CESifo WORKING PAPERS

6720
2017

November 2017

Ethnic Geography: Measurement and Evidence

Roland Hodler, Michele Valsecchi, Alberto Vesperoni

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editors: Clemens Fuest, Oliver Falck, Jasmin Gröschl
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 6720*
Category 6: Fiscal Policy, Macroeconomics and Growth

Ethnic Geography: Measurement and Evidence

Abstract

The effects of ethnic geography, i.e., the distribution of ethnic groups across space, on economic, political and social outcomes are not well understood. We develop a novel index of ethnic segregation that takes both ethnic and spatial distances between individuals into account. Importantly, we can decompose this index into indices of spatial dispersion, generalized ethnic fractionalization, and the alignment of spatial and ethnic distances. We use maps of traditional ethnic homelands, historical population density data, and language trees to compute these four indices for more than 150 countries. We apply these indices to study the relation between historical ethnic geography and current economic, political and social outcomes. Among other things, we document that countries with higher historical alignment, i.e., countries where ethnically diverse individuals lived far apart, have higher-quality government, higher incomes and higher levels of trust.

JEL-Codes: C430, D630, O100, Z130.

Keywords: ethnic diversity, ethnic geography, segregation, fractionalization, quality of government, economic development, trust.

Roland Hodler
Department of Economics
University of St. Gallen / Switzerland
roland.hodler@unisg.ch

Michele Valsecchi
New Economic School
Moscow / Russia
mvalsecchi@new.ru

Alberto Vesperoni
Department of Economics
University of Klagenfurt / Austria
alberto.vesperoni@gmail.com

October 15, 2017
We acknowledge helpful comments by Magnus Hatlebakk, Mario Jametti, Nadine Ketel, Maria Petrova, Marta Reynal-Querol, Måns Söderbom, Ragnar Torvik, participants at the 2016 CESifo Workshop on Political Economy and the 2017 ASWEDE conference, and seminar participants at IEB Barcelona, CMI Bergen, NHH Bergen, Universitat Pompeu Fabra, University of Gothenburg, University of Lugano and University of St.Gallen.

1 Introduction

There is a vast literature on how a country's ethnic diversity affects economic, political and social outcomes. This literature provides evidence for negative effects of ethnic diversity on, e.g., peace, public goods provision, redistribution, the quality of government, and economic development in general. In these studies, ethnic diversity is typically quantified by indices based on the different ethnic groups' country-wide population shares.[1] By definition, these indices ignore ethnic geography, i.e., the distribution of ethnic groups across space.

Ethnic geography may however play an important role. Consider first a country that is ethnically diverse in all locations. The spatial proximity of ethnically diverse individuals could be a cause of friction and mutual distrust, making cooperation at the local level hard to achieve and possibly leading to dysfunctional communities and local governments.[2] As a result of weak social cohesion and poor governance in most locations, this country might well end up with poor governance and poor economic performance at the national level.

Alternatively, consider a country that is equally ethnically diverse (based on the different ethnic group's country-level population shares), but in which all locations are ethnically homogeneous, as the different ethnic groups are separated from one another. In this country, individual communities may be more functional and local governance better. However, at the country level, divisions may be larger and a sense of community harder to achieve, among other things, because the less cumbersome cooperation and preference aggregation at the local level may make it easier for ethnic groups to recruit resources to fight (peacefully or violently) for their own interests at the national level.

These two hypothetical countries suggest that the effects of ethnic geography on governance at the national level are unclear from a theoretical perspective. The notion that the second (more segregated) country would be worse-off at the national level is consistent with the findings of Alesina and Zhuravskaya (2011), who make an important first step towards taking ethnic geography into account. They construct an 'a-spatial' index of ethnic segregation, i.e., an index based on the various ethnic groups' population shares in different subnational units.[3] They find that the quality of government is lower in more ethnically segregated countries.

We contribute to the literature on ethnic diversity by proposing a set of indices that

[1] Prominent examples are the index of ethnic fractionalization (e.g., Easterly and Levine 1997, Alesina et al. 2003, Desmet et al. 2012) and the indices of ethnic polarization (e.g., Esteban and Ray 1994, Montalvo and Reynal-Querol 2005). See Alesina and La Ferrara (2005) for a review of the early literature on ethnic diversity and economic performance.

[2] Studies exploiting within-country variation indeed show that higher local ethnic diversity goes hand-in-hand with lower local public goods provision, less trust, less social capital, less cooperation, weaker social norms, and weaker social sanctioning (e.g., Alesina and La Ferrara 2000, 2002, Miguel and Gugerty 2005, Algan et al. 2016, Gershman and Rivera 2017).

[3] Reardon and Firebaugh (2002) and Reardon and O'Sullivan (2004) review a-spatial and spatial segregation measures, respectively.

capture important aspects of ethnic geography. Our first contribution is a methodological one: we derive a new segregation index that is based on both spatial and ethnic distances between pairs of individuals. There is indeed evidence that both these distances matter.[4]

To develop our index, we consider a society divided into ethnic or, more generally, social groups and scattered over a territory. The starting point is a general class of indices that are expressions of the relation between a randomly selected pair of individuals. The basic idea is that the relation of two individuals depends on whether they are (*i*) unlikely to interact personally due to high spatial distance and (*ii*) unlikely to share a common ethnocultural background due to high ethnic distance. We then uniquely characterize an index from this class via a set of axioms that are intuitive properties of a segregation measure. These axioms capture the notions that segregation is higher when individuals in the same locations are more ethnically homogeneous and when ethnically diverse individuals are located farther apart from one another. Our segregation index can be interpreted as the probability that two randomly selected individuals neither interact personally, nor share a common ethnocultural background.[5]

This index has two prominent features. First, it avoids standard problems of a-spatial segregation indices, such as border dependence and the checkerboard problem (White 1983, Reardon and O'Sullivan 2004).[6] Second, it can be decomposed into three (sub-)indices: an index of spatial dispersion, a well-known index of generalized ethnic fractionalization (see below), and a measure of the alignment of spatial and ethnic distances between individuals (i.e., ethno-spatial alignment or, simply, alignment hereinafter). Figure 1 illustrates these components and the corresponding properties of our segregation index (using different tones of gray to represent different ethnic groups).

Figure 1 about here

First consider part (a) of this figure. Our index suggests that the society in the right diagram is less segregated than the society in the left diagram because the spatial distance between individuals from ethnically distinct groups is lower, all else being equal. This feature is captured by the spatial dispersion component of our segregation index. In part (b) our index suggests that the society in the right diagram is less segregated than the society in the left diagram, because the ethnic distance between individuals from spatially

[4]For the spatial dimension, White (1983) shows that rankings of US cities by racial segregation can be reversed when taking measures that are sensitive to spatial clustering instead of standard a-spatial segregation indices. For the ethnic dimension, Desmet et al. (2009) compare measures of ethnic diversity to predict redistribution at the country level, showing that indices based on linguistic distances between ethnic groups are better predictors than indices based on categorical ethnicity data.

[5]Such probabilistic interpretation simply requires that ethnic and spatial distances are normalized to take values in the unit interval.

[6]There are at least two reasons why overcoming these problems, in particular border dependence, is important: First, administrative borders are the result of policy choices that may be endogenous to ethnic geography. Second, border-dependent segregation measures can lead to different rankings of ethnic segregation across countries depending on the administrative units used (e.g., provinces/states versus districts). Online Appendix A illustrates these standard problems of a-spatial segregation indices.

distant locations (as represented by the more similar tones of gray) is lower, all else equal. This is captured by the generalized ethnic fractionalization component. Part (c) illustrates the important role that ethno-spatial alignment plays in our conceptualization. On average, ethnic and spatial distances are identical in the societies in the left and the right diagrams. However, in the society in the left diagram ethno-spatial alignment is high, as individuals that are ethnically most distant are also located furthest apart. Ethno-spatial alignment is lower in the society in the right diagram, where ethnically distant individuals live spatially relatively close to one another, while spatially distant individuals are ethnically relatively close.

Our second contribution is that we compute and provide these four indices of ethnic geography for 159 countries from all over the world.[7] We define as ethnic groups all language groups listed in the Ethnologue (Gordon, 2005), which allows us to rely on the map of these groups' traditional homelands by the World Language Mapping System (WLMS) and the Ethnologue's own language trees to measure spatial and ethnolinguistic distances, respectively. We further use population density data for 1900 from the History Database of the Global Environment (Klein Goldewijk et al. 2010). The combination of using the WLMS ethnographic map of traditional ethnic homelands and population density data for 1900 implies that our indices measure historical ethnic segregation and its three components.

Our third contribution is an application of our indices of ethnic geography. We use them in cross-country regressions to improve our understanding of the role ethnic geography plays in economic, political and social outcomes around the globe. Our indices are well suited to this purpose thanks to the various precautions we took in designing and computing them. First, they are based on spatial distances rather than administrative borders. They are therefore not driven by the drawing of administrative borders, which is a policy choice that may be endogenous to ethnic geography. Second, our indices are computed by using an ethnographic map of traditional ethnic homelands and historical population density data. They are therefore independent of more recent (voluntary or forced) migration and urbanization, which might again be endogenous to ethnic geography. Third, we have computed these indices for many countries, so that we have a sample with almost full global coverage.

We first focus on the associations between our index of ethnic segregation on the one hand, and the quality of government, incomes and generalized trust on the other. We find a negative (but typically not statistically significant) relation between ethnic segregation and the quality of government, similar to Alesina and Zhuravskaya (2011) with their index of a-spatial segregation in their sample of 97 countries. We further find that our index of ethnic segregation tends to be negatively associated with incomes, but positively with

[7] We do not compute our indices for small countries with a current population of less than 250,000 or a land surface area of less than 5,000 km^2.

trust.

More importantly, we study the relation between the three components of historical ethnic segregation and these economic, political and social outcome variables. Ethnic fractionalization tends to be associated with worse outcomes, but this association is not robust when we control for biological, climatic, geographical or historical variables that may shape ethnic diversity and ethnic geography. Spatial dispersion is not associated with the quality of government or incomes, but positively with trust.[8] Most strikingly, we find a positive and statistically significant association between the historical alignment of ethnic and spatial distances between individuals on the one hand, and the quality of government, incomes and trust on the other. Hence, societies in which ethnically diverse people lived far apart in the past are, on average, better governed, richer and more trusting today.

Our work is related to other contributions on the measurement of segregation that incorporate the spatial dimension. Several contributions introduce spatial distances into well-known a-spatial models of segregation (e.g., Jakubs 1981 for the dissimilarity index; White 1983 for the isolation index; Reardon and O'Sullivan 2004 for the dissimilarity index, the Theil index and the interaction index). Moreover, Echenique and Fryer Jr (2007) develop a segregation index based on proximity in networks.[9] To our knowledge, there is, however, no other segregation measure that presents both ethnic/social and spatial distances in the same framework.[10]

Our framework is also related to prominent models of fractionalization and polarization (e.g., Esteban and Ray 1994, Duclos et al. 2004, Bossert et al. 2011), as we introduce ethnic/social distances in the very same way they do. In particular, the generalized ethnic fractionalization component of our ethnic segregation index coincides with the generalized fractionalization index introduced by Greenberg (1956) and later axiomatized by Bossert et al. (2011), which in turn is equivalent to the standard fractionalization index when ethnic distances are binary.[11]

As mentioned earlier, this paper is related to the extensive literature on the relation between ethnic diversity and economic, political and social outcomes. We contribute to

[8]The positive association between spatial dispersion and trust contributes to the positive association between our index of ethnic segregation and trust.

[9]In their model spatial distances are binary, but the degree of isolation of an individual depends on the isolation of every other individual in the network. Blumenstock and Fratamico (2013) also rely on network data for providing a-spatial segregation measures.

[10]Methodologically, our approach is in the tradition of exposure measurement, being loosely based on the isolation-interaction models of Bell (1954), White (1983), and Philipson (1993). Most axiomatic work on segregation focuses on another class of models, known as evenness indices (e.g., Hutchens 2004, Chakravarty and Silber 2007, and Frankel and Volij 2011). While some evenness measures are extended to introduce spatial distances, they do not lend themselves naturally to the introduction of both spatial and ethnic distances.

[11]From a purely mathematical view point, the generalized fractionalization index axiomatized in Bossert et al. (2011) is essentially an unnormalized Gini index. Analogously, our segregation index can be seen as a particular type of multivariate Gini index (see, e.g., Gajdos and Weymark 2005). However, as it violates standard majorization criteria of multivariate inequality measurement, it should not be interpreted as an inequality measure.

this literature by developing, computing and applying our spatial index of ethnic segregation and its three sub-indices – all with global coverage and based on historical data. There are two complementary strands of the literature that also rely on ethnographic maps to study the role of ethnic geography. The first of these strands chooses subnational ethnographic regions as units of analysis. Prominent examples include studies on the relation between the location of ethnic groups and conflict (e.g., Cederman et al. 2009, Weidmann 2009, Michalopoulos and Papaioannou 2016, König et al. 2017), on the effect of pre-colonial and current institutions on development (Michalopoulos and Papaioannou 2013, 2014), and on ethnic favoritism (De Luca et al. 2016). These contributions provide interesting insights into the effect of ethnic geography on within-country variation while our segregation index allows for comparing ethnic geography across countries and understanding the country-level effects of historical ethnic geography.

Just as we do, contributions to the second strand combine ethnographic maps with population density maps to construct country-level measures of ethnic diversity. Matuszeki and Schneider (2006) compute a measure of average subnational ethnic fractionalization, and study how this measure relates to conflict at the country level. Desmet et al. (2016) construct an alternative measure of average local ethnic diversity, which captures the extent to which individuals live in the same location as individuals from other ethnic groups that are widespread at the country level. They study how this measure relates to public goods provision. There are two main differences between these approaches and ours: First, we focus on conceptualizing ethnic segregation, while they extend the fractionalization framework. Matuszeki and Schneider (2006) do so in a straightforward way, and Desmet et al. (2016) by introducing population weights in a non-linear fashion. Second, spatial (and ethnic) distances play a key role in our approach, while Matuszeki and Schneider (2006) and Desmet et al. (2016) treat these distances as binary variables when constructing their measures of average local ethnic diversity. Hence, these measures remain border-dependent despite taking important aspects of ethnic geography into account.[12]

Section 2 presents the theoretical framework, derives our segregation index, and establishes its decomposability into indices of generalized ethnic fractionalization, spatial dispersion, and ethno-spatial alignment. Section 3 explains the data and the methodology used to construct our four indices of historical ethnic geography and offers a first look at these indices. Section 4 reports the cross-country estimates, and Section 5 concludes.

[12]Montalvo and Reynal-Querol (2016) use ethnographic maps to look at ethnic geography by computing ethnic fractionalization in grid cells of different sizes. Alesina et al. (2016) and Guariso and Rogall (2016) use ethnographic maps to measure inequality across ethnic groups and to study the country-level effects of between-group inequality on economic development and conflict, respectively. Due to the focus of these studies, they take neither the spatial distances between individuals from different ethnic homelands nor the linguistic distances between individuals from different ethnic groups into account.

2 Development of indices of ethnic geography

2.1 General model

A population is partitioned into n ethnic or, more generally, social groups $G := \{1, \ldots, n\}$ and distributed over t locations on a territory $T := \{1, \ldots, t\}$, where $n, t \geq 1$. Denote by $\mu_p^g \in [0,1]$ the share of population that corresponds to group $g \in G$ in location $p \in T$. Let $\mu_p := \sum_{g \in G} \mu_p^g$ and $\mu^g := \sum_{p \in T} \mu_p^g$ be the total population shares of location $p \in T$ and group $g \in G$ respectively, where $\sum_{p \in T} \mu_p = \sum_{g \in G} \mu^g = 1$. Then, the $n \times t$ matrix of population shares

$$\mu := \begin{bmatrix} \mu_1^1 & \cdots & \mu_t^1 \\ \vdots & \ddots & \vdots \\ \mu_1^n & \cdots & \mu_t^n \end{bmatrix}$$

defines a mass distribution, where $\mathcal{M}$ is the space of all mass distributions. For any pair of locations $p, q \in T$, let $\lambda_{p,q} \in [0,1]$ be the (normalized) spatial distance between them. A spatial distribution is defined by the $t \times t$ matrix of spatial distances between all pairs of locations

$$\lambda := \begin{bmatrix} \lambda_{1,1} & \cdots & \lambda_{1,t} \\ \vdots & \ddots & \vdots \\ \lambda_{t,1} & \cdots & \lambda_{t,t} \end{bmatrix},$$

where $\mathcal{L}$ is the space of all spatial distributions. For any pair of groups $g, h \in G$, let $\gamma^{g,h} \in [0,1]$ be the (normalized) ethnic distance between them. The $n \times n$ matrix of ethnic distances between all pairs of groups

$$\gamma := \begin{bmatrix} \gamma^{1,1} & \cdots & \gamma^{1,n} \\ \vdots & \ddots & \vdots \\ \gamma^{n,1} & \cdots & \gamma^{n,n} \end{bmatrix}$$

defines an ethnic distribution, and the space of all ethnic distributions is $\mathcal{G}$. Finally, a joint distribution is a triple of mass, spatial and ethnic distributions, and an index is a function $S : (\mathcal{M}, \mathcal{L}, \mathcal{G}) \to \mathbb{R}_+$, where $S(\mu, \lambda, \gamma)$ quantifies some property of the joint distribution $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$.

To give meaning to our framework we now impose some more structure. We assume (a relevant feature of) the relation between each pair of individuals is determined by the distances between their groups and locations.[13] For each pair of individuals that inhabit locations $p, q \in T$ and belong to groups $g, h \in G$, we quantify the relation between them by $\pi(\lambda_{p,q}, \gamma^{g,h})$, where the function $\pi : [0,1]^2 \to \mathbb{R}_+$ is continuous and non-decreasing in each argument and satisfies $\pi(0,0) = 0$. Among the various interpretations of the

[13] For related approaches, see Esteban and Ray (1994), Duclos et al. (2004), and Bossert et al. (2011).

function π, one possibility is to see it as the degree of alienation (i.e., lack of common interests) between a pair of individuals, which naturally increases with their spatial and ethnic distances. Given this, we consider the class of indices that are expression of the relation between a randomly selected pair of individuals, taking the form

$$\mathcal{S}(\mu, \lambda, \gamma) := \sum_{(p,q)\in T^2} \sum_{(g,h)\in G^2} \mu_p^g \mu_q^h \pi(\lambda_{p,q}, \gamma^{g,h}) \tag{1}$$

for each joint distribution $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$.

We will introduce a set of axioms that pin down a particular index (up to positive scalar multiplications) from the class of measures (1) as our segregation index. As function π is generic (e.g., logarithmic, exponential, multiplicative, additive, etc.), class (1) is vast. Nevertheless, the focus on class (1) considerably narrows the set of indices under consideration by taking pairs of individuals as the relevant unit of analysis and by imposing that any pair's contribution to segregation depends on their spatial and ethnic distances only.[14] We are not concerned by these restrictions. First, we think of segregation as a measure of the extent to which ethnically diverse individuals are located far apart, which captures the notion that society becomes more segregated when the interaction between ethnically diverse individuals becomes less likely. Second, we deliberately take spatial (and ethnic) distances as primitives of the model in order to build a segregation measure that is based on continuous distances rather than arbitrary borders between locations (and ethnic groups). As our unit of analysis is the pair of individuals, function π could only be generalized by making it dependent on some elements of the mass distribution μ. However, by introducing some element of μ in function π, we would implicitly assume that the relation between two individuals is discontinuous at some borders between locations (or ethnic groups).[15] Any generalization of function π would therefore (re-)introduce border dependence "through the back door."

2.2 Axiomatization of the segregation index

We now introduce a set of axioms that are desirable properties of a segregation measure. In the statements of the axioms, we write $(\mu, \lambda, \gamma) \prec (\tilde{\mu}, \tilde{\lambda}, \tilde{\gamma})$ to say that a segregation measure should assign to joint distribution (μ, λ, γ) a strictly lower degree of segregation than to joint distribution $(\tilde{\mu}, \tilde{\lambda}, \tilde{\gamma})$. For simplicity of exposition, our axioms define desirable

[14] To see this, one can rewrite $\mathcal{S}$ as a function of distances between pairs of individuals rather than groups and locations. With some abuse of notation, let $\lambda_{i,j}$ and $\gamma^{i,j}$ denote the spatial and ethnic distances between each pair of individuals i, j from a finite population P. Then, $\mathcal{S} = (1/|P|^2) \sum_{(i,j)\in P^2} \pi(\lambda_{i,j}, \gamma^{i,j})$.

[15] As pointed out in Footnote 14, class (1) can be written as a function of spatial and ethnic distances between pairs of individuals. In applications, categorizing individuals in a limited number of locations and ethnicities (i.e., introducing arbitrary borders) is a necessary approximation. Ideally, this should not lead to systematic biases in the approximation of the index. While these biases are minimal for class (1) due to its linearity in each element of μ, they would be magnified if we had some element of μ in function π due to the non-linearity.

properties of segregation through simple examples of distributions with two or three mass points. The first two axioms consider pairs of groups and locations, thereby focusing on obtaining ethnic homogeneity within a location. In particular, segregation should increase when the population becomes ethnically homogeneous in all locations, such that there is no interaction between ethnically diverse individuals within any location. Axiom 1 formalizes this property and, in addition, requires this to hold when the ethnic distance between the two groups is reduced by an arbitrarily small amount.

Axiom 1 (Local ethnic homogeneity and ethnic distances) *Data:* Consider a joint distribution $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$ with two locations $p, q \in T$ and two groups $g, h \in G$ such that

$$\mu_p^g = \mu_p^h = \mu_q^h = 1/3,$$

$$\lambda_{p,q} > \lambda_{p,p} = \lambda_{q,q} \text{ and } \gamma^{g,h} > \gamma^{g,g} = \gamma^{h,h},$$

while letting $\tilde{\mu} \in \mathcal{M}$, $\tilde{\gamma} \in \mathcal{G}$ and $\epsilon \geq 0$ satisfy

$$\tilde{\mu}_p^g = \mu_p^g,\ \tilde{\mu}_q^h = \mu_p^h + \mu_q^h,$$

$$\tilde{\gamma}^{g,g} = \tilde{\gamma}^{h,h} = \gamma^{g,g} \text{ and } \tilde{\gamma}^{g,h} = \gamma^{g,h} - \epsilon.$$

Statement: We require $(\mu, \lambda, \gamma) \prec (\tilde{\mu}, \lambda, \tilde{\gamma})$ for $\epsilon > 0$ arbitrarily small.

Let us discuss Axiom 1, whose distributions are depicted in Figure 2(a). There are two locations (left and right) and two ethnic groups (represented by dark and light tones of gray). Initially, in distribution (μ, λ, γ), two-thirds of the population are in the left location, whose ethnic composition is perfectly balanced (half dark, half light), while the remaining one-third of the population is in the right location and is homogeneously dark. Given this, we transfer all individuals of the dark group into the right location, so that the left location becomes homogeneously light while the right location remains homogeneously dark. Moreover, we reduce the ethnic distance between the light and the dark group by an arbitrarily small amount ϵ (represented by the slightly lighter tone of gray of the dark group in the right diagram). Axiom 1 requires segregation to increase as a consequence of this transformation. Intuitively, the axiom considers a trade off between ethnic homogeneity within locations and the ethnic distance across groups, requiring the former to dominate the trade off when the reduction in ethnic distance is arbitrarily small.

Figure 2 about here

Axiom 2 is very similar to Axiom 1. It is based on the same initial distribution and the same transfer of population from the left to the right location. The only difference is that, instead of reducing the ethnic distance between the light and the dark groups, we

reduce the spatial distance between the left and right locations by an arbitrarily small amount.

Axiom 2 (Local ethnic homogeneity and spatial distances) *Data:* Consider a joint distribution $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$ with two locations $p, q \in T$ and two groups $g, h \in G$ such that

$$\mu_p^g = \mu_p^h = \mu_q^h = 1/3,$$

$$\lambda_{p,q} > \lambda_{p,p} = \lambda_{q,q} \text{ and } \gamma^{g,h} > \gamma^{g,g} = \gamma^{h,h},$$

while letting $\tilde{\mu} \in \mathcal{M}$, $\tilde{\lambda} \in \mathcal{L}$ and $\epsilon \geq 0$ satisfy

$$\tilde{\mu}_p^g = \mu_p^g,\ \tilde{\mu}_q^h = \mu_p^h + \mu_q^h,$$

$$\tilde{\lambda}_{p,p} = \tilde{\lambda}_{q,q} = \lambda_{p,p} \text{ and } \tilde{\lambda}_{p,q} = \lambda_{p,q} - \epsilon.$$

Statement: We require $(\mu, \lambda, \gamma) \prec (\tilde{\mu}, \tilde{\lambda}, \gamma)$ for $\epsilon > 0$ arbitrarily small.

These distributions are depicted in Figure 2(b). Intuitively, this axiom considers a trade off between ethnic homogeneity within locations and the spatial distance across locations, requiring the former to dominate the trade off when the reduction in the spatial distance is arbitrarily small.

The next two axioms are still inspired by the generally desirable property that segregation should increase whenever the interaction between ethnically diverse individuals becomes less likely. However, unlike Axioms 1 and 2, they consider triples of groups and locations, thereby focusing on changes in distributions that foster the alignment of spatial and ethnic distances across pairs of individuals. The basic idea is that, to obtain higher segregation, closely located pairs of individuals should be ethnically closer, while ethnically distant pairs should be spatially further apart. Axioms 3 and 4 formalize this idea.

Axiom 3 (Alignment of ethnic distances) *Data:* Consider any joint distribution $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$ with three locations $p, q, r \in T$ and three groups $g, h, i \in G$ such that

$$\begin{aligned} \mu_p^g = \mu_q^h = \mu_r^i &= 1/3, \\ \lambda_{p,q} > \lambda_{q,r} > \lambda_{p,p} = \lambda_{q,q} = \lambda_{r,r} &\text{ and } \lambda_{p,r} = \lambda_{p,q} + \lambda_{q,r}, \\ \gamma^{g,h} = \gamma^{h,i} = \gamma^{g,i}/2 > \gamma^{g,g} &= \gamma^{h,h} = \gamma^{i,i}, \end{aligned}$$

and let $\tilde{\gamma} \in \mathcal{G}$ and $\epsilon \geq 0$ satisfy

$$\tilde{\gamma}^{g,g} = \gamma^{g,g},\ \tilde{\gamma}^{h,h} = \gamma^{h,h},\ \tilde{\gamma}^{i,i} = \gamma^{i,i},$$

$$\tilde{\gamma}^{g,i} = \gamma^{g,i},\ \tilde{\gamma}^{g,h} = \gamma^{g,h} + \epsilon,\ \tilde{\gamma}^{h,i} = \gamma^{h,i} - \epsilon.$$

Statement: We require $(\mu, \lambda, \gamma) \prec (\mu, \lambda, \tilde{\gamma})$ for all $\epsilon \in (0, \gamma^{h,i} - \gamma^{g,g})$.

Let us discuss Axiom 3, whose distributions are depicted in Figure 2(c). The population mass is uniformly distributed on three locations (left, central and right) and three ethnic groups (represented by dark, medium and light tones of gray), where the left location is homogeneously light, the central location is homogeneously medium and the right location is homogeneously dark. The three locations are on a line, where the central location is closer to the right than to the left. Regarding ethnic distances, the medium group is halfway between the other two groups in the left diagram representing distribution (μ, λ, γ). Axiom 3 requires segregation to increase when we change ethnic distances so that the medium group becomes ethnically closer to the dark group (represented by the darker tone of gray of the middle location in the right diagram). This is intuitive: as the medium group already inhabits a location that is spatially closer to the location of the dark group than to the location of the light group, the interaction between ethnically diverse individuals becomes less likely.

Axiom 4 (Alignment of spatial distances) *Data:* Consider any joint distribution $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$ with three locations $p, q, r \in T$ and three groups $g, h, i \in G$ such that

$$\begin{aligned} \mu_p^g = \mu_q^h = \mu_r^i &= 1/3, \\ \lambda_{p,q} = \lambda_{q,r} = \lambda_{p,r}/2 &> \lambda_{p,p} = \lambda_{q,q} = \lambda_{r,r}, \\ \gamma^{g,h} > \gamma^{h,i} &> \gamma^{g,g} = \gamma^{h,h} = \gamma^{i,i}, \text{ and } \gamma^{g,i} = \gamma^{g,h} + \gamma^{h,i}, \end{aligned}$$

and let $\tilde{\lambda} \in \mathcal{L}$ and $\epsilon \geq 0$ satisfy

$$\tilde{\lambda}_{p,p} = \lambda_{p,p},\ \tilde{\lambda}_{q,q} = \lambda_{q,q},\ \tilde{\lambda}_{r,r} = \lambda_{r,r},$$

$$\tilde{\lambda}_{p,r} = \lambda_{p,r},\ \tilde{\lambda}_{p,q} = \lambda_{p,q} + \epsilon,\ \tilde{\lambda}_{q,r} = \lambda_{q,r} - \epsilon.$$

Statement: We require $(\mu, \lambda, \gamma) \prec (\mu, \tilde{\lambda}, \gamma)$ for all $\epsilon \in (0, \lambda_{q,r} - \lambda_{p,p})$.

Figure 2(d) represents Axiom 4 graphically. Again, there are three locations respectively inhabited by three equally sized ethnic groups. The medium group is ethnically closer to the dark group than to the light, while the central location is halfway between the right and the left location. Axiom 4 requires segregation to increase if the central location is moved closer to the right location. Similarly to the previous axiom, the intuition is that as the spatial distance between ethnically diverse individuals increases, their interaction becomes less likely.

Our four axioms identify our segregation index from the class of measures (1):[16]

Theorem 1 *Let $n, t \geq 3$. An index from class (1) satisfies Axioms 1-4 if and only if it takes the form*

$$S(\mu, \lambda, \gamma) := \sum_{(p,q)\in T^2} \sum_{(g,h)\in G^2} \mu_p^g \mu_q^h \lambda_{p,q} \gamma^{g,h}, \tag{2}$$

up to a positive scalar multiplication.

This theorem implies that our segregation index always provides unambiguous rankings of joint distributions $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$. Further, it implies that ethnic and spatial distances are complementary forces in the determination of the relation of a pair of individuals, so that segregation is high only if pairs of individuals that are ethnically heterogeneous are systematically located apart from each other.

Given $\lambda_{p,q} \in [0,1]$ and $\gamma^{g,h} \in [0,1]$, the function $\pi(\lambda_{p,q}, \gamma^{g,h}) = \lambda_{p,q}\gamma^{g,h}$ always takes a value in $[0,1]$. It can thus be interpreted probabilistically. Intuitively, the relation between two individuals depends on (i) whether they do not interact personally and (ii) whether they do not share a common ethnocultural background. Given this, it is natural to interpret the function π as the probability that *both* these events are realized, where the spatial distance $\lambda_{p,q}$ is the probability of event (i) and the ethnic distance $\gamma^{g,h}$ is the probability of event (ii). Then, our segregation index S represents the probability that two randomly selected individuals neither interact personally nor share an ethnocultural background.

2.3 Decomposition of the segregation index

By construction, our segregation index is strongly related to the fractionalization literature. Let $\mathbf{1}_t \in \mathcal{L}$ be the spatial distribution where the spatial distance between each pair of locations is equal to 1. It is easy to show that, when all locations are equidistant, our index is equivalent to the generalized fractionalization index by Bossert et al. (2011),

$$F(\mu, \gamma) := S(\mu, \mathbf{1}_t, \gamma) = \sum_{(g,h)\in G^2} \mu^g \mu^h \gamma^{g,h}. \tag{3}$$

This generalized fractionalization index represents the average ethnic distance between pairs of individuals, and can be interpreted as the probability that two randomly selected individuals do not share a common ethnocultural background. If we also impose ethnic distances to take value in $\{0,1\}$, our index reduces to the standard fractionalization index, which has been widely applied to measure ethnic fractionalization based on categorical data (see, e.g., Alesina et al. 2004 and references therein).[17]

[16] The proof of Theorem is 1 in the Appendix.

[17] To see this, let $\mathbf{1}_n^0 \in \mathcal{G}$ be the ethnic distribution, where $\gamma^{g,h} = 1$ if $h \neq g$ and $\gamma^{g,g} = 0$ for each $g \in G$, so that $F(\mu, \mathbf{1}_n^0) = S(\mu, \mathbf{1}_t, \mathbf{1}_n^0) = 1 - \sum_{g\in G} (\mu^g)^2$, which is the standard fractionalization index,

Applying the same reasoning to the other dimension, and letting $\mathbf{1}_n \in \mathcal{G}$ be the ethnic distribution where the distance between each pair of groups is 1, we can define the spatial dispersion index as

$$D(\mu, \lambda) := S(\mu, \lambda, \mathbf{1}_n) = \sum_{(p,q) \in T^2} \mu_p \mu_q \lambda_{p,q}. \tag{4}$$

This index measures the average spatial distance between pairs of individuals and can be interpreted as the probability that two randomly selected individuals will not interact personally.

Our segregation index tends to be high if spatial distances between locations and ethnic distances between groups are high, i.e., when F and D are high. Moreover, it also depends on the alignment between spatial and ethnic distances, i.e., on whether a high spatial distance between two individuals tends to go hand-in-hand with a high ethnic distance between them. For each $\mu \in \mathcal{M}$, denote by $\overline{\mu} \in \mathcal{M}$ the uniform mass distribution corresponding to μ, where (i) groups and locations have the same mass as in μ, i.e., $\overline{\mu}^g = \mu^g$ and $\overline{\mu}_p = \mu_p$ for all $g \in G$ and $p \in T$; and (ii) groups are proportionally represented at each location, i.e., $\overline{\mu}^g_p / \overline{\mu}_p = \overline{\mu}^g$ for all $g \in G$ and $p \in T$. We propose as a measure of ethno-spatial alignment

$$A(\mu, \lambda, \gamma) := \begin{cases} S(\mu, \lambda, \gamma) / S(\overline{\mu}, \lambda, \gamma) & \text{if } S(\overline{\mu}, \lambda, \gamma) > 0, \\ 1 & \text{if } S(\overline{\mu}, \lambda, \gamma) = 0. \end{cases} \tag{5}$$

Given our probabilistic interpretation of S, A can be seen as a likelihood ratio: it is the probability that two randomly selected individuals do not interact personally and do not share an ethnocultural background given mass distribution μ, relative to the probability of the same event given mass distribution $\overline{\mu}$, which is identical to μ except that the ethnic composition is the same everywhere. Intuitively, focusing on the likelihood ratio should 'neutralize' the magnitude effects of average spatial and ethnic distances. In fact, $A(\mu, k\lambda, k'\gamma) = A(\mu, \lambda, \gamma)$ for all $k, k' > 0$, while $S(\mu, k\lambda, k'\gamma) = kk' S(\mu, \lambda, \gamma)$ for all $k, k' > 0$. Hence, our measure of alignment satisfies scale invariance with respect to both spatial and ethnic distances, while our segregation index does not. Other properties of our measure of alignment directly follow from the axioms in the previous section, which are all satisfied in the sense that alignment increases whenever segregation increases.

Lastly, we show how the various measures are related to one other:[18]

Proposition 1 *It holds that*

$$S(\mu, \lambda, \gamma) = \begin{cases} F(\mu, \gamma) D(\mu, \lambda) A(\mu, \lambda, \gamma) & \textit{if } F(\mu, \gamma) > 0 \textit{ and } D(\mu, \lambda) > 0, \\ 0 & \textit{if } F(\mu, \gamma) = 0 \textit{ or } D(\mu, \lambda) = 0. \end{cases} \tag{6}$$

i.e., the probability that two randomly selected individuals belong to different ethnic groups.

[18] The proof of Proposition 1 is in the Appendix.

This proposition shows that our segregation index S can be decomposed into the generalized ethnic fractionalization index F, the spatial dispersion index D, and the alignment index A in a multiplicative fashion.[19]

3 Computing our indices of ethnic geography

3.1 Data and computation

We aim at computing our indices of ethnic geography, i.e., the segregation index and its three components, for a large and diverse set of countries from all over the world. For these countries, we need information on locations and ethnic groups, so that we can then derive mass distribution μ, spatial distribution λ, and ethnic distribution γ. These distributions are the inputs required for the computation of our indices.

We therefore combine two data sources. First, we use the Ethnologue (Gordon, 2005), which provides a comprehensive list of the world's known living languages. We consider the language groups listed in the Ethnologue as ethnic groups. It is important to remember that language is more than just a communication device. Common language often implies common ancestry, homeland, cultural heritage, norms, and values.[20] The advantages in relying on the Ethnologue for classifying ethnic groups are fourfold: First, the Ethnologue provides a comprehensive rather than a selective list of ethnolinguistic groups. Second, the Ethnologue provides linguistic trees for the different language families which show the historical relation between all languages. These linguistic trees are thus helpful in measuring linguistic distances between ethnic groups. Third, the World Language Mapping System (WLMS, version 19) provides an ethnographic map representing the homelands of the language groups in the Ethnologue. An ethnographic map allows measuring spatial distances between locations inhabited by different groups. Last, but not least, this ethnographic map focuses on the different groups' traditional homelands, while populations living away from their traditional homelands, e.g., migrations to cities and refugees, are not mapped. This focus on traditional homelands makes this ethnographic map a useful tool for constructing indices of historical ethnic geography.[21]

The second data source is the History Database of the Global Environment (HYDE, version 3.2) by Klein Goldewijk et al. (2010). This database contains historical informa-

[19] We discuss in Online Appendix B how this decomposition relates to the interpretation of our segregation index as a geometric projection and to a decomposition of S based on the Euclidean norms of vectors of spatial and ethnic distances. More specifically, we argue that F is proportional to the Euclidean norm of the vector of ethnic distances, D to the Euclidean norm of the vector of spatial distances, and A to the inner product of these two vectors. Our focus remains on decomposition (6), which is more readily applicable in terms of data availability.

[20] Desmet et al. (2017) find that ethnic identity is an important determinant of responses to many questions on cultural norms, values and preferences in the World Value Surveys.

[21] Notice that while we use many components of the Ethnologue product family, we do not use its population data, which is based on recent population censuses.

tion on population density and land use for grid cells of 0.5×0.5 arc minutes (corresponding to around 9×9 km near the equator).[22] We mainly rely on their population density data for 1900.

The combination of using an ethnographic map of traditional ethnic homelands and population density data for 1900 implies that our indices will measure key dimensions of historical ethnic geography. Hence, our indices are mainly shaped by biological, climatic, geographical and historical forces that shaped the distribution of people in space in times of lower mobility within countries rather than by the more recent mass migration of individuals to cities.[23]

We take as ethnic groups in each country all the language groups with more than 100 native speakers listed in the Ethnologue and with a homeland mapped within this country. The median and average number of ethnic groups per country are 9 and 30, respectively. There is however a lot of variability in the number of groups: Some countries (15 out of 159 in our sample) have only one ethnic group, while Papua New Guinea, Indonesia and Nigeria have 734, 607 and 450 ethnic groups, respectively.

To determine locations, we use the HYDE grid cells and cut them at country borders and at the boundaries between different ethnic homelands. We thereby get "proper" cells of 0.5×0.5 arc minutes as well as smaller "squiggly" cells (due to country borders or ethnic homeland boundaries). We take each of these (proper or squiggly) cells as a location.

To determine the mass distribution μ, we rely on the population density data for 1900 from HYDE. Let m, m_p and m_p^g denote the total population of a country, the population in cell p and the population of language group g in cell p, respectively. Assigning population m_p to proper cells of 0.5×0.5 arc minutes is straightforward. To obtain population m_p for squiggly cells, which are subsets of HYDE grid cells, we assume that population is uniformly distributed across squiggly cells belonging to the same HYDE grid cell.

Figure 3 illustrates the ethnic homelands and the HYDE grid cells for Togo (left) and Benin (right). Moreover, it indicates the historical population in each proper and squiggly cell.[24]

Add Figure 3 around here

Ultimately, we do not need population m_p per cell p, but population m_p^g per cell p and group g. For cells p that are part of a traditional homeland of a single language group g, it is straightforward that $m_p^g = m_p$. The ethnographic map by WMLS indeed suggests that most homelands have only one language group, but other homelands contain more than one and up to seven language groups. We find that 90 percent of our proper and

[22] See Klein Goldewijk (2005) for information on the construction of historical population density for the years 1700-2000.

[23] The urbanization rate increased from below 30 percent to above 50 percent from 1950 to 2000, not least because of a large increase in urbanization rates in poorer countries (Glaeser, 2014).

[24] Figure 3 further provides information on the spatial distribution of different language groups in Togo and Benin. We will make use of this information in our discussion in Section 3.2.

squiggly cells belong to the homeland of a single group. The remaining 10 percent of our cells belong to ethnic homelands of multiple ethnic groups. Let n_p denote the number of ethnic groups whose ethnic homeland includes cell p. We find that for 9 percent of cells $n_p = 2$, while $n_p > 2$ for 1 percent of cells. For these groups and cells, we simply assume $m_p^g = \frac{m_p}{n_p}$.[25] We then compute population shares as $\mu_p^g = \frac{m_p^g}{m}$, where $m = \sum_{p \in T} m_p$.

To derive the spatial distribution λ, we use ArcGIS to determine the centroid of each (proper or squiggly) cell p. We then use the latitude and the longitude of these centroids to compute the geodesic distance $\lambda_{p,q}$ between any two cells p and q of any given country.[26]

To derive the ethnic distribution γ, we rely on the Ethnologue's linguistic trees for the different language families. Linguistic trees characterize each language by a series of nodes and thereby contain information about the evolution of languages and the historical relation between ethnolinguistic groups. Two languages share no common node if they belong to different language families, e.g., the Indo-European and the Uralic language family. Such coarse divisions suggest that the language groups separated early and interacted little. In contrast, languages with many common nodes, e.g., Norwegian and Swedish, suggest that the language groups separated late or interacted regularly. Following Fearon (2003), it has become common practice to calculate linguistic distance between groups as a function of the number of common nodes of their languages and to use the linguistic distance between groups as a proxy for their cultural distance more broadly defined. We follow Putterman and Weil (2010, Appendix C) in defining the ethnic distance between ethnic groups g and h as

$$\gamma^{g,h} := 1 - \sqrt{2\tilde{\eta}^{g,h}/(\eta^g + \eta^h)},$$

where η^i is the number of nodes of language $i \in \{g, h\}$ and $\tilde{\eta}^{g,h}$ the number of common nodes.[27]

Using mass distribution μ, spatial distribution λ, and ethnic distribution γ, we derive our indices of historical ethnic geography for 159 countries with a land surface area of more than 5,000 km^2 and a current population of more than 250,000.[28]

[25] This simple rule may lead us to overestimate the local population of very small language groups, which is the main reason for dropping languages spoken by no more than 100 individuals.

[26] We measure geodesic distances in 1,000 miles or 1,600 km, respectively.

[27] Fearon (2003) proposes a slightly different formula. Online Appendix E (Table E.2) shows that our cross-country results are robust to using his formula.

[28] See Online Appendix C for a list of the 159 countries for which we provide our indices of historical ethnic geography. We view HYDE as unsuitable for small countries due its spatial resolution and its incomplete coverage of small island states. Besides small countries, we also exclude Austria, because the homelands in the ethnographic map cover only a small portion of the area, and Serbia, because of the many changes to its borders in recent years. For the 15 countries with only one traditional ethnic homeland, alignment $A(\mu, \lambda, \gamma)$ is equal to one by definition although it is not very informative. Online Appendix E (Tables E.3–E.5) shows that our cross-country results are robust to dropping these 15 countries.

3.2 A first look at our indices

Table 1 provides some summary statistics for our indices of ethnic geography, and Figure 4 provides scatter plots illustrating the empirical relation between our index of ethnic segregation and its three components.

Add Table 1 and Figure 4 around here

The ten most ethnically segregated countries according to our index of ethnic segregation are (in decreasing order of segregation) India, Peru, Mali, Kazakhstan, Indonesia, Papua New Guinea, China, Nigeria, Democratic Republic of the Congo (DRC), and Canada. The two scatter plots in the top row of Figure 4 show positive correlations between ethnic segregation, on the one hand, and ethnic fractionalization and spatial dispersion, on the other hand. They suggest that Mali, Nigeria, Papua New Guinea, and Peru are among the most ethnically segregated countries mainly because they are highly ethnically fractionalized, while Canada, China, DRC, Indonesia, and Kazakhstan are among the most ethnically segregated countries mainly because they are highly spatially dispersed. India is both highly ethnically fractionalized and highly spatially dispersed.[29]

These two scatter plots also illustrate that neither high ethnic fractionalization, nor high spatial dispersion is sufficient for high ethnic segregation. Good examples are Australia and Belize: Australia is a large country with high spatial dispersion, but is characterized by a high share of English speakers, such that ethnic fractionalization is very low, thus leading to low ethnic segregation. Belize is a country with high linguistic distances between various ethnic groups and, therefore, high generalized ethnic fractionalization. But it is also a rather small country with little spatial dispersion, such that ethnic segregation is relatively low nevertheless.

The scatter plot on the bottom left of Figure 4 shows the relation between our index of ethnic segregation and the alignment between ethnic and spatial distances. It documents an empirically negative relation between ethnic segregation and ethno-spatial alignment. We have seen in Proposition 1 in Section 2 that, all else being equal, segregation increases with ethno-spatial alignment. This scatter plot now shows that, all else not being equal, more aligned countries tend to be less ethnically segregated. The scatter plot on the bottom right of Figure 4 shows that, as we would expect, the relation between ethnic segregation and ethno-spatial alignment becomes positive once we partial out $F \times D$.

Norway is one of the countries with high ethno-spatial alignment. Most people speak Norwegian, which is a language from the Indo-European language family, and they used to live and still live relatively close to one another in the South of the country (e.g., around

[29]The correlation between our spatial index of historical ethnic segregation and the a-spatial index of ethnic segregation by Alesina and Zhuravskaya (2011) is 0.26; and the correlation between our historical index of generalized ethnic fractionalization and their index of ethnic fractionalization is 0.57.

Bergen or Oslo). There are however some small language groups that speak Kven Finnish and Sami. Like Finnish, these languages belong to the Uralic language family. Moreover, the homelands of these language groups are in the far North of Norway. The members of these groups were therefore both linguistically and spatially very far from the Norwegian speakers in the South, such that the linguistic distance of a pair of individuals was a very good predictor of the spatial distance, and vice versa.

Interestingly, there are also countries where alignment is less than one, implying that the ethnic distance between spatially distant pairs of individuals tends to be smaller than the ethnic distance between spatially close pairs of individuals. One example is Turkmenistan, where the Turkmen are the largest language group. Moreover, there are three minority groups, speaking Balochi, Kurdish, and Uzbek. Balochi and Kurdish belong to the Indo-European language family, while Turkmen and Uzbek belong to the Altaic language family. Because the homelands of the two Indo-European languages are in fairly central and densely populated areas, pairs of linguistically diverse individuals lived on average closer to one another than pairs of individuals speaking the same or very similar languages.

Of course, Norway and Turkmenistan differ in many dimensions. Let us therefore look at Benin and Togo, which differ in their ethno-spatial alignment, but are similar along many other dimensions. They are neighboring countries located in West Africa, with comparable climatic, geographic and demographic characteristics. Moreover, they were both French colonies after WWI, became independent in 1960, and started their post-colonial history in tumultuous ways that culminated in coups by French-trained military figures: Mathieu Kérékou in Benin and Gnassingbé Eyedéma in Togo (Meredith, 2005). These autocrats both managed to stay in power for many years. Benin and Togo are also comparable in terms of generalized ethnic fractionalization (0.31 vs 0.27) and spatial dispersion (both 0.13). Ethno-spatial alignment is however considerably higher in Benin than in Togo (1.32 vs 1.11). Figure 3 shows the different ethnic homelands and the main language groups to which these ethnic homelands belong. Ethno-spatial alignment is relatively high in Benin as there is a relatively clear divide between Kwa speaking groups in the south, Defoid speaking groups in the center, Gur speaking groups in the north, and some smaller groups speaking very different languages in the north east. As a result of this divide, linguistically distant individuals tended to live far apart from one another. In contrast, ethno-spatial alignment is relatively low in Togo, mainly because there are Gur and Kwa speaking groups in the country's south, its center and its north. As a result of these large and widespread language groups, linguistically distant individuals often lived relatively close to one another.

4 Cross-country evidence

We now turn to applications of our indices of ethnic geography to see whether they are helpful in understanding cross-country differences in the quality of government and economic outcomes. The use of cross-country regressions is common in the literature on the effects of ethnic heterogeneity, as is the caveat that the estimated coefficients may not necessarily represent causal effects despite efforts to reduce the risk of reverse causality or omitted variable biases. In our case, the risk of reverse causality is reduced by our reliance on traditional ethnic homelands and historical population data in the computation of the indices.

In most specifications we control for absolute latitude and dummy variables for the different continents. These variables proxy for a host of geographical, climatic and (maybe) cultural aspects, and are known to be strong predictors of economic and institutional outcomes. To address omitted variable bias, we control for additional variables that are known determinants of ethnic heterogeneity or ethnic geography, and may have direct effects on current economic and institutional outcomes. We use five groups of additional control variables that relate to a country's climate and geography or its history: First, we add temperature and precipitation to control more explicitly for climate. Nettle (1998) argues that the length of the growing season is a key determinant of the number of ethnic groups in a territory, and he calculates this length based on temperature and precipitation. In addition, climate is known to have more direct effects on economic outcomes as well (e.g., Dell et al., 2012). Second, we control for terrain ruggedness and its interaction with a dummy variable for Africa. Nunn and Puga (2012) argue that rugged terrain generally has negative effects on economic development, although the effects were positive in Africa, as such terrain offered some protection against slave raiders. Nunn (2008) further argues that the slave trade promoted ethnic and political fragmentation and had negative effects on economic development. Third, we control for the mean and standard deviation of both elevation and soil suitability for agriculture. Michalopoulos (2012) shows that geographic variability as proxied by these variables is a key determinant of ethnic diversity across and within countries. At the same time, land productivity is likely to have direct economic effects.

Turning to historical variables, we, fourth, control for the time elapsed since the agricultural transition as well as for the migratory distance to Addis Ababa (Ethiopia) and its squared term. Ahlerup and Olsson (2012) argue that the agricultural transition had strong effects on population density and ethnic heterogeneity; and the biological and geographical factors that led to the early emergence of sedentary agriculture may well have shaped economic development. Migratory distance from the cradle of humankind in East Africa is a predictor for the duration of human settlement. Ahlerup and Olsson (2012) argue that ethnic diversity increases with this duration. In addition, Ashraf and Galor

(2013) show that genetic diversity is a decreasing function of the migratory distance from East Africa, and that economic development is a hump-shaped function of genetic diversity. Fifth, we control for dummy variables indicating whether the country is a former colony and, if so, whether it was a British, French, Spanish or some other colony. There is considerable evidence that the random drawing of borders and divide-and-rule strategies by the colonial powers shaped ethnic heterogeneity and ethnic geography, and had long-term effects on economic and political outcomes (e.g., Michalopoulos and Papaioannou, 2016).[30]

4.1 Ethnic geography and the rule of law

Inspired by Alesina and Zhuravskaya (2011), we first look at the rule of law as a measure of the quality of government. This measure is provided by the World Bank Governance Indicators. By construction, it has a mean of 0 and a standard deviation of 1. In our sample, which excludes many small island states, its 2010 value has a mean of -0.212 and a standard deviation of 0.995. Table 2 shows our results. The columns differ in the set of control variables used. The top panel presents estimates using our index of ethnic segregation, while the bottom panel replaces this index with its three components: ethno-spatial alignment, generalized ethnic fractionalization, and spatial dispersion.

Table 2 around here

We see in column (1) that the rule of law is negatively associated with segregation in the absence of control variables. This negative association is consistent with the findings by Alesina and Zhuravskaya (2011). When decomposing segregation into its three components, we find – again consistent with the previous literature (e.g., Alesina et al., 2003) – that the rule of law is negatively associated with fractionalization. In contrast, we find no statistically significant association between spatial dispersion and the rule of law. More interestingly, we find that the rule of law is positively associated with ethno-spatial alignment. This result is novel, as is the concept of ethno-spatial alignment itself. Hence, given the levels of fractionalization and dispersion, a country has a better rule of law if individuals from very different groups lived far apart from one another.

In column (2), we add our main controls, i.e., absolute latitude and the continental dummy variables. The associations of the rule of law with segregation (in the top panel) and fractionalization (in the bottom panel) remain negative, but become much weaker and are no longer statistically significant. In contrast, the association with alignment remains

[30]See Online Appendix D for more information about the control variables. We take many of the control variables from Ashraf and Galor (2013). Following them and many others, we exclude from our sample the relatively young countries Montenegro and South Sudan as well as Palestine and Taiwan, which are not UN member states, leaving us with a sample of 155 countries with a land surface area of more than 5,000 km^2 and a current population of more than 250,000.

almost unchanged in magnitude and becomes even more precisely estimated. The point estimate suggests that an increase of alignment by one standard deviation is associated with an increase in the rule of law by 17 percent of a standard deviation.

In columns (3)–(7), we add the additional control variables discussed above. We see that the association between alignment and the rule of law is relatively stable in magnitude and remains statistically significant for any of these five additional groups of control variables.[31] We conclude that high historical alignment between ethnic and spatial distances goes hand-in-hand with high quality of government today.

4.2 Ethnic geography and income

We now look at the association between ethnic geography and income, measured by the log of expenditure-side real GDP per capita in USD in 2010 from the Penn World Tables 9.0. Table 3, which shows the results, is organized in the same way as the previous table.

Table 3 around here

The results are similar as well. Ethnic segregation is negatively associated with income, but this association is only statistically significant when we omit all control variables. The same holds true for generalized ethnic fractionalization when segregation is decomposed into its three components. Moreover, the association between spatial dispersion and income is not statistically significant. The association between ethno-spatial alignment and income is however positive and statistically significant in all specifications. The point estimate in column (2) suggests that an increase in alignment by one standard deviation is associated with an increase in income by 24 percent.

Hence, high historical alignment between ethnic and spatial distances goes hand-in-hand with high quality of government as well as high incomes today. This pattern also holds true when comparing Benin and Togo. Remember that these neighboring countries are similar along many dimensions, but ethno-spatial alignment is higher in Benin. Our data show that Benin indeed does better in terms of quality of government (-0.70 vs -0.91) and income per capita (USD 1,728 vs USD 1,214).[32]

4.3 Ethnic geography and trust

These strong associations raise the question about possible mechanisms linking historical ethno-spatial alignment with current quality of government and current incomes. The within-country studies by Alesina and La Ferrara (2000, 2002), Miguel and Gugerty

[31] When all 24 control variables are added jointly, the coefficient on alignment becomes statistically insignificant at the five percent level (as do all other coefficients except the negative one on the dummy variable for Asia and the positive one on mean soil suitability).

[32] The data on trust, introduced in Section 4.3, is missing for Benin and Togo.

(2005), and Algan et al. (2016) document that high local ethnic diversity leads to or is at least associated with low social capital and lack of trust. High ethno-spatial alignment implies that ethnic diversity tends to be low in most locations (conditional on the level of ethnic fractionalization). As a result, trust may be higher in countries with high ethno-spatial alignment.

We use generalized trust from the World Values Surveys in the 1981–2008 time period (taken from Ashraf and Galor, 2013) to look at the role of trust. Generalized trust is measured as the fraction of people answering "most people can be trusted" (as opposed to "can't be too careful") when asked the standard trust question (see Online Appendix D for details). We have coverage for 76 countries, which implies a drop in sample size by around 50 percent. Table 4 presents the associations between our indices of historical ethnic geography and trust.

Table 4 around here

Ethno-spatial alignment is indeed positively associated with generalized trust in all specifications. The point estimate in column (2) suggests that an increase in alignment by one standard deviation is associated with an increase in trust by 28 percent of a standard deviation. In addition, the estimates in the upper panel show that ethnic segregation is positively associated with trust. The reasons are that, besides ethno-spatial alignment, spatial dispersion is also positively associated with trust, while there is no clear relation between generalized ethnic fractionalization and trust.

In Table 5, we further explore the idea that trust could be a possible mechanism explaining why historically more aligned societies are better governed and wealthier today.

Table 5 around here

In column (1), we replicate our main specification for the rule of law (Table 2, column 2), but restrict the sample to the 76 countries for which the trust variable is available. The effect is similar in magnitude as in the full sample and again statistically significant. In column (2), we then add trust as an additional explanatory variable. We see that the point estimate for ethno-spatial alignment drops by more than half (and is no longer statistically significant), while trust itself has a strong positive effect on the rule of law. This pattern is consistent with the idea that historically more aligned societies have a higher quality of government today, partly because they have higher trust, and higher trust improves the quality of government.

In columns (3) and (4), we repeat the same exercise, but use incomes instead of the rule of law as the dependent variable. The emerging pattern is similar, except that the coefficient on ethno-spatial alignment drops only by around one third when trust is controlled for, and that trust itself is only statistically significant at the 10 percent level.

We conclude that despite the limited number of observations, we find relatively strong

evidence that high historical ethno-spatial alignment goes hand-in-hand with high trust today and some tentative evidence that the alignment's association with trust may partly drive its association with good governance and high incomes.

4.4 Robustness

We document in Online Appendix E that the results reported in Tables 2-4 are by and large robust to, among other things, (i) the use of alternative measures for the quality of government and income, (ii) alternative computations of our indices of ethnic geography, (iii) the exclusion of different continents or outliers, and (iv) the use of alternative estimators such as weighted least squares or poisson pseudo-maximum likelihood.

5 Conclusions

To better understand the role of ethnic geography and to mitigate well-known problems of a-spatial segregation measures, we have developed a new segregation index that is based on ethnic distances between groups and spatial distances between locations rather than categorical data on ethnic groups and administrative units. The decomposition of our segregation index reveals that it corresponds to the product of generalized ethnic fractionalization, spatial dispersion, and the alignment between ethnic and spatial distances. This ethno-spatial alignment is a novel concept that captures, broadly speaking, whether ethnically more diverse individuals tend to live farther away from each other. We have computed these four indices using linguistic trees as well as maps of traditional ethnic homelands and historical population data, so that our indices capture key aspects of historical ethnic geography. Using these indices in cross-country regressions suggests, among other things, that countries with higher historical ethno-spatial alignment tend to be better governed, richer, and more trusting today.

We expect our indices to become useful in future work on the role of ethnic geography in shaping economic, political and social outcomes across countries. However, we also hope to speak to the rapidly growing literature that uses ethnic homelands (or pixels) as units of analysis to achieve convincing identification strategies. To this literature, we would like to convey the message that local economic, political or social outcomes in any given ethnic homeland may well depend on the broader ethnic geography of the area or country in which this homeland is located.

Of course, the indices we have developed can also be applied for measuring the ethnic geography of cities. For example, one could use our segregation index instead of a-spatial measures to compare segregation across US metropolitan areas or within metropolitan areas over time. Given that our indices allow for non-categorical ethnicity data, they may be even more attractive in studying the ethnic geography of emerging African mega-cities,

where there is typically great variability in ethnic distances across pairs of individuals.

Finally, we would like to stress that our theoretical framework is not specific to the ethnic dimension. Instead of categorizing individuals by ethnic groups and measuring linguistic distances, future research could focus on other social or socio-economic cleavages that are believed to be salient in a particular setting.

Appendix: Proofs

Proof of Theorem 1: It is easy to verify that our segregation index (2) belongs to class (1) and satisfies Axioms 1-4. Let us show that, if an index belongs to class (1) and satisfies Axioms 1-4, then it must take the form (2) up to a positive scalar multiplication. Take any index from class (1) and let $a, b > 0$ be any scalars, where a is spatial distance and b is ethnic distance in what follows. By Axiom 1, for $\epsilon > 0$ arbitrarily small,

$$\pi(a, b) + \pi(0, b) + \pi(a, 0) < 2\pi(a, b - \epsilon).$$

Letting $a \to 0$, by continuity of π and $\pi(0, 0) = 0$, we obtain at the limit

$$\pi(0, b) \leq \pi(0, b - \epsilon).$$

Then, since π is non-decreasing, $\pi(0, b)$ must be constant in b; and by $\pi(0, 0) = 0$ we must have

$$\pi(0, b) = 0 \text{ for all } b \geq 0. \tag{7}$$

Similarly, by Axiom 2, for $\epsilon > 0$ arbitrarily small,

$$\pi(a, b) + \pi(0, b) + \pi(a, 0) < 2\pi(a - \epsilon, b),$$

so that letting $b \to 0$ by the same arguments we obtain

$$\pi(a, 0) = 0 \text{ for all } a \geq 0. \tag{8}$$

Keeping our interpretation of a as spatial distance and b as ethnic distance, let $c > 0$ be any scalar that represents another spatial distance in the following. By Axiom 3, for all $\epsilon \in (0, b)$

$$\pi(a, b) + \pi(c, b) < \pi(a, b + \epsilon) + \pi(c, b - \epsilon) \text{ if } c < a,$$
$$\pi(a, b) + \pi(c, b) > \pi(a, b + \epsilon) + \pi(c, b - \epsilon) \text{ if } c > a,$$

hence by continuity of π

$$\pi(a, b) + \pi(c, b) = \pi(a, b + \epsilon) + \pi(c, b - \epsilon) \text{ if } c = a.$$

Rearranging terms this leads to

$$\pi(a, b) = \frac{\pi(a, b + \epsilon) + \pi(a, b - \epsilon)}{2} \text{ for all } \epsilon \in (0, b),$$

hence π must be linear in the second argument. Jointly with (7) and (8), this implies $\pi(a,b) = \phi(a)b$ for all $a, b \geq 0$, where $\phi : [0,1] \to \mathbb{R}_+$ is some continuous non-decreasing function that satisfies $\phi(0) = 0$. Similarly, by Axiom 4 (interpreting a as spatial distance, b as ethnic distance and c as another ethnic distance), for all $\epsilon \in (0, b)$

$$\pi(b,a) + \pi(b,c) = \pi(b+\epsilon, a) + \pi(b-\epsilon, c) \text{ if } c = a,$$

hence π must also be linear in the first argument. It follows that $\phi(a) = ka$ for some $k > 0$, and we obtain $\pi(a,b) = kab$ for all $a, b \geq 0$. □

Proof of Proposition 1: It is straightforward that, if $F(\mu,\gamma) = 0$ or $D(\mu,\lambda) = 0$, we must have $S(\mu,\lambda,\gamma) = 0$. To see this, note that $F(\mu,\gamma) = 0$ implies $\gamma^{g,h} = 0$ for all $g, h \in G$. Similarly, $D(\mu,\lambda) = 0$ implies $\lambda_{p,q} = 0$ for all $p, q \in T$. Then, if $F(\mu,\gamma) = 0$ or $D(\mu,\lambda) = 0$, there is either zero spatial distance or zero ethnic distance between each pair of individuals, which implies $S(\mu,\lambda,\gamma) = 0$ by the multiplicative form of p.

We now show that, if $F(\mu,\gamma) > 0$ and $D(\mu,\lambda) > 0$, we must have

$$S(\mu,\lambda,\gamma) = F(\mu,\gamma)D(\mu,\lambda)A(\mu,\lambda,\gamma).$$

By the definition of $A(\mu,\lambda,\gamma)$, this is true if and only if

$$S(\overline{\mu},\lambda,\gamma) = F(\mu,\gamma)D(\mu,\lambda), \tag{9}$$

where the uniform mass distribution $\overline{\mu}$ corresponding to μ is such that (i) $\overline{\mu}^g = \mu^g$ and $\overline{\mu}_p = \mu_p$ for all $g \in G$ and $p \in T$; and (ii) $\overline{\mu}^g_p / \overline{\mu}_p = \overline{\mu}^g$ for all $g \in G$ and $p \in T$. Combining the definition of our index with (ii) we obtain

$$\begin{aligned} S(\overline{\mu},\lambda,\gamma) &= \sum_{(p,q)\in T^2} \sum_{(g,h)\in G^2} \left(\overline{\mu}_p \overline{\mu}^g\right)\left(\overline{\mu}_q \overline{\mu}^h\right) \lambda_{p,q}\gamma^{g,h} \\ &= \left(\sum_{(p,q)\in T^2} \overline{\mu}_p \overline{\mu}_q \lambda_{p,q}\right)\left(\sum_{(g,h)\in G^2} \overline{\mu}^g \overline{\mu}^h \gamma^{g,h}\right), \end{aligned}$$

which together with (i) implies (9). □

References

Ahlerup, Pelle, and Ola Olsson, "The Roots of Ethnic Diversity," *Journal of Economic Growth*, 17 (2012), 71–102.

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

Alesina, Alberto, and Eliana La Ferrara, "Participation in Heterogeneous Communities," *Quarterly Journal of Economics*, 115 (2000), 847–904.

Alesina, Alberto, and Eliana La Ferrara, "Who Trusts Others?" *Journal of Public Economics*, 85 (2002), 207–234.

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

Alesina, Alberto, Stelios Michalopoulos, and Elias Papaioannou, "Ethnic Inequality," *Journal of Political Economy*, 124 (2016), 428–488.

Alesina, Alberto, and Ekaterina Zhuravskaya, "Segregation and the Quality of Government in a Cross Section of Countries," *American Economic Review*, 101 (2011), 1872–1911.

Algan, Yann, Camille Hémet, and David Laitin, "The Social Effects of Ethnic Diversity at the Local Level: A Natural Experiment with Exogenous Residential Allocation," *Journal of Political Economy*, 124 (2016), 696–733.

Ashraf, Quamrul, and Oded Galor, "The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development," *American Economic Review*, 103 (2013), 1–46.

Bell, Wendell, "A Probability Model for the Measurement of Ecological Segregation," *Social Forces*, 32 (1954), 357–364.

Blumenstock, Joshua, and Lauren Fratamico, "Social and Spatial Ethnic Segregation: A Framework for Analyzing Segregation with Large-Scale Spatial Network Data," *Proceedings of the 4th Annual Symposium on Computing for Development*, 4 (2013), 11.

Bossert, Walter, Conchita D'Ambrosio, and Eliana La Ferrara, "A Generalized Index of Fractionalization," *Economica*, 78 (2011), 723–750.

Cederman, Lars-Erik, Halvard Buhaug, and Jan K. Rød, "Ethno-Nationalist Dyads and Civil War: A GIS-based Analysis," *Journal of Conflict Resolution*, 53 (2009), 496–525.

Chakravarty, Satya R., and Jacques Silber, "A Generalized Index of Employment Segregation," *Mathematical Social Sciences*, 53 (2007), 185–195.

De Luca, Giacomo, Roland Hodler, Paul A. Raschky, and Michele Valsecchi, "Ethnic Favoritism: An Axiom of Politics?" CEPR Discussion Paper 11351 (2016).

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken, "Temperature Shocks and Economic Growth: Evidence from the Last Half Century," *American Economic Journal: Macroeconomics*, 4 (2012), 66–95.

Desmet, Klaus, Joseph Gomes, and Ignacio Ortuño-Ortín, "The Geography of Linguistic Diversity and the Provision of Public Goods," CEPR Discussion Paper 11683 (2016).

Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg, "The Political Economy of Linguistic Cleavages," *Journal of Development Economics*, 97 (2012), 322–338.

Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg, "Culture, Ethnicity and Diversity," *American Economic Review*, 107 (2017), 2479–2513.

Desmet, Klaus, Shlomo Weber, and Ignacio Ortuño-Ortín, "Linguistic Diversity and Redistribution," *Journal of the European Economic Association*, 7 (2009), 1291–1318.

Duclos, Jean-Yves, Joan Esteban, and Debraj Ray, "Polarization: Concepts, Measurement, Estimation," *Econometrica*, 72 (2004), 1737–1772.

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

Echenique, Federico, and Roland G. Fryer, Jr., "A Measure of Segregation Based on Social Interactions," *Quarterly Journal of Economics*, 122 (2007), 441–485.

Esteban, Joan, and Debraj Ray, "On the Measurement of Polarization," *Econometrica*, 62 (1994), 819–851.

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

Frankel, David M., and Oscar Volij, "Measuring School Segregation," *Journal of Economic Theory*, 146 (2011), 1–38.

Gajdos, Thibault, and John A. Weymark, "Multidimensional generalized Gini indices," *Economic Theory*, 26 (2005), 471–496.

Gershman, Boris, and Diego Rivera, "Subnational Diversity in Sub-Saharan Africa: Insights from a New Dataset," Mimeo (2017).

Glaeser, Edward L., "A World of Cities: The Causes and Consequences of Urbanization in Poorer Countries," *Journal of the European Economic Association*, 12 (2014), 1154–1199.

Gordon, Raymond G., Jr., *Ethnologue: Languages of the World* (Dallas: SIL International, 2005).

Greenberg, Joseph H., "The Measurement of Linguistic Diversity," *Language*, 32 (1956), 109–115.

Guariso, Andrea, and Thorsten Rogall, "Rainfall Inequality, Political Power, and Ethnic Conflict in Africa," Mimeo (2016)

Hutchens, Robert M., "One Measure of Segregation," *International Economic Review*, 45 (2004), 555–578.

Jakubs, John F., "A Distance-Based Segregation Index," *Socio-Economic Planning Sciences*, 15 (1981), 129–136.

Klein Goldewijk, Kees, "Three Centuries of Global Population Growth: A Spatial Referenced Population (Density) Database for 1700–2000." *Population and Environment*,

26 (2005), 343–367.

Klein Goldewijk, Kees, Arthur Beusen, and Peter Janssen. "Long-term Dynamic Modeling of Global Population and Built-up Area in a Spatially Explicit Way: HYDE 3.1," *The Holocene*, 20 (2010), 565–573.

König, Michael D., Dominic Rohner, Mathias Thoenig, and Fabrizio Zilibotti, "Networks in Conflict: Theory and Evidence from the Great War of Africa." *Econometrica*, 85 (2017), 1093–1132.

Matuszeki, Janina, and Frank Schneider, "Patterns of Ethnic Group Segregation and Civil Conflict," Mimeo (2006).

Meredith, Martin, *The Fate of Africa: A History of the Continent Since Independence* (New York: Free Press, 2005).

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

Michalopoulos, Stelios, and Elias Papaioannou, "Pre-Colonial Ethnic Institutions and Contemporary African Development," *Econometrica*, 81 (2013), 113–152.

Michalopoulos, Stelios, and Elias Papaioannou, "National Institutions and Subnational Development in Africa," *Quarterly Journal of Economics*, 129 (2014), 151–213.

Michalopoulos, Stelios, and Elias Papaioannou, "The Long-Run Effects of the Scramble for Africa," *American Economic Review*, 106 (2016), 1802–1848.

Miguel, Edward, and Mary Kay Gugerty, "Ethnic Diversity, Social Sanctions, and Public Goods in Kenya," *Journal of Public Economics*, 89 (2005), 2325–2368.

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

Montalvo, Jose G., and Marta Reynal-Querol, "Ethnic Diversity and Growth: Revisiting the Evidence," Mimeo (2016).

Nettle, Daniel, "Explaining Global Patterns of Language Diversity," *Journal of Anthropological Archaeology*, 17 (1998), 354–374.

Nunn, Nathan, "The Long-term Effects of Africa's Slave Trades," *Quarterly Journal of Economics*, 123 (2008), 139–176.

Nunn, Nathan, and Diego Puga, "Ruggedness: The Blessing of Bad Geography in Africa," *Review of Economics and Statistics*, 94 (2012), 20–36.

Philipson, Tomas, "Social Welfare and Measurement of Segregation," *Journal of Economic Theory*, 60 (1993), 322–334.

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

Reardon, Sean F., and Glenn Firebaugh, "Measures of Multigroup Segregation," *Sociological Methodology*, 32 (2002), 33–67.

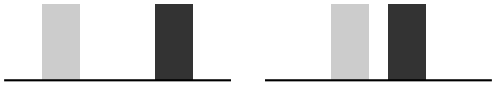
Reardon, Sean F., and David O'Sullivan, "Measures of Spatial Segregation," *Sociological*

Methodology, 34 (2004), 121–162.

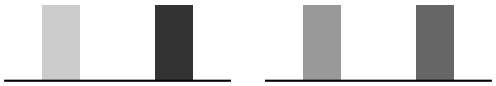
Weidmann, Nils B., "Geography as Motivation and Opportunity: Group Concentration and Ethnic Conflict, *Journal of Conflict Resolution*, 53 (2009), 526–543.

White, Michael J., "The Measurement of Spatial Segregation," *American Journal of Sociology*, 88 (1983), 1008–1018.

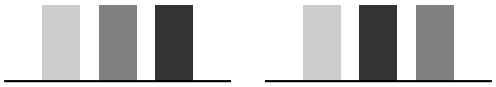
Figures and Tables



(a) Importance of spatial distances



(b) Importance of ethnic distances



(c) Importance of alignment

Figure 1: Illustration of our segregation measure

Notes: The two diagrams of each sub-figure depict two distributions of ethnic groups in space. Each tone of gray indicates a different ethnic group, and ethnic distances between groups are given by differences in tones of gray. Spatial locations are on the horizontal axis, which also measures spatial distances, while the vertical axis measures the population mass at each location.

(a) Distributions of Axiom 1.

(b) Distributions of Axiom 2.

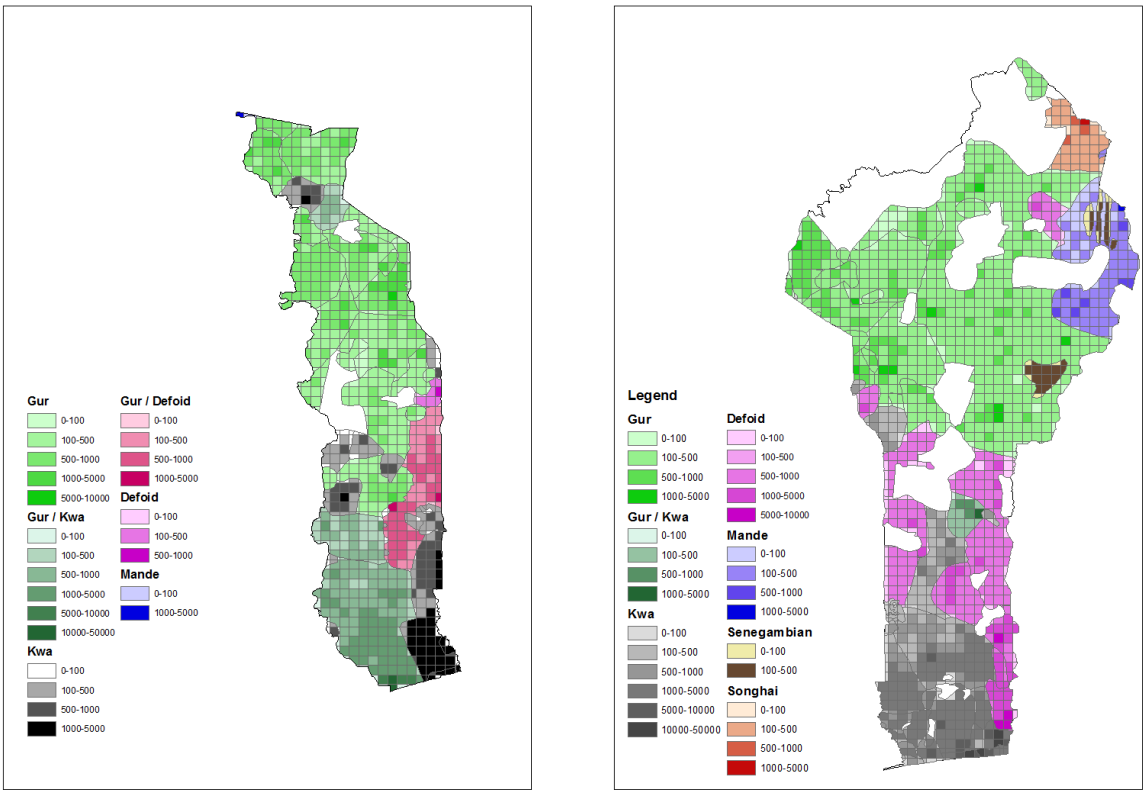
(c) Distributions of Axiom 3.

(d) Distributions of Axiom 4.

Figure 2: Illustration of the distributions of the axiomatization

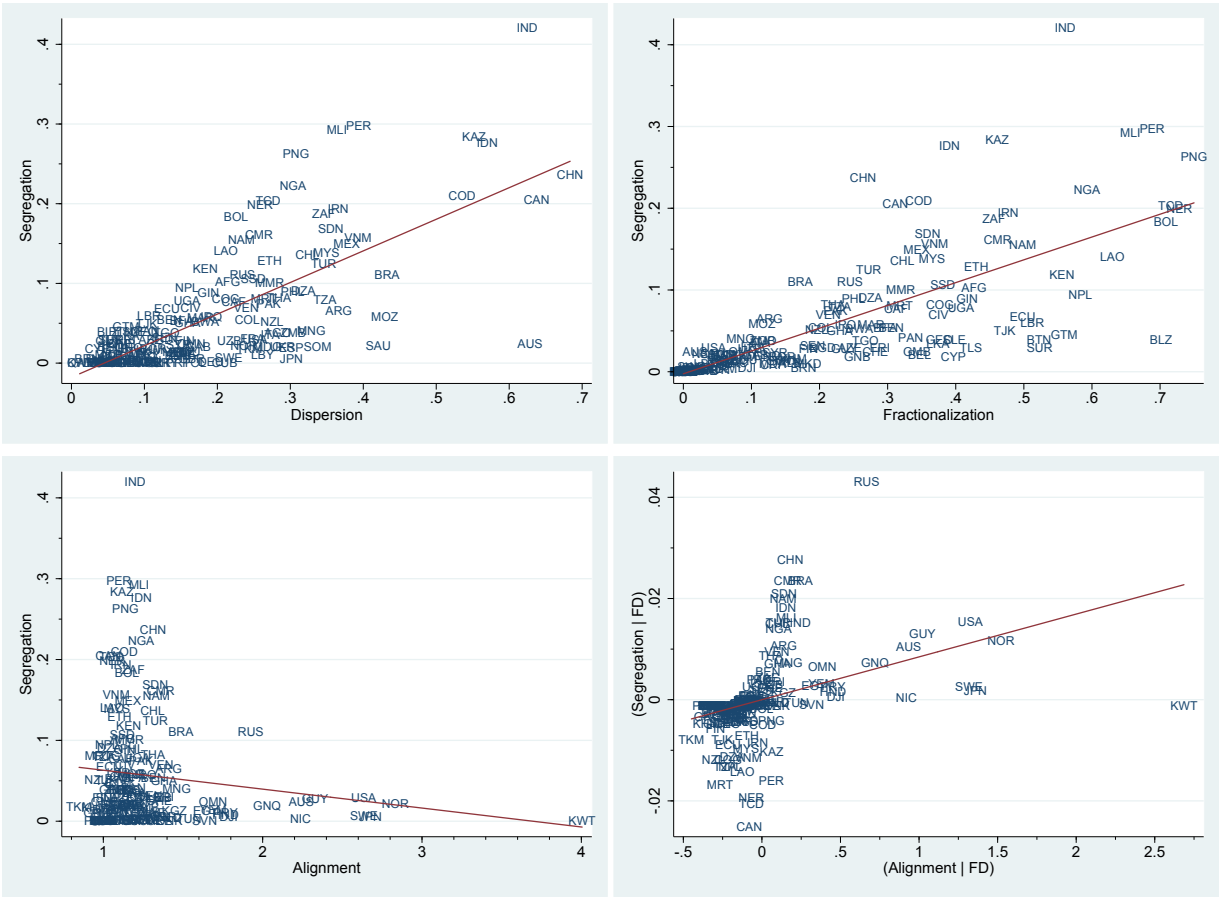
Notes: The two diagrams of each sub-figure depict two distributions of ethnic groups in space. Each tone of gray indicates a different ethnic group, and ethnic distances between groups are given by differences in tones of gray. Spatial locations are on the horizontal axis, which also measures spatial distances, while the vertical axis measures the population mass at each location.

Figure 3: Traditional ethnic homelands and historical population data for Togo and Benin



Notes: Maps of Togo (left) and Benin (right) showing the traditional homelands of language groups according to WLMS and the grid cells in HYDE. Each grid cell constitutes a different location in the computation of our indices, each color indicates that the corresponding grid cell belongs to the traditional homeland of a certain language group (with the relevant language groups given in the legend), and the brightness of this color indicates the size of the population that historically inhabited the grid cell (also given in the legend). The legend entries Gur/Kwa and Gur/Defoid indicate the traditional homelands of multiple language groups, some speaking a Gur language and some a Kwa or Defoid language. WLMS indicates no traditional homelands in the white areas.

Figure 4: Scatter plots illustrating the index of ethnic segregation and its components



Notes: Scatter plots showing the associations between the index of ethnic segregation S and its three components: spatial dispersion (D, top left), generalized ethnic fractionalization (F, top right) and alignment (A, bottom left). Additional scatter plot showing the association between S and A after partialling out F×D from both S and A (bottom right). The (red) lines indicate the best linear fit.

Table 1: Summary statistics for our indices of ethnic geography

	Observations	Mean	Std. Dev.	Min.	Max.
Segregation	159	0.057	0.075	0	0.420
Alignment	159	1.269	0.400	0.848	4.005
Fractionalization	159	0.213	0.201	0	0.750
Dispersion	159	0.188	0.139	0.011	0.685

Table 2: Ethnic geography and the rule of law

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Segregation	-3.04***	-0.78	-0.59	-0.73	-0.08	-1.02	-0.63
	(0.96)	(0.84)	(0.82)	(0.84)	(0.97)	(0.90)	(0.77)
R^2	0.05	0.42	0.43	0.43	0.47	0.42	0.46
Alignment	0.43**	0.43***	0.43***	0.42***	0.36**	0.49***	0.36**
	(0.20)	(0.14)	(0.14)	(0.15)	(0.16)	(0.14)	(0.14)
Fractionalization	-1.51***	-0.37	-0.35	-0.27	-0.14	-0.35	-0.54
	(0.35)	(0.33)	(0.33)	(0.34)	(0.37)	(0.33)	(0.34)
Dispersion	0.25	0.18	0.33	0.12	0.48	-0.06	0.50
	(0.60)	(0.45)	(0.46)	(0.47)	(0.49)	(0.51)	(0.43)
R^2	0.14	0.46	0.47	0.46	0.50	0.47	0.49
Main controls	No	Yes	Yes	Yes	Yes	Yes	Yes
Add. controls	No	No	Climate	Rugged	Geo. var.	Deep hist.	Col. hist.
Countries	155	155	154	154	148	150	155

Notes: The dependent variable is rule of law in 2010 from the World Bank Governance Indicators. Each column presents two OLS regressions with the same set of controls. In the upper panel the main explanatory variable is ethnic segregation, and in the lower panel these are ethno-spatial alignment, generalized ethnic fractionalization and spatial dispersion. These indices are all explained in Sections 2 and 3. Main controls are absolute latitude and continental dummy variables. Additional controls are temperature and precipitation in column (3); terrain ruggedness and its interaction with a dummy variable for Africa in column (4); averages and standard deviations of elevation and land suitability for agriculture in column (5); migratory distance from Addis Ababa, its square term, and the time elapsed since the agricultural transition in column (6); and dummy variables for former British/French/Spanish/other colonies in column (7). Online Appendix D contains more information on dependent and control variables. Robust standard errors. ***, **, * indicate significance at the 1, 5 and 10%-level, respectively.

Table 3: Ethnic geography and income

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Segregation	-4.00***	-0.96	-0.82	-1.18	-0.62	-0.95	-0.77
	(1.26)	(0.90)	(0.97)	(0.93)	(1.09)	(0.92)	(0.90)
R^2	0.06	0.53	0.55	0.56	0.63	0.56	0.56
Alignment	0.59***	0.53***	0.52***	0.50***	0.34**	0.44***	0.46***
	(0.17)	(0.15)	(0.16)	(0.15)	(0.17)	(0.14)	(0.14)
Fractionalization	-2.20***	-0.64	-0.59	-0.46	-0.60	-0.66	-0.75*
	(0.49)	(0.43)	(0.42)	(0.43)	(0.47)	(0.42)	(0.44)
Dispersion	0.52	0.31	0.38	-0.02	0.53	0.31	0.64
	(0.75)	(0.53)	(0.54)	(0.56)	(0.52)	(0.57)	(0.57)
R^2	0.19	0.58	0.58	0.59	0.65	0.59	0.60
Main controls	No	Yes	Yes	Yes	Yes	Yes	Yes
Add. controls	No	No	Climate	Rugged	Geo. var.	Deep hist.	Col. hist.
Countries	146	146	146	145	140	143	146

Notes: The dependent variable is log of expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0. Each column presents two OLS regressions with the same set of controls. In the upper panel the main explanatory variable is ethnic segregation, and in the lower panel these are ethno-spatial alignment, generalized ethnic fractionalization and spatial dispersion. These indices are all explained in Sections 2 and 3. Main controls are absolute latitude and continental dummy variables. Additional controls are temperature and precipitation in column (3); terrain ruggedness and its interaction with a dummy variable for Africa in column (4); averages and standard deviations of elevation and land suitability for agriculture in column (5); migratory distance from Addis Ababa, its square term, and the time elapsed since the agricultural transition in column (6); and dummy variables for former British/French/Spanish/other colonies in column (7). Online Appendix D contains more information on dependent and control variables. Robust standard errors. ***, **, * indicate significance at the 1, 5 and 10%-level, respectively.

Table 4: Ethnic geography and trust

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Segregation	0.15	0.48***	0.52***	0.47***	0.26	0.31**	0.45**
	(0.18)	(0.16)	(0.16)	(0.16)	(0.19)	(0.15)	(0.17)
R^2	0.01	0.40	0.44	0.41	0.51	0.48	0.40
Alignment	0.11***	0.10***	0.09***	0.11***	0.08**	0.09***	0.10***
	(0.04)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.04)
Fractionalization	-0.15*	0.03	0.02	0.08	0.07	0.04	0.02
	(0.08)	(0.08)	(0.07)	(0.08)	(0.09)	(0.07)	(0.09)
Dispersion	0.25**	0.27***	0.30***	0.23**	0.19*	0.18**	0.27***
	(0.11)	(0.09)	(0.09)	(0.09)	(0.11)	(0.09)	(0.09)
R^2	0.23	0.50	0.55	0.52	0.57	0.54	0.51
Main controls	No	Yes	Yes	Yes	Yes	Yes	Yes
Add. controls	No	No	Climate	Rugged	Geo. var.	Deep hist.	Col. hist.
Countries	76	76	76	76	74	75	76

Notes: The dependent variable is generalized trust from the World Value Survey in the 1981-2008 time period (taken from Ashraf and Galor 2013). This is the fraction of people answering "most people can be trusted" (as opposed to "can't be too careful") when asked the standard trust question. Each column presents two OLS regressions with the same set of controls. In the upper panel the main explanatory variable is ethnic segregation, and in the lower panel these are ethno-spatial alignment, generalized ethnic fractionalization and spatial dispersion. These indices are all explained in Sections 2 and 3. Main controls are absolute latitude and continental dummy variables. Additional controls are temperature and precipitation in column (3); terrain ruggedness and its interaction with a dummy variable for Africa in column (4); averages and standard deviations of elevation and land suitability for agriculture in column (5); migratory distance from Addis Ababa, its square term, and the time elapsed since the agricultural transition in column (6); and dummy variables for former British/French/Spanish/other colonies in column (7). Online Appendix D contains more information on dependent and control variables. Robust standard errors. ***, **, * indicate significance at the 1, 5 and 10%-level, respectively.

Table 5: Ethnic geography, trust, rule of law, and income

	(1)	(2)	(3)	(4)
Dependent var.	Rule of law	Rule of law	Income	Income
Alignment	0.48**	0.21	0.38***	0.26
	(0.23)	(0.26)	(0.14)	(0.17)
Fractionalization	-0.36	-0.44	-0.40	-0.44
	(0.74)	(0.73)	(0.60)	(0.61)
Dispersion	0.34	-0.38	0.61	0.28
	(0.67)	(0.74)	(0.53)	(0.55)
Trust		2.71***		1.24*
		(0.85)		(0.68)
Main controls	Yes	Yes	Yes	Yes
Countries	76	76	76	76
R^2	0.47	0.53	0.66	0.67

Notes: OLS regressions. The dependent variable is the rule of law in 2010 from the World Bank Governance Indicators in columns (1) and (2), and expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0 in columns (3) and (4). The sample is restricted to countries for which generalized trust from the World Value Survey in the 1981-2008 time period is available. Main controls are absolute latitude and continental dummy variables. Online Appendix D contains more information on the dependent and control variables, and on generalized trust. Ethno-spatial alignment, generalized ethnic fractionalization and spatial dispersion are explained in Sections 2 and 3. Robust standard errors. ***, **, * indicate significance at the 1, 5 and 10%-level, respectively.

ONLINE APPENDIX

Sections:

A. Shortcomings of a-spatial segregation indices

Border dependence: Border dependence occurs due to the (implicit) assumption of a-spatial segregation measures that the distance between two individuals is zero when they are located in the same subnational unit, and one when located in different subnational units. As a result, the index value of a-spatial segregation measures heavily depends on the type of subnational units used when computing the index values. For example, it may depend on whether provinces or districts are used when relying on administrative units, or on the size of cells or circles when researchers construct "geometric" subnational units.

Figure A.1 illustrates the problem of border dependence: The spatial distribution of individuals from different ethnic groups is identical in the left and the right diagram, however there are four administrative units in the left diagram, but only two in the right diagram. Any a-spatial segregation measure would classify the society in the left diagram as highly segregated, because the population is ethnically homogenous in each administrative unit, but as non-segregated in the right diagram, where the two groups' population shares are the same in each administrative unit.

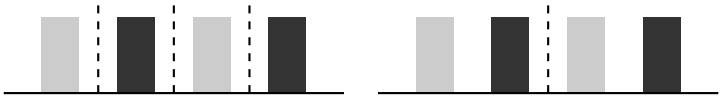


Figure A.1: Illustration of border dependence
Notes: The two diagrams depict two distributions of ethnic groups in space. Each tone of gray indicates a different ethnic group, and ethnic distances between groups are given by differences in tones of gray. Spatial locations are on the horizontal axis, which also measures spatial distances, while the vertical axis measures the population mass at each location. The dotted vertical lines indicate administrative boundaries.

To illustrate that border dependence is a real concern, we use data from the Nigeria Development and Health Survey (DHS) 2013. This survey of more than 38,000 mothers of childbearing age provides information on, among other things, these mothers' self-reported ethnicity and the geo-coordinates of cluster locations. We use these geo-coordinates to assign each cluster (and thereby each mother) to a state and a local government area (LGA). The DHS further groups Nigeria into 6 regions that play no administrative or political role. Table A.1, column (1) shows that, according to the Nigeria DHS 2013, there are 307 different ethnic groups and the population share of the largest group (Hausa) is 24 percent. We then collapse the data at the level of DHS regions, states and LGAs. For each of these levels, we report in columns (2)–(4) the average number of groups, the average population share of the largest group, and the number of subnational units on which these two summary statistics are based. We see an inverse relation between the level of spatial disaggregation and the average ethnic heterogeneity within subnational units. As a result, any a-spatial segregation index would provide markedly different index values for Nigeria in 2013, depending on whether DHS regions, states or LGAs were used

as the relevant subnational units. The index value would be highest for LGAs and lowest for DHS regions.[1]

Table A.1: Ethnic heterogeneity in subnational units in Nigeria

	(1)	(2)	(3)	(4)
	Country	DHS regions	States	LGAs
Number of units	1	6	38	501
Average number of groups	307	98.17	28.29	5.08
Average share of largest group	0.24	0.53	0.59	0.80

Checkerboard problem: The checkerboard problem refers to the impossibility of a-spatial segregation measures to account for the arrangements or relative positions of subnational units in space. It occurs due to the (implicit) assumption of a-spatial segregation measures that the distance between two individuals is one when they are located in different subnational units, no matter how far apart these units are.

Figure A.2 illustrates the problem: A-spatial segregation measures classify the societies in the left and the right diagram as equally segregated, even though the society represented in the left diagram appears more segregated than the one in the right diagram.

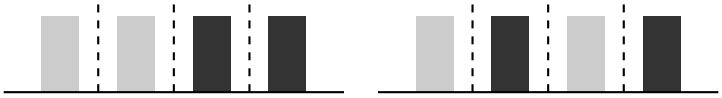


Figure A.2: Illustration of the checkerboard problem
Notes: The two diagrams of each sub-figure depict two distributions of ethnic groups in space. Each tone of gray indicates a different ethnic group, and ethnic distances between groups are given by differences in tones of gray. Spatial locations are on the horizontal axis, which also measures spatial distances, while the vertical axis measures the population mass at each location. The dotted vertical lines indicate administrative boundaries.

[1] Alesina and Zhursavskaysa (2011) use DHS to compute ethnic segregation in various countries, including Nigeria, where they take DHS regions as the relevant subnational units.

B. Geometric interpretation of our segregation index

To illustrate the general properties of our segregation index and its various components, we now provide a geometric interpretation. Suppose the population is finite, where $P := \{1, \ldots, m\}$ is the set of individuals and $m \geq 3$. For each pair of individuals $i, j \in P$, denote by $\lambda_{i,j}$ and $\gamma^{i,j}$ the spatial and ethnic distance between them. Let

$$\Lambda := (\lambda_{1,1}, \ldots, \lambda_{m,m}) \text{ and } \Gamma := (\gamma^{1,1}, \ldots, \gamma^{m,m})$$

be the vectors of spatial and ethnic distances between all unordered pairs of individuals. Then, equation (2) can be written as $S(\mu, \lambda, \gamma) = \frac{4}{m^2}\Lambda \cdot \Gamma$, and by definition of inner product our segregation index can be decomposed into

$$S(\mu, \lambda, \gamma) = \frac{4}{m^2}||\Lambda||_2||\Gamma||_2 \cos[\theta_{\Lambda,\Gamma}], \tag{B.1}$$

where

$$||\Lambda||_2 := \left(\frac{1}{2}\sum_{(i,j)\in P^2}(\lambda_{i,j})^2\right)^{1/2} \text{ and } ||\Gamma||_2 := \left(\frac{1}{2}\sum_{(i,j)\in P^2}(\gamma^{i,j})^2\right)^{1/2}$$

are the Euclidean norms of the two vectors Λ and Γ, and $\theta_{\Lambda,\Gamma}$ is the angle between them.

Since $\cos[0] = 1$, our segregation index is maximized when the two vectors point in the same direction ($\theta_{\Lambda,\Gamma} = 0$), which means that Λ and Γ are linearly dependent, i.e., there is some $k > 0$ such that $\lambda_{i,j} = k\gamma^{i,j}$ for all $i, j \in P$. In this sense, S can be interpreted as a geometric projection. To see an example, consider the two joint distributions in Figure 1(c). Clearly, by S the left distribution is more segregated than the right, as Λ and Γ are co-directional in the left but not in the right distribution, everything else equal. This is in line with our intuition in the Introduction. Another relevant feature of our index is that any increase in the mean of the two vectors, or in their Euclidean norms, also leads to higher segregation. For example, in Figure 1(b) the distribution on the left is more segregated than that on the right as the mean ethnic distance (and the Euclidean norm $||\Gamma||_2$) is higher, everything else being equal. Moreover, any mean-preserving spread of the elements of each of the two vectors Λ and Γ that keeps their alignment constant leads to higher segregation. This can be easily shown by the convexity of the (square of the) Euclidean norms $||\Lambda||_2$ and $||\Gamma||_2$ in the spatial distance and in the ethnic distance between each pair of individuals, respectively.

This geometric interpretation of our segregation index resembles the decomposition in Proposition 1: The generalized social fractionalization index F and the spatial dispersion index D are related to the Euclidean norms of the two respective vectors, and the alignment index A is therefore related to the cosign of the angle between the vectors of ethnic and spatial distances. In particular, it follows from Proposition 1 and Equation (B.1)

that $A(\mu, \lambda, \gamma) \approx \cos[\theta_{\Lambda,\Gamma}]$ if $F(\mu, \gamma)D(\mu, \lambda) \approx 4||\Lambda||_2||\Gamma||_2/m^2$. To see this, it is useful to write

$$F(\mu, \gamma)D(\mu, \lambda) = \left(\frac{2}{m^2}\right)^2 \left(\sum_{(i,j)\in P^2} \gamma^{i,j}\right) \left(\sum_{(i,j)\in P^2} \lambda_{i,j}\right),$$

$$4||\Lambda||_2||\Gamma||_2/m^2 = \left(\frac{2}{m^2}\right) \left(\sum_{(i,j)\in P^2} (\gamma^{i,j})^2\right)^{1/2} \left(\sum_{(i,j)\in P^2} (\lambda_{i,j})^2\right)^{1/2}.$$

Note the proportionality across the two equations for each of the three elements that respectively correspond to population size (m), social distances ($\gamma^{i,j}$) and spatial distances ($\lambda_{i,j}$). Although different, $F(\mu, \gamma)D(\mu, \lambda)$ and $4||\Lambda||_2||\Gamma||_2/m^2$ are closely related, which means that $A(\mu, \lambda, \gamma)$ and the cosign of $\theta_{\Lambda,\Gamma}$ are closely related as well.[2] This relation further justifies our interpretation of A as alignment or co-directionality of spatial and ethnic distances. For the purpose of empirical applications, A has the advantage – compared to the consign of $\theta_{\Lambda,\Gamma}$ – that its computation does not require data at the individual level. Similarly, F and D are related to the Euclidean norms $||\Gamma||_2$ and $||\Lambda||_2$ and have the same empirical advantage compared to them.

[2]One can show that $A(\mu, \lambda, \gamma)$ is a positively-biased proxy of $\cos[\theta_{\Lambda,\Gamma}]$. This follows from $4||\Lambda||_2||\Gamma||_2/m^2 \geq S(\mu, \lambda, \gamma)$ for all $\mu \in \mathcal{M}$ (as $\cos[\theta_{\Lambda,\Gamma}] \in [0, 1]$) and $F(\mu, \gamma)D(\mu, \lambda) = S(\overline{\mu}, \lambda, \gamma)$, which jointly imply $4||\Lambda||_2||\Gamma||_2/m^2 \geq F(\mu, \gamma)D(\mu, \lambda)$. Hence, $A(\mu, \lambda, \gamma) \geq \cos[\theta_{\Lambda,\Gamma}]$.

C. List of countries

We provide our four indices of historical ethnic geography (i.e., ethnic segregation, generalized ethnic fractionalization, spatial dispersion, and ethno-spatial alignment) for the following 159 countries with a current population of more than 250,000 and a land surface area of more than 5,000 km^2: Afghanistan, Albania, Algeria, Angola, Argentina, Armenia, Australia, Azerbaijan, Bangladesh, Belarus, Belgium, Belize, Benin, Bhutan, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Brunei, Bulgaria, Burkina Faso, Burundi, Cambodia, Cameroon, Canada, Central African Republic, Chad, Chile, China, Colombia, Congo, Costa Rica, Cote d'Ivoire, Croatia, Cuba, Cyprus, Czech Republic, Democratic Republic of the Congo, Denmark, Djibouti, Dominican Republic, East Timor, Ecuador, Egypt, El Salvador, Equatorial Guinea, Eritrea, Estonia, Ethiopia, Finland, France, Gabon, Gambia, Georgia, Germany, Ghana, Greece, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, Honduras, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Kuwait, Kyrgyzstan, Laos, Latvia, Lebanon, Lesotho, Liberia, Libya, Lithuania, Macedonia, Madagascar, Malawi, Malaysia, Mali, Mauritania, Mexico, Moldova, Mongolia, Montenegro, Morocco, Mozambique, Myanmar, Namibia, Nepal, Netherlands, New Zealand, Nicaragua, Niger, Nigeria, North Korea, Norway, Oman, Palestine, Pakistan, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Qatar, Romania, Russian Federation, Rwanda, Saudi Arabia, Senegal, Sierra Leone, Slovakia, Slovenia, Somalia, South, Africa, South Korea, South Sudan, Spain, Sri Lanka, Sudan, Suriname, Swaziland, Sweden, Switzerland, Syria, Taiwan, Tajikistan, Tanzania, Thailand, Togo, Tunisia, Turkey, Turkmenistan, Uganda, Ukraine, United Arab Emirates, United Kingdom, United States, Uruguay, Uzbekistan, Venezuela, Viet Nam, Yemen, Zambia, Zimbabwe.

D. Definitions of dependent and control variables

D.1. Dependent variables

D.1.1. Main dependent variables

Rule of law: This is one of six World Bank Governance Indicators (also called Worldwide Governance Indicators) for 2010. These indicators are based on several hundred individual variables from many different organizations measuring perceptions of governance. These individual measures of governance are assigned to categories capturing key dimensions of governance. An unobserved component model is used to construct the six aggregate governance indicators. They are normally distributed with a mean of zero and a standard deviation of one each year of measurement. The rule of law indicator includes several indicators that measure the extent to which agents have confidence in and abide by the rules of society. These include perceptions of the incidence of crime, the effectiveness and predictability of the judiciary, and the enforceability of contracts. This indicator thus measures the success of a society in developing an environment in which fair and predictable rules form the basis for economic and social interactions and the extent to which property rights are protected.

Income (PWT): Logarithm of expenditure-side real GDP per capita in 2010 at chained purchasing power parities (in 2011 US dollars) by Penn World Table, version 9.

Trust: Measure of generalized trust based on World Values Surveys conducted from 1981-2008. It is calculated as the fraction of total respondents who responded with "most people can be trusted" (as opposed to "can't be too careful") when asked: "Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?" Variable taken from Ashraf and Galor (2013).

D.1.2. Additional dependent variables used in Online Appendix E

Control of corruption: This is one of six World Bank Governance Indicators for 2010. It measures perceptions of corruption, including the frequency of bribe payments in the business environment and the extent of political corruption.

Government effectiveness: This is one of six World Bank Governance Indicators for 2010. It measures public service provision, the quality of the bureaucracy, the competence of civil servants, and the independence of the civil service from political pressures.

Political stability: This is one of six World Bank Governance Indicators for 2010. It measures perceptions of the likelihood that the government in power will be destabilized

or overthrown by possibly unconstitutional and/or violent means.

Regulatory quality: This is one of six World Bank Governance Indicators for 2010. It measures the incidence of market-unfriendly policies and perceptions of the burdens imposed by excessive regulation in areas such as foreign trade and business development.

Voice and accountability: This is one of six World Bank Governance Indicators for 2010. It measures various aspects of the political process, civil liberties and political rights to indicate the extent to which citizens of a country are able to participate in the selection of governments.

Quality of government: This indicator from the International Country Risk Guide (ICRG) corresponds to the mean of three ICRG variables in 2010: Corruption, law and order, and bureaucratic quality.

Corruption perception index: This index from Transparency International focuses on perceptions of corruption in the public sector in 2010 and includes both administrative and political corruption. We have rescaled it so that it ranges between zero and one, with higher values implying less corruption.

Income (WDI): Logarithm of GDP per capita in 2010 based on purchasing power parity (in constant 2011 international dollars) from the World Development Indicators.

D.1.3. Summary statistics

Table D.1: Summary statistics for our dependent variables

	Observations	Mean	Std. Dev.	Min.	Max.
Rule of law	155	-0.212	0.995	-2.448	1.977
Income (PWT, in logs)	146	9.032	1.243	6.341	11.708
Trust	76	0.280	0.140	0.049	0.664
Control of corruption	155	-0.186	1.000	-1.739	2.414
Government effectiveness	155	-0.135	0.988	-2.239	2.245
Political stability	155	0.249	0.381	0.000	1.393
Regulatory quality	155	-0.111	0.994	-2.446	1.888
Voice and accountability	155	-0.239	1.007	-2.193	1.637
Quality of government	130	0.523	0.198	0.083	1.000
Corruption perception index	152	0.386	0.206	0.110	0.930
Income (WDI, in logs)	149	9.035	1.255	6.391	11.157

D.2. Control variables

Absolute latitude: The absolute value of the latitude of a country's approximate centroid, as reported by the CIA's World Factbook, taken from Ashraf and Galor (2013).

Temperature: The intertemporal average monthly temperature of a country in degrees Celsius per month over the 1961–1990 time period, calculated using geospatial average monthly temperature data, taken from Ashraf and Galor (2013).

Precipitation: The intertemporal average monthly precipitation of a country in mm per month over the 1961–1990 time, calculated using geospatial average monthly precipitation data, taken from Ashraf and Galor (2013).

Terrain roughness: Terrain Ruggedness Index by Nunn and Puga (2012), which quantifies average local topographic heterogeneity by measuring elevation differences for grid points within 30 arc-seconds.

Average and standard deviation of elevation: Variables based on geospatial elevation data, taken from Michalopoulos (2012).

Average and standard deviation of land suitability: Variables based on a geospatial index of the suitability of land for agriculture based on ecological indicators of climate and soil suitability for cultivation, taken from Michalopoulos (2012).

Migratory distance from Addis Ababa: The great circle distance from Addis Ababa (Ethiopia) to the country's modern capital city along a land-restricted path forced through one or more of five intercontinental waypoints (Cairo, Istanbul, Phnom Penh, Anadyr, and Prince Rupert), taken from Ashraf and Galor (2013).

Time elapsed since the agricultural transition: The number of years elapsed up to the year 2000 CE since the majority of the population residing within a country's modern national borders began practicing sedentary agriculture as the primary mode of subsistence, taken from Ashraf and Galor (2013).

Former colonizer: A variable indicating whether a country is a former British colony, a former French colony, a former Spanish colony, the former colony of another Western colonizer, or not a former Western colony. It is based on the classification of Western overseas colonies in the Authoritarian Regime Dataset.

E. Robustness of cross-country regressions

Table E.1: Alternative measures of the quality of government and incomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent var.	CC (WBGI)	GE (WBGI)	PS (WBGI)	RQ (WBGI)	V&A (WBGI)	QoG (ICRG)	CPI (TI)	Income (WDI)
Alignment	0.38**	0.34**	0.14**	0.32**	0.27*	0.08**	0.07**	0.48***
	(0.16)	(0.16)	(0.06)	(0.14)	(0.15)	(0.03)	(0.04)	(0.16)
Fractionalization	-0.32	-0.19	-0.13	-0.28	0.44	0.01	-0.09	-1.07**
	(0.36)	(0.33)	(0.13)	(0.34)	(0.33)	(0.08)	(0.07)	(0.43)
Dispersion	-0.17	0.27	-0.14	0.05	0.04	0.04	-0.01	0.71
	(0.47)	(0.44)	(0.18)	(0.45)	(0.48)	(0.10)	(0.10)	(0.55)
Main controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.41	0.45	0.37	0.44	0.51	0.48	0.43	0.54
Countries	155	155	155	155	155	130	152	149

Notes: OLS regressions. Dependent variables are control of corruption, government effectiveness, political stability, regulatory quality, and voice and accountability by the World Bank Governance Indicators in columns (1)–(5); quality of government by ICRG in column (6); the corruption perception index by Transparency International in column (7), and the log of real GDP per capita from the World Development Indicators in column (8). All dependent variables refer to 2010. Main controls are absolute latitude and continental dummy variables. Online Appendix D contains more information on dependent and control variables. Alignment, fractionalization and dispersion are explained in Sections 2 and 3. Robust standard errors. ***, **, * indicate significance at the 1, 5 and 10%-level, respectively.

Table E.2: Alternative computations of our indices of ethnic geography

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent var.	Rule of law (WBGI)			Income (PWT)			Trust (WVS)		
Alignment	0.68***	0.78*	0.42***	0.77***	1.09***	0.53***	0.18***	0.16**	0.10***
	(0.22)	(0.41)	(0.14)	(0.20)	(0.33)	(0.15)	(0.05)	(0.07)	(0.03)
Fractionalization	-0.35	-0.36	-0.45	-0.68	-0.62	-0.68	0.13*	0.01	0.03
	(0.34)	(0.37)	(0.34)	(0.43)	(0.43)	(0.44)	(0.07)	(0.07)	(0.08)
Dispersion	0.37	-0.09	0.32	0.57	0.09	0.38	0.24***	0.38***	0.24***
	(0.46)	(0.56)	(0.46)	(0.52)	(0.68)	(0.49)	(0.09)	(0.11)	(0.08)
Main controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Alternative ethnic distances	Yes	No	No	Yes	No	No	Yes	No	No
Alternative spatial distances	No	Yes	No	No	Yes	No	No	Yes	No
Alternative year	No	No	Yes	No	No	Yes	No	No	Yes
R^2	0.46	0.45	0.46	0.58	0.57	0.58	0.52	0.49	0.50
Countries	155	155	155	146	146	146	76	76	76

Notes: OLS regressions. Dependent variables are the rule of law in 2010 by the World Bank Governance Indicators in columns (1)–(3), the log of expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0 in columns (4)–(6), and generalized trust from the World Value Survey in the 1981-2008 time period (Ashraf and Galor 2013) in columns (7)–(9). Main controls are absolute latitude and continental dummy variables. Appendix D contains more information on dependent and control variables. Alignment, fractionalization and dispersion are explained in Sections 2 and 3. However, we compute these indices slightly differently than reported in Section 3. We use ethnolinguistic distances calculated using the formula in Fearon (2003) in columns (1), (4) and (7); spatial distances as the square root of the geodesic distance in columns (2), (5) and (8); and the HYDE population map for 1950 in columns (3), (6) and (9). Robust standard errors. ***, **, * indicate significance at the 1, 5 and 10%-level, respectively.

Table E.3: Ethnic geography and the rule of law in restricted samples

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent var.	Rule of law (WBGI)							
Alignment	0.43***	0.47***	0.34*	0.40**	0.44***	0.40***	0.46***	0.46***
	(0.15)	(0.16)	(0.18)	(0.15)	(0.14)	(0.14)	(0.15)	(0.12)
Fractionalization	-0.68	-0.61	0.09	-0.44	-0.24	-0.25	-0.32	-0.22
	(0.44)	(0.42)	(0.38)	(0.33)	(0.33)	(0.32)	(0.39)	(0.29)
Dispersion	0.52	-0.21	-0.11	0.34	0.13	-0.20	0.30	-0.00
	(0.52)	(0.47)	(0.56)	(0.46)	(0.47)	(0.43)	(0.46)	(0.41)
Main controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Omitted observations	Africa	Americas	Asia	Europe	Oceania	Neo-Europe	$F = 0$	Outliers
R^2	0.41	0.49	0.59	0.25	0.44	0.45	0.47	0.51
Countries	107	129	112	120	152	151	140	147

Notes: OLS regressions. Dependent variable is the rule of law in 2010 by the World Bank Governance Indicators. We omit countries from one continent in each of the columns (1)–(5), the settler colonies Australia, Canada, New Zealand and United States in column (6), the ethnically homogeneous countries in column (7), and outliers as identified by Cook's distance (with a threshold of 4/155) in column (8). Main controls are absolute latitude and continental dummy variables. Online Appendix D contains more information on dependent and control variables. Alignment, fractionalization and dispersion are explained in Sections 2 and 3. Robust standard errors. ***, **, * indicate significance at the 1, 5 and 10%-level, respectively.

Table E.4: Ethnic geography and income in restricted samples

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent var.	Income (PWT)							
Alignment	0.42***	0.59***	0.49*	0.57***	0.54***	0.53***	0.56***	0.51***
	(0.16)	(0.14)	(0.25)	(0.19)	(0.16)	(0.16)	(0.15)	(0.12)
Fractionalization	-0.88*	-1.05*	0.09	-0.72	-0.66	-0.65	-0.68	-0.66
	(0.51)	(0.54)	(0.47)	(0.47)	(0.43)	(0.43)	(0.45)	(0.40)
Dispersion	0.54	-0.08	0.46	0.35	0.34	0.16	0.49	0.77
	(0.55)	(0.62)	(0.69)	(0.56)	(0.55)	(0.57)	(0.53)	(0.49)
Main controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Omitted observations	Africa	Americas	Asia	Europe	Oceania	Neo-Europe	$F = 0$	Outliers
R^2	0.32	0.61	0.69	0.47	0.57	0.56	0.60	0.65
Countries	101	122	106	111	144	142	133	139

Notes: OLS regressions. Dependent variable is the log of expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0. We omit countries from one continent in each of the columns (1)–(5), the settler colonies Australia, Canada, New Zealand and United States in column (6), the ethnically homogeneous countries in column (7), and outliers as identified by Cook's distance (with a threshold of 4/146) in column (8). Main controls are absolute latitude and continental dummy variables. Online Appendix D contains more information on dependent and control variables. Alignment, fractionalization and dispersion are explained in Sections 2 and 3. Robust standard errors. ***, **, * indicate significance at the 1, 5 and 10%-level, respectively.

Table E.5: Ethnic geography and trust in restricted samples

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent var.	Trust (WVS)							
Alignment	0.09***	0.10**	0.11***	0.04	0.11***	0.11***	0.11***	0.09**
	(0.04)	(0.04)	(0.03)	(0.05)	(0.03)	(0.04)	(0.04)	(0.04)
Fractionalization	0.03	0.01	0.05	-0.01	0.00	0.00	0.08	0.01
	(0.10)	(0.09)	(0.07)	(0.07)	(0.08)	(0.08)	(0.09)	(0.08)
Dispersion	0.28***	0.32***	-0.02	0.30***	0.30***	0.28***	0.29***	0.28***
	(0.09)	(0.10)	(0.10)	(0.10)	(0.09)	(0.10)	(0.09)	(0.09)
Main controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Omitted observations	Africa	Americas	Asia	Europe	Oceania	Neo-Europe	$F = 0$	Outliers
R^2	0.42	0.47	0.66	0.59	0.50	0.48	0.49	0.50
Countries	66	65	58	41	74	72	69	71

Notes: OLS regressions. Dependent variable is generalized trust from the World Value Survey in the 1981-2008 time period (Ashraf and Galor 2013). We omit countries from one continent in each of the columns (1)–(5), the settler colonies Australia, Canada, New Zealand and United States in column (6), the ethnically homogeneous countries in column (7), and outliers as identified by Cook's distance (with a threshold of 4/76) in column (8). Main controls are absolute latitude and continental dummy variables. Online Appendix D contains more information on dependent and control variables. Alignment, fractionalization and dispersion are explained in Sections 2 and 3. Robust standard errors. ***, **, * indicate significance at the 1, 5 and 10%-level, respectively.

Table E.6: Weight least squares (WLS)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent var.	Rule of law (WBGI)		Income (PWT)		Trust (WVS)	
Alignment	0.45***	0.43***	0.52***	0.49***	0.10***	0.10***
	(0.15)	(0.16)	(0.15)	(0.15)	(0.03)	(0.03)
Fractionalization	-0.41	-0.41	-0.63	-0.66	0.03	0.03
	(0.34)	(0.35)	(0.43)	(0.43)	(0.08)	(0.08)
Dispersion	0.28	0.31	0.35	0.40	0.27***	0.26***
	(0.44)	(0.45)	(0.52)	(0.52)	(0.09)	(0.09)
Main controls	Yes	Yes	Yes	Yes	Yes	Yes
Weights	Pop.	Area	Pop.	Area	Pop.	Area
R^2	0.46	0.46	0.59	0.59	0.50	0.51
Countries	155	155	146	146	76	76

Notes: WLS regressions. Weights are the log of population size in odd columns and the log of surface area in even columns, both from the World Development Indicators. Dependent variables are the rule of law in 2010 by the World Bank Governance Indicators in columns (1) and (2), expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0 in columns (3) and (4), and generalized trust from the World Value Survey in the 1981-2008 time period (Ashraf and Galor 2013) in columns (5) and (6). Main controls are absolute latitude and continental dummy variables. Online Appendix D contains more information on dependent and control variables. Alignment, fractionalization and dispersion are explained in Sections 2 and 3. Robust standard errors. ***, **, * indicate significance at the 1, 5 and 10%-level, respectively.

Table E.7: Poisson pseudo-maximum likelihood (PPML)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent var.	QoG (ICRG)		Income (PWT)		Trust (WVS)	
Alignment	0.16	0.23**	0.10***	0.10***	0.46**	0.39**
	(0.12)	(0.09)	(0.03)	(0.03)	(0.18)	(0.16)
Fractionalization	-0.02	-0.01	-0.01*	-0.01	0.01	-0.01
	(0.02)	(0.02)	(0.01)	(0.00)	(0.05)	(0.02)
Dispersion	0.01	0.01	0.01	-0.00	0.21***	0.21***
	(0.04)	(0.04)	(0.01)	(0.01)	(0.07)	(0.06)
Main controls	Yes	Yes	Yes	Yes	Yes	Yes
Windsorizing F	No	Yes	No	Yes	No	Yes
R^2	0.46	0.45	0.58	0.56	0.42	0.44
N	118	130	133	146	69	76

Notes: PPML regressions. Dependent variables are the quality of government by ICRG in columns (1) and (2), expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0 in columns (3) and (4), and generalized trust from the World Value Survey in the 1981-2008 time period (Ashraf and Galor 2013) in columns (5) and (6). We use the quality of government by ICRG rather than the rule of law in 2010 by the World Bank Governance Indicators as in most other tables, because PPML requires non-negative dependent variables. This change of the dependent variable leads to a drop in the sample size. Main controls are the log of absolute latitude and continental dummy variables. Alignment, fractionalization and dispersion all enter in logs as well. We thus lose all countries in which fractionalization is zero in odd columns. We add a small constant (0.001) to fractionalization before taking logs in even columns, which allows keeping these countries in the sample. Appendix D contains more information on dependent and control variables. Alignment, fractionalization and dispersion are explained in Sections 2 and 3. Robust standard errors. ***, **, * indicate significance at the 1, 5 and 10%-level, respectively.

Table E.8: Allowing for non-linear effects of fractionalization and dispersion

	(1)	(2)	(3)
Dependent var.	Rule of law (WBGI)	Income (PWT)	Trust (WVS)
Alignment	0.42***	0.55***	0.12***
	(0.15)	(0.16)	(0.04)
Fractionalization	0.04	0.06	-0.08
	(1.05)	(1.21)	(0.22)
Fractionalization^2	-0.30	-1.54	-0.15
	(1.65)	(1.69)	(0.42)
Dispersion	-2.35	-0.19	-0.00
	(1.47)	(1.70)	(0.30)
Dispersion^2	4.87**	0.04	0.03
	(2.44)	(2.26)	(0.41)
Fractionalization	-0.78	1.41	0.90
$\times$ Dispersion	(2.14)	(2.75)	(0.60)
Main controls	Yes	Yes	Yes
R^2	0.48	0.58	0.52
N	155	146	76

Notes: OLS regressions. Dependent variables are the rule of law in 2010 by the World Bank Governance Indicators in column (1), expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0 in column (2), and generalized trust from the World Value Survey in the 1981-2008 time period (Ashraf and Galor 2013) in column (3). The addition of square and interaction terms of fractionalization and dispersion allows showing that the coefficient on alignment is not driven by some non-linearity in the effects of fractionalization or dispersion. Main controls are absolute latitude and continental dummy variables. Appendix D contains more information on dependent and control variables. Alignment, fractionalization and dispersion are explained in Sections 2 and 3. Robust standard errors. ***, **, * indicate significance at the 1, 5 and 10%-level, respectively.