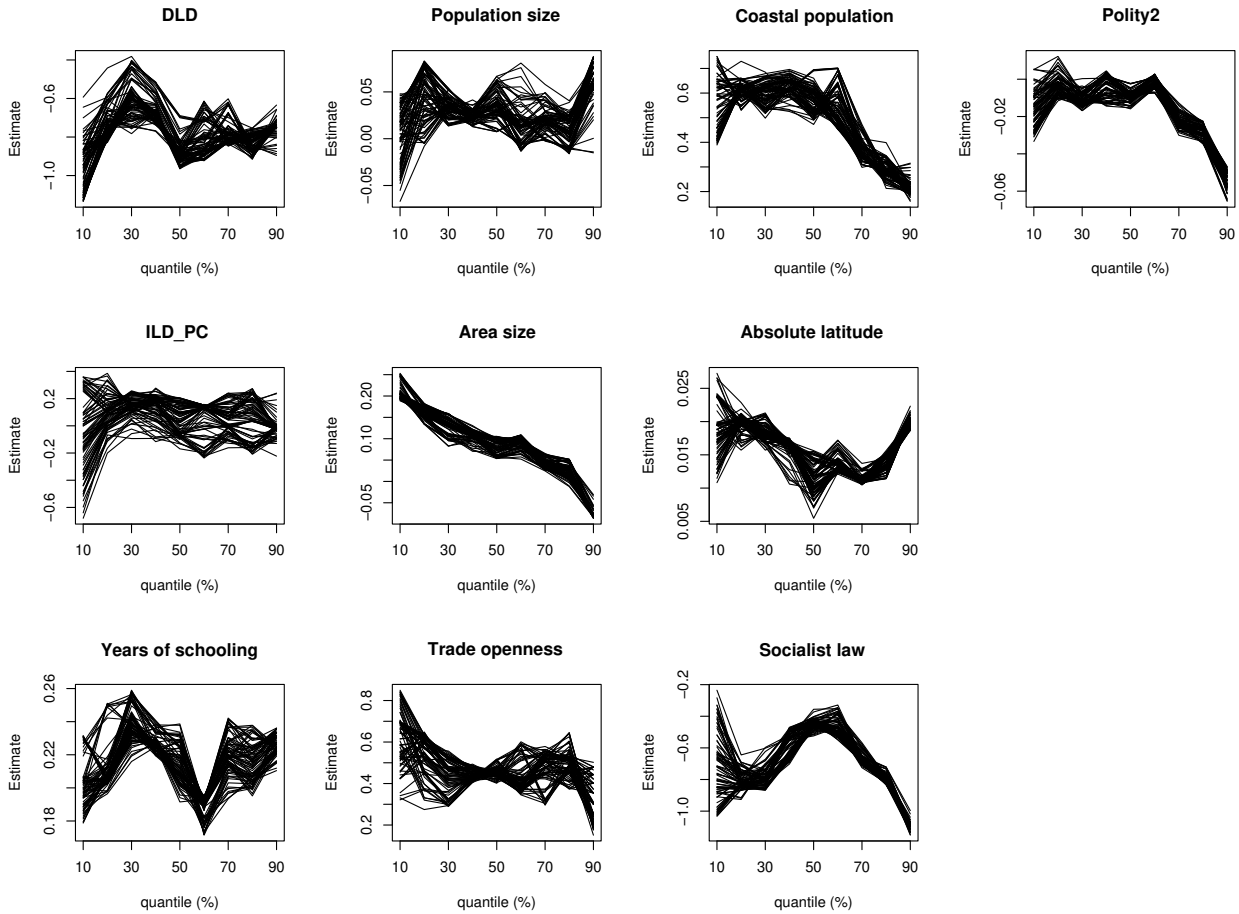

Figure 1: Quantile regression (Baseline model with DLD and ILD_{PC})

Table 4: Effect of temperature and precipitation

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable (log)			GDP per capita			
$DLD(\delta_D = 0.5)$	-0.593***	-0.603***	-0.589***	-0.572***	-0.586**	-0.561**
	(0.21)	(0.22)	(0.22)	(0.21)	(0.24)	(0.23)
$ILD_{PC}(\delta_I = 0.5)$	0.030	0.024	0.044			
	(0.23)	(0.20)	(0.22)			
$ILD_{CC}(\delta_I = 0.5)$				0.040	0.032	0.052
				(0.18)	(0.18)	(0.18)
Temperature	0.022*		0.023*	0.022*		0.023*
	(0.01)		(0.01)	(0.01)		(0.01)
Precipitation		-0.000	0.000		-0.000	0.000
		(0.00)	(0.00)		(0.00)	(0.00)
Years of schooling	0.250***	0.232***	0.250***	0.250***	0.232***	0.251***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Population size	-0.007	0.000	-0.009	-0.008	-0.001	-0.011
	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)
Area size	0.092**	0.090*	0.094*	0.092**	0.090**	0.095**
	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)	(0.05)
Trade openness	0.360**	0.405**	0.352**	0.359**	0.403**	0.349**
	(0.16)	(0.16)	(0.16)	(0.17)	(0.17)	(0.16)
Coastal population	0.540**	0.655**	0.526**	0.541**	0.654**	0.524**
	(0.21)	(0.27)	(0.26)	(0.21)	(0.27)	(0.25)
Absolute latitude	0.025***	0.017***	0.026***	0.025***	0.017***	0.027***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Socialist law	-0.716***	-0.715***	-0.726***	-0.718***	-0.718***	-0.729***
	(0.16)	(0.17)	(0.17)	(0.17)	(0.18)	(0.18)
Polity2	-0.023*	-0.023*	-0.024**	-0.023*	-0.023*	-0.023**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	111	111	111	111	111	111
Adjusted R-squared	0.82	0.82	0.82	0.82	0.82	0.82

Robust standard errors are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 5: Effect of elevation

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable (log)	GDP per capita					
$DLD(\delta_D = 0.5)$	-0.661***	-0.623***	-0.642***	-0.622***	-0.601***	-0.589***
	(0.20)	(0.22)	(0.20)	(0.20)	(0.22)	(0.21)
$ILD_{PC}(\delta_I = 0.5)$	0.085	0.038	0.123			
	(0.22)	(0.23)	(0.22)			
$ILD_{CC}(\delta_I = 0.5)$				0.073	0.041	0.097
				(0.18)	(0.18)	(0.18)
Elevation (average)	-0.230*		-0.375*	-0.231*		-0.380*
	(0.13)		(0.21)	(0.13)		(0.21)
Elevation (std. dev.)		-0.105	0.251		-0.105	0.255
		(0.12)	(0.21)		(0.12)	(0.22)
Years of schooling	0.242***	0.233***	0.245***	0.242***	0.233***	0.245***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Population size	-0.003	0.004	-0.019	-0.005	0.003	-0.022
	(0.05)	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)
Area size	0.082*	0.093**	0.074	0.082*	0.093**	0.073
	(0.05)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)
Trade openness	0.306*	0.377**	0.295*	0.306*	0.376**	0.296*
	(0.17)	(0.16)	(0.17)	(0.17)	(0.17)	(0.17)
Coastal population	0.452**	0.608**	0.409*	0.452**	0.610***	0.408*
	(0.21)	(0.23)	(0.21)	(0.21)	(0.23)	(0.21)
Absolute latitude	0.018***	0.018***	0.017***	0.017***	0.018***	0.017***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Socialist law	-0.790***	-0.739***	-0.803***	-0.789***	-0.740***	-0.800***
	(0.16)	(0.17)	(0.16)	(0.17)	(0.17)	(0.17)
Polity2	-0.027**	-0.025*	-0.026*	-0.027**	-0.025*	-0.026**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	111	111	111	111	111	111
Adjusted R-squared	0.82	0.82	0.83	0.83	0.82	0.83

Robust standard errors are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 6: Effect of agriculture

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable (log)			GDP per capita			
$DLD(\delta_D = 0.5)$	-0.606***	-0.575***	-0.580***	-0.586***	-0.552**	-0.558**
	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)
$ILD_{PC}(\delta_I = 0.5)$	0.041	0.014	0.014			
	(0.23)	(0.22)	(0.22)			
$ILD_{CC}(\delta_I = 0.5)$				0.036	0.048	0.044
				(0.18)	(0.18)	(0.18)
Agricultural suitability (average)	-0.074		-0.050	-0.066		-0.039
	(0.32)		(0.32)	(0.32)		(0.31)
Agricultural suitability (std. dev.)		-0.528	-0.516		-0.538	-0.528
		(0.62)	(0.61)		(0.64)	(0.63)
Years of schooling	0.232***	0.236***	0.236***	0.232***	0.237***	0.237***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Population size	0.004	0.015	0.019	0.002	0.014	0.017
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Area size	0.084	0.082*	0.077	0.085	0.083*	0.079
	(0.06)	(0.05)	(0.06)	(0.06)	(0.05)	(0.06)
Trade openness	0.390**	0.372**	0.369**	0.391**	0.369**	0.367**
	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)	(0.18)
Coastal population	0.636***	0.594***	0.592***	0.637***	0.596***	0.594***
	(0.23)	(0.22)	(0.22)	(0.23)	(0.22)	(0.22)
Absolute latitude	0.018***	0.019***	0.018***	0.018***	0.019***	0.018***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Socialist law	-0.718***	-0.742***	-0.735***	-0.719***	-0.748***	-0.742***
	(0.16)	(0.17)	(0.17)	(0.17)	(0.18)	(0.17)
Polity2	-0.023**	-0.024*	-0.023**	-0.023**	-0.023*	-0.023**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	111	111	111	111	111	111
Adjusted R-squared	0.82	0.82	0.82	0.82	0.82	0.82

Robust standard errors are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 7: Principal component analysis (Infectious disease)

	First principal component	Second principal component
Eigenvalue	1.686	1.211
Difference	0.475	0.299
Proportion	0.281	0.202
Cumulative	0.281	0.483
Loadings:		
Tuberculosis	0.614	0.032
Malaria	0.409	-0.162
Measles	0.024	0.665
Neonatal Tetanus	0.602	-0.003
Pertussis	-0.293	-0.302
Total Rubella	-0.085	0.663

Principal component analysis based on a correlation matrix (unrotated). 70 observations are in the data. Only lists two principal components which show eigenvalues larger than one, based on the Kaiser's rule.

Table 8: Effect of infectious disease

	(1)	(2)	(3)	(4)
Dependent variable (log)	GDP per capita			
$DLD(\delta_D = 0.5)$	-0.895***	-0.659***	-0.802***	-0.603**
	(0.26)	(0.24)	(0.25)	(0.23)
$ILD_{PC}(\delta_I = 0.5)$	0.155	0.188		
	(0.25)	(0.25)		
$ILD_{CC}(\delta_I = 0.5)$			0.212	0.144
			(0.20)	(0.18)
First principal component (infectious disease)		-0.195*		-0.186*
		(0.10)		(0.10)
Second principal component (infectious disease)		-0.104		-0.105
		(0.07)		(0.07)
Years of schooling	0.189***	0.183***	0.187***	0.179***
	(0.04)	(0.04)	(0.04)	(0.03)
Population size	0.005	-0.005	-0.002	-0.009
	(0.06)	(0.06)	(0.07)	(0.06)
Area size	0.071	0.066	0.072	0.066
	(0.06)	(0.05)	(0.06)	(0.05)
Trade openness	0.279	0.337	0.307	0.366
	(0.27)	(0.25)	(0.26)	(0.24)
Coastal population	0.559**	0.307	0.558**	0.303
	(0.27)	(0.23)	(0.26)	(0.22)
Absolute latitude	0.024***	0.019***	0.024***	0.020***
	(0.01)	(0.01)	(0.01)	(0.01)
Socialist law	-0.730***	-0.688***	-0.748***	-0.688***
	(0.21)	(0.20)	(0.22)	(0.20)
Polity2	-0.023	-0.034	-0.022	-0.034
	(0.02)	(0.02)	(0.02)	(0.02)
Observations	70	70	70	70
Adjusted R-squared	0.81	0.83	0.81	0.83

Robust standard errors are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 9: Effect of immigrant population and share

	(1)	(2)	(3)	(4)
Dependent variable (log)	GDP per capita			
$DLD(\delta_D = 0.5)$	-0.613***	-0.698***	-0.571**	-0.677***
	(0.22)	(0.22)	(0.22)	(0.21)
$ILD_{PC}(\delta_I = 0.5)$	0.106	0.213		
	(0.26)	(0.27)		
$ILD_{CC}(\delta_I = 0.5)$			0.073	0.005
			(0.20)	(0.18)
Immigrant population	0.000		0.000	
	(0.00)		(0.00)	
Immigrant population share		0.024*		0.023*
		(0.01)		(0.01)
Years of schooling	0.228***	0.200***	0.228***	0.195***
	(0.03)	(0.03)	(0.03)	(0.03)
Population size	-0.010	0.030	-0.012	0.031
	(0.05)	(0.04)	(0.05)	(0.04)
Area size	0.090**	0.102**	0.089**	0.098**
	(0.04)	(0.04)	(0.04)	(0.04)
Trade openness	0.400**	0.369**	0.402**	0.383**
	(0.17)	(0.15)	(0.17)	(0.15)
Coastal population	0.656***	0.611***	0.656***	0.604***
	(0.24)	(0.21)	(0.24)	(0.21)
Absolute latitude	0.018***	0.016***	0.018***	0.015***
	(0.00)	(0.00)	(0.00)	(0.00)
Socialist law	-0.718***	-0.605***	-0.714***	-0.575***
	(0.17)	(0.16)	(0.17)	(0.15)
Polity2	-0.023*	-0.010	-0.023*	-0.011
	(0.01)	(0.01)	(0.01)	(0.01)
Observations	111	111	111	111
Adjusted R-squared	0.82	0.84	0.82	0.84

Robust standard errors are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 10: Effect of multiple official languages

	(1)	(2)
Dependent variable (log)	GDP per capita	
$DLD(\delta_D = 0.5)$	-0.562**	-0.506**
	(0.22)	(0.22)
$ILD_{PC}(\delta_I = 0.5)$	0.103	
	(0.23)	
$ILD_{CC}(\delta_I = 0.5)$		0.104
		(0.18)
Multiple official languages	-0.239*	-0.248*
	(0.14)	(0.14)
Years of schooling	0.244***	0.244***
	(0.03)	(0.03)
Population size	0.005	0.002
	(0.04)	(0.04)
Area size	0.084*	0.084*
	(0.04)	(0.04)
Trade openness	0.415***	0.416**
	(0.16)	(0.16)
Coastal population	0.565**	0.565**
	(0.22)	(0.22)
Absolute latitude	0.017***	0.017***
	(0.00)	(0.00)
Socialist law	-0.796***	-0.799***
	(0.17)	(0.18)
Polity2	-0.025*	-0.025*
	(0.01)	(0.01)
Observations	111	111
Adjusted R-squared	0.82	0.82

Robust standard errors are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 11: Effect of linguistic diversity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Dependent variable (log)	GDP per capita									
$DLD(\delta_D = 0.5)$	-0.558**	-0.330	-0.595***	-0.648***	-0.588***	-0.514**	-0.273	-0.556**	-0.629***	-0.553**
	(0.23)	(0.27)	(0.22)	(0.23)	(0.22)	(0.23)	(0.26)	(0.22)	(0.23)	(0.22)
$ILD_{PC}(\delta_I = 0.5)$	0.079	0.071	0.081	0.032	0.065					
	(0.23)	(0.23)	(0.24)	(0.22)	(0.24)					
$ILD_{CC}(\delta_I = 0.5)$						0.078	0.089	0.071	0.037	0.063
						(0.19)	(0.17)	(0.19)	(0.18)	(0.19)
GI	-0.318					-0.336				
	(0.35)					(0.36)				
ELF		-0.403					-0.419*			
		(0.24)					(0.24)			
ER			-0.895					-0.930		
			(1.20)					(1.22)		
RQ				0.820					0.817	
				(1.01)					(1.01)	
PH					-0.249					-0.265
					(0.48)					(0.49)
Years of schooling	0.237***	0.233***	0.236***	0.230***	0.235***	0.237***	0.234***	0.237***	0.230***	0.235***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Population size	-0.008	0.008	-0.011	-0.003	-0.008	-0.010	0.006	-0.014	-0.004	-0.010
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Area size	0.102**	0.097**	0.101**	0.092**	0.099**	0.102**	0.097**	0.101**	0.092**	0.099**
	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Trade openness	0.398**	0.425***	0.391**	0.389**	0.397**	0.399**	0.424**	0.392**	0.388**	0.397**
	(0.16)	(0.16)	(0.16)	(0.17)	(0.16)	(0.17)	(0.16)	(0.16)	(0.17)	(0.17)
Coastal population	0.629***	0.624**	0.645***	0.658***	0.642***	0.630***	0.627***	0.646***	0.659***	0.644***
	(0.24)	(0.24)	(0.24)	(0.24)	(0.24)	(0.24)	(0.24)	(0.24)	(0.24)	(0.24)
Absolute latitude	0.017***	0.018***	0.018***	0.017***	0.018***	0.017***	0.018***	0.018***	0.017***	0.018***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Socialist law	-0.720***	-0.713***	-0.718***	-0.731***	-0.722***	-0.719***	-0.715***	-0.717***	-0.732***	-0.722***
	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)
Polity2	-0.024*	-0.024*	-0.025*	-0.023*	-0.025*	-0.024*	-0.024*	-0.024*	-0.022*	-0.024*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	110	110	110	110	110	110	110	110	110	110
Adjusted R-squared	0.81	0.82	0.81	0.81	0.81	0.81	0.82	0.81	0.81	0.81

Robust standard errors are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 12: Effect of genetic distance to the US

	(1)	(2)	(3)	(4)
Dependent variable (log)	GDP per capita			
$DLD(\delta_D = 0.5)$	-0.544**	-0.542**	-0.505**	-0.503**
	(0.21)	(0.21)	(0.22)	(0.22)
$ILD_{PC}(\delta_I = 0.5)$	0.259	0.281		
	(0.23)	(0.25)		
$ILD_{CC}(\delta_I = 0.5)$			0.063	0.056
			(0.19)	(0.20)
Genetic distance to the US (dominant)	-0.000**		-0.000*	
	(0.00)		(0.00)	
Genetic distance to the US (weighted)		-0.000*		-0.000
		(0.00)		(0.00)
Years of schooling	0.227***	0.225***	0.223***	0.221***
	(0.03)	(0.03)	(0.03)	(0.03)
Population size	-0.022	-0.026	-0.020	-0.022
	(0.05)	(0.05)	(0.05)	(0.05)
Area size	0.102**	0.104**	0.099**	0.100**
	(0.04)	(0.04)	(0.04)	(0.04)
Trade openness	0.438***	0.439***	0.445***	0.444***
	(0.16)	(0.16)	(0.16)	(0.17)
Coastal population	0.567**	0.575***	0.573**	0.579***
	(0.22)	(0.21)	(0.22)	(0.22)
Absolute latitude	0.015***	0.015***	0.015***	0.015***
	(0.00)	(0.00)	(0.00)	(0.00)
Socialist law	-0.748***	-0.741***	-0.718***	-0.710***
	(0.17)	(0.17)	(0.16)	(0.17)
Polity2	-0.023*	-0.023*	-0.024*	-0.024*
	(0.01)	(0.01)	(0.01)	(0.01)
Observations	110	108	110	108
Adjusted R-squared	0.83	0.82	0.83	0.82

Robust standard errors are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 13: SAR result

	(1)	(2)	(3)	(4)
Dependent variable (log)	GDP per capita			
Spatial weight matrix	Contiguity		Inverse distance	
$DLD(\delta_D = 0.5)$	-0.589***	-0.569***	-0.530***	-0.520**
	(0.19)	(0.20)	(0.19)	(0.21)
$ILD_{PC}(\delta_I = 0.5)$	0.003		0.054	
	(0.29)		(0.29)	
$ILD_{CC}(\delta_I = 0.5)$		0.044		0.014
		(0.17)		(0.17)
Years of schooling	0.231***	0.232***	0.230***	0.229***
	(0.03)	(0.03)	(0.03)	(0.03)
Population size	-0.014	-0.015	-0.015	-0.015
	(0.05)	(0.05)	(0.05)	(0.05)
Area size	0.097**	0.097**	0.095**	0.094**
	(0.04)	(0.04)	(0.04)	(0.04)
Trade openness	0.372**	0.369**	0.314*	0.317*
	(0.16)	(0.16)	(0.17)	(0.17)
Coastal population	0.693***	0.696***	0.625***	0.624***
	(0.17)	(0.17)	(0.17)	(0.17)
Absolute latitude	0.018***	0.018***	0.015***	0.015***
	(0.00)	(0.00)	(0.00)	(0.00)
Socialist law	-0.725***	-0.732***	-0.735***	-0.729***
	(0.16)	(0.15)	(0.16)	(0.15)
Polity2	-0.027***	-0.026***	-0.025**	-0.025***
	(0.01)	(0.01)	(0.01)	(0.01)
ρ	0.030	0.030	0.296	0.294
	(0.02)	(0.02)	(0.17)	(0.17)
Observations	111	111	111	111
Moran's I (p-value)	0.43	0.39	0.88	0.86

Standard errors are in parentheses. Moran's I statistics are calculated based on the residuals of the OLS regressions of the baseline model.
*** p<0.01, ** p<0.05, * p<0.1

Appendix A Data sources and summary statistics

Table A1: Data sources

Variable name	Definition	Source
Linguistic distance		
Domestic linguistic distance DLD	Weighted average of linguistic distances to the domestic linguistic center as defined in the main text	Own calculation from Ethnologue 16th edition (Lewis, 2009)
International linguistic distance ILD_{PC}	Weighted average of linguistic distances to the international linguistic center as defined in the main text	Own calculation from Ethnologue 16th edition (Lewis, 2009)
International linguistic distance ILD_{CC}	Linguistic distances from the domestic linguistic center(s) to the international linguistic center as defined in the main text	Own calculation from Ethnologue 16th edition (Lewis, 2009)
Language status		
Multiple official languages	Dummy variable taking 1 if a country has multiple status-1 languages	Own calculation from Ethnologue 17th edition (Lewis et al., 2014)
Language status	Status labeled to each language based on its intra- and international usages and importance	Ethnologue 17th edition (Lewis et al., 2014)
Income		
GDP per capita	log of GDP/capita in year 2010 (Output-side real GDP at current PPPs in mil. 2005US$)	Penn World Tables 8.0, Feenstra, Inklaar, and Timmer (2013b)
Population		
Population size	log of population size (in mil.) in year 2005	Penn World Tables 8.0, Feenstra, Inklaar, and Timmer (2013b)
Education		
Years of schooling	Years of schooling, population aged over 25 in year 2005	Barro and Lee (2013)
Trade		
Trade openness	Merchandise exports + imports in share in GDP, at PPP in year 2005	Penn World Tables 8.0, Feenstra, Inklaar, and Timmer (2013b)
Institutions		
Quality of institutions	Combined Polity2 score in year 2005 (-10 for most repressive, 10 for most democratic)	PolityIV database, Marshall and Jaggers (2012)
Legal origin	British law (Dummy takes 1 if country's legal origin is British law), socialist law (dummy takes 1 if socialist law), French law (dummy takes 1 if French law), German law (dummy takes 1 if German law), Scandinavian law (dummy takes 1 if Scandinavian law)	La Porta, Lopez-de Silanes, Shleifer, and Vishny (1999)
Geography & climate		
Absolute latitude	Absolute latitude of capital city	CEPII (2010), Head, Mayer, and Ries (2010)
Land area size	log of country land area size in km^2	CEPII (2010), Head, Mayer, and Ries (2010)
Coastal population	Share in total population within 100 km of ice-free coast to total population in % in year 1995	Gallup, Sachs, and Mellinger (1999)
Average elevation	Average elevation across regions within a country	Michalopoulos (2012)
Standard deviation of elevation	Standard deviation of elevations across regions within a country	Michalopoulos (2012)
Average agricultural suitability	Average land quality across regions within a country	Michalopoulos (2012)
Standard deviation of agricultural suitability	Standard deviation of land quality across regions within a country	Michalopoulos (2012)
Average temperature	Average annual temperatures in deg. C for the period 1961-1999	World Bank (2011)
Average precipitation	Average annual precipitation in mm for the period 1961-1999	World Bank (2011)
Regional dummy	Sub-Saharan Africa	World Bank (2014)
Infectious disease		
Tuberculosis	Incidence of tuberculosis (per mil. people) in year 2005	WHO (2020)
Malaria	Reported cases of malaria in the area at risk (per mil. people) in year 2005	WHO (2020)
Measles	Reported cases of measles (per mil. people) in year 2005	WHO (2020)
Neonatal Tetanus	Reported cases of neonatal tetanus (per mil. people) in year 2005	WHO (2020)
Pertussis	Reported cases of pertussis (per mil. people) in year 2005	WHO (2020)
Total Rubella	Reported cases of the total rubella (per mil. people) in year 2005	WHO (2020)
Diversity/polarization in language		
GI index	Herfindahl index of language group shares (linguistic distance considered)	Desmet et al. (2009)
ELF index	Herfindahl index of language group shares	Desmet et al. (2009)
ER index	Polarization index of language group shares (linguistic distance considered)	Desmet et al. (2009)
RQ index	Polarization index of language group shares	Desmet et al. (2009)
PH index	Peripheral heterogeneity index of linguistic groups (linguistic distance considered)	Desmet et al. (2009)
Immigrants		
Immigrant population	International migrant stock in year 2005	United Nations (2015)
Immigrant population share in %	International migrant stock as a percentage of the total population in year 2005	United Nations (2015)
Genetic distance		
Genetic distance to the US (dominant)	F_{ST} genetic distance index to the US (between dominant groups)	Spolaore and Wacziarg (2009)
Genetic distance to the US (weighted)	F_{ST} population weighted genetic distance index to the US	Spolaore and Wacziarg (2009)

Table A2: Summary statistics

Variable	Obs	Mean	Std.Dev.	Min	Max
GDP per capita (log)	111	8.80	1.25	6.14	11.02
$DLD(\delta_D = 0.1)$	111	0.29	0.36	0	1
$DLD(\delta_D = 0.2)$	111	0.31	0.36	0	1
$DLD(\delta_D = 0.3)$	111	0.33	0.36	0	1
$DLD(\delta_D = 0.4)$	111	0.34	0.36	0	1
$DLD(\delta_D = 0.5)$	111	0.36	0.36	0	1
$DLD(\delta_D = 0.6)$	111	0.37	0.36	0	1
$DLD(\delta_D = 0.7)$	111	0.38	0.36	0	1
$DLD(\delta_D = 0.8)$	111	0.39	0.37	0	1
$DLD(\delta_D = 0.9)$	111	0.39	0.37	0	1
$ILD_{PC}(\delta_I = 0.1)$	111	0.62	0.37	0.01	1
$ILD_{PC}(\delta_I = 0.2)$	111	0.70	0.31	0.01	1
$ILD_{PC}(\delta_I = 0.3)$	111	0.75	0.27	0.01	1
$ILD_{PC}(\delta_I = 0.4)$	111	0.80	0.24	0.01	1
$ILD_{PC}(\delta_I = 0.5)$	111	0.83	0.23	0.01	1
$ILD_{PC}(\delta_I = 0.6)$	111	0.86	0.22	0.01	1
$ILD_{PC}(\delta_I = 0.7)$	111	0.88	0.21	0.01	1
$ILD_{PC}(\delta_I = 0.8)$	111	0.89	0.21	0.01	1
$ILD_{PC}(\delta_I = 0.9)$	111	0.90	0.21	0.01	1
$ILD_{CC}(\delta_I = 0.1)$	111	0.44	0.38	0	1
$ILD_{CC}(\delta_I = 0.2)$	111	0.53	0.34	0	1
$ILD_{CC}(\delta_I = 0.3)$	111	0.60	0.32	0	1
$ILD_{CC}(\delta_I = 0.4)$	111	0.66	0.31	0	1
$ILD_{CC}(\delta_I = 0.5)$	111	0.70	0.31	0	1
$ILD_{CC}(\delta_I = 0.6)$	111	0.73	0.32	0	1
$ILD_{CC}(\delta_I = 0.7)$	111	0.76	0.32	0	1
$ILD_{CC}(\delta_I = 0.8)$	111	0.78	0.33	0	1
$ILD_{CC}(\delta_I = 0.9)$	111	0.79	0.33	0	1
Years of schooling	111	7.56	3.27	1.11	13.13
Population size	111	2.67	1.41	0.10	7.16
Area size	111	12.52	1.59	9.33	16.65
Trade openness	111	0.57	0.39	0.10	2.41
Coastal population	111	0.40	0.35	0	1
Absolute latitude	111	28.64	17.68	0.23	60.13
Polity2	111	4.72	6.05	-10	10
Temperature	111	16.74	8.71	-7.14	28.30
Precipitation	111	1034.10	709.60	32.91	3268.27
Agricultural suitability (average)	111	0.46	0.24	0	0.96
Agricultural suitability (std. dev.)	111	0.19	0.09	0	0.41
Elevation (average)	111	0.60	0.51	0.03	2.52
Elevation (std. dev.)	111	0.37	0.38	0.01	1.91
Multiple official language	111	0.21	0.41	0	1
Immigrant population	111	1525180	4139164	6290	39300000
Immigrant population share	111	6.58	8.85	0.05	58.90
British law	111	0.28	0.45	0	1
French law	111	0.43	0.50	0	1
German law	111	0.05	0.21	0	1
Socialist law	111	0.21	0.41	0	1
Scandinavian law	111	0.04	0.19	0	1
GI	110	0.16	0.17	0	0.65
ELF	110	0.45	0.31	0	0.96
ER	110	0.04	0.05	0	0.21
RQ	110	0.11	0.06	0	0.24
PH	110	0.12	0.13	0	0.50
Genetic distance to the US (dominant)	110	843.13	822.41	0	2288.00
Genetic distance to the US (weighted)	108	957.32	565.62	314.48	2088.01
Tuberculosis	70	1381.11	2280.91	53.90	11538.73
Malaria	70	3554.34	13198.79	0	74308.77
Measles	70	40.02	158.48	0	1062.37
Neonatal Tetanus	70	0.58	1.18	0	4.98
Pertussis	70	52.11	172.80	0	1197.41
Total Rubella	70	67.94	319.39	0	2598.09

Appendix B Note on construction of linguistic distance indices

Table A3: Notes on multiple status-1 language countries

Cameroon	Double status-1 language (English and French), but no L1 speakers either of English or French, so weights 0.5 are assigned to English and French
Israel	Double status-1 language (Hebrew and Standard Arabic), but no L1 speakers of Standard Arabic, so no weight is assigned to Standard Arabic
Lesotho	Double status-1 language (Southern Sotho and English), but no L1 speakers of English, so no weight is assigned to English
Malaysia	Triple status-1 language (Mandarin Chinese, English, and Standard Malay), but no L1 speakers of Mandarin Chinese and Standard Malay, so no weights are assigned to Mandarin Chinese and Standard Malay
Pakistan	Double status-1 language (English and Urdu), but no L1 speakers of English, so no weight is assigned to English
Somalia	Quadruple status-1 language (Somali, Standard Arabic, Italian, and English), but no L1 speakers of Standard Arabic, Italian, and English, so no weights are assigned for Standard Arabic, Italian, and English.
Sudan	Double status-1 language (English and Standard Arabic), but no L1 speakers of English and Standard Arabic, so weights 0.5 are assigned to English and Standard Arabic
Switzerland	Quadruple status-1 language (French, Standard German, and Italian), but no L1 speakers of Standard German, so no weight is assigned for Standard German

This table lists the notes on calculations of linguistic distance indices for countries with multiple status-1 languages which need special treatments. Basically, calculation of linguistic distance indices is conducted as in Section 2. However, for some of the countries with multiple Status-1 languages, Ethnologue reports there are no status-1 language speaker as their L1 languages (mother tongues). For example, Ethnologue reports that no one speaks English or French as L1 language in Cameroon, both of which are labeled as status-1 languages there. In this case, we simply calculate the linguistic distance indices based on the same weights on English and French, that is, $\gamma_{\text{English}} = \gamma_{\text{French}} = 0.5$. Such all special treatments implemented in linguistic distance index calculation are listed in this table.

Appendix C Other estimation results

Table A4: Linguistic distance indices and GDP per capita (full range of (δ_D, δ_I)): Baseline model with ILD_{CC}

$\delta_D \backslash \delta_I$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	-0.685***	-0.668***	-0.639***	-0.610***	-0.587***	-0.570***	-0.559***	-0.551***	-0.546***
	(0.19)	(0.19)	(0.19)	(0.20)	(0.20)	(0.20)	(0.20)	(0.20)	(0.20)
0.2	-0.691***	-0.674***	-0.647***	-0.619***	-0.598***	-0.582***	-0.572***	-0.565***	-0.560***
	(0.19)	(0.20)	(0.20)	(0.20)	(0.21)	(0.21)	(0.21)	(0.21)	(0.21)
0.3	-0.683***	-0.667***	-0.641***	-0.616***	-0.597***	-0.583***	-0.573***	-0.567***	-0.563**
	(0.20)	(0.20)	(0.20)	(0.21)	(0.21)	(0.21)	(0.21)	(0.21)	(0.22)
0.4	-0.669***	-0.652***	-0.629***	-0.606***	-0.588***	-0.576***	-0.568**	-0.562**	-0.558**
	(0.20)	(0.20)	(0.21)	(0.21)	(0.21)	(0.22)	(0.22)	(0.22)	(0.22)
0.5	-0.651***	-0.635***	-0.613***	-0.592***	-0.576***	-0.565**	-0.558**	-0.553**	-0.549**
	(0.20)	(0.21)	(0.21)	(0.21)	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)
0.6	-0.631***	-0.616***	-0.596***	-0.577***	-0.563**	-0.553**	-0.546**	-0.542**	-0.539**
	(0.20)	(0.21)	(0.21)	(0.21)	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)
0.7	-0.613***	-0.598***	-0.579***	-0.562**	-0.549**	-0.540**	-0.534**	-0.530**	-0.528**
	(0.21)	(0.21)	(0.21)	(0.21)	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)
0.8	-0.595***	-0.581***	-0.564***	-0.548**	-0.536**	-0.528**	-0.522**	-0.519**	-0.516**
	(0.21)	(0.21)	(0.21)	(0.21)	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)
0.9	-0.579***	-0.566***	-0.549**	-0.535**	-0.524**	-0.516**	-0.511**	-0.508**	-0.506**
	(0.21)	(0.21)	(0.21)	(0.21)	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)

This table shows coefficients of DLD on a full range of linguistic distance index parameters, δ (δ_D and δ_I). ILD_{CC} as the international linguistic distance index. GDP/capita is the dependent variable. All results include the vector of baseline control variables. Robust standard errors are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table A5: Quantile regression result (baseline control with ILD_{CC})

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent variable (log)	GDP per capita								
Quantile	10%	20%	30%	40%	50%	60%	70%	80%	90%
$DLD(\delta_D = 0.5)$	-1.071***	-0.721**	-0.621*	-0.833***	-0.953***	-0.852***	-0.817**	-0.723*	-0.761*
	(0.29)	(0.36)	(0.36)	(0.30)	(0.30)	(0.29)	(0.38)	(0.37)	(0.42)
$ILD_{CC}(\delta_I = 0.5)$	-0.165	-0.037	0.051	-0.063	-0.016	0.046	0.003	0.223	0.024
	(0.22)	(0.27)	(0.27)	(0.23)	(0.23)	(0.24)	(0.29)	(0.28)	(0.37)
Years of schooling	0.200***	0.199***	0.235***	0.238***	0.215***	0.193***	0.218***	0.233***	0.225***
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)
Population size	0.045	0.061	0.039	0.038	0.043	-0.012	0.027	0.006	0.054
	(0.05)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Area size	0.190***	0.155**	0.112*	0.072	0.066	0.102	0.036	0.035	-0.073
	(0.05)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)
Trade openness	0.676***	0.549**	0.414*	0.445*	0.427*	0.429*	0.543*	0.475*	0.312
	(0.22)	(0.25)	(0.25)	(0.23)	(0.23)	(0.23)	(0.28)	(0.28)	(0.30)
Coastal population	0.464**	0.619**	0.579**	0.533*	0.503*	0.553*	0.301	0.274	0.207
	(0.23)	(0.28)	(0.27)	(0.27)	(0.30)	(0.30)	(0.32)	(0.32)	(0.31)
Absolute latitude	0.017***	0.020***	0.018**	0.011*	0.010	0.012*	0.011*	0.011	0.020***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Socialist law	-0.667***	-0.724***	-0.745***	-0.496**	-0.447*	-0.466**	-0.699***	-0.824***	-1.073***
	(0.22)	(0.27)	(0.26)	(0.23)	(0.23)	(0.23)	(0.24)	(0.24)	(0.23)
Polity2	-0.017	-0.003	-0.006	-0.012	-0.011	0.001	-0.025	-0.026	-0.047**
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Observations	111	111	111	111	111	111	111	111	111
GOF	0.63	0.66	0.67	0.67	0.66	0.65	0.64	0.61	0.56

Standard errors based on 10,000 bootstrap replications are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

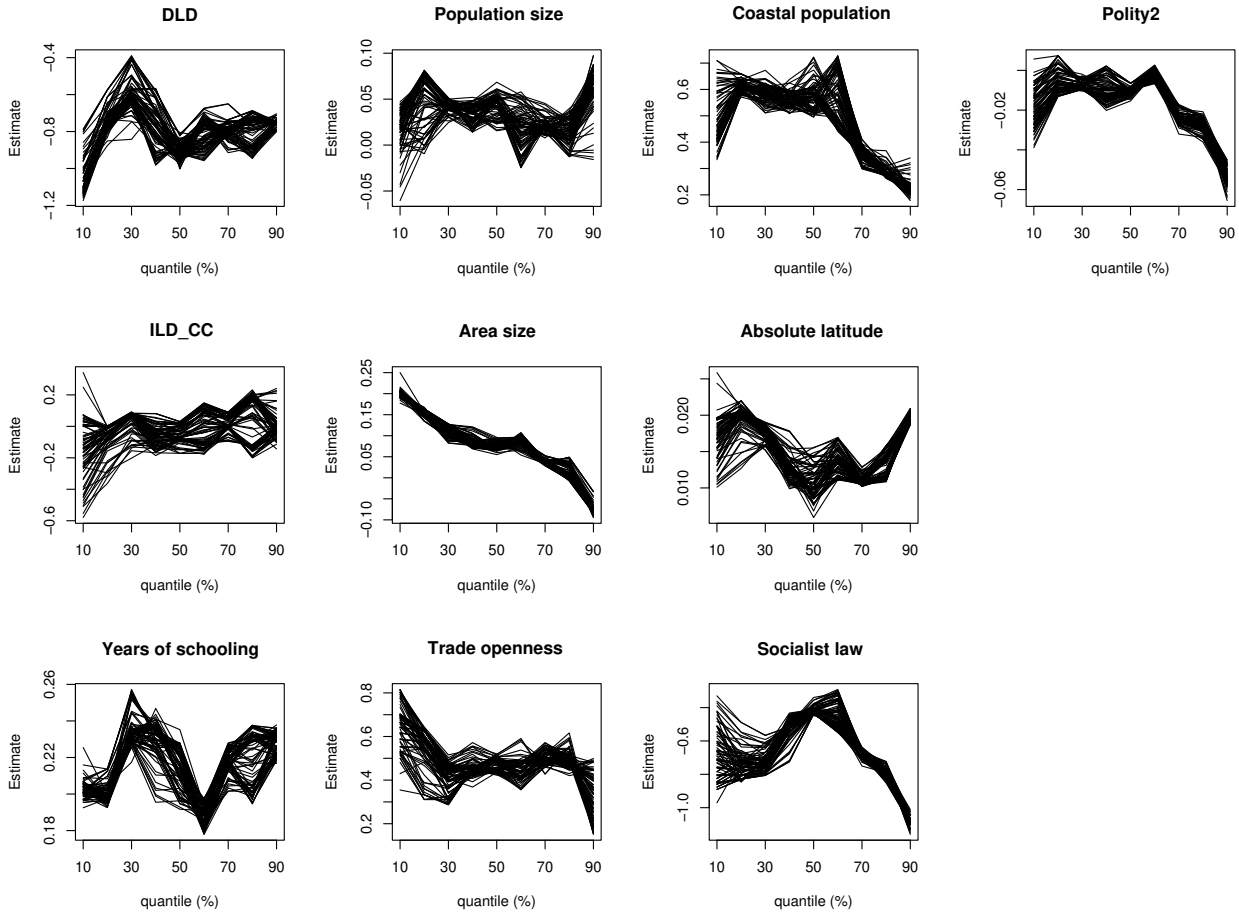

Figure A1: Quantile regression (Baseline model with DLD and ILD_{CC})

Table A6: Kaiser-Meyer-Olkin measure of sampling adequacy

Infectious disease variable	KMO measure
Tuberculosis	0.57
Malaria	0.73
Measles	0.51
Neonatal Tetanus	0.57
Pertussis	0.72
Total Rubella	0.52
Overall	0.59

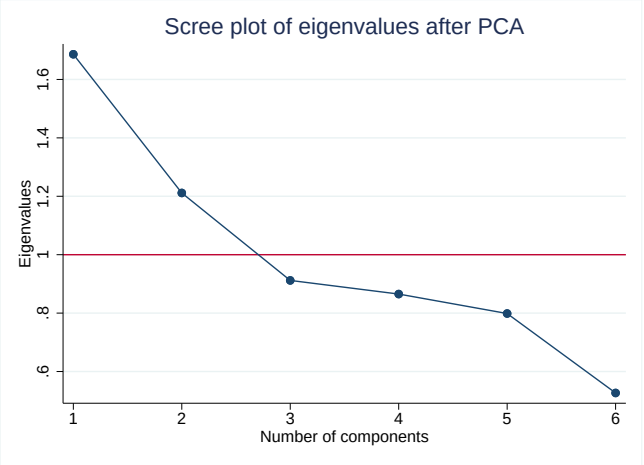

Figure A2: Scree plot of eigenvalues after PCA