Journal of Public Economics 200 (2021) 104446

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Ethnic geography: Measurement and evidence

Roland Hodler [a,b,c,*], Michele Valsecchi [d], Alberto Vesperoni [e]

[a] Department of Economics, University of St.Gallen, Switzerland
[b] CEPR, London, United Kingdom
[c] CESifo, Munich, Germany
[d] New Economic School, Moscow, Russia
[e] Department of Political Economy, King's College London, United Kingdom

ARTICLE INFO

Article history:
Received 13 March 2020
Revised 22 April 2021
Accepted 13 May 2021
Available online 8 June 2021

JEL classification:
C43
D63
O10
Z13

Keywords:
Measurement theory
Ethnic diversity
Ethnic geography
Segregation
Fractionalization
Comparative development

ABSTRACT

We know little about how ethnic geography, i.e., the distribution of ethnic groups across space, shapes comparative economic, political and social development. To make progress and to harness the growing availability of spatially explicit data, we need indices summarizing key aspects of ethnic geography. We develop and axiomatize a novel index of ethnic segregation that takes both ethnic and spatial distances between individuals into account. We can decompose this index into indices of generalized ethnic fractionalization, spatial dispersion, and the alignment of spatial and ethnic distances. For our application, we compute different country-level versions of the segregation index and its components based on either ethnographic maps or geo-referenced survey data. Reassuringly, the different versions of the segregation index are highly correlated. We explore the relation of our indices to (i) existing measures of ethnic segregation and diversity; (ii) climatic and geographical factors; and (iii) the quality of government, economic development, and trust.

1. Introduction

There is a vast literature on how ethnic diversity affects economic, political and social outcomes. This literature provides evidence for negative effects of country-level ethnic diversity on, e.g., public good provision, redistribution, the quality of government, peace, and economic development in general. In these studies, ethnic diversity is typically measured by the standard index of ethnic fractionalization (e.g., Easterly and Levine, 1997, Alesina et al., 2003, Desmet et al., 2012) or indices of ethnic polarization (e.g., Esteban and Ray, 1994, Montalvo and Reynal-Querol, 2005).[1] These indices are based on the different ethnic groups' country-wide population shares. By definition, they ignore ethnic geography, i.e., the distribution of ethnic groups across space. Ethnic geography, however, may well shape comparative economic, political and social development, but we do not know how. To make progress and to harness the growing availability of spatially explicit data, we need indices summarizing key aspects of ethnic geography.

We contribute to the literature on ethnic diversity by proposing a set of indices that capture key aspects of ethnic geography. Our main contribution is a methodological one: we derive a new segregation index that is based on both spatial and ethnic distances between pairs of individuals instead of population shares only. There is indeed evidence that both these distances matter for economic, political and social outcomes (see, e.g., White, 1983, for spatial distances and Desmet et al., 2009, for ethnolinguistic distances). For conceptual clarity we provide an axiomatic characterization. Starting from a general class of indices that are expressions of the relation between a randomly selected pair of individuals, we uniquely characterize the proposed segregation index via a set of axioms that are intuitive properties of a segregation measure.

* Corresponding author at: Department of Economics, University of St.Gallen, Switzerland.
E-mail addresses: roland.hodler@unisg.ch (R. Hodler), mvalsecchi@nes.ru (M. Valsecchi), alberto.vesperoni@kcl.ac.uk (A. Vesperoni).

[1] Alesina and La Ferrara (2005) review the early literature on ethnic diversity and economic performance.

https://doi.org/10.1016/j.jpubeco.2021.104446
0047-2727/

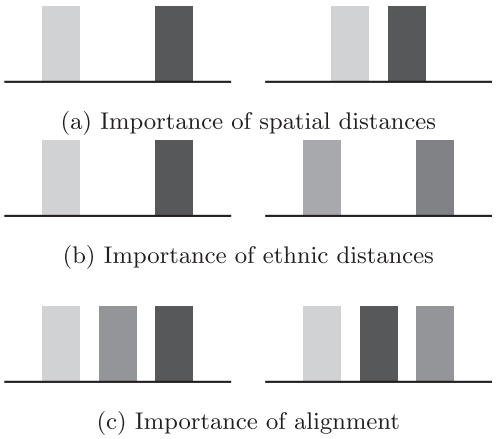

Fig. 1. Illustration of our segregation measure.
Notes: The two diagrams of each sub-figure depict two distributions of ethnic groups in space. Each tone of gray indicates a different ethnic group, and ethnic distances between groups are given by differences in tones of gray. Spatial locations are on the horizontal axis, which also measures spatial distances, while the vertical axis measures the population mass at each location.

The proposed segregation index has three prominent features. The first is that, if we abstract from the spatial dimension (by assuming the same spatial distance between an individual and anyone else, including oneself), our segregation index coincides with Greenberg's (1956) generalized fractionalization index. If ethnic distances between individuals are further assumed to be binary, this latter index becomes equivalent to the standard fractionalization index used by Alesina et al. (2003) and many others. Hence, our segregation index can be seen as a straightforward spatial extension of the most commonly used measure of ethnic diversity. Symmetrically, our segregation index reduces to a simple index of spatial dispersion if we abstract from the ethnic dimension.

The second (and closely related) prominent feature is that our segregation index can be decomposed into three (sub-) indices: the generalized fractionalization index, the index of spatial dispersion, and a measure of the alignment of spatial and ethnic distances between pairs of individuals (hereinafter simply ethno-spatial alignment). Fig. 1 illustrates the role of these three components. In all parts, the society to the left is more segregated than the one to the right, but for different reasons: in part (a), it is due to higher spatial distance; in part (b), it is due to higher ethnic distance (as indicated by more different tones of gray); and in part (c), it is due to higher ethno-spatial alignment, as spatially very distant pairs of individuals are ethnically very distant too in the society to the left but not in the society to the right.

To discuss the third prominent feature of our segregation index, we borrow the terminology from Reardon and O'Sullivan (2004). They call segregation measures "a-spatial" if they are based on population shares in administrative units, and "spatial" if they are based on spatial distances between individuals.[2] Our index is a spatial segregation measure and, therefore, avoids standard problems of a-spatial segregation measures. An important problem is border dependence, i.e., the property of a-spatial segregation measures that the index value depends on the way the government draws subnational borders and on the type of subnational units (e.g., provinces versus districts) used in its computation. Related issues, including the checkerboard problem (White 1983), follow from the disregard of patterns of spatial distances between such units.[3]

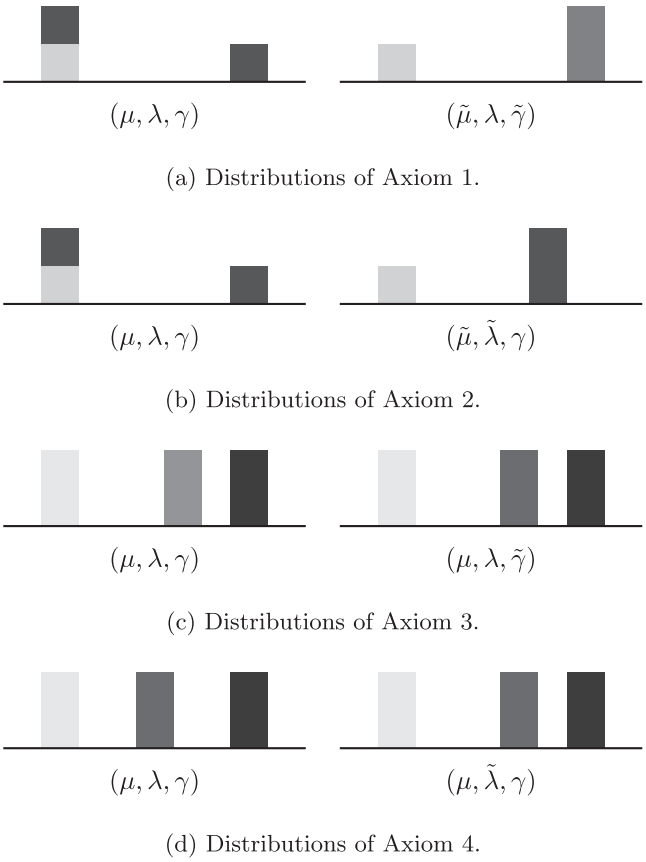

Fig. 2. Illustration of the distributions of the axiomatization.
Notes: The two diagrams of each sub-figure depict two distributions of ethnic groups in space and are to be interpreted in the same way as in Fig. 1.

The application of our indices requires data with ethnic and spatial information. But once appropriate data is available, they can be computed for different types of spatial units, including continents, countries, provinces, districts, cities, or grid cells. Hence, our indices can be used to study the relation between ethnic geography and economic, political or social development at the macro, the meso, or the micro level. Furthermore, our indices could be applied to measure and study segregation and human geography along dimensions of identity other than ethnicity. Prominent alternatives include caste and religion, but one could also consider different types of occupations, party affiliations, or cultural values.[4]

In this paper, we illustrate how our segregation index and its components can be computed at the country level, using different types of data that are readily available for many countries and contain ethnic and spatial information. On the one hand, we compute our indices based on ethnographic maps featuring the traditional homelands of ethnolinguistic groups. We use the ethnographic

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

[3] Online Appendix A illustrates border dependence and the checkerboard problem in some detail.

[4] One could also consider applications in which either the ethnic or the spatial dimension in our setup were replaced by income, wealth, or some other economic variable. Notice, however, that our axiomatic characterization treats the two dimensions symmetrically as distances in an abstract space, while more specific arguments may apply when distances represent differences in endowments. Hodler et al. (2020) propose a framework analogous to ours for the measurement of ethnic stratification based on ethnic and economic distances between pairs of individuals. Thereby, they rely on specific axioms for the economic dimension that are based on progressive/regressive transfers of wealth in the tradition of inequality measurement.

map by the World Language Mapping System (WLMS, version 19), which represents the homelands of the language groups listed in the Ethnologue (Gordon, 2005), to compute our indices for 161 countries from across the world, and Murdock's (1959) map of pre-colonial ethnicities to compute our indices for 48 African countries. On the other hand, we compute our indices for around half of these African countries based on geo-coded survey data from the Demographic and Health Surveys ICF (1986-2019) and Afrobarometer (BenYishay et al., 2017). Compared to survey data, the main advantage of ethnographic maps is that they offer better spatial coverage. A potential disadvantage is that we may attribute internal migrants (e.g., those who left their homeland to move to a large metropolitan area) to the wrong ethnic group when relying on ethnographic maps. Reassuringly, the four different versions of our indices are highly correlated despite these differences, with the correlation coefficients between the different versions of the segregation index ranging from 0.82 to 0.95.

We document how our indices relate to other prominent measures of ethnic segregation and diversity. We also show that the variation in climatic and geographical factors explain a considerable share of the variation in ethnic segregation, generalized ethnic fractionalization, and spatial dispersion, but a somewhat lower share of the variation in ethno-spatial alignment.

Finally, we study how our indices are associated to prominent measures of comparative economic, political and social development in our global Ethnologue-based sample. These relations are interesting even though the standard caveat applies that the estimated coefficients may not represent causal effects. The segregation index is negatively related to the rule of law and GDP per capita (but not generalized trust). However, these negative relations disappear once we control for climatic and geographical factors. In contrast, ethno-spatial alignment is positively related to the rule of law, GDP per capita and generalized trust even if we control for climatic and geographical factors. This latter finding suggests that, on average, countries tend to be better governed, richer and more trusting if ethnically diverse people live far apart.

Our theoretical 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; or 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.[5] To our knowledge, there is, however, no other segregation measure that presents both ethnic/social and spatial distances in the same framework.[6] Our framework is also related to prominent models of fractionalization and polarization (e.g., Rao, 1982, Esteban and Ray, 1994, Duclos et al., 2004, Bossert et al., 2011), in particular the generalized fractionalization index introduced by Greenberg (1956), as explained above.

Our application is related to other contributions that measure ethnic geography and study its effect on economic, political and social outcomes at the country level. Alesina and Zhuravskaya (2011) compute an a-spatial index of ethnic segregation and find that the quality of government is lower in more ethnically segregated countries.[7] Matuszeki and Schneider (2006) compute a measure of average subnational ethnic fractionalization and study how this measure relates to conflict. Desmet et al. (2020) study public good provision and develop a measure that captures the average exposure of an individual to members of the country's different ethnic groups with an emphasis on weighting this exposure according to the representation of these groups at the individual's location. There are two main differences between these contributions and ours. First, we focus on conceptualizing spatial segregation and introducing the novel concept of ethno-spatial alignment, while they either compute a-spatial segregation (Alesina and Zhuravskaya, 2011) or extend the fractionalization framework (Matuszeki and Schneider, 2006, Desmet et al., 2020). Second, spatial (and ethnic) distances are continuous in our approach, but binary in their contributions. We thus see our spatial segregation index as complementary to their measures, which capture alternative important aspects of ethnic geography.[8]

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 presents our applications, and Section 4 concludes. The Appendix contains the proofs of our theoretical results, and the Online Appendix additional information and further results.

2. Development of indices of ethnic geography

2.1. General model

A population $P \subset \mathbb{N}$ comprehending an arbitrarily large number of individuals is partitioned into $n \in \mathbb{N}$ ethnic or, more generally, social groups $G := \{1, \dots, n\}$ and distributed over $t \in \mathbb{N}$ locations on a territory $T := \{1, \dots, t\}$. We generally assume $t \geqslant n \geqslant 3$ so that (*i*) there is significant ethnic heterogeneity; (*ii*) there are at least as many locations as groups, so that it is possible that no individuals of different groups share the same location.

Denoting by $\mu_p^g \in [0,1]$ the share of population that corresponds to group $g \in G$ in location $p \in T$, we 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, and the space of all mass distributions $\mathcal{M}$ is the subset of $[0,1]^{t \times n}$ such that the restrictions above are satisfied.

[5] Blumenstock and Fratamico (2013) also rely on network data for providing a-spatial segregation measures.

[6] 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.

[7] Relatedly, Ejdemyr et al. (2018) and Tajima et al. (2018) use census data to compute segregation measures for subnational administrative units in Malawi and Indonesia, respectively. The latter measure a-spatial segregation while the former employ the spatial dissimilarity index, which acknowledges spatial distances but disregards ethnic distances by construction. Despite their differences, both contributions find that higher ethnic segregation leads to higher local public goods provision. Hence, their findings and ours point in the same direction of there being some advantages of ethnically diverse individuals not living close to one another.

[8] Many other contributions studying economic, political and social effects of ethnic geography rely on ethnographic maps as well, but do not choose a measurement-based approach. 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, 2017), on the effect of pre-colonial and current institutions on economic development (e.g., Michalopoulos and Papaioannou, 2013, Michalopoulos and Papaioannou, 2014), and on ethnic favoritism (e.g., De Luca et al., 2018).

For any pair of locations $p, q \in T$, let $\lambda_{p,q} \in [0,1]$ be the spatial distance between them, where we generally assume $\lambda_{p,q} = 0$ if $p = q$ and $\lambda_{p,q} = \lambda_{q,p}$. A spatial distribution is defined by the $t \times t$ matrix of spatial distances between all pairs of locations

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

and the space of all spatial distributions $\mathcal{L}$ is the subset of $[0,1]^{t \times t}$ such that the restrictions above are satisfied.

For any pair of groups $g, h \in G$, let $\gamma^{g,h} \in [0,1]$ be the ethnic distance between them, where we generally assume $\gamma^{g,h} = 0$ if $g = h$ and $\gamma^{g,h} = \gamma^{h,g}$. 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 $\mathcal{G}$ is the subset of $[0,1]^{n \times n}$ such that the restrictions above are satisfied.

Finally, a joint distribution is a triple of mass, spatial and ethnic distributions, and an index is a function $S : \left([0,1]^{t \times n}, [0,1]^{t \times t}, [0,1]^{n \times n}\right) \to \mathbb{R}_+$, 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. For each pair of individuals $i, j \in P$ that inhabit locations $p, q \in T$ and belong to groups $g, h \in G$, we quantify the relation between them by $r_{i,j} = \pi\left(\lambda_{p,q}, \gamma^{g,h}\right)$, 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 function π, one possibility is to see it as the degree of alienation (i.e., lack of common interest) between a pair of individuals, which naturally increases with their spatial and ethnic distances. The definition of an index requires to aggregate all these pairwise relations into a scalar S which describes a relevant feature of the population as a whole. Following the axiomatic foundations in Rao (1982) and Bossert et al. (2011), we consider the class of indices that are expression of the expected relation between a randomly selected pair of individuals, $S = \frac{1}{|P|} \sum_{(i,j) \in P} r_{i,j}$. Then, given our assumption $r_{i,j} = \pi\left(\lambda_{p,q}, \gamma^{g,h}\right)$, in our framework these indices take the form

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

for each joint distribution $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$ and any function π that satisfies the above restrictions. In the next section we will introduce a set of axioms that pin down a particular index (up to positive scalar multiplication) from class (1) as our segregation index.

2.2. Axiomatization of the segregation index

We now introduce a set of axioms that we see as desirable properties of a segregation measure. For simplicity of exposition, these properties are defined through 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.

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

$\mu_p^g = \mu_p^h = \mu_q^h = 1/3$ with $\lambda_{p,q} > 0$ and $\gamma^{g,h} > 0$, and let $\tilde{\mu} \in \mathcal{M}, \tilde{\gamma} \in \mathcal{G}$ and $\epsilon > 0$ satisfy

$\tilde{\mu}_p^g = \mu_p^g$ and $\tilde{\mu}_q^h = \mu_p^h + \mu_q^h$ with $\tilde{\gamma}^{g,h} = \gamma^{g,h} - \epsilon$.

Statement: We require $S(\mu, \lambda, \gamma) < S(\tilde{\mu}, \lambda, \tilde{\gamma})$ for ϵ arbitrarily small.

Let us discuss Axiom 1, whose distributions are depicted in Fig. 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 increasing ethnic homogeneity within locations and decreasing ethnic distance across groups, requiring the former to dominate the effect on segregation when the latter is arbitrarily small.

Axiom 2 is very similar to Axiom 1. As shown in Fig. 2(b), 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

$\mu_p^g = \mu_p^h = \mu_q^h = 1/3$ with $\lambda_{p,q} > 0$ and $\gamma^{g,h} > 0$, and let $\tilde{\mu} \in \mathcal{M}, \tilde{\lambda} \in \mathcal{L}$ and $\epsilon > 0$ satisfy

$\tilde{\mu}_p^g = \mu_p^g$ and $\tilde{\mu}_q^h = \mu_p^h + \mu_q^h$ with $\tilde{\lambda}_{p,q} = \lambda_{p,q} - \epsilon$.

Statement: We require $S(\mu, \lambda, \gamma) < S\left(\tilde{\mu}, \tilde{\lambda}, \gamma\right)$ for ϵ 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.

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

$$\mu_p^g = \mu_q^h = \mu_r^i = 1/3,$$
$$\lambda_{p,q} > \lambda_{q,r} > 0,\ \lambda_{p,r} = \lambda_{p,q} + \lambda_{q,r},$$
$$\gamma^{g,h} = \gamma^{h,i} = \gamma^{g,i}/2 > 0,$$

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

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

Statement: We require $S(\mu, \lambda, \gamma) < S(\mu, \lambda, \tilde{\gamma})$ for all $\epsilon \in \left(0, \gamma^{h,i}\right)$.

Let us discuss Axiom 3, whose distributions are depicted in Fig. 2(c). The population mass is uniformly distributed on three locations (left, central and right) and three ethnic groups (repre-

sented 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

$$\mu_p^g = \mu_q^h = \mu_r^i = 1/3,\ \lambda_{p,q} = \lambda_{q,r} = \lambda_{p,r}/2 > 0,\ \gamma^{g,h} > \gamma^{h,i} > 0,\ \gamma^{g,i} = \gamma^{g,h} + \gamma^{h,i},$$

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

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

Statement: We require $S(\mu, \lambda, \gamma) < S\left(\mu, \tilde{\lambda}, \gamma\right)$ for all $\epsilon \in (0, \lambda_{q,r})$.

Fig. 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):

Theorem 1. 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. 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.

For any $\lambda_{p,q} \in [0, 1]$ and $\gamma^{g,h} \in [0, 1]$, the function $\pi\left(\lambda_{p,q}, \gamma^{g,h}\right) = \lambda_{p,q}\gamma^{g,h}$ always takes 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. To see this, let us assume that space "does not matter" by replacing the spatial distribution $\mathcal{L}$ with the "quasi-spatial" distribution $\mathbf{1}_t$, where the spatial distance between each pair of locations is equal to 1 (including the spatial distance between a location and itself, implying $\mathbf{1}_t \notin \mathcal{L}$). In this case our index becomes equivalent to the generalized fractionalization index by Greenberg (1956):

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

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., 2003, and references therein).

Symmetrically, we can assume that ethnicity "does not matter" by replacing the ethnic distribution $\mathcal{G}$ by the "quasi-ethnic" distribution $\mathbf{1}_n$, where the distance between each pair of groups is 1 (implying $\mathbf{1}_n \notin \mathcal{G}$). We can then define the spatial dispersion index

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

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 benchmark 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., $\overline{\mu}_p^g / \overline{\mu}_p = \overline{\mu}^g$ for all $g \in G$ and $p \in T$. Accordingly, we refer to $S(\overline{\mu}, \lambda, \gamma)$ as the benchmark segregation of $S(\mu, \lambda, \gamma)$, and 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 the corresponding benchmark mass distribution $\overline{\mu}$. Intuitively, focusing on the likelihood ratio should "neutralize" the magnitude effects of average spatial and ethnic distances. In fact, $A(\mu, k\lambda, k\prime\gamma) = A(\mu, \lambda, \gamma)$ for all $k, k\prime > 0$, while $S(\mu, k\lambda, k\prime\gamma) = kk\prime S(\mu, \lambda, \gamma)$ for all $k, k\prime > 0$.

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

Proposition 1. It holds that

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

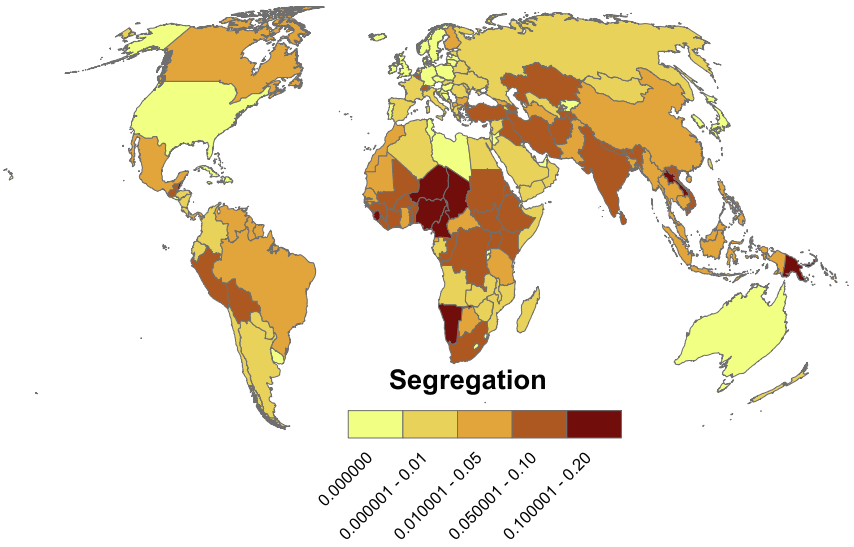

Fig. 3. Ethnic segregation across the globe.
Notes: This map shows variation in our Ethnologue-based segregation index (see Section 3.1 and Online Appendix C.1 for details). The map is projected using Eckert VI.

Table 1
Summary statistics for our indices of ethnic geography.

	Obs.	Mean	Std. Dev.	Min.	Max.
Panel A: Ethnologue-based indices					
Segregation	161	0.067	0.065	0 (many)	0.253 (NGA)
Alignment	161	1.274	0.383	0.801 (TKM)	3.176 (NOR)
Fractionalization	161	0.210	0.195	0 (many)	0.748 (PNG)
Dispersion	161	0.270	0.067	0.029 (RUS)	0.415 (SLE)
Panel B: Murdock-based indices					
Segregation	48	0.100	0.073	0.002 (SWZ)	0.269 (MLI)
Alignment	48	1.276	0.271	0.944 (DJI)	2.833 (EGY)
Fractionalization	48	0.266	0.192	0.004 (SWZ)	0.721 (TCD)
Dispersion	48	0.304	0.052	0.168 (GNQ)	0.411 (SLE)
Panel C: DHS-based indices					
Segregation	23	0.144	0.051	0.043 (GAB)	0.252 (NGA)
Alignment	23	1.137	0.083	1.052 (CIV)	1.361 (CMR)
Fractionalization	23	0.396	0.148	0.097 (GAB)	0.615 (TCD)
Dispersion	23	0.328	0.051	0.232 (MLI)	0.418 (GAB)
Panel D: Afrobarometer-based indices					
Segregation	27	0.108	0.069	0.000 (BDI)	0.232 (CMR)
Alignment	27	1.135	0.087	1.002 (TZA)	1.295 (BWA)
Fractionalization	27	0.292	0.180	0.001 (BDI)	0.563 (NER)
Dispersion	27	0.330	0.048	0.228 (MLI)	0.418 (SLE)

Notes: Summary statistics for our indices computed using the Ethnologue map (panel A), the Murdock map (panel B), the DHS (panel C), and the Afrobarometer surveys (panel D). Section 3.1 and Online Appendix C provide more information on these data sources and the computation of our indices.

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 ethno-spatial alignment index A in a multiplicative fashion.[9]

3. Applications

3.1. Data and computation of our indices

To compute our segregation index S and its components, we need the mass distribution μ, the spatial distribution λ, and the ethnic distribution γ. To get these distributions, we need information on locations and ethnic groups. At the most general level, we can rely on two types of data for this information: ethnographic maps or geo-coded survey (or census) data. These types of data both have their advantages and disadvantages.

For illustrative purposes, we compute two versions of our indices for each type of data. In what follows, we describe the data sources and sketch how we get the distributions μ, λ, and γ. Thereby, we discuss the main advantages and disadvantages of each data source and, therefore, each version of our indices. Online Appendix C provides more detailed information about the data and the computation of these four versions of our indices.

Ethnographic maps: Ethnographic maps typically show the traditional homelands of ethnic groups in space. Given the focus on traditional homelands, these maps do not typically provide any information about people living away from their traditional homelands, such as internal migrants now living in large metropolitan areas. This is a disadvantage if the goal is to measure current ethnic geography. However, it can be an advantage if researchers are interested in historical ethnic geography per se or if they prefer measures of ethnic geography that are pre-determined to recent economic or political developments and, therefore, less vulnerable

[9] Online Appendix B shows how this decomposition relates to the interpretation of S as a geometric projection.

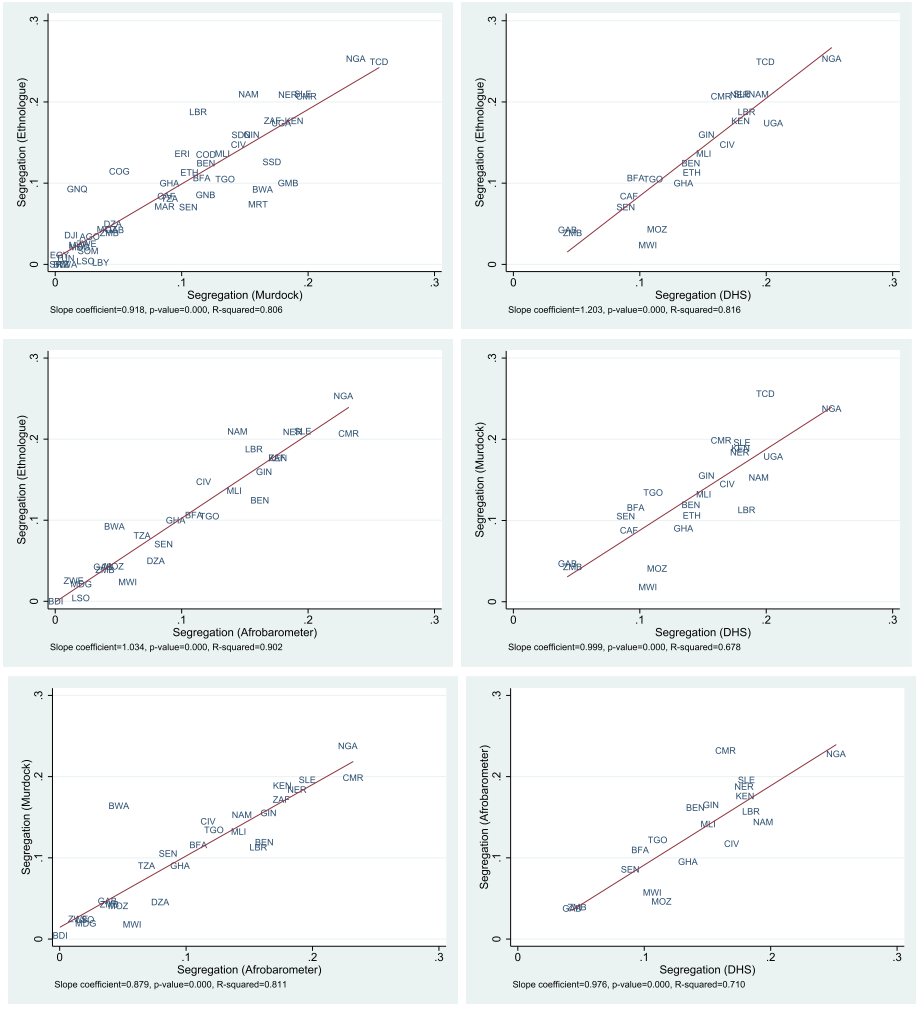

Fig. 4. Comparing our index of ethnic segregation across different data sources.
Notes: The scatter plots show the associations between the indices of ethnic segregation computed based on the Ethnologue map (and current population data), the Murdock map (and historical population data), the DHS, and the Afrobarometer surveys. Section 3.1 and Online Appendix C provide more information on these data sources and the computation of our indices.

to concerns of reverse causality. Another key advantage of ethnographic maps is that they typically offer broad spatial coverage along two dimensions. First, they provide ethnicity information for each (populated) location shown on a map. Second, they are available for an entire continent or even the entire world.

We mainly rely on the ethnographic map provided by the World Language Mapping System (WLMS, version 19). This map is based on the Ethnologue (Gordon, 2005) and represents "the region within each country, which is the traditional homeland of each indigenous language" (WLMS, version 19, n.p.). For brevity, we subsequently call it the "Ethnologue map." Relying on the Ethnologue map has several advantages. First, it has global coverage. Second, the Ethnologue provides a comprehensive rather than a selective list of language groups. We treat these language groups as the relevant ethnic groups.[10] Third, the Ethnologue provides linguistic trees for the different language families. These trees show the historical relation between languages and can be used to compute ethnolinguistic distances between groups.

We overlay the ethnographic map with small grid cells, which we take as our locations. To get the spatial distribution λ of a given country, we compute the geodesic distances between the centroids of any two grid cells and normalize them by the maximum distance between any two grid cells within this country. We use the Ethnologue's linguistic trees to derive the ethnic distribution γ. More specifically, we follow Putterman and Weil (2010) and let the ethnic distance between groups g and h be $\gamma^{g,h} = 1 - \sqrt{2\tilde{\eta}^{g,h}/(\eta^g + \eta^h)}$, where η^i is the number of nodes of language $i \in \{g, h\}$ and $\tilde{\eta}^{g,h}$ the number of common nodes. Finally, to get the mass distribution μ, we use the population density map from the Gridded Population of the World (GPW, version 4), which is based on recent population census tables and provided by CIESIN (2016). We compute our Ethnologue-based indices for 161 countries with a land surface area of more than 5,000 km^2 and a current population of more than 250,000.

We also compute our indices based on Murdock's map of precolonial ethnicities (Murdock, 1959). Relative to the Ethnologue map, the main advantage is the clearer reference to a particular time period, i.e., the times around 1900 (Michalopoulos and Papaioannou 2016, p. 1811). The main disadvantage is that it covers only Africa.

We derive the spatial distribution λ as we did for the Ethnologue map. To get the ethnic distribution γ, we merge Murdock's

[10] Common language often implies common ancestry, homeland, cultural heritage, norms, and values. Desmet et al. (2017) show that ethnolinguistic identity is indeed an important determinant of responses to many questions on cultural norms, values and preferences asked in the World Value Surveys.

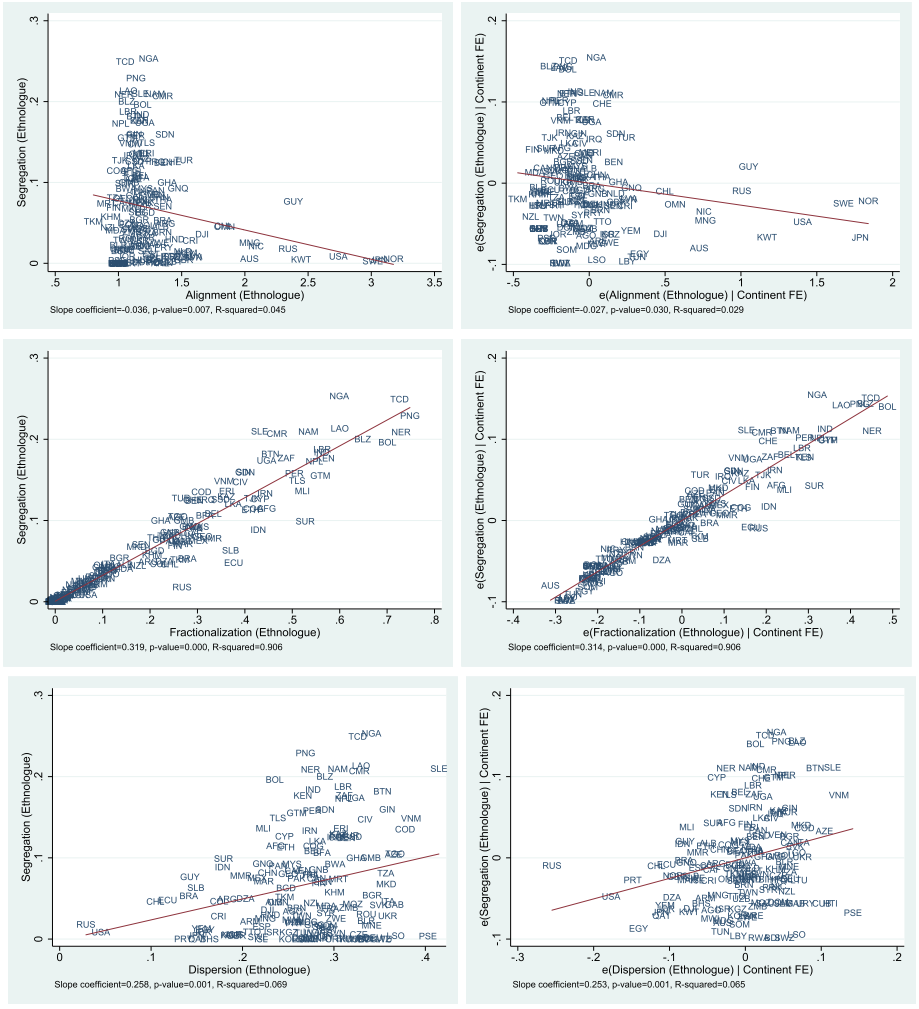

Fig. 5. The segregation index and its components.
Notes: The scatter plots on the left show the associations between the Ethnologue-based index of ethnic segregation and its three components: ethno-spatial alignment, generalized ethnic fractionalization, and spatial dispersion. The scatter plots on the right show the same associations when partialling out continent fixed effects.

ethnicities to the Ethnologue's language groups and again apply the Putterman and Weil (2010) formula to measure the ethnic distances between these groups. For these two steps, we use the Linking Ethnic Data from Africa (LEDA) software package by Carl et al. (forthcoming). Finally, we leverage the main advantage of the Murdock map and combine it with the population density map for 1900 from the History Database of the Global Environment (HYDE, version 3.2) by Klein Goldewijk et al. (2010) to derive the mass distribution μ. We compute these Murdock-based indices for 48 African countries that satisfy the area and population thresholds introduced above.

Survey data: Geo-coded survey data with information on the respondents' ethnicity (or language) have the advantage that they allow capturing the *current* distribution of ethnic groups in space. That is, they take internal migration into account. They however have some disadvantages. First, their coverage is typically not global. Second, even within countries, they only provide information for relatively few survey locations (often called clusters or enumeration areas). Hence, spatial coverage is typically much sparser.

We use the Demographic and Health Surveys (ICF (1986-2019), henceforth DHS) that are geo-coded, ask about the respondents' ethnicity, and were conducted in African countries. In total, we use information from 1,204,181 respondents of 88 surveys in 23 African countries. To get the spatial distribution λ, we compute the geodesic distance between cluster locations and normalize them by the maximum distance between any two cluster locations within the given country. To get the ethnic distribution γ, we use the LEDA software package to merge the DHS respondents' ethnicities to the language groups in the Ethnologue and to measure the ethnic distances between these groups. The mass distribution μ is based on all respondents in our final sample.

We also compute our indices using 84 geo-coded Afrobarometer surveys from 27 African countries (BenYishay et al., 2017). However, compared to the DHS, Afrobarometer surveys have fewer clusters per country (of which some are not precisely geo-referenced) and fewer respondents per cluster. As a result, we can only use information from 70,408 respondents (as opposed to more than 1.2 million in case of the DHS).

3.2. Descriptive statistics

The map in Fig. 3 shows the global distribution of our Ethnologue-based segregation index, and Table 1 provides summary statistics for all four versions of the segregation index and their components.[11]

[11] Online Appendix E provides the corresponding maps for the Murdock-, DHS- and Afrobarometer-based segregation indices as well as for the three Ethnologue-based components of the segregation index.

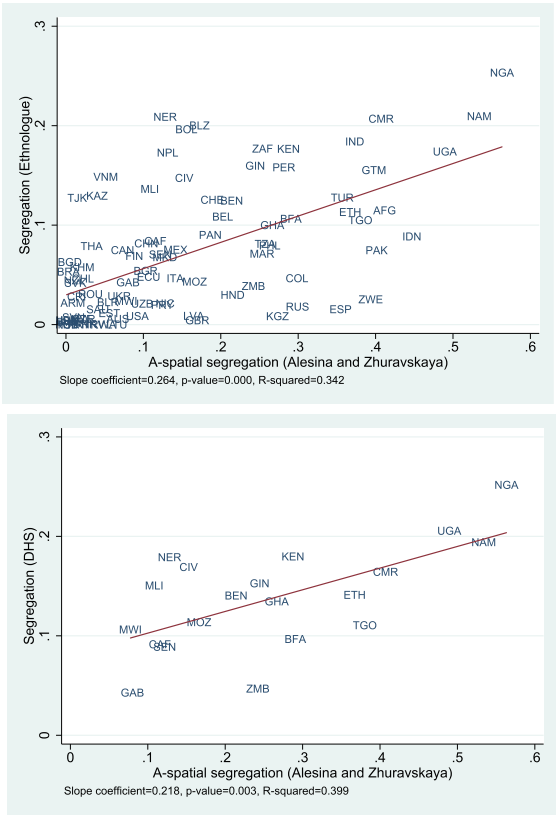

Fig. 6. Comparing our segregation index to Alesina and Zhuravskaya's (2011) a-spatial segregation index.
Notes: The scatter plots show the associations between our Ethnologue- and DHS-based segregation indices and the a-spatial segregation index by Alesina and Zhuravskaya (2011), which is based on the population shares of different ethnic groups in different subnational units rather than ethnic and spatial distances.

Nigeria is the most segregated country according to the Ethnologue and the DHS data, and the second and third most segregated country according to the Afrobarometer and the Murdock data, respectively. Its segregation index is in the range of 0.23–0.25 in all four instances. In the global Ethnologue-based sample, ethnic segregation (and generalized ethnic fractionalization) is zero in the 15 countries that have only one ethnic homeland. These countries include three from Africa: Burundi, Rwanda, and Swaziland. Burundi and Swaziland are also the least segregated countries according to the Afrobarometer and the Murdock data, respectively.[12]

Norway has the highest ethno-spatial alignment according to the Ethnologue data and is a useful example to illustrate this novel concept. Most Norwegian citizens speak Norwegian, which belongs to the Indo-European language family, and live relatively close to one another around Oslo and elsewhere in the southern parts of the country. There are, however, some small groups of Kven Finnish and Sami speakers in the far north of Norway. These languages belong to the Uralic language family. Therefore, large spatial distances between individuals predict large ethnolinguistic distances, and vice versa. That is exactly why ethno-spatial alignment is high in Norway.

In Section 3.1, we discussed the advantages and disadvantages of the different data sources that we used to compute the four different versions of our indices. Fig. 4 now provides scatter plots comparing the four versions of our segregation index.

We observe positive and fairly strong associations between any two of these segregation indices and, therefore, similar rankings across data sources. The corresponding correlation coefficients are all fairly high and range from 0.77 to 0.95 (see Table E.1 in Online Appendix E).[13] Interestingly, the correlations between the Ethnologue-based segregation index and the three other segregation indices are all higher than the correlations between any two of these other segregation indices, including the two survey-based segregation indices. These findings are reassuring. They imply that the type of data used to compute our segregation index (and its components) may not be as crucial as one may have thought. In particular, the reliance on ethnographic maps, which have larger country coverage but may induce us to attribute internal migrants to the wrong ethnic group, should lead to the same pattern of results as one would get with survey-based indices if surveys had been available for as many countries. Therefore, we mainly focus on our global Ethnologue-based indices in the subsequent analysis and relegate results for the Murdock-, DHS- and Afrobarometer-based indices to the Online Appendix.

Next, we investigate how the segregation index is related to its components. The scatter plots on the left-hand side of Fig. 5 illustrate these relations for our Ethnologue-based indices.

We see that the segregation index is positively related to generalized ethnic fractionalization and spatial dispersion and negatively to ethno-spatial alignment. The scatter plots on the right-hand side show that these relations remain very similar when partialling out continental fixed effects. Hence, these relations are not just the result of cross-continental differences but hold within continents.[14]

Finally, we compare our indices to other prominent indices of ethnic diversity. The top scatter plot in Fig. 6 illustrates the empirical relation between our Ethnologue-based spatial segregation index and the a-spatial segregation index by Alesina and Zhuravskaya (2011). The association is positive and statistically significant, but much weaker than the relations between the different versions of our spatial segregation index (shown in Fig. 4). The second scatter plot uses our DHS-based segregation index, as Alesina and Zhuravskaya (2011) use DHS data to compute their a-spatial segregation index for some of the countries in their sample. The association remains relatively weak, which is not surprising given the important conceptual differences between spatial and a-spatial segregation indices (discussed in the Introduction and illustrated in Online Appendix A).

Table 2 reports correlation coefficients between our Ethnologue-based indices of ethnic geography and various prominent indices of ethnic diversity. Panel A again looks at Alesina and Zhuravskaya's (2011) a-spatial segregation index. Panel B compares our indices to the commonly used standard fractionalization indices that Alesina et al. (2003) computed based on ethnicity, language, and religion. Unsurprisingly, our index of generalized fractionalization, which is based on language trees, is most strongly correlated to the language-based fractionalization index and uncorrelated with the one based on religion. Panel C compares our indices to the indices of standard fractionalization, generalized fractionalization, and polarization by Esteban et al. (2012). Unsurprisingly, our index of generalized ethnic fractionalization is most

[12] The DHS sample includes none of these three African countries.

[13] Tables E.2–E.4 in Online Appendix E show the correlation coefficients between the different versions of the components of the segregation index. These correlation coefficients are fairly high as well. They are in the range of 0.42–0.88 for ethno-spatial alignment, 0.86–0.94 for generalized ethnic fractionalization, and 0.69–0.90 for spatial dispersion.

[14] Table E.5 in Online Appendix E shows the correlation coefficients between the segregation index and its components are similar across continents, but differ somewhat across data sources used for the computation of the indices.

Table 2
Correlations between our indices of ethnic geography and alternative indices.

	(1)	(2)	(3)	(4)	(5)
	S	*A*	*F*	*D*	Obs.
Panel A: Alesina/Zhuravskaya (2011)					
A-spatial segregation	0.585	−0.129	0.541	0.073	90
Panel B: Alesina et al., (2003)					
Standard fractionalization (ethnicity)	0.515	−0.203	0.487	0.176	156
Standard fractionalization (language)	0.603	−0.235	0.541	0.365	154
Standard fractionalization (religion)	0.115	−0.005	0.060	0.185	158
Panel C: Esteban et al., (2012)					
Standard fractionalization	0.575	−0.216	0.569	0.112	133
Generalized fractionalization	0.591	−0.081	0.631	−0.016	133
Polarization	0.381	−0.049	0.441	−0.068	133

Notes: Cells in columns (1)–(4) report correlation coefficients between established indices of ethic diversity and our Ethnologue-based indices of ethnic segregation (*S*), ethno-spatial alignment (*A*), generalized ethnic fractionalization (*F*), or spatial dispersion (*D*). The established indices are: the a-spatial segregation used by Alesina and Zhuravskaya (2011); standard fractionalization, which is the fractionalization index based on categorical data; generalized fractionalization, which is based on (non-binary) ethnic distances and sometimes called the Greenberg-Gini index; and polarization, which is the polarization index by Duclos et al. (2004). Column (5) reports the number of observations on which the correlation coefficients are based.

Table 3
Climate, geography, and our indices of ethnic geography.

	(1)	(2)	(3)	(4)	(5)
Dependent variable:	*S*	*S*	*A*	*F*	*D*
Americas	−0.042***	−0.061***	0.224*	−0.095*	−0.108***
	(0.015)	(0.017)	(0.119)	(0.050)	(0.018)
Asia	−0.027*	−0.030*	0.080	−0.049	−0.056***
	(0.014)	(0.016)	(0.102)	(0.042)	(0.015)
Europe	−0.071***	−0.027	−0.038	−0.024	−0.039
	(0.012)	(0.024)	(0.211)	(0.061)	(0.029)
Oceania	−0.013	−0.009	−0.168	0.064	−0.052*
	(0.045)	(0.038)	(0.342)	(0.121)	(0.030)
Absolute latitude		0.002	−0.012	0.002	0.004**
		(0.001)	(0.010)	(0.003)	(0.001)
Temperature		0.002	−0.032*	0.005	0.004
		(0.002)	(0.019)	(0.006)	(0.003)
Precipitation		0.003*	0.000	0.007*	0.000
		(0.001)	(0.010)	(0.004)	(0.002)
Access to sea		−0.045***	0.211*	−0.128***	−0.051***
		(0.016)	(0.112)	(0.047)	(0.015)
Terrain roughness		−0.003	0.037	−0.013	0.003
		(0.006)	(0.036)	(0.016)	(0.006)
Mean elevation		0.002	−0.311	−0.002	0.032
		(0.027)	(0.227)	(0.077)	(0.033)
St. dev. elevation		0.074***	−0.021	0.222***	0.012
		(0.019)	(0.079)	(0.056)	(0.020)
Mean land suitability		−0.006	−0.645***	−0.087	0.140***
		(0.021)	(0.172)	(0.061)	(0.023)
St. dev. land suitability		0.087	0.173	0.342**	−0.153***
		(0.054)	(0.344)	(0.161)	(0.056)
Malaria suitability		0.133**	−0.320	0.326*	0.131**
		(0.061)	(0.622)	(0.193)	(0.064)
R^2	0.162	0.435	0.242	0.454	0.437
Countries	161	147	147	147	147

Notes: Dependent variables are indicated in the top row. They are the Ethnologue-based indices of ethnic segregation (*S*), ethno-spatial alignment (*A*), generalized ethnic fractionalization (*F*), and spatial dispersion (*D*). OLS regressions with continent fixed effects. Africa is the omitted category. Online Appendix D.1 provides more information on the climatic and geographical variables, including summary statistics in Table D.1. Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

strongly correlated to the generalized fractionalization index by Esteban et al. (2012).[15] Finally, notice the low correlation between our novel index of ethno-spatial alignment and all these established indices.

3.3. Climate, geography, and our indices of ethnic geography

Previous contributions argue that climate and geography are key determinants of local ethnic diversity (e.g., Nettle, 1998, Nunn, 2008, Michalopoulos, 2012, Cervellati et al., 2019). Table 3 shows how climatic and geographical factors shape ethnic geography as measured by our Ethnologue-based indices.

Column (1) regresses the segregation index on dummy variables for the different continents (with Africa being the omitted category). The results imply that African countries are significantly more segregated than countries in the Americas or Europe. Column (2) adds a large set of climatic and geographical variables: Absolute latitude, temperature and precipitation, access to the sea, terrain roughness, mean and standard deviation of elevation, mean and standard deviation of land suitability for agriculture, and malaria suitability (see Online Appendix D.1 for definitions and data sources). We see that American countries are less segregated than

[15] The correlation between the generalized fractionalization index by Esteban et al. (2012) and our survey-based generalized fractionalization indices are even higher: 0.77 for the Afrobarometer surveys and 0.76 for the DHS.

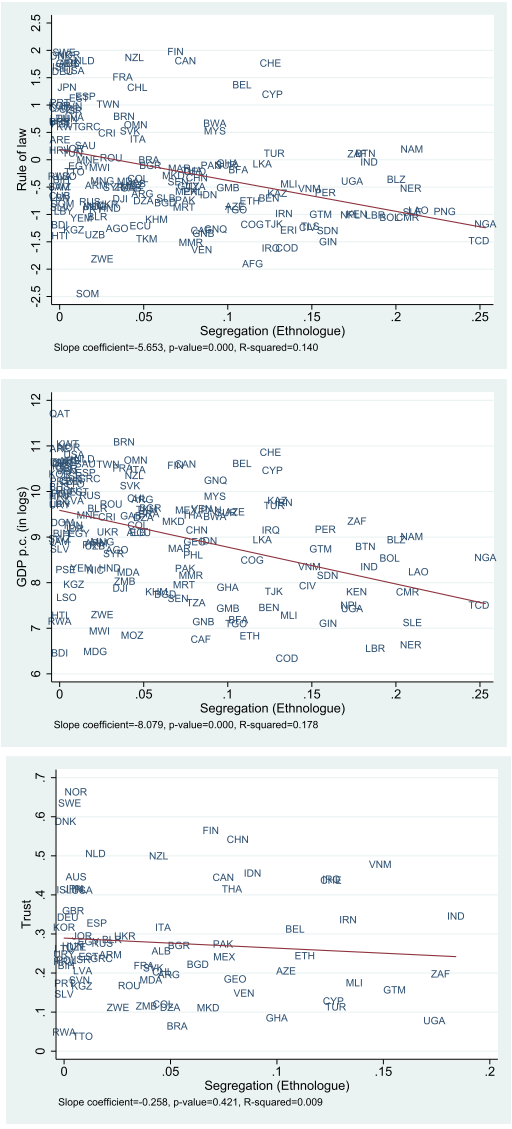

Fig. 7. Ethnic segregation and comparative development.
Notes: The scatter plots show the association between the Ethnologue-based index of ethnic segregation and three measures of comparative development: The rule of law in 2010 from the World Bank Governance Indicators in the top graph; the log of GDP per capita in 2010 from the Penn World Tables in the middle graph; and generalized trust from the World Value Surveys 1981–2008 in the bottom graph. Online Appendix D.2 provides more information on these measures.

African ones even when controlling for these climatic and geographical factors (while the same does not hold true for European countries). Access to the sea, limited variability in elevation, and a climate unsuitable for malaria are also related to low levels of ethnic segregation. Taken together, all the climatic and geographical variables explain 44 percent of the variation in our segregation index.

Columns (3)–(5) present the results for the three components of our segregation index. These climatic and geographical variables also explain 44–45 percent of the variation in generalized ethnic fractionalization and spatial dispersion, and 24 percent of the variation in ethno-spatial alignment.[16]

3.4. Ethnic geography and comparative development

In a last step, we study the relation between our Ethnologue-based indices of ethnic geography and prominent measures of comparative political, economic and social development. These measures are the rule of law from the World Bank Governance Indicators (following Alesina and Zhuravskaya, 2011), the log of GDP per capita in 2010 from the Penn World Tables 9.0, and generalized trust from the World Values Surveys 1981–2008 (taken from Ashraf and Galor, 2013), which is only available for around half of the countries in our sample. Fig. 7 shows scatter plots between the segregation index and these measures of comparative development.[17]

We see that more segregated societies have substantially weaker rule of law and substantially lower GDP. Trust levels tend to be slightly lower too, but no significantly so.

Next, we turn to cross-country regressions. The use of cross-country regressions is common in the literature on the economic, political and social effects of ethnic diversity, as is the caveat that the estimated coefficients may not represent causal effects. We try to address omitted variable bias by using continent fixed effects and controlling for climatic and geographical factors; and we try to alleviate concerns of reverse causality by using our indices based on ethnographic maps of traditional homelands. Nevertheless, we abstain from a causal interpretation of our results.

The upper panel of Table 4 presents the results of linear regressions of our measures of comparative development on the Ethnologue-based segregation index. The dependent variable is the rule of law in columns (1)–(2), log GDP per capita in columns (3)–(4), and generalized trust in columns (5)–(6). We include continent fixed effects in all columns and the climatic and geographical variables introduced in Section 3.3 in even columns. Columns (1) and (3) show that the rule of law and GDP are negatively associated with ethnic segregation even within continents. However, the results in columns (2) and (4) suggest that this negative association may be an artefact of cross-country differences in climate and geography.

In the lower panel of Table 4, we replace the segregation index by its three components, i.e., the indices of ethno-spatial alignment, generalized ethnic fractionalization, and spatial dispersion. Consistent with the previous literature (e.g., Easterly and Levine, 1997; Alesina et al., 2003; Alesina and La Ferrara, 2005), we find that the rule of law and incomes are negatively associated with ethnic fractionalization. However, this negative association too becomes insignificant when we control for climatic and geographical factors. There is no consistent pattern for the association between spatial dispersion and our measures of comparative development.

The most interesting result in Table 4 concerns ethno-spatial alignment. Higher alignment is associated with better rule of law, higher levels of trust, and, arguably, higher GDP even if we control for climatic and geographical factors. Hence, countries where diverse individuals live farther apart tend to be better governed, richer, and more trusting. This result is novel, as is the concept of ethno-spatial alignment.[18]

[16] The positive effects of the standard deviations of elevation and land suitability on ethnic segregation and fractionalization corroborate the findings of Michalopoulos (2012); and the positive effects of malaria suitability the finding of Cervellati et al. (2019). Online Appendix F presents the effects of climate and geography on our Murdock-, DHS-, and Afrobarometer-based indices in Tables F.1–F.3.

[17] Online Appendix D.2 provides definitions and data sources for these measures; and Online Appendix G presents the corresponding scatter plots for our Murdock-, DHS- and Afrobarometer-based indices in Figure G.1.

[18] Online Appendix G presents the results for our Murdock-, DHS-, and Afrobarometer-based indices in Table G.1 and robustness tests for our Ethnologue-based indices in Tables G.2–G.10.

Table 4
Ethnic geography and comparative development.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent var.:	Rule of law		Log GDP p.c.		Generalized trust	
Segregation	−2.864***	0.482	−2.996**	0.345	0.060	0.411
	(1.056)	(1.174)	(1.223)	(1.286)	(0.319)	(0.330)
R^2	0.375	0.560	0.517	0.683	0.252	0.565
Countries	159	147	151	139	77	74
Alignment	0.459***	0.428**	0.295*	0.235*	0.121***	0.081**
	(0.160)	(0.171)	(0.173)	(0.129)	(0.038)	(0.033)
Fractionalization	−0.852**	0.021	−0.776*	−0.017	0.106	0.137
	(0.332)	(0.389)	(0.407)	(0.588)	(0.089)	(0.087)
Dispersion	−1.380	0.941	−3.459**	0.407	0.065	0.428
	(1.378)	(1.461)	(1.355)	(1.368)	(0.254)	(0.260)
R^2	0.435	0.579	0.568	0.687	0.388	0.606
Countries	159	147	151	139	77	74
Continent FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes

Notes: Dependent variables are indicated in the top row. They are the rule of law in 2010 from the World Bank Governance Indicators in columns (1) and (2); the log of GDP per capita in 2010 from the Penn World Tables in columns (3) and (4); and generalized trust from the World Value Surveys 1981–2008 in columns (5) and (6). Each column presents two OLS regressions. The main explanatory variable(s) is the Ethnologue-based index of ethnic segregation in the upper panel, and the Ethnologue-based indices of ethno-spatial alignment, generalized ethnic fractionalization and spatial dispersion in the lower panel. All specifications include continent fixed effects and those in even columns also control for the climatic and geographical variables used in Table 3. Online Appendices D.1 and D.2 provide more information on the control and the dependent variables. Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

4. Conclusions

We have developed a novel index of ethnic segregation based on ethnic distances between groups and spatial distances between locations. We have provided an axiomatic characterization and have shown that our segregation index is decomposable into three sub-indices: generalized ethnic fractionalization, spatial dispersion, and the alignment between ethnic and spatial distances.

We have computed our indices for a large sample of countries using either ethnographic maps or geo-coded survey data as the main input. While the different types of data have different advantages and disadvantages, we have shown that the resulting segregation indices are highly correlated. We have documented that variations in climate and geography explain a large share of the variation in ethnic geography as measured by our indices. Further cross-country regressions have revealed that the negative association between ethnic segregation and current economic and political development is mainly due to differences in climate and geography. In contrast, countries with higher ethno-spatial alignment tend to be more successful even when controlling for climatic and geographical factors.

The indices we have developed can be applied in many more ways than just for measuring country-level ethnic geography. First, they can be used to measure ethnic geography at the level of alternative spatial units, such as continents, provinces, districts, cities, or grid cells. Second, they can be used to measure spatial segregation or human geography along dimensions of identity other than ethnicity, such as caste, religion, types of occupation, party affiliations, or even cultural values. Therefore, we hope that these indices become a useful tool for researchers keen to improve our understanding of how ethnic and human geography shape comparative economic, political and social development.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful for the clear and insightful guidance provided by the editor, Thomas Fujiwara, and the helpful comments by two anonymous referees, Magnus Hatlebakk, Mario Jametti, Nadine Ketel, Stelios Michalopoulos, Maria Petrova, Marta Reynal-Querol, Måns Söderbom, Ragnar Torvik, David Yanagizawa-Drott, Ekaterina Zhuravskaya, and other seminar participants at IEB Barcelona, CMI Bergen, NHH Bergen, Deakin University, Lancaster University, Monash University, Universitat Pompeu Fabra, University of Gothenburg, University of Lugano, University of St.Gallen, University of Zurich, the CESifo Workshop on Political Economy, the ASWEDE conference, and the NES CSDSI International Conference “Towards Effective and Equitable Development: the Role of Institutions and Diversity.” Steve Berggreen-Clausen and Noémie Zurlinden provided excellent research assistance.

Appendix A

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

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

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

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

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

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

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 $a,b\geqslant 0$, 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 $a,b\geqslant 0$. □

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 $g,h\in G$ with $\mu^g,\mu^h>0$. Similarly, $D(\mu,\lambda)=0$ implies $\lambda_{p,q}=0$ for all $p,q\in T$ with $\mu_p,\mu_q>0$. 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 π.

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). □

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jpubeco.2021.104446.

References

Alesina, Alberto, Devleeschauwer, Arnaud, Easterly, William, Kurlat, Sergio, Wacziarg, Romain, 2003. Fractionalization. J. Econ. Growth 8, 155–194.

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

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

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

Bell, Wendell, 1954. A Probability Model for the Measurement of Ecological Segregation. Soc. Forces 32, 357–364.

BenYishay, A., R. Rotberg, J. Wells, Z. Lv, S. Goodman, L. Kovacevic, and D. Runfola, Geocoding Afrobarometer Rounds 1–6: Methodology & Data Quality (AidData, 2017).

Blumenstock, Joshua, Fratamico, Lauren, 2013. In: Proceedings of the 4th Annual Symposium on Computing for Development, p. 11.

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

Cederman, Lars-Erik, Buhaug, Halvard, Rød, Jan K., 2009. Ethno-Nationalist Dyads and Civil War: A GIS-based Analysis. J. Conflict Resolut. 53, 496–525.

Center for International Earth Science Information Network at Columbia University (CIESIN), Gridded Population of the World, Version 4, Palisades, NY (2016).

Cervellati, Matteo, Chiovelli, Giorgio, Esposito, Elena, 2019. Bite and Divide: Malaria and Ethnolinguistic Diversity. CEPR Discussion Paper 13437.

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

Luca, De, Giacomo, Roland Hodler, Raschky, Paul A., Valsecchi, Michele, 2018. Ethnic Favoritism: An Axiom of Politics?. J. Dev. Econ. 132, 115–129.

Desmet, Klaus, Gomes, Joseph, Ortuño-Ortín, Ignacio, 2020. The Geography of Linguistic Diversity and the Provision of Public Goods. J. Dev. Econ. 143, 102384.

Desmet, Klaus, Ortuño-Ortín, Ignacio, Wacziarg, Romain, 2012. The Political Economy of Linguistic Cleavages. J. Dev. Econ. 97, 322–338.

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

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

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

Easterly, William, Levine, Ross, 1997. Africa's Growth Tragedy: Policies and Ethnic Divisions. Quart. J. Econ. 112, 1203–1250.

Echenique, Federico, Fryer Jr., Roland G., 2007. A Measure of Segregation Based on Social Interactions. Quart. J. Econ. 122, 441–485.

Ejdemyr, Simon, Kramon, Eric, Lea Robinson, Amanda, 2018. Segregation, Ethnic favoritism, and the Strategic Targeting of Local Public Goods. Comparative Political Studies 51, 1111–1143.

Esteban, Joan, Mayoral, Laura, Ray, Debraj, 2012. Ethnicity and Conflict: An Empirical Study. American Economic Review 102, 1310–1342.

Esteban, Joan, Ray, Debraj, 1994. On the Measurement of Polarization. Econometrica 62, 819–851.

Frankel, David M., Volij, Oscar, 2011. Measuring School Segregation. Journal of Economic Theory 146, 1–38.

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

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

Hodler, Roland, Srisuma, Sorawoot, Vesperoni, Alberto, Zurlinden, Noémie, 2020. Measuring ethnic stratification and its effect on trust in Africa. J. Dev. Econ. 146, 102475.

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

ICF, "Demographic and Health Surveys" (various), 1986-2016.

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

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

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

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

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

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

Michalopoulos, Stelios, Papaioannou, Elias, 2014. National Institutions and Subnational Development in Africa. Quart. J. Econ. 129, 151–213.

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

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

Müller-Crepon, Carl, Yannick Pengl, and Nils-Christian Bormann, "Linking Ethnic Data from Africa," Journal of Peace Research, forthcoming.

Murdock, George P., 1959. Africa: Its Peoples and Their Culture History. McGraw-Hill, New York, NY.

Nettle, Daniel, 1998. Explaining Global Patterns of Language Diversity. J. Anthropol. Archaeol. 17, 354–374.

Nunn, Nathan, 2008. The Long-term Effects of Africa's Slave Trades. Quart. J. Econ. 123, 139–176.

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

Putterman, Louis, Weil, David N., 2010. Post-1500 Population Flows and The Long-Run Determinants of Economic Growth and Inequality. Quart. J. Econ. 125, 1627–1682.

Rao, C. Radhakrishna, 1982. Diversity and Dissimilarity Coefficients: A Unified Approach. Theor. Popul. Biol. 21, 24–43.

Reardon, Sean F., Firebaugh, Glenn, 2002. Measures of Multigroup Segregation. Sociol. Methodol. 32, 33–67.

Reardon, Sean F., O'Sullivan, David, 2004. Measures of Spatial Segregation. Sociol. Methodol. 34, 121–162.
Tajima, Yuhki, Samphantharak, Krislert, Ostwald, Kai, 2018. Ethnic Segregation and Public Goods: Evidence from Indonesia. American Political Science Review 112, 637–653.
Weidmann, Nils B., 2009. Geography as Motivation and Opportunity: Group Concentration and Ethnic Conflict. J. Conflict Resolut. 53, 526–543.
White, Michael J., 1983. The Measurement of Spatial Segregation. Am. J. Sociol. 88, 1008–1018.