THE HEALTH COSTS OF ETHNIC DISTANCE: EVIDENCE FROM SUB-SAHARAN AFRICA

Joseph Flavian Gomes

DISCUSSION PAPER | 2020 / 05

UCLouvain

IRES | LIDAM Louvain Institute of Data Analysis and Modeling in economics and statistics

The Health Costs of Ethnic Distance: Evidence from Sub-Saharan Africa

Joseph Flavian Gomes*

January 2020

Abstract

This paper shows that children of mothers who are ethnically more distant from their neighbours have worse health outcomes. I combine individual-level micro data from DHS surveys for 14 sub-Saharan African countries with a novel high-resolution dataset on the spatial distribution of ethnic groups at the 1 km × 1 km level. I measure ethnic distance using linguistic distance and construct the spatial distribution of ethnic groups using an iterative proportional fitting algorithm. Using a time-varying ethnicity fixed effects framework to curb unobserved heterogeneity across ethnic groups, I show that children whose mothers are linguistically more distant from their neighbours face higher mortality rates and are shorter in stature. The pernicious effects of linguistic distance are more pronounced in areas where malaria is endemic. I argue that higher linguistic distance impedes the transmission of information. Consistent with this interpretation, mothers who are linguistically more distant from their neighbours are less likely to receive health-related information. Linguistic distances driven by splits that occurred thousands of years ago are more relevant than more recent splits.

Keywords: ethnic distance, ethnic diversity, ethnic networks, child mortality, African development.
JEL Codes: I14, O10, O15, Z10, Z13

*IRES/LIDAM, UCLouvain, Louvain-la-Neuve, Belgium; & CEPR, UK; E-mail: joseph.gomes@uclouvain.be; Tel. +32 10 47 3990. The author is grateful to Klaus Desmet, Ignacio Ortuño-Ortin, Ulrich Wagner, Jesus Carro, Sonia Bhalotra, Irma Clots-Figueras, Damian Clarke, J. Andrew Harris, Juan Jose Dolado, Joan Maria Esteban, Jim Fearon, James Fenske, Jed Friedman, Paola Giuliano, Oded Galor, Saumitra Jha, Eliana La Ferrara, Yuya Kudo, Edward Leamer, Matilde Machado, Stelios Michalopoulos, Ricardo Mora, Owen Ozier, Dan Posner, Diego Puga, Richard Scheffler, Alessandro Tarozzi, Nico Voigtlaender, Romain Wacziarg and all seminar/conference participants at the 6th European Conference on Networks, Oxford CSAE conference, World Bank ABCDE conference, NOVAFRICA conference, ISI Delhi ACEGD Conference, NCID Research Workshop, Warwick Summer Workshop in Economic Growth, UC3M, ISER, Nottingham, Essex Govt Dept, UPNA, and Kent for their comments and suggestions; to Jim Fearon for generously sharing the data on ethnicity and languages matching; to Gordon McCord for generously making the data on malaria suitability available; and to UCLA Anderson and UPF Barcelona for their hospitality.

1 Introduction

What are the individual-level consequences of living in ethnically diverse societies? A vast body of literature has established the role of ethnic heterogeneity in economic performance (Alesina and Ferrara, 2005). However, the focus has often been on investigating the consequences of ethnic heterogeneity at an aggregate level (such as country, district, or city) using aggregate indices of diversity, rather than on an individual level. The most commonly used aggregate indices include fractionalization (Easterly and Levine, 1997; Alesina et al., 2003), polarization (Esteban and Ray, 1994, 1999; Montalvo and Reynal-Querol, 2005), and genetic diversity (Ashraf and Galor, 2013b). For instance, ethnic fractionalization describes the probability that two randomly selected individuals from a given region belong to two different ethnic groups. All individuals, regardless of their ethnicity, face the same level of ethnic fractionalization in a region. This approach disregards the possibility of individual-specific heterogeneity in the consequences of living in diverse regions. In contrast, this paper focuses on the concept of individual-level ethnic distance to identify precisely which individuals lose out within a particular region, even if they face the same level of ethnic diversity in the aggregate. Ethnic distance measures how ethnically different an individual is from others living in the same region. Unlike aggregate measures, it is specific to an individual's ethnicity and geographic location within a region.[1]

In this study, I combine high-quality individual-level micro data from the Demographic and Health Surveys (DHS) for fourteen sub-Saharan African countries with a novel dataset on the spatial distribution of ethnic groups at the level of approximately 1 km $\times$ 1 km. The spatial distribution data are constructed using an iterative proportional fitting algorithm recently developed by Desmet et al. (2020). Exploiting the individual mother's location, ethnicity (which I map to languages from the Ethnologue database) and the spatial distribution of language groups, I construct individual-level ethnic distances of the mothers from people living around them. Following a burgeoning trend in the literature, I measure the ethnic distance between any given pair of ethnic groups by the degree of difference between the languages spoken by the two groups.[2] The distance metric is based on the number of branches shared between any two languages according to Ethnologue language trees. Instead of taking a stand on what the

[1] This builds on a well-established macro literature that demonstrates how ethnic distances impede trade and the diffusion of innovation and technology, thereby negatively affecting economic development (Guiso et al., 2009; Spolaore and Wacziarg, 2016, 2009).

[2] See, for instance, Fearon (2003), Desmet et al. (2012), Desmet et al. (2009), Esteban et al. (2012a), Esteban et al. (2012b) and Laitin and Ramachandran (2016). Since I measure ethnic distance using linguistic distance, the terms ethnic distance and linguistic distance will be used interchangeably throughout the paper.

appropriate neighbourhood or region for calculating these distances ought to be, I calculate the distances by drawing circles of different radii around the mothers. Then, using a time-varying ethnicity fixed effects framework, I investigate how the mother's ethnic distance from her neighbours affects her children's health outcomes. This allows me to curb the effects of unobserved heterogeneity across ethnic groups such as ethnic advantage or dominance of certain groups (Kramon and Posner, 2016; Burgess et al., 2015; Franck and Rainer, 2012), as well as within-group genetic diversity (Arbatlı et al., 2015).

My primary finding is that the children of mothers who are ethnically distant from their neighbours experience higher mortality rates. This effect is stronger for individuals who have never migrated from their village of birth. Considering a circle of 50 km in radius around the mother, a one SD increase in linguistic distance leads to 14 additional child deaths (approximately 3.3% SD) for the non-migrant group. The corresponding number for the migrant group is four additional child deaths (1% SD). Next, using recently constructed disaggregated spatial data on malaria suitability (Kiszewski et al., 2004; McCord and Anttila-Hughes, 2017), I find that malaria suitability worsens the pernicious effects of linguistic distance on child mortality. For the non-migrant sample, a one SD increase in linguistic distance when malaria suitability is one SD above its average value leads to 34 additional child deaths compared to the 14 additional child deaths for the average level of malaria suitability.

Following the ethnic networks literature (Larson and Lewis, 2017; Fisman et al., 2017; Pongou, 2009), I hypothesize that information does not flow smoothly across ethnic lines; thus, individuals who are ethnically distant from their neighbours suffer adverse effects. I uncover several pieces of evidence in support of this hypothesis. Delving deeper into the heterogeneity by migration status, I find that linguistically distant individuals who have never migrated are less likely to have heard of oral rehydration for treating children with diarrhoea, to have received tetanus injections and iron tablets during pregnancy, to have washed hands before making their meals, and to know when to seek medical help for their children.

The increased strength of the results for individuals who have never migrated might itself indicate that individual ethnic distances impose barriers to accessing information related to health. Individuals who have moved from other places are more likely to have acquired health information in their previous place of residence, which I cannot observe. However, individuals who have never moved have acquired knowledge related to health in their current place of residence. Hence, linguistic distance affects them more profoundly than migrants.[3] The more

[3]Furthermore, individuals with the possibility to move might choose to relocate to more favourable locations.

profound effects of linguistic distance on child health in malaria-endemic areas might also indicate an information channel. For instance, it may suggest that linguistically distant mothers do not have access to best-practice information about preventing or treating malaria, resulting in higher child mortality among them in malaria-endemic areas. Furthermore, maternal linguistic distance also has strong and robust negative effects on child height, which is an important marker of child health.This lends further credence to the hypothesis that linguistic distance impedes the flow of information, as information is crucial for child height (Thomas et al., 1991).[4] Finally, using survey data collected by Larson and Lewis (2017) as part of an experiment that seeded information in different Ugandan villages, I provide evidence that information is less likely to pass on to individuals who are linguistically more distant from their neighbours, even in an experimental setting.[5]

I investigate several alternative channels through which linguistic distance from one's neighbours might be injurious to health. First, using data on access to public goods such as education, water, and electricity, I investigate whether linguistically distant individuals face more discrimination in accessing these public goods. I find no evidence supporting a direct discrimination channel. Next, using the share of coethnics living in a mother's vicinity, I investigate whether linguistic distance captures the role of kin networks and resource sharing within close-knit ethnic communities rather than imposing information barriers. I do not unearth any evidence in favour of this hypothesis either. Finally, in the spirit of Buckles and Hungerman (2013), I investigate whether selection can provide an alternative explanation of why linguistic distance worsens health outcomes, but find no evidence to support this. More specifically, I do not find any evidence suggesting that women who are more linguistically distant from their neighbours and have children are different from those who do not have children.

Another possible explanation could be that individuals speaking different languages do not understand each other. Public health papers have underscored that minorities such as Hispanics

However, I find that migrants generally have higher linguistic distance from their neighbours in addition to experiencing higher child mortality rates.

[4]Linguistic distance might impede access to other types of information including information on feeding-practices, which are crucial for child health (Malhotra, 2012; UNICEF, 2012), but which I cannot measure in the current study.

[5]More generally, a large number of papers underscore that easier information flows within ethnic groups is one of the main reasons behind the ability of ethnic groups to act collectively. For instance, the coethnic advantage of socially sanctioning other group members (Fearon and Laitin, 1996; Miguel and Gugerty, 2005) draws from the coethnic informational advantage of learning when a group member defects and also on how to locate the defectors to mete out punishment. The ease of information dissemination among coethnics also facilitates the spread of incendiary rumours (Varshney, 2003), viable rebel group formation (Larson and Lewis, 2018), access to credit (Fafchamps, 2000; Fisman et al., 2017), fidelity decisions (Pongou, 2009), and access to knowledge networks (Romani, 2004).

in the US (Singleton and Krause, 2009; Flores, 2006) or asylum seekers in Europe (Bischoff and Denhaerynck, 2010) face communication barriers when seeking medical attention because they speak a different language from the medical staff. An analysis that varies the values of a parameter that determines how fast the distance between any two languages declines as the number of shared branches increases suggests the irrelevance of mutual intelligibility of two languages in explaining the findings. Languages become unintelligible quite quickly as they move a few branches away from each other in the language tree. Rather, I uncover evidence suggesting the importance of deeper cleavages arising from splits that occurred thousands of years ago.[6] Building on the ethnic networks literature, I argue that individuals are reluctant to pass on information to others who are linguistically distant from them.[7]

The empirical analysis relies on controlling for a rich set of control variables and fixed effects.[8] All specifications include controls for aggregate levels of ethnic diversity such as fractionalization or polarization.[9] Controlling for ethnicity- and religion-specific fixed effects purge heterogeneity in unobservable characteristics across ethnic or religious groups. The inclusion of time-varying ethnicity fixed effects allows me to rule out explanations based on political ethnic favouritism (Kramon and Posner, 2016; Burgess et al., 2015; Franck and Rainer, 2012; Kudamatsu, 2009) as well as purging the effects of within-group genetic diversity (Arbatlı et al., 2015). The inclusion of time-varying region fixed effects rules out the possibility of region-specific transfers benefiting certain ethnic groups at the cost of others (De Luca et al., 2017; Dickens, 2016). Further, using the heuristics of Altonji et al. (2005) and incorporating insights from Oster (2017), I show that the results are unlikely to be driven by selection on unobservable variables. If anything, selection on unobservable variables drives the main coefficient of interest away from zero.

Finally, I construct additional variables to rule out several potential confounders. First, I show that the effects of linguistic distance are not explained by differences in individual-level

[6] This is in line with some existing papers that find that coarser divisions matter more for civil conflicts (Desmet et al., 2012) and market integration (Fenske et al., 2017).

[7] Several papers have established the correlation of ethnicity with networks. For instance, Larson and Lewis (2017) show that individuals in Ugandan villages form networks on the basis of ethnicity. Fafchamps (2000) shows how ethnicity affects access to trade and bank credit through network effects in the form of socialization and information sharing in Kenyan and Zimbabwean firms. Again, Romani (2004) shows that ethnic minorities in Ivory Coast are less likely to access and benefit from extension services because they have lower access to information networks.

[8] The controls include several birth-specific variables such as child gender and birth order; mother-specific variables such as education and wealth; religion fixed effects; and time-varying region and ethnicity fixed effects.

[9] The inclusion or exclusion of aggregate measures of ethnic diversity, with or without weighting by distance, does not affect the results. While data unavailability prevents the inclusion of controls for genetic diversity, recent papers have highlighted that genetic diversity spans measures of ethnic fractionalization (Ashraf and Galor, 2013a, 2018). Furthermore, while Ashraf and Galor (2018) find that the effect of genetic diversity trumps the effect of genetic distance, I find that linguistic distance trumps linguistic diversity.

genetic distance (Spolaore and Wacziarg, 2016), or cultural distance (Giuliano and Nunn, 2018). I also construct an index of distance from the dominant group in the region and pit it against my measure of average linguistic distance that considers all individuals residing in the region when constructing linguistic distance.[10] Average linguistic distance comes out to be both economically and statistically more significant compared to distance from the dominant group.

In addition to the ethnic networks literature mentioned previously, this paper contributes to two other strands of the literature. First, I contribute to the literature that demonstrates the role of ethnic and cultural distances in economic outcomes (Spolaore and Wacziarg, 2009, 2016; Guiso et al., 2009; Desmet et al., 2017). This literature is largely agnostic about specific mechanisms and how exactly ethnic barriers operate. My micro-level analysis demonstrates how ethnic distance might act as a barrier to health information, leading to higher child mortality rates among linguistically distant groups. A related nascent strand of the micro literature highlights the ways in which ethnic distance affects economic development through human capital accumulation (Laitin and Ramachandran, 2016; Shastry, 2012), trade flows (Isphording and Otten, 2013), literacy and labour market outcomes of immigrants (Isphording, 2014), market integration (Fenske et al., 2017), and the effectiveness of counterinsurgency policies (Armand et al., 2017). I introduce a spatial dimension at the individual level and highlight an information channel, both of which are novel in the literature.

Second, this paper contributes to a sizeable body of literature investigating the effects of ethnic diversity on different political economy outcomes.[11] While most of the literature is at the cross-country level, there has been a recent surge in the number of studies investigating political economy outcomes at the local level, such as: Alesina et al. (1999) (U.S. cities); Dahlberg et al. (2012) (Swedish municipalities); Munshi and Rosenzweig (2015) (wards in India); Algan et al. (2016) (apartment blocks in France); Montalvo and Reynal-Querol (2017) ($1^\circ \times 1^\circ$ pixels in Africa); and Desmet et al. (2020) (local interaction at the 5 km $\times$ 5 km cell level and public goods outcomes at the national level). In contrast, I focus on individual-level ethnic distances, controlling for ethnic diversity at the local level in addition to a rich set of other controls.[12]

[10] This tests whether the relevant dimension of distance is that between the centre and the periphery, rather than that between the peripheral groups themselves (Desmet et al., 2017). Following Francois et al. (2015) I assign dominance to the most populous group.

[11] See Easterly and Levine (1997), Ashraf and Galor (2013b), Miguel and Gugerty (2005), Habyarimana et al. (2007), and Alesina et al. (2003). See also Desmet et al. (2012), Desmet et al. (2009), and Esteban et al. (2012a,b) for aggregate cross-country diversity measures incorporating ethnic distances.

[12] Another strand of the literature highlights the pernicious effects of ethnic inequality, defined as the inequality in well-being across ethnic groups that coexist, on economic growth (Alesina et al., 2016), public goods (Baldwin and Huber, 2010), and civil conflicts (Mitra and Ray, 2014; Gomes, 2015). This paper shows how ethnic distance may exacerbate ethnic inequality in health outcomes in Africa.

2 Data

This paper aims to estimate the effects of the ethnic distance of a mother from the people living around her on her children's health outcomes. For this purpose, I require the mother's GPS location, her ethnicity, and the spatial distribution of ethnic groups around her. This section explains how the different variables were constructed and discusses their data sources.[13]

2.1 Spatial Distribution of Ethnic Groups

For the spatial distribution of ethnic groups, I rely on a high-resolution global database recently constructed by Desmet et al. (2020). They combine two sources of data using an iterative proportional fitting (IPF) algorithm to construct their final high-resolution (5 km $\times$ 5 km) dataset. For the spatial distribution of population, they rely on the LandScan database. At a resolution of 1 km $\times$ 1 km (approximately at the equator), LandScan is the finest resolution global population distribution database currently available. For information on ethnic groups they use the 17^{th} edition of the Ethnologue database (Lewis et al., 2014), from the World Language Mapping System (WLMS), which maps 6,905 distinct linguistic groups worldwide and is the most comprehensive database on linguistic groups currently available. The linguistic groups are represented in the form of polygons across space, where each polygon represents the traditional homeland of a particular linguistic group. Areas where multiple languages are spoken are represented by overlapping polygons. Ethnologue also provides the total population of each linguistic group within a particular political boundary.[14]

Following Desmet et al. (2020), I combine the above two sources of data using an IPF algorithm to generate a distribution of languages at a 1 km $\times$ 1 km grid-cell level for the 14 countries in my sample. Using the IPF algorithm, which is widely used in statistics, ensures that while allocating languages to cells, the total population of each country, the population of each of the cells, and the population speaking each of the languages in every country add up to precise, consistent totals.[15] Alternative attempts in the literature to generate the spatial distribution of languages, such as that of Matuszeski and Schneider (2006), do not ensure the consistency of population totals. Furthermore, they also disregard languages that are considered widespread

[13] The sample comprises fourteen countries (Appendix Figure M1). See Appendix A.1 for more details.

[14] One alternative to the Ethnologue data would be the Geo-Referencing of Ethnic Groups (GREG) database (Weidmann et al., 2010) based on the Atlas Narodov Mira. However, these data are far less detailed, containing information on only 929 language groups compared to Ethnologue's nearly 7000 groups.

[15] Section A.2 provides details of the IPF algorithm.

by Ethnologue, nor languages for which Ethnologue only provides a point as the location rather than a polygon. Finally, other sources of sub-national data on ethnic diversity including Alesina and Zhuravskaya, 2011 (district level for 92 countries), and Gershman and Rivera, 2018 (around 400 first level administrative regions in 36 countries of sub-Saharan Africa) are not available at the disaggregated cell-level.[16]

2.2 Linguistic Distance

I measure the ethnic distance between any given pair of ethnic groups by the degree of difference between the languages spoken by the two groups. I first match the ethnicity of each mother, which is provided by the DHS, to the unique language spoken by her ethnic group.[17] Then, following a wide stream of papers, I use a distance metric based on the number of branches shared between any two languages from tree diagrams based on the Ethnologue database.[18] The distance between two languages i and v, using this approach is defined as:

$$\tau_{iv} = 1 - \left(\frac{l}{m}\right)^{\delta} \tag{1}$$

where l is the number of shared branches between languages i and v; m is the maximum number of branches between any two languages; and δ is the decay factor, which is a parameter that determines how fast the distance declines as the number of shared branches increases. Not all languages in the Ethnologue database have the same number of branches connecting them to their proto-language. Following the empirical literature, I assume that all languages went through intermediate states, starting from the proto-languages of their respective families, before reaching their current form. This approach adds $m - x$ branches to any language which has a distance of x to its proto-language, before calculating distances.[19]

The decay factor δ measures, "how much more distant [one should] consider two languages from different families to be relative to languages that belong to the same family" (Desmet et al.,

[16] Desmet et al. (2020) demonstrate high correlations of IPF-based diversity measures with the census-based measures of Gershman and Rivera (2018): 0.80 at the region level and 0.95 at the country level.

[17] Appendix A.1 provides the list of countries and DHS surveys used in the paper. Appendix A.3 provides the exact procedure used to map languages to ethnic groups.

[18] See, for instance, Fearon (2003), Desmet et al. (2012), Desmet et al. (2009), Esteban et al. (2012a), Esteban et al. (2012b), Laitin and Ramachandran (2016), and Gershman and Rivera (2018) for a similar approach.

[19] Please refer to Desmet et al. (2012) for a more detailed discussion of the issue. There are other ways of measuring linguistic distances, for instance, using the proportion of cognates in any two languages (Dyen et al., 1992; Isphording, 2014). However, the advantage of using the Ethnologue language trees is that my spatial distribution of languages data are based on the same Ethnologue database.

2009). The literature contains no consensus on what the value of δ should be. While in their empirical exploration Desmet et al. (2009) find that values of δ between 0.04 and 0.10 perform well and select a δ of 0.05, Fearon (2003) uses a δ of 0.5. In the absence of a theoretical basis for choosing one value of δ over another, I examine the data and find that lower values of δ perform better than higher values. I consequently fix δ at 0.0025. As I will later show, choosing a δ of 0.05 like Desmet et al. (2009) leads to qualitatively similar results. However, a high value of δ, like that chosen by Fearon (2003), leads to insignificant results. I discuss the implications of this finding in more detail in Section 4.2.

To understand what the different values of δ imply in practice, consider the two Bantu languages of Gikuyu and Kiembu, both of which are spoken in Kenya. Both belong to the Niger-Congo language family and have the following language family structure: Niger-Congo, Atlantic-Congo, Volta-Congo, Benue-Congo, Bantoid, Southern, Narrow Bantu, Central, E, Kikuyu-Kamba (Lewis et al., 2014). Taking a δ of 0.5 following Fearon (2003), the distance between them is 0.2254. Now consider the distance between Gikuyu and the Nilotic language Dholuo, which is also spoken in Kenya, but belongs to the Nilo-Saharan language family and has the following language family structure: Nilo-Saharan, Eastern Sudanic, Nilotic, Western, Luo, Southern, Luo-Acholi, Luo (Lewis et al., 2014). The distance between them is 1. On the other hand, considering a δ of 0.05 following Desmet et al. (2009), the distance between Gikuyu and Kiembu becomes 0.00252, whereas that between Gikuyu and Dholuo continues to be 1. Finally, choosing a δ of 0.0025 implies a distance of 0.0013 between Gikuyu and Kiembu while that between Gikuyu and Dholuo remains 1.[20]

The final analysis requires calculating the average linguistic distance of each mother in the sample from others living around her. Instead of taking a stand on what region or geographic aggregation should be more appropriate, I calculate linguistic distance by drawing circles of different radii around the mothers. The final spatial distribution dataset from Section 2.1 provides the total population and the shares of individuals belonging to different ethnic groups for every 1 km $\times$ 1 km grid-cell for each of the 14 countries in the dataset. Using the GPS location of each mother provided by DHS, I draw circles around each mother and obtain the estimated population size and ethnicity shares for each circle. For instance, consider an individual living in the south-east corner of Mali, where four different languages are spoken (see Figure 1). In this context, for circles around any mother who resides in the region, the final dataset provides

[20] Appendix Figure M8 provides simulations of how distances between any two languages change as they share different numbers of branches ranging from 0 to 15, for different values of δ.

the total population n, in addition to the numbers p_1 (Mamara Senoufo speakers), p_2 (Northern Bobo Madare speakers), p_3 (Maasina Fulfulde speakers) and p_4 (Bamanankan speakers), where $p_j \geqslant 0 \; \forall j = 1(1)4$; and $\sum_j p_j = n$.

The linguistic distance LD_i, for mother i (who speaks language i) from all other individuals in the circle is given by:

$$LD_i = \frac{1}{n} \sum_{v=1}^{n} \tau_{iv} \tag{2}$$

where n individuals live in the circle and v represents the language groups of each of those n individuals. The function τ_{iv} is defined by formula (1). While the baseline specifications use the average linguistic distance of each mother from her neighbours defined by equation (2) as the main variable of interest, the empirical section explores other alternatives (sections 5.4 and 6.2).[21]

2.3 Linguistic Diversity

I measure ethnic diversity using either an index of ethno-linguistic fractionalization (ELF) (Alesina et al., 2003), or an index of ethno-linguistic polarization (ELP) (Esteban and Ray, 1994, 1999; Montalvo and Reynal-Querol, 2005). ELF gives the probability that two randomly selected individuals from a given region speak two different languages. ELP, on the other hand, measures how far the distribution of the linguistic groups in a given region is from the bipolar distribution (i.e. the (1/2, 0, 0, ... , 0, 1/2) distribution) which represents the highest level of polarization (Montalvo and Reynal-Querol, 2005).

To address the issue of which linguistic groups should be used as primitives in the calculation of the diversity indices, I follow the recent literature (Desmet et al., 2012, 2020; Gershman and Rivera, 2018) and calculate these indices at different levels of aggregation of the Ethnologue language trees. There are 15 possible levels, with Level 1 being the most disaggregated. Formally, the two measures of ELF and ELP, in region j and at linguistic aggregation level k, are defined as follows:

[21] As evident from the Mali example above, multiple linguistic homelands overlap or border each other in the Ethnologue-based spatial distribution data (e.g. Figure 1). This leads to heterogeneity at the local level and subsequently individuals have non-zero linguistic distance from people around them even if they reside in their ethnic homeland. However, some individuals possibly reside in the linguistic homelands of others either because they are themselves migrants or are children of migrants.

$$\text{Fractionalization: } ELF_j^k = 1 - \sum_{i=1}^{n} [s_{i(j)}^k]^2. \tag{3}$$

$$\text{Polarization: } ELP_j^k = 4 \sum_{i=1}^{n} [s_{i(j)}^k]^2 [1 - s_{i(j)}^k]. \tag{4}$$

where $s_{i(j)}^k$ is the proportion of the population speaking language i at linguistic aggregation level k in the geographic region j. As in the case of linguistic distance, I calculate these measures of diversity at the circle level by drawing circles of different radii around the mothers.[22]

2.4 Genetic and Cultural Distance

I measure individual-level genetic distance using the FST index (Spolaore and Wacziarg, 2016) based on the Pemberton et al. (2013) data. Combining eight different datasets covering 645 common microsatellite loci, Pemberton et al. (2013) construct a single dataset on genetic distance for 267 worldwide populations.[23] I map each group from Pemberton et al. (2013) to an Ethnologue group for all of Africa. I linearly interpolate and extrapolate information on genetic distance for missing groups using geographic distance.

I measure cultural distance using the extended version of Murdock's ethnographic atlas, extended by Giuliano and Nunn (2018). First, I extend a mapping of groups from Murdock (1967) to the Ethnologue groups by Giuliano and Nunn (2018) for all ethnic groups in my sample. Then I construct a cultural distance variable measuring the proportion of variables from Murdock's atlas which are different across any two groups.[24]

2.5 Individual-level Data on Health and Other Characteristics

The individual-level child health data are based on the Demographic and Health Surveys (DHS). Funded by the U.S. Agency for International Development (USAID), the DHS has been conducting surveys in developing countries since the 1980s. By interviewing a nationally representative sample of women of childbearing age (15 to 49), the DHS collect data on all the children these

[22] Furthermore, Appendix H.4 calculates distance-weighted measures of diversity inspired by Greenberg (1956), Duclos et al. (2004) and Esteban and Ray (2011).

[23] The other possibility was using the Cavalli-Sforza et al. (1994) dataset. However, the coverage of the Pemberton et al. (2013) dataset is much broader. See discussion in Spolaore and Wacziarg (2016).

[24] See Desmet et al. (2017) for a similar approach.

women have ever given birth to, including those who did not survive until the time of the interview. The standardized components of the DHS questionnaires can be used to compile cross country micro datasets.

2.5.1 Child Mortality

Child mortality is the death of a child under the age of five. If a child dies before reaching one year of age, it is termed infant mortality. If a child fails to survive the first month after its birth, it is termed neonatal mortality. Appendix Figure M6 plots the 28,993 DHS clusters that show the geographic locations of the 208,898 individual mothers whose children's survival outcomes I use in this study. Figure 2 shows the locations of the individual mothers in the case of Mali, along with the 25 km circles around the mothers' locations and the language groups in the background.

2.5.2 Other Health Outcomes

I use a host of additional child- and mother-level variables from the DHS data as either outcome or control variables. Outcome variables include the height-for-age z-score (HAZ), the weight-for-age z-score (WAZ), whether the child is stunted (defined as the child being less than 2 standard deviations of HAZ), immunizations received (polio, DPT, measles, tetanus, BCG, and full immunization), whether the mother received iron tablets during pregnancy, antenatal visits, and if the delivery was done by a doctor or a nurse (i.e. skilled birth attendance). Section 3 provides the full list of control variables.

2.5.3 Migration

The DHS make available a variable that gives the "number of years the respondent has lived in the village, town, or city where she was interviewed." Exploiting this question enables me to determine which individuals have always lived in the DHS cluster where they were interviewed and which individuals have moved there from elsewhere. The migration status variable is available for 13 of the 14 countries and 25 of the 30 surveys used in the study.[25] Of the 208,898 mothers in the sample, the migrant status variable is available for 167,130, of which another 2,822 mothers are identified as temporary visitors rather than residents and consequently dropped from the

[25]The variable is missing from one of the four surveys for Burkina Faso, one of the three surveys for Ethiopia, one of the two surveys for Guinea, one of the two surveys for Senegal, and from the only survey used for Uganda.

sample. Hence, I have information on the migrant status of 164,209 mothers.

2.5.4 Access to Information and Public Goods

The DHS surveys ask each respondent whether they have either heard of or used an oral rehydration product to treat children with diarrhoea. Using the responses to this question, I create a 0-1 binary variable called ORS, which takes the value 1 if the individual has either heard of or used an oral rehydration solution to treat children with diarrhoea, or 0 if not. This question serves as a test for access to health-related knowledge or information.

I also exploit information on whether the respondent's household has access to electricity, whether the household has access to water (defined as requiring less than 30 minutes to reach a water source), the individual's educational attainment, and whether the individual is literate or not. Among these, electricity access, water access, and literacy are binary variables, taking the value 1 if the individual has access to electricity or water or is literate, and 0 if not. Educational attainment is a categorical variable taking the values 0 (no education), 1 (incomplete primary education), 2 (completed primary education), 3 (incomplete secondary education), 4 (completed secondary education), or 5 (higher education). These variables allow me to measure access to public goods in general.

2.6 Malaria Suitability

I measure malaria suitability using a malaria stability index originally constructed by Kiszewski et al. (2004). Their index provides a time-invariant measure of predicted historical malaria exposure. More recently, McCord and Anttila-Hughes (2017) made these data publicly available at a 5 km $\times$ 5 km grid-cell level raster format. I first extract these data for the 14 countries that constitute my sample and then standardize the index to construct a standardized index of malaria suitability for the countries in my sample.[26]

3 Econometric Specification

To measure the effect of maternal linguistic distance (LD) from her neighbours on her children's health outcomes, I estimate the following model as the baseline specification:

$$y_{iet} = \alpha_w + \alpha_r + \alpha_{et} + \alpha_{Rt} + \beta_1 LD_{ie} + \beta_2 ELF_i + \beta_3 X_{it} + \beta_4 X_i + \epsilon_{iet} \quad (5)$$

[26] See Appendix Figure M7 for a spatial representation of these data for the 14 countries in my sample.

where y_{iet} is the mortality outcome of child i born to a mother belonging to ethnicity e in the year t. It is a binary variable, taking the value of 1 if the child dies before reaching the age of five and 0 if the child survives until at least the age of five. In some of the analyses, other child health variables such as infant mortality, neonatal mortality, HAZ, stunting, and WAZ, replace child mortality as the dependent variable.

The LD_{ie} variable is the primary variable of interest and provides the linguistic distance of the mother of child i belonging to ethnicity e from people living within circles of different radii around her (see Section 2.3). The ELF_i variable gives the linguistic fractionalization in the circles of different radii around the mother. ELF_i is later replaced with alternative measures of diversity as part of the robustness checks. To calculate the linguistic distance, and the diversity measures such as ELF, I use circles of different radii, namely 25, 50, 75, 100, 125, 150, 175, 200, and 250 km around the mother.

The variables X_{it} and X_i come from the literature on child mortality and have been found to be important for child mortality.[27] X_{it} includes birth-specific variables, namely a female child dummy, mother's age at birth, mother's age at birth squared, multiple birth indicator, birth order, birth order squared, short birth spacing prior to the birth, and short birth spacing after the birth. X_i includes mother-specific variables, namely the location of the mother in the form of an urban dummy, and dummies for her educational attainment and her family's wealth index.[28] X_i also includes the mother's geographic distance from the capital (to control for isolation) and the logged population in the circle (to control for population density).

The inclusion of the time-varying region fixed effects, α_{Rt}, purge the effects of geographic and environmental advantages of some regions, region-specific shocks such as conflict and natural calamities, and region-specific transfers from the centre that benefit certain ethnic groups at the cost of others (De Luca et al., 2017; Dickens, 2016). Religion-specific fixed effects, represented by α_r, control for differences in religious beliefs and practices among different individuals. The inclusion of the time-varying ethnicity fixed effects, α_{et}, controls for unobserved heterogeneity across ethnic groups. This allows me to identify the effect of ethnic distance on child mortality that is not driven by ethnicity-specific characteristics such as the ethnic dominance of certain groups or cultural differences leading to differences in health practices between different groups. Moreover, since these ethnic group-specific fixed effects are time varying, having a coethnic at

[27] See, for example, Kudamatsu (2009), Baird et al. (2011), and Franck and Rainer (2012).

[28] The wealth index is a categorical variable taking values from 1 (lowest wealth level) to 5 (highest wealth level).

the helm of the country does not affect my results (Kramon and Posner, 2016). Finally, α_w controls for the survey-wave specific fixed effects.

I use a linear probability model to estimate equation (5) and cluster standard errors at the regional level for the 109 regions in the sample.[29] The main coefficient of interest β_1 gives the effect of maternal linguistic distance from her neighbours on the probability of her children dying before reaching the age of five. Due to various possible endogeneity concerns, giving a causal interpretation to β_1 is not straightforward. Furthermore, I cannot use mother-specific fixed effects since the ethnic distance variable does not vary across time for the same mother. However, I am able to control for a host of maternal and birth characteristics, which alleviate endogeneity concerns to a great extent. Moreover, I later use insights from Altonji et al. (2005) and Oster (2017) to demonstrate that the results are not driven by selection on unobservable variables, allowing for a more causal interpretation of β_1.

Appendix Tables A2–A7 provide the descriptive statistics of the variables used in the study. Appendix Table A8 provides the correlations between the aggregate diversity measures of ELF and ELP (at four different levels of aggregation), and the LD variables (for the three alternative values of δ) at the individual mother level. Appendix Table A9 provides the correlation between ELF and ELP at different levels of aggregation. The final sample comprises 14 countries and a total of 30 surveys with information on the births and deaths of over 860,000 children of 206,076 mothers.[30] For the child mortality variable I consider only those children who have already reached the age of five by the day of sampling, since I cannot know whether those younger than five would subsequently survive until the age of five. Hence, the child mortality sample contains information on 654,672 children.[31]

4 Results

4.1 Mother's Ethnic Distance and Child Mortality

Table 1 presents the first set of results. It provides estimates of the effect of maternal linguistic distance from the people living around her on child mortality, while controlling for overall ELF and a host of other variables. I consider a radius of 50 km around the mother to compute the LD and ELF variables, and a decay factor δ of 0.0025 for the LD variable. Column 1 provides a

[29] Appendix I shows that results are robust to clustering standard errors by either ethnicity or both ethnicity and region together instead of region alone.

[30] See Appendix A.1 for a list of countries and DHS surveys used.

[31] I also exclude mothers who are identified as temporary visitors from the sample.

parsimonious specification, controlling only for the survey wave, and time-varying region fixed effects.[32] Column 2 adds ethnic group fixed effects. Column 3 adds individual-level controls listed in Section 3 (also listed in the table notes). Column 4 adds the logged population of the circle and logged distance to the capital. Finally, column 5 adds time-varying ethnic group fixed effects and represents my most complete and hence most preferred specification.[33]

Table 1 demonstrates that LD significantly increases the probability of child death, which effect is robust to a host of controls. ELF, if anything, has a negative effect on child mortality. This implies that, on the one hand, the children of mothers who are linguistically distant from others living around them have a higher mortality rate. On the other hand, the children of mothers living in more linguistically fractionalized localities face lower rates of mortality. However, while LD has a significant effect on child mortality, ELF does not.

In Table 1, I calculate the linguistic distance of the mother from all individuals living in a circle with a 50 km radius around her. Using the complete specification from column 5 in Table 1, Table 2 presents results for alternative radii ranging from 25 km to 250 km. The results remain relatively similar, although the effect size shows a marginal increase for higher radii. Depending on the radius of the circle, a one SD increase in LD increases child mortality by 1.6–2.6% SD. This implies that a one SD increase in LD leads to 6.6–10.5 additional child deaths per 1000 live births. Considering a radius of 50 km, as is the case in Table 1, a one SD increase in LD leads to approximately 8.2 additional deaths per 1000 live births, an SD of approximately 2%.[34]

My analysis includes all births in the entire maternal history of the mother. One possible concern with using retrospective data is recall bias. This stems from the fact that women might be less likely to accurately remember more distant births and deaths. To minimize recall bias I replicate my baseline results using births and deaths occurring in the 10 years preceding the date of the survey (following Baird et al., 2011 and Kudamatsu et al., 2012). The results remain qualitatively similar.[35]

As discussed in Section 2.1, the spatial distribution of linguistic groups is based on an IPF algorithm combining two different datasets. Appendix H.3 computes linguistic distance and

[32] Comparing two women residing in the same DHS cluster but who have different levels of linguistic distance from the majority could be the ideal natural experiment. However, I choose to use region fixed effects rather than DHS cluster fixed effects because of the lack of sufficient variability in the data within each DHS cluster. Appendix Table F5 presents results with DHS-cluster fixed effects.

[33] I use a consistent sample in all columns of Table 1. Allowing for different samples in different columns based on the availability of different variables does not affect the results (see Appendix Table B2).

[34] Appendix Table C1 provides the marginal effects for each of the nine circles of alternative radii.

[35] Results not provided and are available from the author upon request.

ELF using cluster-level information on ethnic groups from the DHS survey data. In particular, Appendix Table H5 shows that using measures of LD and ELF based on data from the DHS cluster in which the mother lives, along with nearby clusters within the specified circular radius around her (50km in this case), yield economically and statistically comparable results. Further, the correlation between the IPF-based LD and the DHS cluster-level LD is 0.8.[36]

4.2 Varying the Decay Factor δ

For Tables 1 and 2, I calculate linguistic distance using a decay factor δ of 0.0025. Appendix Tables B1 and B3, replicate the results from Tables 1 and 2 for three alternative values of δ, namely $\delta = 0.0025$ (Panel 1), $\delta = 0.05$ *à la* Desmet et al. (2009) (Panel 2) and $\delta = 0.50$ *à la* Fearon (2003) (Panel 3). Appendix Tables B1 and B3 demonstrate that the results are a lot more robust for $\delta = 0.0025$ compared to $\delta = 0.05$, which in turn leads to more robust results compared to $\delta = 0.50$. My choice of a lower $\delta = 0.0025$ for the main analysis is based on this result. The following paragraphs explore the possible implications of this finding in more detail.

As explained in Section 2.3, the decay factor δ is a parameter that determines how fast the distance between any two languages declines as the number of shared branches increases. Under lower values of δ, as soon as two languages share a single branch, their distance falls more rapidly than under higher values of δ. However, subsequently, as the number of shared branches increases, the decline in distance is not as drastic, being comparable to higher values of δ even though the actual magnitudes of the distances differ.[37] This implies that my results are driven by the divisions in broad language families. In other words, splits that occurred thousands of years ago are more relevant than more recent splits. This is in line with Desmet et al. (2012), who show that, for explaining civil conflicts, higher values of aggregation matter more than lower values of aggregation of ELF. In the same vein, Fenske et al. (2017) find that the coarse divisions between languages explains market integration in colonial India.[38]

The relevance of lower rather than higher values of δ provides two insights. First, small

[36] For the sake of transparency, Appendix Table H5 also shows results based on using only the cluster-level information. However, as discussed in Appendix H.3 these do not yield high quality measures.

[37] Appendix Figure M8 provides simulations of how distances between any two languages change as they share different numbers of branches ranging from 0 to 15, for different values of δ.

[38] The assumption that coarse linguistic divisions represent splits going back thousands of years follows a long established body of literature (Darwin, 1859; Cavalli-Sforza et al., 1988; Gray and Atkinson, 2003; Belle and Barbujani, 2007; Desmet et al., 2012). Furthermore, language trees such as the ones from Ethnologue were constructed by linguists primarily to capture the time that has passed since the populations speaking these languages split from each other (Desmet et al., 2012). However, the assumption that linguistic division at the highest levels go back thousands of years is not crucial to my results.

differences across dialects of the same language or minor differences in closely related languages do not matter. Second, the issue is not one of the mutual intelligibility of two languages. Languages become unintelligible quite quickly as they move a few branches away from each other in the language tree. Thus, if it were a matter of intelligibility, then varying the δ would not have mattered. Rather, the importance of deeper cleavages arising from splits that occurred thousands of years ago might indicate other potential explanations. I interpret it as a sign of linguistic distance acting as a barrier to information between individuals who are linguistically very distant. I hypothesize that individuals are less willing to pass on information to others belonging to groups that speak languages that split from each other thousands of years ago. It is possible that such individuals belong to different networks and hence do not interact with each other, even if they live in close proximity. It is also possible that even if such individuals interact they are more reluctant to pass on information to each other. While I cannot directly confirm this hypothesis using the DHS data, it is in line with a vast body of literature on ethnic networks that argues that ethnicity imposes a barrier to information transmission and that individuals are less likely to pass on information to persons belonging to other ethnic groups (Larson and Lewis, 2017). Following up on this hypothesis, in Section 6, I demonstrate that ethnically distant individuals are less likely to receive information in an experimental setting.

For the analyses that follow, I always use the most comprehensive specification contained in column 5 of Table 1, fix the decay factor δ at 0.0025 and calculate the LD and ELF variables using a radius of 50 km around the mother, unless otherwise specified.

4.3 Other Health Outcomes

So far, my focus has been on child mortality, or the death of a child before the age of five. Other relevant variables include infant mortality (defined as the child dying before reaching the age of one) and neonatal mortality (the child dying before reaching one month of age). Columns 1 and 2 of Table 3 provide the results for infant and neonatal mortality. In general, the results are similar, with LD significantly increasing both infant and neonatal mortality. ELF continues to have a negative effect on mortality outcomes and is significant at the 10% level for the neonatal mortality variable.[39]

Columns 3–5 of Table 3 investigate the impact of LD on the child's HAZ, the probability of the child being stunted, and the child's WAZ. The results demonstrate that linguistic distance

[39]In results not provided, I find that the ELF variable is not robust to changing the circle radius. These results are available from the author upon request.

has a strong and significant effect on child height, whether measured by HAZ or the stunting status of the child. LD also reduces child weight as measured by WAZ, but the effect is not statistically significant. ELF continues to have a benign effect on the different variables and significantly improves WAZ.

Appendix Table D1 investigates whether the mother received tetanus injections during pregnancy, whether the child received measles immunization, polio vaccination, or DPT immunization, and whether the mother received iron tablets during pregnancy. Among these variables, LD significantly (at the 10% level) reduces the probability of the mother taking iron tablets during pregnancy. LD does not have a significant effect on any of the other variables.[40]

4.4 Heterogeneous Effects

4.4.1 Migration

A possible concern in estimating the effects of ethnic distance on child mortality is spatial sorting. If individuals realize that being linguistically distant to one's neighbours is a disadvantage, they might try to sort themselves into neighbourhoods where they are less linguistically distant from others. Given various barriers to movement (e.g., transportation costs), perfect sorting is not observed in reality. Rather, in spite of population movements, ethnic populations tend to reside in their respective historical homelands (Michalopoulos and Papaioannou, 2014). Even in the face of large-scale population displacements caused by civil wars, individuals try to return to their historical ethnic homelands (Glennerster et al., 2013).[41] However, if individuals are actually able to move to places where they are less distant from others, then, if anything, I am underestimating the effects of ethnic distance on child mortality. This is borne out by the data.

Table 6 investigates the heterogeneity of the results by migrant status. Column 1 of Table 6 first uses a 0–1 binary variable indicating migrant status as the dependent variable. It shows that being a migrant reduces the effect of LD on child mortality. In other words, the effects of being linguistically distant are worse for children of mothers who have never moved from their village of birth. Column 2 directly checks for heterogeneity by the continuous variable

[40] Please refer to Appendix D.3 for additional variables. Appendix L reports p-values adjusted for multiple comparisons following seven alternative methods.

[41] Almost 55% of the Afrobarometer Survey respondents lived in their ethnic group's ancestral homeland at the time of the survey (Nunn and Wantchekon, 2009). Again, Gershman and Rivera (2018) show how sub-national ethnic diversity is stable across several decades in sub-Saharan Africa. More importantly, they find that changes in diversity at the sub-national level are not correlated with changes in economic conditions (Gershman and Rivera, 2018).

that measures how many years the mother has lived in her current village of residence. This again shows that the effects of LD on child mortality are much stronger for individuals who have resided in their current village of residence for longer.[42]

Finally, columns 3 and 4 restrict the sample to individuals who have moved and individuals who have never moved from their village of residence, respectively. These two columns demonstrate that the results are driven by non-migrants rather than migrants. The coefficient on LD is much bigger and more statistically significant for the non-migrant sample. For a circle with a radius of 50 km around the mother, a one SD increase in LD leads to around 14 additional child deaths per 1000 live births, or an approximately 3.3% SD in the non-migrant sample. This stands in sharp contrast to the approximately four additional child deaths per 1000 live births (around 1% SD) in the migrant sample for a similar one SD increase in LD. The corresponding figures for the full sample are 8.2 deaths per 1000 live births, which is 2% of the SD.[43]

The two panels of Appendix Tables D2 and D3 provide the differences in the effects of LD on other variables by splitting the sample by migrants and non-migrants. In particular, Table D2 shows the impact of LD on infant mortality, neonatal mortality, the HAZ, whether the child is stunted, and the WAZ. Table D3 examines the effects of LD on whether the mother received tetanus injections during pregnancy, whether the child received measles immunization, polio vaccination, or DPT immunization, and finally whether the mother received iron tablets during her pregnancy. I find that LD consistently worsens the various health outcomes in the non-migrant sample, rather than the migrant sample. These results indicate that linguistic distance has a more detrimental effect on the health outcomes of mothers who have never moved from their village of birth, and that my results are driven by the non-migrant sample.

It is conceivable that migrants choose to relocate to places where they are less linguistically distant from others. Hence, the migrant sample might, in general, have a lower average linguistic distance from their neighbours than the non-migrant sample. Appendix Table D5 investigates the correlates of migrant status and shows the opposite: migrants have a higher linguistic distance from their neighbours than non-migrants.[44] Clearly, if anything, migration biases my results away from zero; without migration my results would have been much stronger.

[42] Appendix Table G4 shows that the heterogeneity by years lived in village of residence possibly picks up the effect of an individual's age. While the heterogeneity by migrant status remains robust, the heterogeneity by years lived disappears when controlling for an interaction of LD with age. Furthermore, the pernicious effects of LD are further exacerbated for older individuals in both the migrant and non-migrant samples.

[43] For circles of alternative radii ranging from 25 km to 250 km, a one SD increase in LD leads to approximately 11.5–19.2 additional child deaths per 1000 live births in the non-migrant sample (3.3–4.2 additional deaths in the migrant sample). Please refer to Appendix Table C1 for more details.

[44] Moreover, migrants tend to be more concentrated in urban areas, and are wealthier.

4.4.2 Malaria Suitability

In recent research, Cervellati et al. (2016) argue that ancestral exposure to malaria might have increased the benefits of isolation and thereby encouraging interaction in small groups. This in turn might have reinforced ethnicities in Africa leading to the persistence of ethnic diversity in the continent. This section investigates the implications of these findings for the current study.

First, column 1 of Table 7 shows that the results are unaffected by adding a control for malaria suitability.This rules out that the LD variable captures the effects of malaria suitability, which can itself directly affect child mortality. Next, column 2 unearths evidence of heterogeneity in the effects of LD on child mortality by malaria suitability, but not in the direction suggested by Cervellati et al. (2016). Following their hypothesis, one might expect some positive effects of being linguistically distant in the presence of malaria. However, Table 7 uncovers evidence that malaria suitability exacerbates the pernicious effects of linguistic distance on child mortality. A one SD increase in LD when malaria suitability is one SD above its average value leads to 25 additional child deaths compared to the eight additional child deaths for the average level of malaria suitability. The corresponding number for the non-migrant sample is approximately 34 additional child deaths.[45]

Malaria might be correlated with other variables such as Tse Tse suitability. Appendix J, however, establishes the robustness of heterogeneity by malaria suitability to the inclusion of controls for LD $\times$ Tse Tse suitability, LD $\times$ crop suitability, LD $\times$ Population and LD $\times$ Urban Residence in the same specification. Again, if child mortality is greater in malaria-prone areas then the effect of LD might be driven by the larger variance of the dependent variable in these areas. Column 1 of Table 7 shows that child mortality is not significantly higher in areas with higher levels of malaria suitability, allaying such concerns.

The results are not necessarily contradictory to the findings of Cervellati et al. (2016). Historically there might have been some advantages to interacting in small groups in malaria-prone areas. However, in present times, isolation might impede access to modern medical practices for treating and preventing malaria.[46] In particular, linguistically distant mothers might not have access to best-practice information about malaria, leading to higher child mortality among them in malaria-prone areas.[47] Furthermore, 57% of child deaths from malaria are due to

[45] Appendix Table C2 provides the full set of marginal effects and standardized βs.

[46] The Kiszewski et al. (2004) measure is a time-invariant measure of malaria suitability (see Section 2.6). Areas where malaria was more prevalent in the past are also areas where malaria is more prevalent in the present.

[47] Appendix Table J4 shows that linguistically distant individuals living in malaria endemic areas are more likely

under-nutrition (Bryce et al., 2005). Lack of knowledge on nutrition among linguistically distant mothers can exacerbate the negative effects of malaria by worsening nutritional outcomes for their children.[48] Hence, while there might have been some benefits from ancestral exposure to malaria for linguistically distant mothers, in present circumstances they seem to be at a disadvantage.

4.4.3 Heterogeneity by other variables

Appendix Table G1 explores the possibility of heterogeneity in the effects of linguistic distance by other observable variables such as child's gender, place of residence (urban or rural), mother's educational attainment, ELF, ELP, population, distance from the capital, and wealth. I find no evidence of heterogeneity by any of the aforementioned variables.

5 Robustness Checks and Alternative Explanations

5.1 Are the Results Driven by Ethnic Diversity?

Ethnic diversity, which is usually measured by ethnolinguistic fractionalization (ELF), has often been found to have a negative effect on different socio-economic outcomes.[49] In contrast to the LD variable, the circle-level ELF variable seems to have a more benign effect on health outcomes. However, the effect is almost never statistically significant. Recent literature has underscored the importance of the level at which the linguistic groups enter the ELF calculations (Desmet et al., 2012). In order to incorporate this insight, I follow the recent literature (Desmet et al., 2012, 2020; Gershman and Rivera, 2018), and calculate ELF at different levels of aggregation based on the Ethnologue language trees, with 15 possible levels.To avoid making arbitrary decisions about the appropriate level of aggregation, I consider a range of levels of aggregation, including Levels 15, 10, 5, and 2. This ensures that I have a high level of aggregation (given by Level 2), a medium level of aggregation (given by Level 5), and a lower level of aggregation (given by Level 10). I also include results for the most disaggregated level of ELF (given by Level 15), which is also the basic ELF used in the previous tables.

to possess and use bednets. Hence, the lack of access or use of bednets is unlikely to be the channel. However, Appendix Table K3 finds linguistically distant mothers (non-migrant sample) to be less likely to be able to decide by themselves whether their child should be taken for medical treatment when child is seriously ill. This could be a potential channel through which malaria affects such individuals more adversely.

[48] Table 3 shows that linguistic distance worsens nutrition-dependent anthropometric outcomes such as stunting.

[49] See, for example, Alesina et al. (2003), Alesina et al. (1999), and Easterly and Levine (1997).

Columns 1–4 of Panel 1 of Table 4 show that the results do not change regardless of the level of aggregation at which ELF enters the specification. Column 5 shows that ELF in general is not significant, even if I do not include LD in the specification. However, LD continues to be significant, regardless of the level of aggregation at which ELF is calculated, and its absolute magnitude barely changes. Column 6 includes a quadratic term for ELF, following Ashraf and Galor (2013b) and Gomes (2019), who argue that diversity has a hump shaped effect on economic development. LD continues to have a significant effect on child mortality, whereas ELF does not have any significant effects. Finally, column 7 shows that the results are qualitatively unchanged even if I do not control for ELF.

While ELF has traditionally been used to measure ethnic diversity, some papers have highlighted the relevance of ethnic polarization (ELP) rather than fractionalization, particularly in the context of intergroup conflict (Montalvo and Reynal-Querol, 2005).[50] Panel 2 of Table 4 reruns the estimations described for Panel 1, but using ELP at different levels of aggregation instead of ELF. Columns 1–4 control for ELP at different levels of aggregation. Column 5 includes ELP by omitting LD. Column 6 includes a quadratic term for ELP and, finally, column 7 includes a specification controlling for both ELF and ELP together (following Montalvo and Reynal-Querol, 2005). LD continues to have a significant and robust effect on child mortality.[51]

5.2 Are the Results Explained by Ethnic favouritism?

Kramon and Posner (2016) show that having a coethnic as president during one's school-age years leads to better schooling outcomes for children. Franck and Rainer (2012) provide evidence of similar ethnic favouritism for the educational and child mortality outcomes of ethnic groups in 18 sub-Saharan African countries. The inclusion of time-varying ethnicity fixed effects rules out the possibility of such ethnic favouritism driving the results in the current context.[52] The ethnicity fixed effects also purge any possible effects of within-group genetic diversity (Arbatli et al., 2015).

The recent literature has also discussed region-specific transfers from the centre that benefit certain ethnic groups at the cost of others (De Luca et al., 2017; Dickens, 2016). Burgess et al. (2015) show that, during less democratic periods in Kenya, there is ethnic favouritism in road

[50] Appendix Table A9 demonstrates a high correlation between ELF and ELP.

[51] Appendix H.4 demonstrates robustness to controlling for linguistic distance-weighted ELF and ELP measures following Greenberg (1956), Duclos et al. (2004), and Esteban and Ray (2011).

[52] This is line with Kudamatsu (2009), who found no evidence of ethnic favouritism on infant mortality in Guinea.

building in regions that share the ethnicity of the president. I include region-specific year fixed effects in all specifications to rule out the confounding effects of such region-specific transfers.

5.3 Genetic and Cultural Distance

Table 5 investigates whether the results are about language per se, or whether linguistic distance is a proxy for broader cultural and genetic distances. Table 5 controls for cultural distance and genetic distance entering either separately or together in the specification with linguistic distance. While linguistic distance continues to have a significant effect on child mortality, neither cultural distance nor genetic distance has any significant effects on child mortality regardless of whether linguistic distance enters the specification. In specifications that include either one or both of cultural and genetic distance together with linguistic distance, linguistic distance clearly dominates cultural and genetic distance.[53]

5.4 Distance from the Dominant Group

My LD variable (equation (2)) measures the average linguistic distance of an individual from all other individuals living around her in circles of different radii. If information were indeed the channel, then this approach implicitly assumes that the best-practice knowledge about health resides randomly in the population in some group(s). Hence, the higher an individual's average distance is from the population around her, the less likely she is to have access to the best-practice information. Another equally plausible alternative is that the best-practice information resides with the dominant group. In this case, the linguistic distance from the dominant group would be the relevant metric.[54] In order to explore this second possibility, Appendix E calculates a variable measuring distance from the dominant group (DD) and compares it with the LD variable. In line with the findings of Francois et al. (2015), who show that political power in Africa is proportional to group size, I assign dominance to the group with the largest size within the circle. While my results using DD are similar to those using LD, I find LD to be both statistically and economically more significant than DD.[55]

[53] These results are consistent with Fenske et al. (2017), who also find that including genetic distance in the same specification as linguistic distance does not remove the effects of linguistic distance.

[54] In this case the relevant dimension of distance would be between the centre and the periphery, but not between the peripheral groups themselves (Desmet et al., 2017).

[55] Given the conceptual appeal of the DD measure, Appendix E replicates all the main tables from the paper using DD instead of LD.

5.5 Selection on Unobservables

My identification strategy relies on controlling for a rich set of observable control variables and fixed effects. In order to understand how selection on unobservable variables might be driving my results, I turn to the methodology developed by Altonji et al. (2005) and Bellows and Miguel (2009), who present new estimation strategies that can be used when strong prior information regarding the exogeneity of the variable of interest is unavailable. Following their heuristics, I check for coefficient stability while moving from a specification with a parsimonious set of controls to the full set of controls.

I find that the coefficients become substantially larger when controlling for more observables, which implies that selection on unobservables pushes the estimates away from zero. Following Oster (2017), I also verify that the R^2 becomes substantially larger when moving from the restricted to unrestricted regressions. See, for instance, the movement in the coefficient for LD and R^2 while moving from columns 1 to 5 in Table 1. Hence, if I could have controlled for the unobserved variables that might bias my results, my estimated beta coefficients would have become much larger and my results would have been further strengthened.[56]

6 Channels

6.1 Linguistic Distance as a Barrier to Information

One possible explanation for why linguistic distance worsens health outcomes could be that it acts as a barrier to health-related information. For instance, linguistically distant mothers might not receive the information on best practices about how to rear their children, perhaps due to lack of communication with groups who are very different to them. Another, though not the only other, possibility is that linguistically distant mothers have worse access to public goods in general, arising from, for instance discrimination, which harms their children's health. This section provides some evidence in favour of the former hypothesis.

In order to understand whether linguistic distance acts as a barrier to information, I exploit the DHS question about whether the respondent has heard of the oral rehydration product (ORS) for treating children with diarrhoea. Diarrhoea is the second leading cause of child mortality causing 1.9 million child deaths every year.[57] Oral rehydration therapy is the cornerstone of

[56] The full set of results from this section are not provided and are available from the author upon request.
[57] See Rehydration Project (23/04/2014) and WHO (02/05/2017).

treatment for diarrhoea (Victora et al., 2000). To measure access to public goods in general, I exploit information on access to four different public goods: access to electricity, access to water, the individual's educational attainment, and whether the individual is literate or not. For instance, lower levels of literacy or educational attainment among linguistically distant mothers might indicate lower levels of access to schools.[58]

The previous sections uncovered evidence of stronger effects for non-migrants. Moreover, given the possibility that migrants might have acquired their knowledge on ORS in a location other than the one in which they currently reside, I split the sample by migrant status. Table 8 examines whether linguistic distance impedes educational attainment, literacy, access to water, electricity, and, finally, knowledge about ORS. The two panels present results for the sample split by migrant status. The LD variable does not have a significant effect on any of the variables, except for the ORS variable in the non-migrant sample. Hence, while LD does not impede general access to public goods, it poses a barrier to information about ORS, in particular for individuals who have never moved from their place of birth.[59]

The above findings suggest that individuals who are linguistically distant from others living around them have less access to information, leading to higher rates of mortality for their children. However, linguistically distant individuals do not necessarily face lower levels of access to public goods in general.[60] Furthermore, my results are driven by individuals who have never moved from their place of residence rather than individuals who have migrated from their place of birth. This further supports my interpretation of access to information being the channel. If ethnic distance is a barrier to knowledge and information about how to take care of one's children, it is important to understand where individuals might acquire such information. If individuals have moved from some other place to their current place of residence, then it is likely that they have already acquired such information elsewhere. Hence, linguistic distance in the place of their current residence would do little to affect their children's health outcomes unless LD affected discrimination in general rather than imposing information barriers.[61]

[58] These variables are outcome variables and not direct measures of public good provision. Hence, these results are amenable to other interpretations.

[59] Appendix K.3 examines the effects of linguistic distance on additional variables: hand washing behaviour (i.e. whether the respondent washed their hands before preparing their previous meal), knowledge about when to seek medical treatment, knowledge about where to seek medical treatment, knowledge about the curability of tuberculosis. I find that linguistic distance reduces the probability that individuals: washed their hands before making their last meal; and know when to seek medical help by themselves, in the non-migrant sample.

[60] Since region-specific transfers might benefit certain ethnic groups (Burgess et al., 2015; Dickens, 2016), there might exist more across-region variation in ethnic group specific public good access than within-region variation. However, removing region fixed effects does not change the primary results from this section (Appendix F.1).

[61] Appendix G.2 shows that within the migrant sample the pernicious effects of LD are not dependent on either how long the migrants have lived in their current place of residence or the age at which they migrated, even

Next, to investigate further the information barriers faced by linguistically-different minorities, I use survey data from Ugandan villages collected by Larson and Lewis (2017). They establish that ethnicity is a barrier to information and that individuals are less likely to pass on information to persons belonging to other ethnic groups. They use an experimental setting that seeds identical information in two different villages of Uganda. "The seeded information was that in three days an event would be held at which all adults in attendance would receive a valuable block of soap in exchange for taking a survey" Larson and Lewis (2017). Following up on their results, I use their individual-level data and show that individuals who are linguistically distant from their neighbours are less likely to have heard about the advertised event (Appendix Table K1). This lends further support to the information channel theory.

6.2 The Role of Kin Networks and Resource Sharing

Above, I have argued that higher linguistic distance acts as a barrier to health-related information. Alternatively, rather than information barriers, linguistic distance could be capturing the role of kin networks and resource sharing within close-knit ethnic communities. I investigate this possibility by using the share or the count of other coethnic mothers or other coethnics in general within the radius of interest. Appendix Table H1 shows that variables measuring the presence of coethnics do not have any statistically significant effects on child mortality. Furthermore, controlling for these variables does not affect the coefficients of my linguistic distance variable, nor is there evidence of heterogeneity by these variables.[62] Hence, overall the evidence points more towards LD representing information barriers, rather than resource sharing within close-knit ethnic communities.[63]

6.3 The Role of Selection

Selection can provide another possible explanation of why linguistic distance worsens health outcomes. Women who are more linguistically distant to their neighbours and have children might be different from those who are linguistically more distant and do not have children. Appendix Table K2 tests this possibility by running regressions with mother-level characteristics

though older migrants are less likely to know about ORS and face higher child mortality rates.

[62] Appendix Table H1 further shows that the results are similar using the geographic distance from coethnics instead of the number or proportion of coethnics.

[63] In the same vein, Appendix Table H2 demonstrates that using a continuous LD measure yields more significant results compared to a 0-1 binary measure. This highlights that the intensive margin is more relevant than the extensive margin.

on the left hand side (in the spirit of Buckles and Hungerman, 2013). I do not find evidence of selection.

7 Conclusion

Child mortality rates are still unacceptably high, particularly in sub-Saharan Africa. Nineteen thousand children die worldwide every day before reaching the age of five. The highest rates of child mortality are still concentrated in sub-Saharan Africa, where one in every nine children die before reaching the age of five. This is not only more than 16 times the average for developed regions (1 in 152) but also substantially higher than in South Asia (1 in 16), which has the second-highest rates of child mortality (UNICEF, 2012). Not surprisingly, reducing child mortality was part of the Millennium Developmental Goals and is currently part of the Sustainable Development Goals.

In this paper, I created a high-quality individual-level micro database from the Demographic and Health Surveys and combined it with a novel dataset on the spatial distribution of ethnic groups at the level of approximately 1 km $\times$ 1 km for 14 sub-Saharan African countries. I mapped individual-level ethnicities to languages and calculated how ethnically distant an individual is from her neighbours. Subsequently, I demonstrated that the children of mothers who are ethnically distant from their neighbours face a higher probability of dying before reaching the age of five and that those who survive are shorter in stature. Further, I demonstrated that children of ethnically distant mothers fare even worse in malaria-endemic areas and are less likely to be aware of oral rehydration therapy, which can be of critical importance to their children. Finally, using experimental data, I established that information is less likely to flow to linguistically distant individuals more generally.

One clear policy implication from my paper is that, in order to reduce child mortality rates in Africa, policy-makers need to target ethnic minorities, who may be losing out solely because they speak a language distant from that spoken by their neighbours. Ensuring the dissemination of health information to ethnic minorities, who currently appear to not have access to such information, could help to achieve the Sustainable Development Goal of reducing child mortality.

References

Alesina, A., R. Bakir, and W. Easterly (1999). Public goods and ethnic divisions. *The Quarterly Journal of Economics, MIT Press 114(4) November*, 1243–1284.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003). Fractionalization. *Journal of Economic Growth 8, no. 2, June*, 155–194.

Alesina, A. and E. L. Ferrara (2005). Ethnic diversity and economic performance. *Journal of economic literature 43*(3), 762–800.

Alesina, A., S. Michalopoulos, and E. Papaioannou (2016). Ethnic inequality. *Journal of Political Economy 124*(2), 428–488.

Alesina, A. F. and E. Zhuravskaya (2011). Segregation and the quality of government in a cross section of countries. *American Economic Review vol. 101(5), August.*

Algan, Y., C. Hémet, and D. D. Laitin (2016). The social effects of ethnic diversity at the local level: A natural experiment with exogenous residential allocation. *Journal of Political Economy 124*(3), 696–733.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy Vol. 113, No. 1, February*, 151–184.

Arbatlı, C. E., Q. H. Ashraf, O. Galor, and M. Klemp (2015). Diversity and conflict. *NBER*.

Armand, A., P. Atwell, and J. Gomes (2017). The reach of radio: Ending civil conflict through rebel demobilization. *HICN Working Paper*.

Ashraf, Q. and O. Galor (2013a). Genetic diversity and the origins of cultural fragmentation. *American Economic Review 103*(3), 528–33.

Ashraf, Q. and O. Galor (2013b). The 'out of africa' hypothesis, human genetic diversity, and comparative economic development. *American Economic Review 103*(1), 1–46.

Ashraf, Q. H. and O. Galor (2018). The macrogenoeconomics of comparative development. *Journal of Economic Literature 56*(3), 1119–55.

Baird, S., J. Friedman, and N. Schady (2011). Aggregate income shocks and infant mortality in the developing world. *Review of Economics and Statistics 93*(3), 847–856.

Baldwin, K. and J. D. Huber (2010). Ecnomic versus cultural differences: Forms of ethnic diversity and public good provision. *American Political Science Review 104, No. 4, November*.

Belle, E. M. and G. Barbujani (2007). Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on dna diversity. *American journal of physical anthropology 133*(4), 1137–1146.

Bellows, J. and E. Miguel (2009). War and local collective action in sierra leone. *Journal of Public Economics 93*, 1144—-1157.

Bischoff, A. and K. Denhaerynck (2010). What do language barriers cost? an exploratory study among asylum seekers in switzerland. *BMC Health Services Research 10*(1), 248.

Bishop, Y., S. Fienberg, and P. Holland (1975). Discrete multivariate analysis: Theory and practicemit press. *Cambridge, Massachusetts*.

Bryce, J., C. Boschi-Pinto, K. Shibuya, R. E. Black, W. C. H. E. R. Group, et al. (2005). Who estimates of the causes of death in children. *The Lancet 365*(9465), 1147–1152.

Buckles, K. S. and D. M. Hungerman (2013). Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics 95*(3), 711–724.

Burgess, R., R. Jedwab, E. Miguel, A. Morjaria, et al. (2015). The value of democracy: evidence from road building in kenya. *The American Economic Review 105*(6), 1817–1851.

Cavalli-Sforza, L. L., L. Cavalli-Sforza, P. Menozzi, and A. Piazza (1994). *The history and geography of human genes*. Princeton university press.

Cavalli-Sforza, L. L., A. Piazza, P. Menozzi, and J. Mountain (1988). Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences 85*(16), 6002–6006.

Cervellati, M., G. Chiovelli, and E. Esposito (2016). Bite and divide: Ancestral exposure to malaria and the emergence and persistence of ethnic diversity in africa. Technical report, mimeo, University of Bologna.

Dahlberg, M., K. Edmark, and H. Lundqvist (2012). Ethnic diversity and preferences for redistribution. *Journal of Political Economy 120*(1), 41–76.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection Or the Preservation of Favoured Races in the Struggle for Life.* H. Milford; Oxford University Press.

De Luca, G., R. Hodler, P. A. Raschky, and M. Valsecchi (2017). Ethnic favoritism: An axiom of politics? *Journal of Development Economics*.

Deming, W. E. and F. F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics 11*(4), 427–444.

Desmet, K., J. Gomes, and I. Ortuño (2020). The geography of linguistic diversity and the provision of public goods. *Journal of Development Economics 143 (2020) 102384*.

Desmet, K., I. Ortuno-Ortin, and R. Wacziarg (2012). The political economy of linguistic cleavages. *Journal of Development Economics 97*, 322–338.

Desmet, K., I. Ortuno-Ortin, and I. Weber (2009). Linguistic diversity and redistribution. *Journal of European Economic Association 7(6)*, 1291–1318.

Desmet, K., I. Ortuño-Ortín, and S. Weber (2017). Peripheral diversity: transfers versus public goods. *Social Choice and Welfare 49*(3-4), 787–823.

Desmet, K., I. Ortuño-Ortín, and R. Wacziarg (2017). Culture, ethnicity, and diversity. *American Economic Review 107*(9), 2479–2513.

Dickens, A. (2016). Ethnolinguistic favoritism in african politics. *American Economic Journal: Applied Economics*.

Duclos, J.-Y., J. Esteban, and D. Ray (2004). Polarization: concepts, measurement, estimation. *Econometrica 72*(6), 1737–1772.

Dyen, I., J. B. Kruskal, and P. Black (1992). An indo-european classification, a lexicostatistical experiment. 1. *Transactions of the American Philosophical Society 82*, 1–132.

Easterly, W. and R. Levine (1997). Africa's growth tragedy: Policies and ethnic divisions. *Quarterly Journal of Economics 112, no.4 November*, 1203–1250.

Esteban, J., L. Mayoral, and D. Ray (2012a). Ethnicity and conflict: An empirical study. *American Economic Review 102, No.4*, 1310–1342.

Esteban, J., L. Mayoral, and D. Ray (2012b). Ethnicity and conflict: Theory and facts. *Science 336, 858*.

Esteban, J. and D. Ray (1999). Conflict and distribution. *Journal of Economic Theory 87(2)*, 379–415.

Esteban, J. and D. Ray (2011). Linking conflict to inequality and polarization. *American Economic Review 101*(4), 1345–74.

Esteban, J.-M. and D. Ray (1994). On the measurement of polarization. *Econometrica: Journal of the Econometric Society*, 819–851.

Fafchamps, M. (2000). Ethnicity and credit in african manufacturing. *Journal of Development economics 61*(1), 205–235.

Fearon, J. and D. Laitin (1996). Explaining interethnic cooperation. *American Political Science Review 90 (4)*, 715–35.

Fearon, J. D. (2003). Ethnic and cultural diversity by country. *Journal of Economic Growth 8*(2), 195–222.

Fenske, J., N. Kala, et al. (2017). Linguistic distance and market integration in india. Technical report, Competitive Advantage in the Global Economy (CAGE).

Fienberg, S. E. (1970). An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 907–917.

Fisman, R., D. Paravisini, and V. Vig (2017). Cultural proximity and loan outcomes. *American Economic Review 107*(2), 457–92.

Flores, G. (2006). Language barriers to health care in the united states. *New England Journal of Medicine 355*(3), 229–231.

Franck, R. and I. Rainer (2012). Does the leader's ethnicity matter? ethnic favoritism, education and health in sub-saharan africa. *American Political Science Review 106*(2, May).

Francois, P., I. Rainer, and F. Trebbi (2015). How is power shared in africa? *Econometrica 83*(2), 465–503.

Gershman, B. and D. Rivera (2018). Subnational diversity in sub-saharan africa: Insights from a new dataset. *Journal of Development Economics, Elsevier 133*, 231–263.

Giuliano, P. and N. Nunn (2018). Ancestral characteristics of modern populations. *Economic History of Developing Regions 33*(1), 1–17.

Glennerster, R., E. Miguel, and A. D. Rothenberg (2013). Collective action in diverse sierra leone communities. *The Economic Journal 123*(568), 285–316.

Gomes, J. F. (2015). The political economy of the maoist conflict in india: an empirical analysis. *World Development 68*, 96–123.

Gomes, J. F. (2019). Linguistic fractionalization and health information in sub-saharan africa. *The World Bank Economic Review*.

Gray, R. D. and Q. D. Atkinson (2003). Language-tree divergence times support the anatolian theory of indo-european origin. *Nature 426*(6965), 435.

Greenberg, J. H. (1956). The measurement of linguistic diversity. *Language 32*(1), 109–115.

Guiso, L., P. Sapienza, and L. Zingales (2009). Cultural biases in economic exchange? *The Quarterly Journal of Economics 124*(3), 1095–1131.

Habyarimana, J., M. Humphreys, D. N. Posner, and J. M. Weinstein (2007). Why does ethnic diversity undermine public goods provision? *American Political Science Review 101*(04), 709–725.

Isphording, I. E. (2014). Disadvantages of linguistic origin—evidence from immigrant literacy scores. *Economics Letters 123*(2), 236–239.

Isphording, I. E. and S. Otten (2013). The costs of babylon—linguistic distance in applied economics. *Review of International Economics 21*(2), 354–369.

Kiszewski, A., A. Mellinger, A. Spielman, P. Malaney, S. E. Sachs, and J. Sachs (2004). A global index representing the stability of malaria transmission. *The American journal of tropical medicine and hygiene 70*(5), 486–498.

Kramon, E. and D. N. Posner (2016). Ethnic favoritism in education in kenya. *Quarterly Journal of Political Science 11*(1).

Kudamatsu, M. (2009). *Ethnic Favoritism: Micro Evidence from Guinea.* unpublished.

Kudamatsu, M., T. Persson, and D. Strmberg (2012). Weather and infant mortality in africa. *CEPR Discussion Paper No. DP9222*.

Laitin, D. D. and R. Ramachandran (2016). Language policy and human development. *American Political Science Review 110*(3), 457–480.

Larson, J. M. and J. I. Lewis (2017). Ethnic networks. *American Journal of Political Science 61*(2), 350–364.

Larson, J. M. and J. I. Lewis (2018). Rumors, kinship networks, and rebel group formation. *International Organization 72*(4), 871–903.

Lewis, M. P., G. F. Simons, C. D. Fennig, et al. (2014). *Ethnologue: Languages of the world*, Volume 17. SIL international Dallas, TX.

Malhotra, N. (2012). Inadequate feeding of infant and young children in india: lack of nutritional information or food affordability? *Public Health Nutrition 1*(1), 1–9.

Matuszeski, J. and F. Schneider (2006). Patterns of ethnic group segregation and civil conflict. *unpublished, Harvard University*.

McCord, G. C. and J. K. Anttila-Hughes (2017). A malaria ecology index predicted spatial and temporal variation of malaria burden and efficacy of antimalarial interventions based on african serological data. *The American journal of tropical medicine and hygiene 96*(3), 616–623.

Michalopoulos, S. and E. Papaioannou (2014). National institutions and subnational development in africa. *The Quarterly Journal of Economics 129*(1), 151–213.

Miguel, E. and M. K. Gugerty (2005). Ethnic diversity, social sanctions, and public goods in kenya. *Journal of Public Economics 89*(11), 2325–2368.

Mitra, A. and D. Ray (2014). Implications of an economic theory of conflict: Hindu-muslim violence in india. *Journal of Political Economy 122*(4), 719–765.

Montalvo, J. and M. Reynal-Querol (2005). Ethnic polarization, potential conflict and civil war. *The American Economic Review 95(3) June*, 796–816.

Montalvo, J. G. and M. Reynal-Querol (2017). Ethnic diversity and growth: Revisiting the evidence. *Barcelona GSE Working Paper Series No. 992*.

Munshi, K. and M. Rosenzweig (2015). Insiders and outsiders: local ethnic politics and public goods provision. Technical report, National Bureau of Economic Research.

Murdock, G. P. (1967). Ethnographic atlas: a summary. *Ethnology 6*(2), 109–236.

Newson, R. B. (2010). Frequentist q-values for multiple-test procedures. *The Stata Journal 10*(4), 568–584.

Nunn, N. and L. Wantchekon (2009). The slave trade and the origins of mistrust in africa. *American Economic Review*.

Oster, E. (2017). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 1–18.

Pemberton, T. J., M. DeGiorgio, and N. A. Rosenberg (2013). Population structure in a comprehensive genomic data set on human microsatellite variation. *G3: Genes, Genomes, Genetics 3*(5), 891–907.

Pongou, R. (2009). Anonymity and infidelity: Ethnic identity, strategic cross-ethnic sexual network formation, and hiv/aids in africa. *Unpublished paper, Department of Economics, Brown University*.

Romani, M. (2004). Love thy neighbour? evidence from ethnic discrimination in information sharing within villages. Technical report, Research Paper, UNU-WIDER, United Nations University (UNU).

Shastry, G. K. (2012). Human capital response to globalization education and information technology in india. *Journal of Human Resources 47*(2), 287–330.

Singleton, K. and E. Krause (2009). Understanding cultural and linguistic barriers to health literacy. *OJIN: The online journal of issues in nursing 14*(3).

Spolaore, E. and R. Wacziarg (2009). The diffusion of development. *The Quarterly Journal of Economics 124*(2), 469–529.

Spolaore, E. and R. Wacziarg (2016). Ancestry and development: New evidence. *Discussion Papers Series, Department of Economics, Tufts University 820*.

Thomas, D., J. Strauss, and M.-H. Henriques (1991). How does mother's education affect child height? *Journal of human resources*, 183–211.

UNICEF (2012). Levels trends in child mortality: Report 2012. *UN Inter-agency Group for Child Mortality Estimation, United Nations Children's Fund, UNICEF*.

Varshney, A. (2003). *Ethnic conflict and civic life: Hindus and Muslims in India*. Yale University Press.

Victora, C. G., J. Bryce, O. Fontaine, and R. Monasch (2000). Reducing deaths from diarrhoea through oral rehydration therapy. *Bulletin of the World Health Organization 78*(10), 1246–1255.

Weidmann, N. B., J. K. Rød, and L.-E. Cederman (2010). Representing ethnic groups in space: A new dataset. *Journal of Peace Research 47*(4), 491–499.

Tables

Table 1: Mother's Linguistic Distance and Child mortality: 50 km Radius

	(1)	(2)	(3)	(4)	(5)
$\delta = 0.0025$					
Linguistic Distance 50 KM	0.0270***	0.0403**	0.0431***	0.0438***	0.0435***
	(0.00916)	(0.0158)	(0.0135)	(0.0135)	(0.0133)
ELF 50 KM	-0.00315	-0.00406	-0.00548	-0.00662	-0.00739
	(0.0101)	(0.00971)	(0.00894)	(0.00830)	(0.00823)
Observations	653666	653666	653666	653666	653666
R^2	0.089	0.091	0.145	0.145	0.154
Survey-wave FE	Y	Y	Y	Y	Y
Region × Year FE	Y	Y	Y	Y	Y
Ethnicity FE	N	Y	Y	Y	N
Religion FE	N	N	Y	Y	Y
Individual Controls	N	N	Y	Y	Y
Geographic isolation	N	N	N	Y	Y
Ethnicity × Year FE	N	N	N	N	Y

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. The numbers after linguistic distance and ELF indicate the radius of the circle around the mother in which these variables have been calculated. The individual controls include female child dummy, mother's age at birth, mother's age at birth squared, multiple birth indicator, birth order, birth order squared, short birth spacing prior to the birth, short birth spacing after the birth, the location of the mother in the form of an urban dummy, dummies for her educational attainment and her families' wealth index. Geographical isolation controls include the distance of the mother's location from the capital and the logged population in the circle.

Table 2: Mother's Linguistic Distance and Child mortality: Alternative radii

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	25 km	75 km	100 km	125 km	150 km	175 km	200 km	250 km
$\delta = 0.0025$								
Linguistic Distance	0.0347***	0.0474***	0.0487**	0.0528**	0.0537**	0.0540**	0.0543**	0.0522**
	(0.00964)	(0.0177)	(0.0192)	(0.0205)	(0.0226)	(0.0229)	(0.0227)	(0.0236)
ELF	-0.00372	-0.00893	-0.0112	-0.00540	0.00393	0.0104	0.0134	-0.0135
	(0.00620)	(0.00996)	(0.0122)	(0.0131)	(0.0138)	(0.0156)	(0.0184)	(0.0258)
Observations	653666	653666	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. The numbers in the column headings indicate the radius of the circle around the mother in which the linguistic distance and ELF have been calculated. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table 1.

Table 3: Mother's Linguistic Distance and Other Child Health Variables

	(1) infant	(2) neonatal	(3) HAZ	(4) stunted	(5) WAZ
Linguistic Distance	0.0204***	0.00725***	-0.0875**	0.0313**	-0.0497
	(0.00601)	(0.00235)	(0.0381)	(0.0132)	(0.0317)
ELF	-0.00291	-0.00357*	0.0233	-0.00605	0.101**
	(0.00445)	(0.00194)	(0.0504)	(0.0169)	(0.0409)
Observations	815267	861386	141475	141475	141475
R^2	0.097	0.069	0.205	0.153	0.161

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The column headings indicate the individual child-level dependent variable for each specification. These are: infant mortality, neonatal mortality, height-for-age Z-score (HAZ), stunting, and the weight-for-age Z-score (WAZ). A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table 1.

Table 4: Mother's Linguistic Distance and child mortality: Robustness for aggregate diversity

	(1) ELFL15	(2) ELFL10	(3) ELFL5	(4) ELFL2	(5) ONLYELF	(6) ELFSQ	(7) NOELF
Linguistic Distance	0.0435***	0.0426***	0.0430***	0.0435***		0.0428***	0.0414***
	(0.0133)	(0.0133)	(0.0125)	(0.0119)		(0.0131)	(0.0143)
ELF	-0.00739	-0.00403	-0.00584	-0.00733	-0.00272	0.00339	
	(0.00823)	(0.00757)	(0.0115)	(0.0135)	(0.00955)	(0.0213)	
ELF squared						-0.00956	
						(0.0254)	
Observations	653666	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154	0.154
	ELPL15	ELPL10	ELPL5	ELPL2	ONLYELP	ELPSQ	Both
Linguistic Distance	0.0410***	0.0417***	0.0425***	0.0442***		0.0411***	0.0444***
	(0.0140)	(0.0140)	(0.0136)	(0.0131)		(0.0140)	(0.0131)
ELP	0.00177	-0.00170	-0.00369	-0.00697	0.00396	-0.00679	0.0146*
	(0.00699)	(0.00630)	(0.00688)	(0.00572)	(0.00745)	(0.0201)	(0.00854)
ELP squared						0.00929	
						(0.0192)	
ELF							-0.0201**
							(0.0101)
Observations	653666	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154	0.154

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. In Panel 1 (Panel 2): column 1 controls for ELF (ELP) at aggregation Level 15; column 2 for ELF (ELP) at aggregation Level 10; column 3 for ELF (ELP) at aggregation Level 5; column 4 for ELF (ELP) at aggregation Level 2; column 5 for ELF (ELP) at aggregation Level 15, without LD; column 6 for ELF (ELP) at aggregation Level 15,and its square term. In column 7 of Panel 1, I do not control for ELF or ELP. In column 7 of Panel 2, I include both ELF and ELP. A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table 1.

Table 5: Mother's Linguistic Distance, Cultural Distance, Genetic Distance and Child mortality

	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic Distance	0.0435***		0.0485***		0.0446***	0.0486***
	(0.0133)		(0.0146)		(0.0137)	(0.0142)
Cultural Distance		0.0113	-0.0543			-0.0567
		(0.0480)	(0.0412)			(0.0503)
Genetic Distance				-0.0600	-0.227	0.0254
				(0.228)	(0.246)	(0.293)
Observations	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. The column heading indicates the robustness test undertaken. A circle of radius 50 km has been considered for calculating the linguistic, cultural and genetic distance variables. Linguistic distance is based on author's own calculations from the Ethnologue dataset. Genetic distance is measured by the FST index (Spolaore and Wacziarg, 2016) constructed by the author using genetic distance data from Pemberton et al. (2013). Cultural distance was constructed by the author, based on matching Ethnologue linguistic groups with Murdock's Ethnographic atlas, extending a mapping by Giuliano and Nunn (2018). All columns include controls survey-wave FE, region × year FE, ethnicity × year FE, religion FE, ELF in the circle, individual controls and geographic isolation controls described in the notes of Table 1.

Table 6: Mother's Linguistic Distance and Child mortality: Heterogeneity by Migration Status

	(1)	(2)	(3)	(4)
	HetMigrant	HetYearsLived	Migrants	NMigrants
Linguistic Distance	0.0577***	0.0341***	0.0198**	0.0758***
	(0.0148)	(0.0114)	(0.00935)	(0.0137)
Het. Variable	0.00610***	-0.000348***		
	(0.00166)	(0.0000718)		
Linguistic Distance × Het. Variable	-0.0185**	0.000555**		
	(0.00819)	(0.000270)		
Observations	521217	521217	278952	241309
R^2	0.163	0.163	0.177	0.167

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. In column 1 the Het. Variable refers to the 0-1 migrant status of the mother; in column 2 it refers to the continuous variable indicating how many years the mother has been living in the village where she was interviewed. In column 3 (column 4), I restrict the sample to only children of mothers who are migrants (non-migrants). A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table 1.

Table 7: Mother's Linguistic Distance, Malaria Suitability and Child mortality

	(1)	(2)	(3)	(4)
	Full Sample		Migrants	NMigrants
Linguistic Distance	0.0435***	0.0408***	0.0196**	0.0642***
	(0.0133)	(0.00845)	(0.00874)	(0.00948)
Malaria Suitability	0.00377	0.00213	0.00181	0.000896
	(0.00426)	(0.00420)	(0.00580)	(0.00410)
ELF	-0.00753	-0.00752	-0.0166*	-0.00687
	(0.00828)	(0.00821)	(0.00887)	(0.0101)
Linguistic Distance × Malaria Suitability		0.0173***	0.00419	0.0218***
		(0.00560)	(0.00500)	(0.00715)
Observations	653666	653666	278952	241309
R^2	0.154	0.154	0.177	0.167

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. Malaria Suitability is measured by the malaria stability index originally constructed by Kiszewski et al. (2004). Columns 1 and 2 use the full sample of mothers. In column 3 (column 4), I restrict the sample to only children of mothers who are migrants (non-migrants). A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table 1.

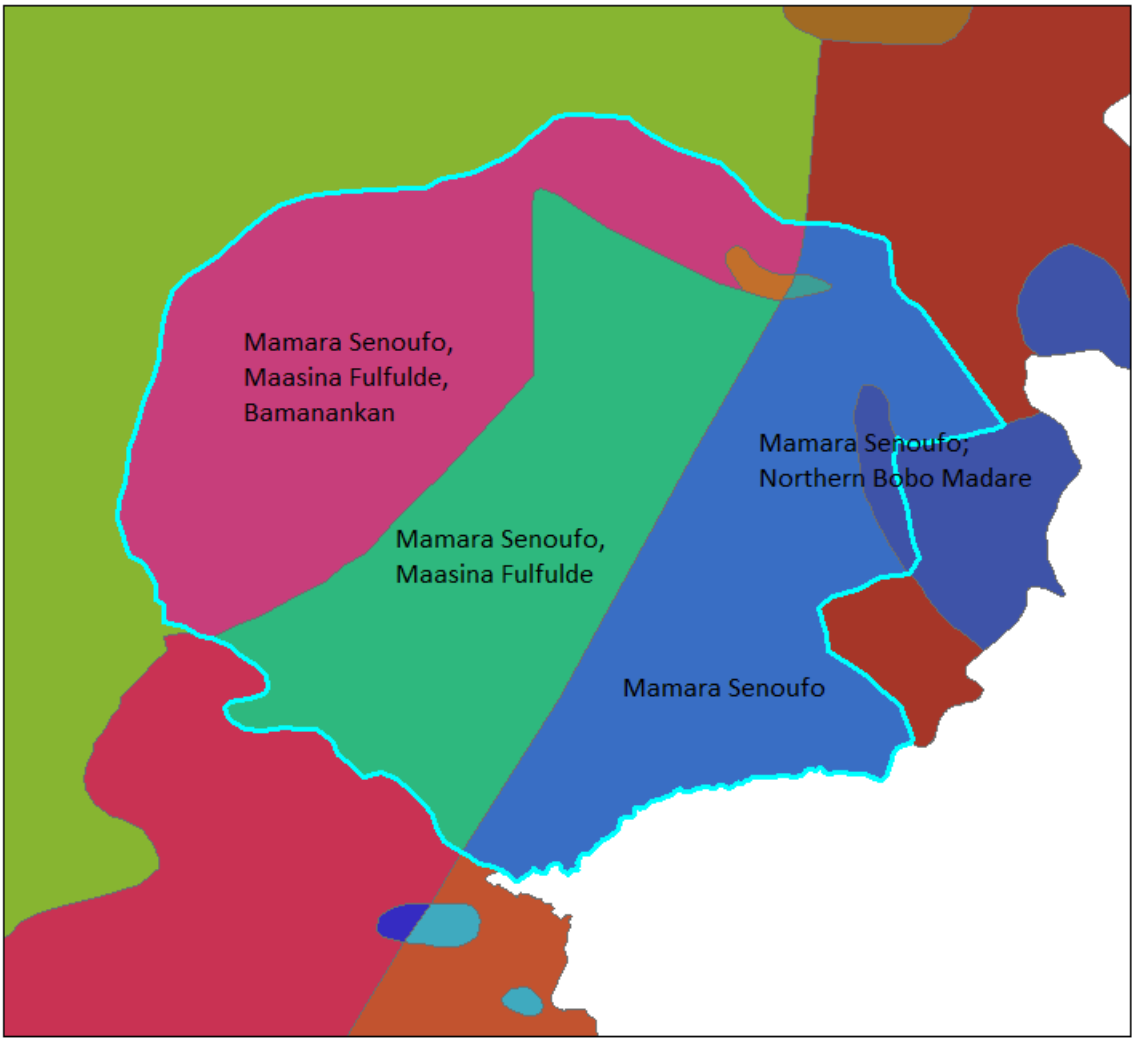
Table 8: Public Goods, Information and Linguistic Distance: Migrants vs. Non-Migrants

	(1)	(2)	(4)	(5)	(6)
	education	literacy	water	electricity	ORS
Migrants					
Linguistic Distance	0.0277	0.0101	-0.0352	0.0312	0.00154
	(0.0795)	(0.0362)	(0.0271)	(0.0190)	(0.0178)
ELF	-0.0198	-0.0278**	0.0565*	-0.0171	-0.0190
	(0.0311)	(0.0138)	(0.0294)	(0.0127)	(0.0330)
Observations	90456	71929	74748	89536	88542
R^2	0.458	0.423	0.112	0.548	0.206
Non-Migrants					
Linguistic Distance	0.000477	0.0121	-0.0352	0.00731	-0.0841**
	(0.0667)	(0.0265)	(0.0271)	(0.0186)	(0.0374)
ELF	0.0704	0.0561*	0.0565*	-0.00774	-0.00560
	(0.0545)	(0.0294)	(0.0294)	(0.00942)	(0.0308)
Observations	73681	60445	74748	72936	72648
R^2	0.474	0.391	0.112	0.546	0.242

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The column headings indicate the individual mother-level dependent variable for each specification. These are: educational attainment, literacy, access to water, access to electricity and knowledge about ORS. A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region FE, ethnicity FE, religion FE, year of birth FE, dummies for wealth index, and geographic isolation controls described in the notes of Table 1.

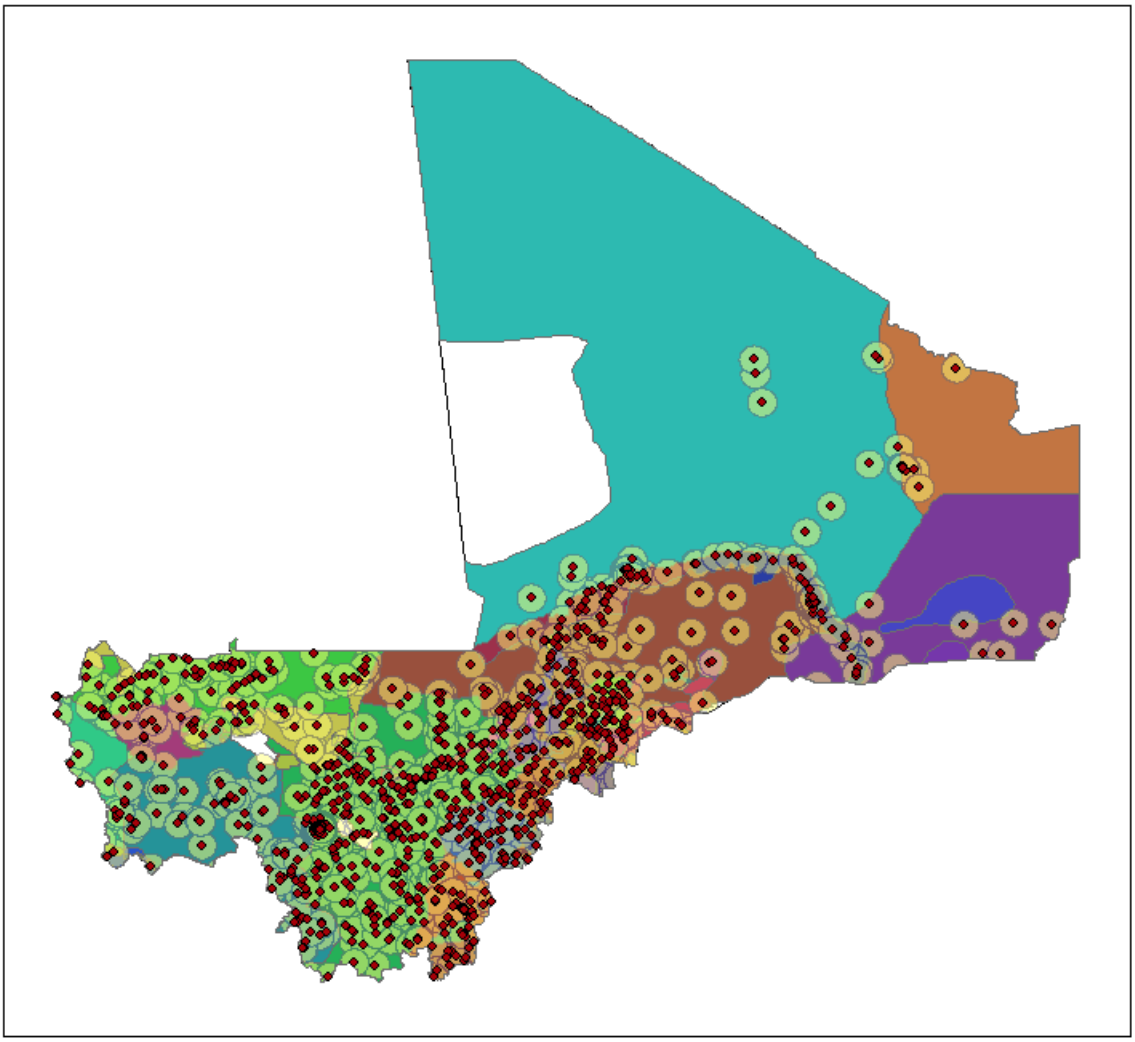
Figures

Figure 1: Example of Overlapping Language Polygons



NOTES: This map plots the linguistic groups in the south-eastern region of Mali from the Ethnologue database. Polygons of different colours represent the different linguistic areas. The polygon highlighted in blue demarcates the linguistic homeland of the Mamara Senoufo language speakers. In the light blue shaded polygon in the south-east corner of the map, there are no other languages spoken apart from Mamara Senoufo. In the polygon with a darker shade of blue, just north of this area, both Mamara Senoufo and Northern Bobo Madare are spoken. In the green shaded polygon in the centre of the map, Mamara Senoufo and Maasina Fulfulde are spoken. Finally, in the pink shaded polygon in the west, Mamara Senoufo is spoken with two other languages, namely Maasina Fulfulde and Bamanankan.

Figure 2: Example of DHS Clusters and circles for Mali



NOTES: This map plots the linguistic groups of Mali from the Ethnologue database (See Appendix Figure M3). The red dots represent the locations of the mother's (DHS clusters) for Mali and the circles around them represent 25 km circles around the mothers.

FOR ONLINE PUBLICATION

APPENDIX for "The Health Costs of Ethnic Distance: Evidence from Sub-Saharan Africa"

A Data Appendix

A.1 DHS Countries and Surveys Used

In this study, I use 30 DHS surveys from 14 sub-Saharan African countries. These are listed in Table A1. These countries and surveys were chosen based on the availability of the GPS coordinates and ethnicities of the mothers, and other covariates. In particular, countries and surveys for which a one to one matching for a large number of the ethnicities was not possible had to be omitted from the sample.

Table A1: Study Sample

Country	DHS surveys used
Benin	1996, 2001
Burkina Faso	1993, 1998-99, 2003, 2010
Ethiopia	2000, 2005, 2011
Ghana	1993, 2998, 2003, 2008
Guinea	1999, 2005
Kenya	2003, 2008
Malawi	2000, 2004, 2010
Mali	1995-96, 2001, 2006
Namibia	2000
Niger	1998
Senegal	2005, 2010-11
Sierra Leone	2008
Uganda	2011
Zambia	2007

For example, Cameroon has multiple DHS surveys (namely, 1991, 1998, 2004 and 2011). For the 1991 survey, the only two ethnicities provided are Cameroonian and others. Hence, for my purposes, this survey is unusable. So is the 1998 survey, which did not collect GPS data. The 2011 survey has a more disaggregated division of ethnic groups, but they are still too broad to allow a one to one matching with languages. For instance, some of the ethnicities consist of three or four groups combined together, for instance, "arab-choa/peulh/haoussa/kanuri", "cotier/ngoe/oroko" or "beti/bassa/mbam." This makes a one-to-one ethnicity-to-language map-

ping impossible. The 2004 survey contained more disaggregated data on ethnic groups, but a sizeable proportion of the respondents remained unmapped. Hence, Cameroon had to be discarded altogether. Several other countries and surveys had to be discarded for similar reasons. A full list is available from the author upon request.

A.2 The Iterative Proportional Fitting Algorithm

This section explains the iterative proportional fitting (IPF) algorithm used to construct the spatial distribution of language groups at the 1 km $\times$ 1 km resolution grid-cell level. The procedure was initially developed by Desmet et al. (2020), to whom the following explanation is heavily indebted.

First, consider an illustrative example. Figure M2 plots the polygons representing the linguistic regions in the 14 countries in my sample, based on the Ethnologue database. The polygons of different colours represent the different language groups. There are many regions in these countries where multiple languages are spoken, which are represented by overlapping polygons. Unfortunately, these overlapping polygons are not distinguishable in the map. In order to illustrate this possibility consider an example from Mali. Figure M3 gives the linguistic map of Mali. The area outlined in blue in the south-eastern corner of the country is the linguistic homeland of the Mamara Senoufo speakers. Figure 1 from the main paper zooms into this region. Notice that, while Mamara Senoufo is spoken in the entire area bordered in blue, there are various possible overlaps with other languages. First, in the light blue–shaded polygon in the south-eastern corner of the map, Mamara Senoufo is the only language spoken. In the polygon shaded a darker blue, just north of this area, both Mamara Senoufo and Northern Bobo Madare are spoken. In the green-shaded polygon in the centre of the map, Mamara Senoufo and Maasina Fulfulde are spoken. Finally, in the pink-shaded polygon in the west, Mamara Senoufo is spoken along with two other languages, namely Maasina Fulfulde and Bamanankan. The IPF algorithm takes into account all of these possibilities.

I now expand on this example from Mali to explain the process of generating the final spatial distribution of language groups in more detail. While Figure M3 provides the map of linguistic regions in Mali, Figure M4 plots Mali's population distribution at the 1 km $\times$ 1 km level using data from LandScan. Figure M5 overlays the language polygons on the population distribution for Mali. Based on the data generated from this combined map, and the information on total populations pertaining to each language group in the country (provided by Ethnologue), the

IPF algorithm allocates languages to each 1 km $\times$ 1 km grid-cell following the steps listed in the next paragraphs. I repeat this exercise for each of the 14 countries in my sample to construct the spatial distribution of linguistic groups for these countries.

Start by considering a country with M linguistic groups and K grid cells of 30 arc seconds $\times$ 30 arc seconds resolution (which is approximately 1 km $\times$ 1 km at the equator). The objective is to allocate each of these M linguistic groups to each of the K grid cells, such that the total population per cell and the total population per language correspond to their actual values. Following Desmet et al. (2020), I exploit three pieces of information: the number of people living in each grid cell (from LandScan); the number of speakers of each language in each country (from Ethnologue); and whether a language is spoken or not in any given grid cell (obtained by rasterizing the digitized version of the Ethnologue database).

The information described above yield three matrices corresponding to the three distinct pieces of information. First, using the LandScan data leads to matrix $\mathcal{N}_{K\times 1}$, the elements of which give the total population in each of the K cells. The Ethnologue dataset allows the construction of two matrices: matrix $\mathcal{L}_{1\times M}$, the elements of which give the number of speakers of each language in the country and the binary matrix $\mathcal{B}_{K\times M}$ the elements of which take the values 1 or 0 denoting whether or not the language corresponding to the column is spoken in the cell corresponding to the row. The following steps from Desmet et al. (2020) describe the iterative proportional fitting algorithm that exploits this information:

1. "*Step 0.* Define $\mathcal{T}^{(0)} = \mathcal{B}$.
2. *Step 1.* For each location ℓ, assign a share $\mathcal{T}^{(2n-2)}(\ell, i)/\sum_j \mathcal{T}^{(2n-2)}(\ell, j)$ to language i. Hence,
$$\mathcal{T}^{(2n-1)}(\ell, i) = \frac{\mathcal{T}^{(2n-2)}(\ell, i)}{\sum_j \mathcal{T}^{(2n-2)}(\ell, j)} \mathcal{N}(\ell, 1),$$
where $n = 1, 2, \ldots$ refers to the times the algorithm has iterated through Steps 1 and 2.
3. *Step 2.* For each language i, assign a share $\mathcal{T}^{(2n-1)}(\ell, i)/\sum_k \mathcal{T}^{(2n-1)}(k, i)$ to cell ℓ. Hence,
$$\mathcal{T}^{(2n)}(\ell, i) = \frac{\mathcal{T}^{(2n-1)}(\ell, i)}{\sum_k \mathcal{T}^{(2n-1)}(k, i)} \mathcal{L}(1, i)$$
4. *Step 3.* Go through Step 1 and Step 2 until $\mathcal{T}^{(2n-1)}(\ell, i)$ converge to $\mathcal{T}^{(2n)}(\ell, i)$ for all ℓ and i" (Desmet et al., 2020).

Step 1 ensures that the allocation satisfies the marginals on cell populations, that is, $\sum_j \mathcal{T}^{(2n-1)}(\ell, j) = \mathcal{N}(\ell, 1)$. Step 2, on the other hand, ensures that the allocation satisfies the marginals on the language populations, that is, $\sum_k \mathcal{T}^{(2n)}(k, i) = \mathcal{L}(1, i)$. For instance, the first time the algorithm undertakes Step 1, each cell's population is divided equally between the different languages that are spoken in that cell. For instance, consider the example that four languages are spoken in a cell, then the first time the algorithm undertakes Step 1, each language is assigned 25% of that cell's population. Similarly, Step 2 ensures that the sum of the population allocated to a language is equal to the actual total number of speakers of that language (Desmet et al., 2020).

The iterative proportional fitting algorithm described above provides the cell level allocation of language speakers: $\mathcal{T}^{(2n)}(\ell, i)$. The algorithm is guaranteed to converge if the three matrices $\mathcal{L}, \mathcal{N}$ and $\mathcal{B}$ are fully consistent with each other. However, in the presence of small inconsistencies between the three matrices due to minor imprecision in the data from the different sources, the algorithm need not necessarily converge. For instance, one likely source of imperfection arises from the possibility that the Ethnologue language polygons are not entirely accurate and that there are minor imperfections in the locations of the language borders. Following Desmet et al. (2020), I replace the 0 values in the binary matrix $\mathcal{B}$ by 0.000001 to allow for such imprecision. This modification ensures that the iterative proportional fitting algorithm converges despite minor inconsistencies in the matrices (Fienberg, 1970).[2]

A.3 Ethnicity and Language Matching

This section lists the steps undertaken to match ethnic groups from the DHS to languages from Ethnologue. A full list of ethnicity and language matching is available upon request.

- If the name of an ethnicity from the DHS is identical to a language name from Ethnologue, then I already have the required language and no further mapping is needed. For instance, Kalenjin is both an ethnicity and a language spoken in Kenya.
- If the name of an ethnicity from the DHS is an alternative name for a language group from Ethnologue, the former is simply renamed to the latter. This provides the required language, and no further mapping is needed. For instance, Kissi is the name of an ethnic group in Kenya, in addition to being an alternative name for the Ekgussii language.

[2]Please refer to Desmet et al. (2020) for a more detailed discussion of these issues. For further reading on the iterative proportional fitting algorithm, please refer to Bishop et al. (1975), Deming and Stephan (1940), and Fienberg (1970).

Similarly, Peulh is an alternative name for the Borgu Fulfulde language spoken in Benin.

- Some of the names of ethnic groups contained in the DHS are also the names of languages from Ethnologue, but the spellings differ across the two sources. In this case, the language assignation is trivial. For instance, the Afar and Amharic language groups from Ethiopia are spelt as Affar and Amhara, respectively, in the DHS.

- In some instances, the DHS provides macro language groups in the ethnicity field. In these cases, I assign one of the actual languages that form part of the macro language group to the entire group. Since distances are based on the number of shared branches, assigning a different language from the same group does not change the actual distance. For instance, to the Luhya group in Kenya, I assign the Lubukusu language.

- For some groups I follow Jim Fearon's classifications (originally from Fearon (2003)). For example, the San group in Namibia is assigned the Haikom language. Similarly, the Diola group in Senegal is assigned the Jola-Fonyi language.

- In a very small number of cases the same ethnicity name from the DHS referred to one of several closely related languages listed in Ethnologue. For instance, Limba in Sierra Leone could refer to either East Limba or West-Central Limba. I randomly assign it to East Limba. But since both East Limba and West-Central Limba are closely related and share precisely the same number of branches with any other language, this should not make a difference in the actual linguistic distance calculations.

A.4 Descriptive Statistics

Table A2: Child-Level Summary Statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
Child Death	0.228	0.42	0	1	654672
Infant Death	0.12	0.325	0	1	816268
Neonatal Death	0.055	0.229	0	1	862358
Height-for-Age Score	-1.571	1.746	-6	5.99	141673
Weight-for-Age Score	-1.182	1.304	-5.98	4.97	141673
Stunting	0.409	0.492	0	1	141673
Tetanus Vaccine	0.703	0.457	0	1	154814
Measles	0.844	0.810	0	3	196789
Polio Vaccine	0.278	0.448	0	1	182048
DPT Vaccine	0.466	0.499	0	1	161085
Iron Tablets	0.697	0.459	0	1	115590
Migrant	0.539	0.498	0	1	686231
Years lived in cluster	23.078	14.461	0	50	686231
Urban Residence	0.224	0.417	0	1	862358
Female Child	0.49	0.5	0	1	862358
Age At Birth	25	6.425	8	50	862358
Age At Birth Squared	666.292	348.533	64	2500	862358
Multiple Birth	0.032	0.177	0	1	862358
Birth Order Number	3.448	2.317	1	18	862358
Birth Order Number Squared	17.256	22.557	1	324	862358
Short Birth Spacing Prior	0.209	0.407	0	1	862358
Short Birth Spacing Post	0.209	0.407	0	1	862358
Highest educational level	0.424	0.664	0	3	862352
Educational Attainment	0.582	1.02	0	5	862352
Years of Education	1.997	3.405	0	26	862028
Log(Distance to the Capital)	5.13	1.217	-2.614	7.221	862358
Wealth Index	2.867	1.399	1	5	862358
Child's Birth Year	1992.274	9.366	1955	2011	862358
Mother's Birth Year	1968.416	9.572	1943	1996	862358
Log(Population) in 25 km	12.09	1.307	3.849	15.238	862358
Log(Population) in 50 km	13.272	1.164	6.137	15.665	862358
Log(Population) in 75 km	13.958	1.093	6.971	16.011	862358
Log(Population) in 100 km	14.428	1.045	7.467	16.346	862358
Log(Population) in 125 km	14.771	1.007	8.176	16.57	862358
Log(Population) in 150 km	15.036	0.974	8.582	16.87	862358
Log(Population) in 175 km	15.258	0.943	8.893	17.046	862358
Log(Population) in 200 km	15.437	0.915	9.332	17.188	862358
Log(Population) in 250 km	15.728	0.871	10.161	17.433	862358

Table A3: Mother Level Summary Statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
ORS Knowledge	0.761	0.426	0	1	205794
Educational Attainment	0.782	1.189	0	5	208896
Water Access	0.656	0.475	0	1	182482
Electricity Access	0.186	0.389	0	1	204920

Table A4: Summary Statistics for LD variables

Variable	Mean	Std. Dev.	Min.	Max.
$\delta = 0.0025$				
Linguistic distance in 25 km	0.073	0.191	0	1
Linguistic distance in 50 km	0.077	0.189	0	1
Linguistic distance in 75 km	0.081	0.19	0	1
Linguistic distance in 100 km	0.083	0.191	0	1
Linguistic distance in 125 km	0.087	0.194	0	1
Linguistic distance in 150 km	0.089	0.196	0	1
Linguistic distance in 175 km	0.092	0.197	0	1
Linguistic distance in 200 km	0.095	0.199	0	1
Linguistic distance in 250 km	0.1	0.202	0	1
$\delta = 0.05$ *à la* Desmet et al. (2012)				
Linguistic distance in 25 km	0.094	0.178	0	1
Linguistic distance in 50 km	0.1	0.176	0	1
Linguistic distance in 75 km	0.105	0.177	0	1
Linguistic distance in 100 km	0.109	0.179	0	1
Linguistic distance in 125 km	0.113	0.182	0	1
Linguistic distance in 150 km	0.117	0.184	0	1
Linguistic distance in 175 km	0.12	0.186	0	1
Linguistic distance in 200 km	0.124	0.188	0	1
Linguistic distance in 250 km	0.131	0.192	0	1
$\delta = 0.50$ *à la* Fearon (2003)				
Linguistic distance in 25 km	0.277	0.247	0	1
Linguistic distance in 50 km	0.294	0.235	0	1
Linguistic distance in 75 km	0.31	0.228	0	1
Linguistic distance in 100 km	0.323	0.224	0	1
Linguistic distance in 125 km	0.337	0.221	0	1
Linguistic distance in 150 km	0.349	0.219	0	1
Linguistic distance in 175 km	0.359	0.216	0	1
Linguistic distance in 200 km	0.369	0.213	0	1
Linguistic distance in 250 km	0.388	0.208	0.002	1
N		862358		

Table A5: Summary statistics for ELF variables

Variable	Mean	Std. Dev.	Min.	Max.
ELF at Level 2 in 25 km	0.179	0.187	0	0.786
ELF at Level 2 in 50 km	0.207	0.193	0	0.792
ELF at Level 2 in 75 km	0.224	0.198	0	0.793
ELF at Level 2 in 100 km	0.238	0.199	0	0.802
ELF at Level 2 in 125 km	0.251	0.2	0	0.776
ELF at Level 2 in 150 km	0.262	0.2	0	0.779
ELF at Level 2 in 175 km	0.272	0.201	0	0.772
ELF at Level 2 in 200 km	0.281	0.201	0	0.773
ELF at Level 2 in 250 km	0.295	0.202	0	0.758
ELF at Level 5 in 25 km	0.245	0.236	0	0.871
ELF at Level 5 in 50 km	0.285	0.242	0	0.868
ELF at Level 5 in 75 km	0.313	0.247	0	0.87
ELF at Level 5 in 100 km	0.333	0.25	0	0.873
ELF at Level 5 in 125 km	0.353	0.251	0	0.866
ELF at Level 5 in 150 km	0.37	0.251	0	0.864
ELF at Level 5 in 175 km	0.384	0.25	0	0.854
ELF at Level 5 in 200 km	0.398	0.249	0	0.84
ELF at Level 5 in 250 km	0.42	0.248	0.001	0.838
ELF at Level 10 in 25 km	0.384	0.26	0	0.917
ELF at Level 10 in 50 km	0.457	0.25	0	0.906
ELF at Level 10 in 75 km	0.507	0.238	0	0.9
ELF at Level 10 in 100 km	0.542	0.225	0	0.898
ELF at Level 10 in 125 km	0.572	0.213	0	0.907
ELF at Level 10 in 150 km	0.596	0.199	0	0.913
ELF at Level 10 in 175 km	0.618	0.185	0	0.918
ELF at Level 10 in 200 km	0.637	0.17	0	0.921
ELF at Level 10 in 250 km	0.667	0.144	0.008	0.921
ELF at Level 15 in 25 km	0.408	0.27	0	0.918
ELF at Level 15 in 50 km	0.487	0.262	0	0.921
ELF at Level 15 in 75 km	0.54	0.25	0	0.937
ELF at Level 15 in 100 km	0.576	0.236	0	0.932
ELF at Level 15 in 125 km	0.607	0.222	0	0.938
ELF at Level 15 in 150 km	0.631	0.207	0	0.941
ELF at Level 15 in 175 km	0.653	0.192	0	0.944
ELF at Level 15 in 200 km	0.672	0.178	0	0.944
ELF at Level 15 in 250 km	0.702	0.152	0.008	0.938
N		862358		

Table A6: Summary statistics for ELP variables

Variable	Mean	Std. Dev.	Min.	Max.
ELP at Level 2 in 25 km	0.329	0.33	0	1
ELP at Level 2 in 50 km	0.377	0.339	0	1
ELP at Level 2 in 75 km	0.407	0.341	0	1
ELP at Level 2 in 100 km	0.428	0.339	0	1
ELP at Level 2 in 125 km	0.45	0.337	0	1
ELP at Level 2 in 150 km	0.469	0.335	0	1
ELP at Level 2 in 175 km	0.485	0.333	0	1
ELP at Level 2 in 200 km	0.5	0.331	0	1
ELP at Level 2 in 250 km	0.521	0.328	0	1
ELP at Level 5 in 25 km	0.372	0.327	0	1
ELP at Level 5 in 50 km	0.426	0.326	0	1
ELP at Level 5 in 75 km	0.456	0.321	0	1
ELP at Level 5 in 100 km	0.476	0.312	0	1
ELP at Level 5 in 125 km	0.495	0.305	0	0.999
ELP at Level 5 in 150 km	0.512	0.299	0	0.998
ELP at Level 5 in 175 km	0.526	0.295	0	0.994
ELP at Level 5 in 200 km	0.538	0.293	0	0.995
ELP at Level 5 in 250 km	0.555	0.289	0.002	0.994
ELP at Level 10 in 25 km	0.504	0.292	0	1
ELP at Level 10 in 50 km	0.567	0.253	0	1
ELP at Level 10 in 75 km	0.599	0.219	0	1
ELP at Level 10 in 100 km	0.616	0.189	0	1
ELP at Level 10 in 125 km	0.626	0.161	0	0.99
ELP at Level 10 in 150 km	0.633	0.142	0	0.986
ELP at Level 10 in 175 km	0.64	0.133	0	0.978
ELP at Level 10 in 200 km	0.646	0.129	0	0.975
ELP at Level 10 in 250 km	0.651	0.128	0.016	0.974
ELP at Level 15 in 25 km	0.504	0.286	0	1
ELP at Level 15 in 50 km	0.555	0.246	0	1
ELP at Level 15 in 75 km	0.577	0.217	0	1
ELP at Level 15 in 100 km	0.587	0.193	0	0.994
ELP at Level 15 in 125 km	0.592	0.173	0	0.982
ELP at Level 15 in 150 km	0.595	0.163	0	0.979
ELP at Level 15 in 175 km	0.597	0.161	0	0.978
ELP at Level 15 in 200 km	0.599	0.162	0	0.975
ELP at Level 15 in 250 km	0.596	0.164	0.016	0.974
N		862358		

Table A7: Summary statistics for Cultural and Genetic Distance variables

Variable	Mean	Std. Dev.	Min.	Max.
Cultural distance in 25 km	0.065	0.05	0	0.254
Cultural distance in 50 km	0.068	0.047	0	0.254
Cultural distance in 75 km	0.072	0.044	0	0.251
Cultural distance in 100 km	0.075	0.043	0	0.244
Cultural distance in 125 km	0.077	0.041	0	0.235
Genetic distance in 25 km	0.008	0.007	0	0.185
Genetic distance in 50 km	0.009	0.006	0	0.185
Genetic distance in 75 km	0.009	0.006	0	0.185
Genetic distance in 100 km	0.009	0.006	0	0.185
Genetic distance in 125 km	0.01	0.006	0	0.185
N		862358		

Table A8: Correlations of Linguistic Distance and diversity (206,076 observations (mothers))

	Correlation of LD with ELF					Correlation of LD with ELP				
	25 km	50 km	75 km	100 km	125 km	25 km	50 km	75 km	100 km	125 km
Aggregation										
Level 2	0.34	0.37	0.38	0.39	0.40	0.31	0.32	0.33	0.33	0.34
Level 5	0.26	0.28	0.30	0.31	0.32	0.27	0.28	0.28	0.28	0.28
Level 10	0.15	0.15	0.16	0.17	0.18	0.14	0.11	0.07	0.03	-0.01
Level 15	0.13	0.13	0.13	0.14	0.16	0.13	0.09	0.05	0.01	-0.03

	25 km	50 km	75 km	100 km	125 km	25 km	50 km	75 km	100 km	125 km
Level 2	0.39	0.42	0.44	0.44	0.44	0.35	0.37	0.37	0.37	0.36
Level 5	0.33	0.36	0.38	0.38	0.38	0.32	0.32	0.32	0.31	0.30
Level 10	0.23	0.23	0.24	0.24	0.24	0.19	0.15	0.09	0.03	-0.04
Level 15	0.22	0.22	0.22	0.22	0.22	0.17	0.12	0.05	-0.01	-0.08

$\delta = 0.5$	25 km	50 km	75 km	100 km	125 km	25 km	50 km	75 km	100 km	125 km
Level 2	0.58	0.61	0.63	0.64	0.65	0.57	0.59	0.61	0.63	0.63
Level 5	0.58	0.60	0.63	0.64	0.65	0.57	0.58	0.59	0.60	0.60
Level 10	0.47	0.47	0.47	0.47	0.47	0.42	0.38	0.32	0.25	0.17
Level 15	0.47	0.47	0.47	0.47	0.48	0.40	0.33	0.24	0.13	0.02

Table A9: Correlations of ELF and ELP (28,839 DHS clusters)

	25 km	50 km	75 km	100 km	125 km
Level 1	0.99	1.00	0.99	0.99	0.99
Level 2	0.98	0.98	0.97	0.97	0.96
Level 3	0.98	0.97	0.97	0.96	0.96
Level 4	0.95	0.95	0.94	0.94	0.93
Level 5	0.94	0.94	0.93	0.93	0.92
Level 6	0.94	0.93	0.92	0.91	0.91
Level 7	0.92	0.90	0.89	0.88	0.87
Level 8	0.91	0.89	0.88	0.86	0.85
Level 9	0.87	0.80	0.73	0.65	0.56
Level 10	0.84	0.76	0.65	0.52	0.37
Level 11	0.83	0.74	0.62	0.49	0.32
Level 12	0.83	0.75	0.62	0.49	0.32
Level 13	0.83	0.75	0.62	0.49	0.32
Level 14	0.83	0.75	0.62	0.49	0.32
Level 15	0.78	0.63	0.44	0.25	0.03

B Varying Parameter Values

B.1 Varying the values of δ

Table B1: Mother's Linguistic Distance and Child mortality: 50 km Radius

	(1)	(2)	(3)	(4)	(5)
$\delta = 0.0025$					
Linguistic Distance 50 KM	0.0270***	0.0403**	0.0431***	0.0438***	0.0435***
	(0.00916)	(0.0158)	(0.0135)	(0.0135)	(0.0133)
ELF 50 KM	-0.00315	-0.00406	-0.00548	-0.00662	-0.00739
	(0.0101)	(0.00971)	(0.00894)	(0.00830)	(0.00823)
Observations	653666	653666	653666	653666	653666
R^2	0.089	0.091	0.145	0.145	0.154
$\delta = 0.05$ *à la* Desmet et al. (2012)					
Linguistic Distance 50 KM	0.0199*	0.0351*	0.0389**	0.0402**	0.0398**
	(0.0103)	(0.0183)	(0.0158)	(0.0159)	(0.0157)
ELF 50 KM	-0.00337	-0.00482	-0.00649	-0.00771	-0.00847
	(0.00993)	(0.00936)	(0.00878)	(0.00805)	(0.00800)
Observations	653666	653666	653666	653666	653666
R^2	0.089	0.091	0.145	0.145	0.154
$\delta = 0.50$ *à la* Fearon (2003)					
Linguistic Distance 50 KM	-0.0131	-0.00692	0.00886	0.0100	0.01000
	(0.00959)	(0.0140)	(0.0123)	(0.0124)	(0.0123)
ELF 50 KM	0.00511	0.00298	-0.00449	-0.00597	-0.00678
	(0.00999)	(0.00976)	(0.00883)	(0.00811)	(0.00815)
Observations	653666	653666	653666	653666	653666
R^2	0.089	0.091	0.145	0.145	0.154
Survey-wave FE	Y	Y	Y	Y	Y
Region x Year FE	Y	Y	Y	Y	Y
Ethnicity FE	N	Y	Y	Y	N
Religion FE	N	N	Y	Y	Y
Individual Controls	N	N	Y	Y	Y
Geographic isolation	N	N	N	Y	Y
Ethnicity × Year FE	N	N	N	N	Y

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. The numbers after linguistic distance and ELF indicate the radius of the circle around the mother in which these variables have been calculated. The three panels use three different decay factors δ for calculating LD as indicated in the panel headings. The individual controls include female child dummy, mother's age at birth, mother's age at birth squared, multiple birth indicator, birth order, birth order squared, short birth spacing prior to the birth, short birth spacing after the birth, the location of the mother in the form of an urban dummy, dummies for her educational attainment and her families' wealth index. Geographical isolation controls include the distance of the mother's location from the capital and the logged population in the circle.

Table B2: Mother's Linguistic Distance and Child mortality: 50 km Radius

	(1)	(2)	(3)	(4)	(5)
$\delta = 0.0025$					
Linguistic Distance 50 KM	0.0271***	0.0402**	0.0430***	0.0437***	0.0435***
	(0.00908)	(0.0158)	(0.0135)	(0.0135)	(0.0133)
ELF 50 KM	-0.00303	-0.00396	-0.00540	-0.00653	-0.00739
	(0.0101)	(0.00969)	(0.00893)	(0.00829)	(0.00823)
Observations	654506	654502	654237	654237	653666
R^2	0.090	0.092	0.146	0.146	0.154
$\delta = 0.05$ *à la* Desmet et al. (2012)					
Linguistic Distance 50 KM	0.0201*	0.0349*	0.0388**	0.0400**	0.0398**
	(0.0102)	(0.0183)	(0.0158)	(0.0159)	(0.0157)
ELF 50 KM	-0.00327	-0.00471	-0.00640	-0.00762	-0.00847
	(0.00989)	(0.00934)	(0.00877)	(0.00805)	(0.00800)
Observations	654506	654502	654237	654237	653666
R^2	0.090	0.092	0.146	0.146	0.154
$\delta = 0.50$ *à la* Fearon (2003)					
Linguistic Distance 50 KM	-0.0130	-0.00712	0.00876	0.00989	0.01000
	(0.00958)	(0.0140)	(0.0123)	(0.0124)	(0.0123)
ELF 50 KM	0.00521	0.00315	-0.00438	-0.00585	-0.00678
	(0.00995)	(0.00974)	(0.00882)	(0.00810)	(0.00815)
Observations	654506	654502	654237	654237	653666
R^2	0.090	0.092	0.146	0.146	0.154
Survey-wave FE	Y	Y	Y	Y	Y
Region x Year FE	Y	Y	Y	Y	Y
Ethnicity FE	N	Y	Y	Y	N
Religion FE	N	N	Y	Y	Y
Individual Controls	N	N	Y	Y	Y
Geographic isolation	N	N	N	Y	Y
Ethnicity × Year FE	N	N	N	N	Y

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. The numbers after linguistic distance and ELF indicate the radius of the circle around the mother in which these variables have been calculated. The three panels use three different decay factors δ for calculating LD as indicated in the panel headings. The individual controls include female child dummy, mother's age at birth, mother's age at birth squared, multiple birth indicator, birth order, birth order squared, short birth spacing prior to the birth, short birth spacing after the birth, the location of the mother in the form of an urban dummy, dummies for her educational attainment and her families' wealth index. Geographical isolation controls include the distance of the mother's location from the capital and the logged population in the circle.

B.2 Alternative radii

Table B3: Mother's Linguistic Distance and Child mortality: Alternative radii

	(1) 25 km	(2) 75 km	(3) 100 km	(4) 125 km	(5) 150 km	(6) 175 km	(7) 200 km	(8) 250 km
$\delta = 0.0025$								
Linguistic Distance	0.0347***	0.0474***	0.0487**	0.0528**	0.0537**	0.0540**	0.0543**	0.0522**
	(0.00964)	(0.0177)	(0.0192)	(0.0205)	(0.0226)	(0.0229)	(0.0227)	(0.0236)
ELF	-0.00372	-0.00893	-0.0112	-0.00540	0.00393	0.0104	0.0134	-0.0135
	(0.00620)	(0.00996)	(0.0122)	(0.0131)	(0.0138)	(0.0156)	(0.0184)	(0.0258)
Observations	653666	653666	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154
$\delta = 0.05$ *à la* Desmet et al. (2012)								
Linguistic Distance	0.0299**	0.0439**	0.0445**	0.0476*	0.0465*	0.0463*	0.0449	0.0466*
	(0.0118)	(0.0204)	(0.0223)	(0.0240)	(0.0268)	(0.0276)	(0.0279)	(0.0276)
ELF	-0.00407	-0.0104	-0.0125	-0.00695	0.00257	0.00943	0.0132	-0.0136
	(0.00596)	(0.00956)	(0.0117)	(0.0126)	(0.0132)	(0.0149)	(0.0177)	(0.0251)
Observations	653666	653666	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154
$\delta = 0.50$ *à la* Fearon (2003)								
Linguistic Distance	0.00553	0.0119	0.0129	0.0128	0.0107	0.0105	0.0112	0.0240
	(0.00987)	(0.0143)	(0.0156)	(0.0179)	(0.0206)	(0.0223)	(0.0243)	(0.0262)
ELF	-0.00227	-0.00868	-0.0110	-0.00445	0.00629	0.0140	0.0182	-0.0119
	(0.00571)	(0.00970)	(0.0121)	(0.0134)	(0.0142)	(0.0158)	(0.0189)	(0.0255)
Observations	653666	653666	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. The numbers in the column headings indicate the radius of the circle around the mother in which the linguistic distance and ELF have been calculated. The three panels use three different decay factors δ for calculating LD as indicated in the panel headings. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table B1.

C Marginal Effects

Table C1: Marginal Effects

	Full Sample		Migrants		Non-Migrants	
Circle Radius	Child Deaths	% of SD	Child Deaths	% of SD	Child Deaths	% of SD
25 km	6.62	1.58%	3.31	0.79%	11.49	2.67%
50 km	8.22	1.96%	4.19	1.00%	14.09	3.28%
75 km	9.00	2.14%	4.15	0.99%	16.37	3.80%
100 km	9.30	2.22%	4.26	1.01%	16.86	3.92%
125 km	10.21	2.43%	4.08	0.97%	18.62	4.33%
150 km	10.48	2.50%	3.82	0.91%	18.96	4.41%
175 km	10.61	2.53%	3.54	0.84%	19.20	4.46%
200 km	10.75	2.56%	3.97	0.95%	18.67	4.34%
250 km	10.51	2.50%	4.23	1.01%	17.35	4.03%

Notes: This table provides the marginal effects for the most comprehensive specification indicated in Table 1 for circles of alternative radii around the mother. The leftmost panel includes the full sample of mothers, the middle panel restricts the sample to only migrant mothers, while the last panel restricts the sample to only non-migrant mothers.

Table C2: Marginal Effects accounting for Malaria Suitability

	Full Sample		Migrants		Non-Migrants	
Circle Radius	Child Deaths	% of SD	Child Deaths	% of SD	Child Deaths	% of SD
25 km	19.77	4.71%	7.87	1.88%	23.27	5.41%
50 km	25.05	5.97%	8.33	1.99%	33.71	7.83%
75 km	30.71	7.32%	10.34	2.46%	43.54	10.12%
100 km	32.23	7.68%	13.75	3.28%	44.07	10.24%
125 km	33.13	7.89%	13.18	3.14%	46.42	10.79%
150 km	33.95	8.09%	14.26	3.40%	48.28	11.22%
175 km	33.39	7.95%	14.15	3.37%	46.58	10.83%
200 km	31.32	7.46%	13.19	3.14%	43.54	10.12%
250 km	28.01	6.67%	11.56	2.76%	40.08	9.31%

Notes: This table provides the marginal effects a one SD increase in LD when the malaria stability index is one SD above its average value for circles of alternative radii around the mother. The actual specification is estimated in Table 7. The leftmost panel includes the full sample of mothers (corresponding to Column 2 of Table 7), the middle panel restricts the sample to only migrant mothers (corresponding to Column 3 of Table 7), while the last panel restricts the sample to only non-migrant mothers (corresponding to Column 4 of Table 7).

D Other Health Variables

D.1 Immunizations Full Sample

Table D1: Mother's Linguistic Distance and Other Variables:

	(2) tetanus	(3) measles	(4) polio	(5) dpt	(6) iron
Linguistic Distance	0.0103	0.0233	0.0113	0.0232	-0.0449*
	(0.0214)	(0.0279)	(0.0152)	(0.0173)	(0.0248)
ELF	0.00982	0.0341	-0.0239	-0.0266	0.000799
	(0.0212)	(0.0242)	(0.0209)	(0.0234)	(0.0198)
Observations	154650	196627	181890	160914	115498
R^2	0.264	0.329	0.249	0.366	0.360

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The column headings indicate the individual-level dependent variable for each specification. These are: tetanus vaccination, measles immunization, polio vaccination, DPT vaccination, and if the mother received iron tablets during pregnancy. A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region FE, ethnicity FE, religion FE, year of birth, dummies for wealth index, and geographic isolation controls described in the notes of Table B1.

D.2 Heterogeneity by Migrant Status: Additional Variables

Table D2: Mother's Linguistic Distance and Other Variables 1: Migrants vs. Non-Migrants

	(1) infant	(2) neonatal	(3) HAZ	(4) stunted	(5) WAZ
Migrants					
Linguistic Distance	0.00574	0.00297	-0.0468	0.0159	-0.0437
	(0.00502)	(0.00312)	(0.0761)	(0.0252)	(0.0465)
ELF	-0.00561	-0.00510*	0.00662	-0.00842	0.0644
	(0.00512)	(0.00290)	(0.0548)	(0.0189)	(0.0422)
Observations	348759	368880	66230	66230	66230
R^2	0.107	0.079	0.214	0.162	0.177
Non-Migrants					
Linguistic Distance	0.0396***	0.0134***	-0.129**	0.0426***	-0.0465
	(0.00869)	(0.00329)	(0.0619)	(0.0134)	(0.0440)
ELF	-0.00315	-0.00351	0.106*	-0.0218	0.163***
	(0.00530)	(0.00282)	(0.0582)	(0.0220)	(0.0580)
Observations	299028	315688	52477	52477	52477
R^2	0.111	0.082	0.218	0.169	0.169

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The column headings indicate the individual child-level dependent variable for each specification. These are: infant mortality, neonatal mortality, height-for-age Z-score (HAZ), stunting, and the weight-for-age Z-score (WAZ). Panel 1 (Panel 2) restricts the sample to only migrants (non-migrants). A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table B1.

Table D3: Mother's Linguistic Distance and Other Variables 2: Migrants vs. Non-Migrants

	(1) tetanus	(2) measles	(3) polio	(4) dpt	(5) iron
Migrants					
Linguistic Distance	0.0620**	0.00764	0.0252	0.0387*	0.0271
	(0.0261)	(0.0292)	(0.0199)	(0.0225)	(0.0256)
ELF	0.00164	0.0209	-0.0336	-0.0513*	0.00384
	(0.0260)	(0.0281)	(0.0255)	(0.0279)	(0.0255)
Observations	67000	83952	77421	69444	48693
R^2	0.249	0.336	0.217	0.359	0.310
Non-Migrants					
Linguistic Distance	-0.0639*	-0.0166	-0.00954	0.00510	-0.108***
	(0.0323)	(0.0288)	(0.0251)	(0.0227)	(0.0343)
ELF	0.0196	0.0224	-0.0170	-0.0178	-0.00390
	(0.0272)	(0.0370)	(0.0215)	(0.0274)	(0.0249)
Observations	54496	69350	64774	58565	39198
R^2	0.300	0.340	0.239	0.384	0.351

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The column headings indicate the individual-level dependent variable for each specification. These are: tetanus vaccination, measles immunization, polio vaccination, DPT vaccination, and if the mother received iron tablets during pregnancy. Panel 1 (Panel 2) restricts the sample to only migrants (non-migrants). A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table B1.

D.3 Additional Health Variables

Table D4: Additional Health Variables

	(1) BCG	(2) Antenatal Visits	(3) Full Immunization	(4) Skilled Birth Attd.
Full Sample				
Linguistic Distance	0.0218	-0.0129	0.00394	0.0516*
	(0.0235)	(0.0185)	(0.0116)	(0.0287)
ELF	0.0127	0.00151	-0.0196	0.00886
	(0.0228)	(0.0222)	(0.0171)	(0.0154)
Observations	199113	158482	183371	196254
R^2	0.182	0.408	0.251	0.374
Migrants				
Linguistic Distance	0.0227	0.0353	0.0156	0.0492
	(0.0260)	(0.0265)	(0.0168)	(0.0398)
ELF	-0.0310	-0.00500	-0.0259	-0.0317
	(0.0256)	(0.0251)	(0.0216)	(0.0212)
Observations	85071	68941	78105	84528
R^2	0.182	0.381	0.221	0.355
Non-Migrants				
Linguistic Distance	0.00474	-0.0771***	-0.0154	0.0505*
	(0.0297)	(0.0252)	(0.0167)	(0.0276)
ELF	0.00108	-0.00620	-0.0104	0.0238
	(0.0361)	(0.0269)	(0.0153)	(0.0169)
Observations	70483	56454	65278	64530
R^2	0.200	0.440	0.233	0.359

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The column headings indicate the individual-level dependent variable for each specification. These are: BCG vaccination, Antenatal visits, whether the child received full immunization, and if the delivery was done by a doctor or a nurse (i.e. skilled birth attendance). A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region FE, ethnicity FE, religion FE, year of birth, dummies for wealth index, and geographic isolation controls described in the notes of Table B1. Panel 1 uses the full sample, Panel 2 uses the migrant sample and Panel 3 uses the non-migrant sample.

D.4 Correlates of the Migrant Status

Table D5: Correlates of Migrant status

	(1) 25 km	(2) 50 km	(3) 75 km	(4) 100 km	(5) 125 km
Linguistic Distance	0.0543***	0.0646***	0.0758***	0.0827***	0.0874***
	(0.0196)	(0.0220)	(0.0250)	(0.0265)	(0.0275)
ELF	0.00873	0.0234	0.00887	-0.00671	-0.0103
	(0.0217)	(0.0208)	(0.0294)	(0.0405)	(0.0526)
urban	0.0616***	0.0575***	0.0569***	0.0570***	0.0570***
	(0.0162)	(0.0160)	(0.0161)	(0.0161)	(0.0161)
population	-0.0104	-0.00225	0.00381	0.00287	-0.00140
	(0.00631)	(0.00725)	(0.00871)	(0.0101)	(0.0111)
Log(Distance to capital)	-0.0106	-0.00717	-0.00459	-0.00469	-0.00594
	(0.0113)	(0.0115)	(0.0114)	(0.0111)	(0.0108)
wealth index=2	0.0116	0.0110	0.0108	0.0108	0.0109
	(0.0105)	(0.0104)	(0.0105)	(0.0104)	(0.0104)
wealth index=3	0.0280**	0.0274**	0.0274**	0.0275**	0.0276**
	(0.0126)	(0.0124)	(0.0125)	(0.0124)	(0.0124)
wealth index=4	0.0693***	0.0681***	0.0679***	0.0681***	0.0682***
	(0.0177)	(0.0174)	(0.0174)	(0.0173)	(0.0174)
wealth index=5	0.159***	0.158***	0.158***	0.158***	0.158***
	(0.0359)	(0.0353)	(0.0353)	(0.0353)	(0.0354)
Education: incomplete primary	0.00526	0.00480	0.00484	0.00501	0.00506
	(0.0115)	(0.0114)	(0.0114)	(0.0113)	(0.0114)
Education: complete primary	0.00289	0.00229	0.00223	0.00232	0.00233
	(0.0192)	(0.0189)	(0.0189)	(0.0189)	(0.0190)
Education: incomplete secondary	-0.00331	-0.00374	-0.00387	-0.00388	-0.00389
	(0.0253)	(0.0251)	(0.0251)	(0.0251)	(0.0251)
Education: complete secondary	0.0346	0.0341	0.0342	0.0344	0.0344
	(0.0386)	(0.0383)	(0.0383)	(0.0383)	(0.0383)
Education: higher	0.0194	0.0186	0.0184	0.0186	0.0187
	(0.0337)	(0.0335)	(0.0335)	(0.0335)	(0.0334)
Observations	164141	164141	164141	164141	164141
R^2	0.122	0.122	0.122	0.122	0.122

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the migrant status of the mother. The column headings indicate the radii of the circles around the mother used to construct the linguistic distance, ELF and population variables. "No education" is the excluded category for the educational attainment variable. All columns include controls for survey-wave FE, region FE, ethnicity FE, religion FE, year of birth, and geographic isolation controls described in the notes of Table B1.

E Distance from Dominant Group

Table E1: Summary statistics for Distance from Dominant Group

Variable	Mean	Std. Dev.	Min.	Max.
Distance from Dominant Group in 25 km	0.054	0.222	0	1
Distance from Dominant Group in 50 km	0.056	0.227	0	1
Distance from Dominant Group in 75 km	0.058	0.23	0	1
Distance from Dominant Group in 100 km	0.061	0.236	0	1
Distance from Dominant Group in 125 km	0.064	0.242	0	1
Distance from Dominant Group in 150 km	0.066	0.246	0	1
Distance from Dominant Group in 175 km	0.066	0.245	0	1
Distance from Dominant Group in 200 km	0.064	0.241	0	1
Distance from Dominant Group in 250 km	0.064	0.241	0	1
N		862358		

Table E2: Correlation of Distance from Dominant Group with Linguistic Distance

	Linguistic Distance								
	25 km	50 km	75 km	100 km	125 km	150 km	175 km	200 km	250 km
Dominant Group distance 25 km	0.9077	0.8604	0.8087	0.7678	0.727	0.6917	0.6624	0.6364	0.6006
Dominant Group distance 50 km	0.8792	0.8718	0.8265	0.7822	0.7415	0.7062	0.6769	0.6498	0.6134
Dominant Group distance 75 km	0.8092	0.8508	0.8582	0.8299	0.7942	0.7614	0.732	0.7051	0.6648
Dominant Group distance 100 km	0.758	0.8108	0.846	0.8571	0.8287	0.7975	0.7677	0.7394	0.6938
Dominant Group distance 125 km	0.7177	0.7772	0.8246	0.8529	0.8596	0.8388	0.8118	0.7847	0.7364
Dominant Group distance 150 km	0.6961	0.7565	0.8044	0.8364	0.8577	0.8577	0.8369	0.8108	0.7635
Dominant Group distance 175 km	0.6682	0.7286	0.7766	0.8147	0.8431	0.8553	0.8516	0.8313	0.7878
Dominant Group distance 200 km	0.6168	0.6821	0.7362	0.7832	0.824	0.8452	0.8524	0.8482	0.8135
Dominant Group distance 250 km	0.601	0.6535	0.7006	0.7502	0.796	0.8258	0.8434	0.8513	0.8514

Table E3: Mother's Linguistic Distance, Dominant Distance and Child mortality

	(1)	(2)	(3)	(4)
Linguistic Distance	0.0435***	0.0458*		0.0811*
	(0.0133)	(0.0250)		(0.0452)
Dominant Distance		-0.00187	0.0224***	0.0412
		(0.0120)	(0.00671)	(0.0289)
Linguistic Distance × Dominant Distance				-0.0849
				(0.0604)
Observations	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. Linguistic Distance measures the average distance of the mother from all individuals living around her in the circle. Dominant Distance measures her distance from the dominant (largest in size) group in the circle. A circle of radius 50 km has been considered for calculating the Linguistic and Dominant distance variables. All columns control for: female child dummy, mother's age at birth, mother's age at birth squared, multiple birth indicator, birth order, birth order squared, short birth spacing prior to the birth, short birth spacing after the birth, the location of the mother in the form of an urban dummy, dummies for her educational attainment and her families' wealth index, ELF in the circle, distance of the mother's location from the capital and the logged population in the circle.

Table E4: Public Goods, Information and Linguistic Distance vs. Dominant Distance

	(1) education	(2) literacy	(3) water	(4) electricity	(5) ORS
Dominant Distance	-0.00151	0.0178	0.0322	0.00256	-0.0480**
	(0.0488)	(0.0148)	(0.0658)	(0.0117)	(0.0204)
Observations	73681	60445	74748	72936	72648
R^2	0.474	0.391	0.112	0.546	0.242
Linguistic Distance	0.00757	-0.0295	-0.0339	0.0141	-0.0884
	(0.0787)	(0.0404)	(0.0961)	(0.0383)	(0.0592)
Dominant Distance	-0.00591	0.0348*	0.0322	-0.00564	0.00361
	(0.0607)	(0.0206)	(0.0658)	(0.0247)	(0.0237)
Observations	73681	60445	74748	72936	72648
R^2	0.474	0.391	0.112	0.546	0.242
Linguistic Distance	-0.0158	-0.0405	-0.0339	-0.0138	-0.188
	(0.127)	(0.0624)	(0.0961)	(0.0378)	(0.113)
Dominant Distance	-0.0382	0.0188	0.0322	-0.0444	-0.136*
	(0.133)	(0.0616)	(0.0658)	(0.0496)	(0.0754)
Linguistic Distance × Dominant Distance	0.0600	0.0296	-0.0402	0.0719	0.258*
	(0.261)	(0.115)	(0.151)	(0.0605)	(0.155)
Observations	73681	60445	74748	72936	72648
R^2	0.474	0.391	0.112	0.546	0.243

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The column headings indicate the individual mother-level dependent variable for each specification. These are: educational attainment, literacy, access to water, access to electricity and knowledge about ORS. Dominant Distance measures her distance from the dominant (largest in size) group in the circle. A circle of radius 50 km has been considered for calculating the Linguistic and Dominant distance variables. This table uses the sample of non-migrant mothers. All columns include controls for survey-wave FE, region FE, ethnicity FE, religion FE, year of birth FE, dummies for wealth index, ELF in the circle, distance of the mother's location from the capital and the logged population in the circle.

Table E5: Marginal Effects for Distance from Dominant Group

	Full Sample		Migrants		Non-Migrants	
Circle Radius	Child Deaths	% of SD	Child Deaths	% of SD	Child Deaths	% of SD
25 km	4.17	0.99%	1.70	0.41%	7.88	1.83%
50 km	5.10	1.21%	3.18	0.76%	8.75	2.03%
75 km	6.15	1.47%	4.42	1.05%	10.23	2.38%
100 km	5.66	1.35%	4.74	1.13%	10.15	2.36%
125 km	6.39	1.52%	4.17	0.99%	10.97	2.55%
150 km	5.93	1.41%	2.34	0.56%	11.27	2.62%
175 km	5.28	1.26%	2.17	0.52%	10.82	2.51%
200 km	4.62	1.10%	1.28	0.30%	9.91	2.30%
250 km	4.06	0.97%	0.97	0.23%	7.56	1.76%

Notes: This table provides the marginal effects for the most comprehensive specification indicated in Table 1 of the main paper for circles of alternative radii around the mother but using Dominant Distance instead of Linguistic Distance as the main independent variable (See Table E6 below). Dominant Distance is measured as distance from the dominant (largest in size) group within the circle. The leftmost panel includes the full sample of mothers, the middle panel restricts the sample to only migrant mothers, while the last panel restricts the sample to only non-migrant mothers.

Table E6: Mother's Dominant Distance and Child mortality: 50 km Radius

	(1)	(2)	(3)	(4)	(5)
Dominant Distance 50 km	0.0153**	0.0198***	0.0219***	0.0221***	0.0224***
	(0.00605)	(0.00729)	(0.00671)	(0.00681)	(0.00671)
ELF 50 km	-0.00167	-0.00126	-0.00254	-0.00358	-0.00442
	(0.0105)	(0.0109)	(0.00959)	(0.00909)	(0.00897)
Observations	653666	653666	653666	653666	653666
R^2	0.089	0.091	0.145	0.145	0.154
Survey-wave FE	Y	Y	Y	Y	Y
Region × Year FE	Y	Y	Y	Y	Y
Ethnicity FE	N	Y	Y	Y	N
Religion FE	N	N	Y	Y	Y
Individual Controls	N	N	Y	Y	Y
Geographic isolation	N	N	N	Y	Y
Ethnicity × Year FE	N	N	N	N	Y

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. The numbers after dominant distance and ELF indicate the radius of the circle around the mother in which these variables have been calculated. The individual controls include female child dummy, mother's age at birth, mother's age at birth squared, multiple birth indicator, birth order, birth order squared, short birth spacing prior to the birth, short birth spacing after the birth, the location of the mother in the form of an urban dummy, dummies for her educational attainment and her families' wealth index. Geographical isolation controls include the distance of the mother's location from the capital and the logged population in the circle.

Table E7: Mother's Dominant Distance and Child mortality: Alternative radii

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	25 km	75 km	100 km	125 km	150 km	175 km	200 km	250 km
Dominant Distance	0.0187***	0.0268***	0.0240**	0.0265***	0.0242**	0.0216**	0.0193	0.0170
	(0.00617)	(0.00903)	(0.01000)	(0.00974)	(0.0103)	(0.0104)	(0.0139)	(0.0143)
ELF	-0.00133	-0.00597	-0.00764	-0.00121	0.00867	0.0164	0.0206	-0.00646
	(0.00664)	(0.0111)	(0.0135)	(0.0150)	(0.0161)	(0.0183)	(0.0204)	(0.0266)
Observations	653666	653666	653666	653666	653666	653666	653140	649890
R^2	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. The numbers in the column headings indicate the radius of the circle around the mother in which the dominant distance and ELF have been calculated. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table B1.

Table E8: Mother's Dominant Distance and Other Child Health Variables

	(1) infant	(2) neonatal	(3) HAZ	(4) stunted	(5) WAZ
Dominant Distance	0.0111***	0.00489***	-0.0205	0.0116	-0.0167
	(0.00345)	(0.00123)	(0.0253)	(0.00889)	(0.0168)
ELF	-0.00156	-0.00316	0.0141	-0.00315	0.0966**
	(0.00466)	(0.00196)	(0.0500)	(0.0170)	(0.0399)
Observations	815267	861386	141475	141475	141475
R^2	0.097	0.069	0.205	0.153	0.161

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The column headings indicate the individual child-level dependent variable for each specification. These are: infant mortality, neonatal mortality, height-for-age Z-score (HAZ), stunting, and the weight-for-age Z-score (WAZ). A circle of radius 50 km has been considered for calculating the dominant distance and ELF variables. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table B1.

Table E9: Mother's Dominant Distance and child mortality: Robustness for aggregate diversity

	(1) ELFL15	(2) ELFL10	(3) ELFL5	(4) ELFL2	(5) ONLYELF	(6) ELFSQ	(7) NOELF
Dominant Distance	0.0224***	0.0221***	0.0221***	0.0221***		0.0222***	0.0220***
	(0.00671)	(0.00662)	(0.00637)	(0.00606)		(0.00655)	(0.00698)
ELF	-0.00442	-0.000963	-0.00113	-0.000532	-0.00272	0.00419	
	(0.00897)	(0.00840)	(0.0130)	(0.0157)	(0.00955)	(0.0212)	
ELF squared						-0.00663	
						(0.0254)	
Observations	653666	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154	0.154
	ELPL15	ELPL10	ELPL5	ELPL2	ONLYELP	ELPSQ	BOTH
Dominant Distance	0.0218***	0.0220***	0.0222***	0.0226***		0.0218***	0.0225***
	(0.00688)	(0.00688)	(0.00676)	(0.00655)		(0.00687)	(0.00672)
ELP	0.00298	-0.000248	-0.00140	-0.00388	0.00396	-0.00326	0.0132
	(0.00723)	(0.00659)	(0.00743)	(0.00657)	(0.00745)	(0.0201)	(0.00868)
ELP squared						0.00677	
						(0.0191)	
ELF							-0.0158
							(0.0112)
Observations	653666	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154	0.154

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. In Panel 1 (Panel 2): column 1 controls for ELF (ELP) at aggregation Level 15; column 2 for ELF (ELP) at aggregation Level 10; column 3 for ELF (ELP) at aggregation Level 5; column 4 for ELF (ELP) at aggregation Level 2; column 5 for ELF (ELP) at aggregation Level 15, without LD; column 6 for ELF (ELP) at aggregation Level 15,and its square term. In column 7 of Panel 1, I do not control for ELF or ELP. In column 7 of Panel 2, I include both ELF and ELP. A circle of radius 50 km has been considered for calculating the dominant distance and ELF variables. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table B1.

Table E10: Mother's Dominant Distance and Child mortality: Heterogeneity by Migration Status

	(1) HetMigrant	(2) HetYearsLived	(3) Migrants	(4) NMigrants
Dominant Distance	0.0327***	0.0169**	0.0122**	0.0402***
	(0.00708)	(0.00643)	(0.00557)	(0.00593)
ELF	-0.00758	-0.00758	-0.0153*	-0.00102
	(0.0100)	(0.0100)	(0.00908)	(0.0120)
Het Var.	0.00555***	-0.000325***		
	(0.00150)	(0.0000656)		
Dominant Distance × Het. Variable	-0.0140**	0.000325*		
	(0.00547)	(0.000175)		
Observations	521217	521217	278952	241309
R^2	0.163	0.163	0.177	0.167

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. In column 1 the Het. Variable refers to the 0-1 migrant status of the mother; in column 2 it refers to the continuous variable indicating how many years the mother has been living in the village where she was interviewed. In column 3 (column 4), I restrict the sample to only children of mothers who are migrants (non-migrants). A circle of radius 50 km has been considered for calculating the dominant distance and ELF variables. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table B1.

Table E11: Mother's Dominant Distance, Malaria Suitability and Child mortality

	(1)	(2)	(3)	(4)
	Full Sample		Migrants	NMigrants
Dominant Distance	0.0224***	0.0228***	0.0126**	0.0354***
	(0.00670)	(0.00483)	(0.00552)	(0.00550)
Malaria Suitability	0.00375	0.00307	0.00206	0.00208
	(0.00424)	(0.00424)	(0.00586)	(0.00423)
ELF	-0.00455	-0.00475	-0.0154*	-0.00132
	(0.00901)	(0.00888)	(0.00909)	(0.0119)
Dominant Distance × Malaria Suitability		0.0102***	0.00230	0.0107**
		(0.00343)	(0.00373)	(0.00444)
Observations	653666	653666	278952	241309
R^2	0.154	0.154	0.177	0.167

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. Malaria Suitability is measured by the malaria stability index originally constructed by Kiszewski et al. (2004). Columns 1 and 2 uses the full sample of mothers. In column 3 (column 4), I restrict the sample to only children of mothers who are migrants (non-migrants). A circle of radius 50 km has been considered for calculating the dominant distance and ELF variables. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table B1.

Table E12: Public Goods, Information and Dominant Distance: Migrants vs. Non-Migrants

	(1) education	(2) literacy	(4) water	(5) electricity	(6) ORS
Migrants					
Dominant Distance	-0.0186	-0.00337	-0.0120	0.0120	-0.00406
	(0.0442)	(0.0183)	(0.0153)	(0.0118)	(0.0111)
ELF	-0.0157	-0.0265*	0.0540*	-0.0148	-0.0186
	(0.0319)	(0.0139)	(0.0290)	(0.0129)	(0.0328)
Observations	90456	71929	74748	89536	88542
R^2	0.458	0.423	0.112	0.548	0.206
Non-Migrants					
Dominant Distance	-0.00151	0.0178	0.0322	0.00256	-0.0480**
	(0.0488)	(0.0148)	(0.0658)	(0.0117)	(0.0204)
ELF	0.0707	0.0558*	0.0554*	-0.00701	-0.0118
	(0.0533)	(0.0287)	(0.0281)	(0.00892)	(0.0323)
Observations	73681	60445	74748	72936	72648
R^2	0.474	0.391	0.112	0.546	0.242

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The column headings indicate the individual mother-level dependent variable for each specification. These are: educational attainment, literacy, access to water, access to electricity and knowledge about ORS. A circle of radius 50 km has been considered for calculating the dominant distance and ELF variables. All columns include controls for survey-wave FE, region FE, ethnicity FE, religion FE, year of birth FE, dummies for wealth index, and geographic isolation controls described in the notes of Table B1.

F Alternative Fixed Effects

F.1 No regional FE

Table F1: Public Goods, Information and Linguistic Distance: Migrants vs. Non-Migrants - No Region FE

	(1) education	(2) literacy	(4) water	(5) electricity	(6) ORS
Migrants					
Linguistic Distance	0.0112	0.00697	-0.0261	0.0104	-0.0253
	(0.0765)	(0.0334)	(0.0277)	(0.0311)	(0.0200)
ELF	0.0146	0.0135	0.0427*	0.0105	-0.0110
	(0.0397)	(0.0200)	(0.0252)	(0.0247)	(0.0286)
Observations	90456	71929	74748	89536	88542
R^2	0.451	0.416	0.099	0.527	0.191
Non-Migrants					
Linguistic Distance	-0.0844	-0.00493	-0.0261	-0.00411	-0.104***
	(0.102)	(0.0301)	(0.0277)	(0.0222)	(0.0323)
ELF	0.132*	0.0687**	0.0427*	0.0150	0.000966
	(0.0737)	(0.0280)	(0.0252)	(0.0155)	(0.0320)
Observations	73681	60445	74748	72936	72648
R^2	0.456	0.382	0.099	0.526	0.225

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region FE, ethnicity FE, religion FE, year of birth FE, dummies for wealth index, and geographic isolation controls described in the notes of Table B1.

Table F2: Public Goods, Information and Linguistic Distance: Full Sample - No Region FE

	(1) education	(2) literacy	(4) water	(5) electricity	(6) ORS
Linguistic Distance	-0.0316	-0.00381	0.00275	-0.00446	-0.0530**
	(0.0740)	(0.0290)	(0.0294)	(0.0271)	(0.0242)
ELF	0.0550	0.0318	0.0266	0.0242	0.00678
	(0.0515)	(0.0216)	(0.0196)	(0.0200)	(0.0270)
Observations	205986	168365	179965	203565	202924
R^2	0.445	0.394	0.125	0.544	0.182

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The column headings indicate the individual mother-level dependent variable for each specification. These are: educational attainment, literacy, access to water, access to electricity and knowledge about ORS. A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region FE, ethnicity FE, religion FE, year of birth FE, dummies for wealth index, and geographic isolation controls described in the notes of Table B1.

Table F3: Within vs Across-Region Correlations: Public Goods, Information and Linguistic Distance

	Migrant Sample			Non-Migrant Sample		
	Overall	Across-Region	Within-Region	Overall	Across-Region	Within-Region
education	0.18	0.41	0.01	0.11	0.43	0.02
literacy	0.13	0.50	0.00	0.05	0.44	0.01
water	-0.02	-0.12	-0.01	0.02	0.00	0.00
electricity	0.15	0.27	0.06	0.12	0.25	0.05
ORS	0.02	0.16	0.01	-0.03	0.22	0.01

Notes: This table provides the overall, across-region and within-region (unconditional) correlation coefficients between linguistic distance and the four different public goods variables: educational attainment, literacy, access to water, access to electricity; and the ORS variable. The left (right) panel restricts the data to the migrant (non-migrant) sample. The column headings indicate the type of correlation coefficient and the row headings indicate the variable with which the correlation has been calculated. The overall correlation coefficient refers to the simple correlation between the two variables of interest using the individual-level observations. To calculate the across-region correlation coefficients I take averages at the region level for the 109 regions and then calculate the correlation coefficient for the two variables of interest using the region-level observations. Finally, to calculate the within-region correlation coefficient I calculate the correlations within each region using the individual-level observations. This leads to approximately 109 correlation coefficients. I then take the average (mean value) of these correlation coefficients to calculate the final within-region correlation coefficient.

Table F4: Within vs Across-Region Correlations: Public Goods, Information and Linguistic Distance (Full Sample)

	Full Sample		
	Overall	Across-Region	Within-Region
education	0.16	0.42	0.01
literacy	0.11	0.44	0.00
water	0.01	0.10	-0.02
electricity	0.11	0.24	0.05
ORS	0.00	0.22	0.01

Notes: This table provides the overall, across-region and within-region (unconditional) correlation coefficients between linguistic distance and the four different public goods variables: educational attainment, literacy, access to water, access to electricity; and the ORS variable for the full sample. The column headings indicate the type of correlation coefficient and the row headings indicate the variable with which the correlation has been calculated. See notes to Table F3 for definitions of the alternative concepts of correlation used in this table.

F.2 DHS Cluster FE

Table F5: Mother's Linguistic Distance and Child mortality: DHS cluster FE

	(1)	(2)	(3)	(4)	(5)
	25 km	50 km	75 km	100 km	125 km
Linguistic Distance	0.00373	0.00985	0.0140	0.0161*	0.0187*
	(0.00771)	(0.00848)	(0.00914)	(0.00946)	(0.00984)
Observations	653646	653646	653646	653646	653646
R^2	0.182	0.182	0.182	0.182	0.182

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the DHS cluster level. The dependent variable is the individual child-level mortality outcome. The column headings indicate the radii of the circles around the mother used to construct the linguistic distance variable. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls described in the notes of Table B1. The inclusion of DHS cluster FE implies that controls for ELF, urban dummy, population and geographic isolation drop out.

G Heterogeneity

G.1 Heterogeneity by Additional Variables

Table G1: Mother's Linguistic Distance and Child mortality: Heterogeneity

	(1) female	(2) urban	(3) education	(4) ELF	(5) ELP	(6) population	(7) lndist2cap	(8) wealth
Linguistic Distance	0.0439***	0.0470***	0.0454***	0.0365***	0.0364***	0.0456*	0.0245	0.0598**
	(0.0150)	(0.0173)	(0.0156)	(0.0114)	(0.0136)	(0.0263)	(0.0182)	(0.0292)
Het. Variable	-0.0180***	-0.0148***	-0.00398***	-0.00827	0.00140	0.00581**	0.00405*	-0.00911***
	(0.00136)	(0.00302)	(0.000405)	(0.00797)	(0.00700)	(0.00260)	(0.00219)	(0.000998)
Interaction Term	-0.000938	-0.00826	-0.000624	0.0139	0.00802	-0.000171	0.00356	-0.00477
	(0.00572)	(0.0141)	(0.00166)	(0.0312)	(0.0197)	(0.00247)	(0.00247)	(0.00583)
Observations	653666	653666	653413	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. The column headings indicate the Het. Variable included in the specification. The individual controls include female child dummy, mother's age at birth, mother's age at birth squared, multiple birth indicator, birth order, birth order squared, short birth spacing prior to the birth, short birth spacing after the birth, the location of the mother in the form of an urban dummy, dummies for her educational attainment and her families' wealth index. Geographical isolation controls include the distance of the mother's location from the capital and the logged population in the circle.

G.2 Heterogeneity by Migration years and Age of Migrant

Table G2: Mother's Linguistic Distance and Child mortality: Heterogeneity by Age Migrated and Residence Years

	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic Distance	0.0196**	0.0196**	0.00536	0.0196**	0.0181*	0.0123
	(0.00933)	(0.00934)	(0.00992)	(0.00933)	(0.00948)	(0.0114)
Residence Years	-0.00774***	-0.00792***				
	(0.00109)	(0.00120)				
Linguistic Distance × Residence Years		0.00185				
		(0.00428)				
Veteran Migrant			-0.0250***			
			(0.00226)			
Linguistic Distance × Veteran Migrant			0.0232***			
			(0.00709)			
Age Migrated				0.00732***	0.00674***	
				(0.00103)	(0.00112)	
Linguistic Distance × Age Migrated					0.00577	
					(0.00401)	
Old Migrant						0.00880***
						(0.00236)
Linguistic Distance × Old Migrant						0.0121
						(0.00774)
Observations	278952	278952	278952	278952	278952	278952
R^2	0.177	0.177	0.177	0.177	0.177	0.177

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. A circle of radius 50 km has been considered for calculating the linguistic distance variable. This table uses the sample of migrant mothers. All columns control for: a female child dummy, mother's age at birth, mother's age at birth squared, multiple birth indicator, birth order, birth order squared, short birth spacing prior to the birth, short birth spacing after the birth, the location of the mother in the form of an urban dummy, dummies for her educational attainment, her families' wealth index, distance of the mother's location from the capital, the logged population in the circle and ELF in the circle. The Residence Years variable measures the standardized value of the number of years the migrant has lived in her current place of residence. The Veteran Migrant variable is a binary 0-1 indicator variable which takes the value 1 if the migrant has lived in her current village of residence for more than the median number of years in the sample. The Age Migrated variable measures the standardized value of age at which migrant migrated to her current place of residence. The Old Migrant variable is a binary 0-1 indicator variable which takes the value 1 if the migrant's age at migration was more than the median number of years in the sample.

Table G3: ORS and Linguistic Distance: Heterogeneity by Age Migrated and Residence Years

	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic Distance	-0.0244	-0.0254	-0.0205	-0.0244	-0.0262	-0.0495
	(0.0200)	(0.0201)	(0.0204)	(0.0200)	(0.0201)	(0.0313)
Residence Years	0.00716**	0.00797**				
	(0.00298)	(0.00319)				
Linguistic Distance × Residence Years		-0.00739				
		(0.00775)				
Veteran Migrant			0.00192			
			(0.00718)			
Linguistic Distance × Veteran Migrant			-0.0623**			
			(0.0268)			
Age Migrated				-0.00684**	-0.00791***	
				(0.00285)	(0.00273)	
Linguistic Distance × Age Migrated					0.00914	
					(0.00930)	
Old Migrant						-0.00813
						(0.00566)
Linguistic Distance × Old Migrant						0.0272
						(0.0246)
Observations	88542	88542	88542	88542	88542	88542
R^2	0.191	0.191	0.191	0.191	0.191	0.191

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual mother-level ORS knowledge variable. A circle of radius 50 km has been considered for calculating the linguistic distance variable. This table uses the sample of migrant mothers. All columns include controls for survey-wave FE, region FE, ethnicity FE, religion FE, year of birth FE, dummies for wealth index, and geographic isolation controls described in the notes of Table B1. The Residence Years variable measures the standardized value of the number of years the migrant has lived in her current place of residence. The Veteran Migrant variable is a binary 0-1 indicator variable which takes the value 1 if the migrant has lived in her current village of residence for more than the median number of years in the sample. The Age Migrated variable measures the standardized value of age at which migrant migrated to her current place of residence. The Old Migrant variable is a binary 0-1 indicator variable which takes the value 1 if the migrant's age at migration was more than the median number of years in the sample.

G.3 Heterogeneity by age

Table G4: Mother's Linguistic Distance and Child mortality: Heteorgeneity by Migration & Age at birth

	(1) HetYearsLived	(2) HetYearsLived2	(3) Migrants	(4) NMigrants	(5) Migrants_age	(6) NMigrants_age
Linguistic Distance	0.0341***	0.0393***	0.0198**	0.0758***	0.0222**	0.0767***
	(0.0114)	(0.0116)	(0.00935)	(0.0137)	(0.00938)	(0.0138)
Residence Years	-0.000348***	-0.000331***				
	(0.0000718)	(0.0000720)				
Linguistic Distance × Residence Years	0.000555**	0.000390				
	(0.000270)	(0.000292)				
Age	-0.0349***	-0.0359***	-0.0332***	-0.0395***	-0.0347***	-0.0402***
	(0.00218)	(0.00223)	(0.00210)	(0.00282)	(0.00209)	(0.00287)
Age squared	0.0160***	0.0160***	0.0150***	0.0171***	0.0151***	0.0172***
	(0.000846)	(0.000843)	(0.000962)	(0.00110)	(0.000954)	(0.00110)
Linguistic Distance × Age		0.0113**			0.0160**	0.00898*
		(0.00496)			(0.00624)	(0.00512)
Observations	521217	521217	278952	241309	278952	241309
R^2	0.163	0.163	0.177	0.167	0.177	0.167

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. A circle of radius 50 km has been considered for calculating the linguistic distance variable. All columns control for: a female child dummy, mother's age at birth, mother's age at birth squared, multiple birth indicator, birth order, birth order squared, short birth spacing prior to the birth, short birth spacing after the birth, the location of the mother in the form of an urban dummy, dummies for her educational attainment, her families' wealth index, distance of the mother's location from the capital, the logged population in the circle and ELF in the circle. The age variables are standardized. The Residence Years variable measures the number of years the migrant has lived in her current place of residence.

H Alternative measures

H.1 Coethnics

Table H1: Mother's Linguistic Distance and Child mortality: Kin Networks

	(1) nCEMothers	(2) pCEMothers	(3) pCoethnic	(4) nCoethnic	(5) MotherDist	(6) CoethnicDist
Panel A: Coethnics without Linguistic Distance						
Coethnic Var.	0.00420	0.00225	-0.159	0.00363	-0.00187	0.000795
	(0.0268)	(0.00630)	(0.258)	(0.00522)	(0.00183)	(0.00197)
Observations	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154
Panel B: Coethnics with Linguistic Distance						
Linguistic Distance	0.0436***	0.0457***	0.0435***	0.0468***	0.0475***	0.0456***
	(0.0133)	(0.0130)	(0.0133)	(0.0134)	(0.0142)	(0.0138)
Coethnic Var.	0.00693	0.00635	0.00359	0.00827	-0.00387*	-0.00150
	(0.0269)	(0.00553)	(0.261)	(0.00536)	(0.00218)	(0.00180)
Observations	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154
Panel C: Interaction of Coethnics with Linguistic Distance						
Linguistic Distance	0.0417***	0.0448***	0.0435***	0.0465***	0.0458***	0.0417***
	(0.0137)	(0.0165)	(0.0135)	(0.0135)	(0.0164)	(0.0156)
Coethnic Var.	0.00597	0.00612	0.00313	0.00789	-0.00407*	-0.00205
	(0.0268)	(0.00598)	(0.282)	(0.00556)	(0.00221)	(0.00217)
Linguistic Distance × Coethnic Var.	0.186	0.00333	0.00904	0.0132	0.00100	0.00261
	(0.191)	(0.0199)	(1.541)	(0.0370)	(0.00339)	(0.00378)
Observations	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. The column heading indicates the coethnic variable used in the specification. They are: nCEMothers: number of coethnic mothers in the circle; pCEMothers: proportion of coethnic mothers in the circle; nCoethnic: number of coethnic individuals in the circle; pCoethnic: proportion of coethnic individuals in the circle; MotherDist: average geographic (Euclidean) distance from coethnic mothers in the country; CoethnicDist: average geographic (Euclidean) distance from all individuals belonging to the same ethnicity in the country. All the measures based on the DHS data are weighted by the DHS sample weights. A circle of radius 50 km has been considered for calculating the linguistic distance and coethnic variables. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, ELF in the circle, individual controls and geographic isolation controls described in the notes of Table B1.

H.2 Measure of Simple LD

Table H2: Mother's Linguistic Distance and Child mortality: Extensive vs. Intensive Margin

	(1)	(2)	(3)	(4)
	Binary LD variable where LD = 0 if			
	LD = 0	LD < 0.001	$LD > 0$	$LD > 0.001$
LD	-0.00172	-0.00102	0.0433***	0.0388***
	(0.00463)	(0.00549)	(0.0135)	(0.0118)
Observations	653666	653666	577897	532199
R^2	0.154	0.154	0.157	0.158

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. The column heading indicates the robustness test undertaken. Columns 1 and 2 use a binary measure of LD with LD equal to zero if either $LD = 0$ (column 1) or $LD < 0.001$ (column 2). Otherwise LD takes the value 1. Columns 3–4 use a continuous measure of LD as used in the rest of the paper, but restricting the population to mothers who have $LD > 0$ (column 3) and $LD > 0.001$ (column 4). A circle of radius 50 km has been considered for calculating the linguistic distance variable. All columns include controls survey-wave FE, region × year FE, ethnicity × year FE, religion FE, ELF in the circle, individual controls and geographic isolation controls described in the notes of Table B1.

H.3 Cluster-based Spatial Distribution

As discussed in Section 2.1, the spatial distribution of linguistic groups is based on an iterative proportional fitting algorithm combining different datasets. This section calculates LD and ELF using cluster-level information on the ethnic groups from DHS survey data. I follow two approaches. The first approach uses information from each DHS cluster along with nearby clusters to predict the spatial distribution of ethnic groups within circles of different radii. This is my preferred approach. The second approach uses only the information from the DHS cluster of each mother without taking into consideration other nearby clusters. Appendix Table H3 provides the summary statistics and Appendix Table H4 provides the correlation of the IPF-based LD and ELF variables with the cluster data-based LD and ELF variables. The correlation of LD based on the iterative proportional fitting algorithm with the cluster-level LD based on the first approach is approximately 0.8 and that based on the second approach is approximately 0.6 (for a circle of 50km radius). The high correlation of LD based on the IPF-based spatial distribution and DHS survey data-based spatial distribution increases my faith on the IPF algorithm-based data.

Panel A of Table H5 shows that using measures of LD and ELF using data from cluster of the

mother's location and the nearby clusters with the specified circular radius (50km in this case) yield economically and statistically comparable results. Furthermore, the correlation between the IPF based LD and the cluster level LD is of 0.8 (see Table H4). Panel B of Table H5 on the other hand presents results based on using only the cluster-level information. The correlation of LD based on this method with the IPF based LD is of 0.6 (see Table H4).

Measures of linguistic distance based on the spatial distribution of ethnic groups constructed using the IPF algorithm combing the Ethnologue and the LandScan databases are better than the DHS based measures due to several reasons. First, the DHS data are based on nationally-representative surveys comparable across countries. Individual-level observations from these data can be used as outcome variables. However, DHS is not the most appropriate source to calculate the distribution of ethnic groups at a geographically disaggregated level, as it is not representative at that level. The average DHS cluster does not have enough variation in terms of number of groups. On the other hand, at the aggregate level Desmet et al. (2020) show very high correlation of the IPF based data with the other census based data of Gershman and Rivera (2018). Focusing on the regions for which Gershman and Rivera (2018) use census data, Desmet et al. (2020) find correlations in local diversity of 0.80 at the regional level and 0.95 at the country level.

Table H3: Summary statistics of cluster-based LD and ELF

Variable	Mean	Std. Dev.	Min.	Max.
LD (M1 un-weighted)	0.055	0.158	0	0.999
LD (M1 weighted)	0.047	0.162	0	1
LD (M2 un-weighted)	0.03	0.121	0	0.983
LD (M2 weighted)	0.024	0.125	0	1
ELF (M1 un-weighted)	0.467	0.221	0	0.884
ELF (M1 weighted)	0.245	0.213	0	0.815
ELF (M2 weighted)	0.142	0.204	0	0.857
ELF (M2 un-weighted)	0.268	0.254	0	0.875
N		206076		

Notes: LD (IPF) and ELF (IPF) refer to linguistic distance and ELF calculated using spatial distribution of language groups based on the IPF algorithm. M1 refers to linguistic distance and ELF calculated using data from the cluster of the mother's location along with nearby clusters within 50km around the mother. M2 refers to using linguistic distance ELF calculated using using data from the cluster of the mother's location alone without considering nearby clusters. Weighted and un-weighted indicate whether the observations have been weighted by DHS-provided sample weights.

Table H4: Cross-correlation of IPF with DHS

Variables	LD (IPF)
LD (M1 unweighted)	0.773
LD (M1 weighted)	0.744
LD (M2 unweighted)	0.582
LD (M2 weighted)	0.527
Variables	**ELF (IPF)**
ELF (M1 unweighted)	0.543
ELF (M1 weighted)	0.491
ELF (M2 unweighted)	0.194
ELF (M2 weighted)	0.165

Notes: See notes to Table H3 for variable details.

Table H5: Mother's Linguistic Distance and Child mortality: Cluster-based Measures

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Cluster-based spatial distribution						
Linguistic Distance	0.0334*	0.0416**	0.0334***	0.0416***	0.0334**	0.0416**
	(0.0181)	(0.0203)	(0.00746)	(0.00889)	(0.0167)	(0.0188)
ELF	-0.00140	0.00368	-0.00140	0.00368	-0.00140	0.00368
	(0.00865)	(0.00989)	(0.00494)	(0.00539)	(0.00685)	(0.00796)
Observations	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154
Panel B: LD, ELF at the cluster Level						
Linguistic Distance	0.0125	0.0166*	0.0125*	0.0166*	0.0125	0.0166*
	(0.0131)	(0.00901)	(0.00741)	(0.00901)	(0.0110)	(0.00901)
ELF	-0.000409	-0.00646	-0.000409	-0.00646	-0.000409	-0.00646
	(0.00747)	(0.00395)	(0.00477)	(0.00395)	(0.00654)	(0.00395)
Observations	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154
Weighted measures	Yes	No	Yes	No	Yes	No
SE Cluster	Region	Region	DHS cluster	DHS cluster	Ethnicity	Ethnicity

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at level mentioned in the row labelled SE Cluster. The dependent variable is the individual child-level mortality outcome. A circle of radius 50 km has been considered for calculating the linguistic distance, ELF and population variables. The LD and ELF variables in Panel A have been calculated using cluster-level information on the ethnic groups from DHS survey data combining data from the cluster of the mother's location as well as nearby clusters within the specified circular radius (50km in this case). The LD and ELF variables in Panel B have been calculated using cluster-level information on the ethnic groups from DHS survey data exploiting data solely from the cluster of the mother's location. The odd-numbered columns weight the measures of LD and ELF with DHS-provided sample weights. The even-numbered columns do not weight the measures of LD and ELF with DHS-provided sample weight. All columns include controls survey-wave FE, region × year FE, ethnicity × year FE, religion FE, ELF in the circle, individual controls and geographic isolation controls described in the notes of Table B1.

H.4 Measures of Diversity incorporating Distances

This section establishes the robustness of linguistic distance to controlling for distance-weighted diversity indices. In particular, I calculate the Greenberg index of linguistic diversity (Greenberg, 1956), and an index of polarization incorporating based on Duclos et al. (2004) and Esteban and Ray (2011). The Greenberg index of linguistic diversity (GI_j)and the polarization index (PI_j) following Duclos et al. (2004) and Esteban and Ray (2011) are defined as follows:

$$GI_j = \sum_{i=1}^{n} \sum_{v=1}^{n} s_{i(j)} s_{v(j)} \tau_{iv} \tag{5}$$

$$PI_j = \sum_{i=1}^{n} \sum_{v=1}^{n} [s_{i(j)}]^2 s_{v(j)} \tau_{iv} \tag{6}$$

where $s_{i(j)}$ and $s_{v(j)}$ are the population shares of language groups i and v, and τ_{iv} is the distance between languages i and v defined by formula (1). The Greenberg index can be thought of an ELF index that incorporates continuous distances directly into the index and can be interpreted as the expected linguistic distance between any two randomly selected individuals in the region (Desmet et al., 2009). Likewise, the polarization index in equation (6) is similar to that defined in equation (4), but incorporates the linguistic distance between the groups. Table H6 provides the summary statistics for the new indices (for a circle of 50km radius around the mother) and Table H7 provides the regression results controlling for the indices.

Table H6: Summary statistics for GI and PI

Variable	Mean	Std. Dev.	Min.	Max.
GI ($\delta = 0.0025$)	0.075	0.131	0	0.634
GI ($\delta = 0.05$)	0.102	0.129	0	0.65
GI ($\delta = 0.5$)	0.269	0.177	0	0.758
PI ($\delta = 0.0025$)	0.02	0.041	0	0.242
PI ($\delta = 0.05$)	0.028	0.04	0	0.242
PI ($\delta = 0.05$)	0.075	0.047	0	0.243
N		862358		

Table H7: Mother's Linguistic Distance and Child mortality: Diversity with distances

	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic Distance	0.0300**	0.0328***	0.0424***	0.0322**	0.0329**	0.0395***
	(0.0115)	(0.0108)	(0.0118)	(0.0142)	(0.0139)	(0.0136)
GI ($\delta = 0.0025$)	0.0406					
	(0.0284)					
GI ($\delta = 0.05$)		0.0306				
		(0.0293)				
GI ($\delta = 0.5$)			-0.00371			
			(0.0151)			
PI ($\delta = 0.0025$)				0.125*		
				(0.0670)		
PI ($\delta = 0.05$)					0.118*	
					(0.0674)	
PI ($\delta = 0.05$)						0.0310
						(0.0418)
Observations	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at region level. The dependent variable is the individual child-level mortality outcome. A circle of radius 50 km has been considered for calculating the linguistic distance, GI, PI and population variables. GI refers to the Greenberg index of linguistic diversity (Greenberg, 1956) defined in equation (5) and was calculated using three alternative values of δ as highlighted by the row labels. PI refers to the Polarization index of linguistic diversity (Duclos et al., 2004; Esteban and Ray, 2011) defined in equation (6) and was calculated using three alternative values of δ as highlighted by the row labels. All columns include controls survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table B1.

I Alternative clustering of Standard Errors

Table I1: Mother's Linguistic Distance and Child mortality: Alternative clustering of SE

	(1)	(2)	(3)	(4)	(5)
	25 km	50 km	75 km	100 km	125 km
Panel A: Cluster SE at ethnicity Level					
Linguistic Distance	0.0347***	0.0435***	0.0474***	0.0487***	0.0528***
	(0.0103)	(0.0136)	(0.0170)	(0.0184)	(0.0200)
ELF	-0.00372	-0.00739	-0.00893	-0.0112	-0.00539
	(0.00544)	(0.00735)	(0.00862)	(0.00996)	(0.0118)
Observations	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154
Panel A: Cluster SE at ethnicity and region Level					
Linguistic Distance	0.0347***	0.0435***	0.0474**	0.0487**	0.0528**
	(0.0117)	(0.0160)	(0.0207)	(0.0224)	(0.0240)
ELF	-0.00372	-0.00739	-0.00893	-0.0112	-0.00540
	(0.00626)	(0.00848)	(0.0103)	(0.0124)	(0.0140)
Observations	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the ethnicity level in Panel A; and ethnicity and region level in Panel B. The dependent variable is the individual child-level mortality outcome. A circle of radius 50 km has been considered for calculating the LD, ELF and population variables.All columns include controls survey-wave FE, region × year FE, ethnicity × year FE, religion FE, ELF in the circle, individual controls and geographic isolation controls described in the notes of Table B1.

J Malaria Robustness

Table J1: Mother's Linguistic Distance, Malaria Suitability and Child mortality (Robustness)

	(1) est1	(2) est2	(3) est3	(4) est4	(5) est5	(6) est6	(7) est7
Linguistic Distance	0.0408***	0.0412***	0.0407***	0.0409***	0.0411***	0.0443***	0.0455**
	(0.00845)	(0.00910)	(0.00930)	(0.00988)	(0.0105)	(0.0119)	(0.0185)
Malaria Suitability	0.00213	0.00191	0.00164	0.00229	0.00212	0.00211	0.00177
	(0.00420)	(0.00416)	(0.00417)	(0.00423)	(0.00420)	(0.00420)	(0.00419)
Linguistic Distance × Malaria Suitability	0.0173***	0.0175***	0.0178***	0.0172***	0.0174***	0.0174***	0.0177***
	(0.00560)	(0.00553)	(0.00564)	(0.00588)	(0.00591)	(0.00563)	(0.00624)
ELF	-0.00752	-0.00791	-0.00799	-0.00776	-0.00751	-0.00777	-0.00850
	(0.00821)	(0.00846)	(0.00832)	(0.00820)	(0.00826)	(0.00810)	(0.00822)
Urban Residence	-0.0154***	-0.0153***	-0.0153***	-0.0154***	-0.0154***	-0.0147***	-0.0145***
	(0.00282)	(0.00282)	(0.00281)	(0.00280)	(0.00282)	(0.00305)	(0.00307)
Population	0.00708**	0.00716**	0.00718**	0.00689**	0.00703**	0.00711**	0.00688**
	(0.00296)	(0.00298)	(0.00298)	(0.00292)	(0.00291)	(0.00300)	(0.00287)
Tse Tse No. Species		0.00142					
		(0.00214)					
Linguistic Distance × Tse Tse No. Species		-0.000899					
		(0.00382)					
Tse Tse Suitability			0.00585				0.00601
			(0.00461)				(0.00469)
Linguistic Distance × Tse Tse Suitability			-0.000584				-0.00118
			(0.00934)				(0.0113)
Crop Suitability				-0.00126			-0.00136
				(0.00105)			(0.00104)
Linguistic Distance × Crop Suitability				-0.000223			0.000243
				(0.00388)			(0.00348)
Linguistic Distance × Population					0.000386		0.00118
					(0.00365)		(0.00489)
Linguistic Distance × Urban Residence						-0.00825	-0.00907
						(0.0139)	(0.0156)
Observations	653666	653666	653666	653666	653666	653666	653666
R^2	0.154	0.154	0.154	0.154	0.154	0.154	0.154

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. Malaria Suitability is measured by the malaria stability index originally constructed by Kiszewski et al. (2004). Different columns undertake different robustness tests by introducing interactions of LD with: No. of Tse Tse species in the area (column 2), general Tse Tse suitability (a binary 0-1 variable) in the area (column 3), soil suitability for crops (column 4); population density (column 5), and an urban dummy (column 6). Finally, column 7 includes all the different interactions together (for Tse Tse suitability only one of the measures is used). A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table B1. Data on Tse Tse suitability have been downloaded from the FAO (27/03/2019). Soil suitability is measured by the crop suitability index estimated for low input level rain-fed cereals downloaded from the FAO (28/03/2019).

Table J2: Mother's Linguistic Distance, Malaria Suitability and Child mortality: Migrant sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Linguistic Distance	0.0196**	0.0166*	0.0154	0.0196**	0.0199**	0.0180	0.0121
	(0.00874)	(0.00965)	(0.0103)	(0.00960)	(0.00976)	(0.0109)	(0.0161)
Malaria Suitability	0.00181	0.00192	0.00150	0.00205	0.00180	0.00184	0.00179
	(0.00580)	(0.00573)	(0.00572)	(0.00580)	(0.00579)	(0.00580)	(0.00571)
Linguistic Distance × Malaria Suitability	0.00419	0.00368	0.00395	0.00410	0.00417	0.00413	0.00380
	(0.00500)	(0.00531)	(0.00534)	(0.00499)	(0.00497)	(0.00505)	(0.00523)
Urban Residence	-0.0160***	-0.0160***	-0.0159***	-0.0159***	-0.0160***	-0.0163***	-0.0163***
	(0.00315)	(0.00316)	(0.00312)	(0.00312)	(0.00315)	(0.00320)	(0.00316)
Population	0.0110***	0.0109***	0.0110***	0.0108***	0.0110***	0.0110***	0.0109***
	(0.00277)	(0.00279)	(0.00279)	(0.00275)	(0.00282)	(0.00279)	(0.00281)
Tse Tse No. Species		-0.00152					
		(0.00220)					
Linguistic Distance × Tse Tse No. Species		0.00460					
		(0.00339)					
Tse Tse Suitability			0.000305				0.000255
			(0.00583)				(0.00590)
Linguistic Distance × Tse Tse Suitability			0.0127				0.0139
			(0.0112)				(0.0120)
Crop Suitability				-0.00124			-0.00129
				(0.00152)			(0.00154)
Linguistic Distance × Crop Suitability				0.00000137			0.000185
				(0.00428)			(0.00403)
Linguistic Distance × Population					0.000499		-0.00116
					(0.00373)		(0.00441)
Linguistic Distance × Urban Residence						0.00319	0.00433
						(0.0132)	(0.0143)
Observations	278952	278952	278952	278952	278952	278952	278952
R^2	0.177	0.177	0.177	0.177	0.177	0.177	0.177

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. Malaria Suitability is measured by the malaria stability index originally constructed by Kiszewski et al. (2004). Different columns undertake different robustness tests by introducing interactions of LD with: No. of Tse Tse species in the area (column 2), general Tse Tse suitability (a binary 0-1 variable) in the area (column 3), soil suitability for crops (column 4); population density (column 5), and an urban dummy (column 6). Finally, column 7 includes all the different interactions together (for Tse Tse suitability only one of the measures is used). A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table B1. Data on Tse Tse suitability have been downloaded from the FAO (27/03/2019). Soil suitability is measured by the crop suitability index estimated for low input level rain-fed cereals downloaded from the FAO (28/03/2019).

Table J3: Mother's Linguistic Distance, Malaria Suitability and Child mortality: Non-Migrant sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Linguistic Distance	0.0642***	0.0697***	0.0680***	0.0687***	0.0698***	0.0680***	0.0856***
	(0.00948)	(0.00937)	(0.00943)	(0.0131)	(0.0153)	(0.0131)	(0.0270)
Malaria Suitability	0.000895	0.000805	0.000587	0.000909	0.000791	0.000936	0.000568
	(0.00410)	(0.00412)	(0.00412)	(0.00412)	(0.00408)	(0.00411)	(0.00413)
Linguistic Distance × Malaria Suitability	0.0218***	0.0202***	0.0213***	0.0229***	0.0222***	0.0216***	0.0222***
	(0.00715)	(0.00664)	(0.00677)	(0.00838)	(0.00815)	(0.00715)	(0.00826)
Urban Residence	-0.0259***	-0.0258***	-0.0258***	-0.0259***	-0.0260***	-0.0249***	-0.0245***
	(0.00438)	(0.00442)	(0.00439)	(0.00438)	(0.00438)	(0.00486)	(0.00497)
Population	0.0100**	0.0104**	0.0104**	0.00977**	0.00938**	0.0101**	0.00936**
	(0.00403)	(0.00401)	(0.00397)	(0.00395)	(0.00367)	(0.00409)	(0.00359)
Tse Tse No. Species		0.00200					
		(0.00255)					
Linguistic Distance × Tse Tse No. Species		-0.00982					
		(0.00670)					
Tse Tse Suitability			0.00688				0.00740
			(0.00572)				(0.00578)
Linguistic Distance × Tse Tse Suitability			-0.0174				-0.0256
			(0.0191)				(0.0240)
Crop Suitability				-0.000464			-0.000620
				(0.00139)			(0.00139)
Linguistic Distance × Crop Suitability				-0.00593			-0.00439
				(0.00742)			(0.00713)
Linguistic Distance × Population					0.00465		0.00661
					(0.00731)		(0.00928)
Linguistic Distance × Urban Residence						-0.0108	-0.0132
						(0.0166)	(0.0198)
Observations	241309	241309	241309	241309	241309	241309	241309
R^2	0.167	0.167	0.167	0.167	0.167	0.167	0.167

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable is the individual child-level mortality outcome. Malaria Suitability is measured by the malaria stability index originally constructed by Kiszewski et al. (2004). Different columns undertake different robustness tests by introducing interactions of LD with: No. of Tse Tse species in the area (column 2), general Tse Tse suitability (a binary 0-1 variable) in the area (column 3), soil suitability for crops (column 4); population density (column 5), and an urban dummy (column 6). Finally, column 7 includes all the different interactions together (for Tse Tse suitability only one of the measures is used). A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table B1. Data on Tse Tse suitability have been downloaded from the FAO (27/03/2019). Soil suitability is measured by the crop suitability index estimated for low input level rain-fed cereals downloaded from the FAO (28/03/2019).

Table J4: Bednet Possession and Use

	Bednet Possession			Bednet Use		
Linguistic Distance	0.0327	0.0324	0.0345*	0.0635*	0.0636*	0.0534*
	(0.0300)	(0.0298)	(0.0207)	(0.0350)	(0.0350)	(0.0281)
Malaria Malaria Suitability		0.0267	0.0230		-0.00652	-0.00980
		(0.0192)	(0.0197)		(0.0170)	(0.0176)
Linguistic Distance × Malaria Suitability			0.0349**			0.0311*
			(0.0141)			(0.0169)
Observations	159084	159084	159084	106656	106656	106656
R^2	0.322	0.322	0.322	0.195	0.195	0.195

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The dependent variable in columns 1–3 is bednet possession at the mother level, which is based on the answer to the question "have bednet for sleeping." The dependent variable in columns 4–6 is bednet use at the mother level, which is based on the answer to the question "slept under bednet last night". Malaria Suitability is measured by the malaria stability index originally constructed by Kiszewski et al. (2004). A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for survey-wave FE, region FE, ethnicity FE, religion FE, year of birth FE, dummies for wealth index, and geographic isolation controls described in the notes of Table B1.

K Channels

K.1 Information Treatment from Larson and Lewis (2017)

Table K1: Linguistic Distance and Information

	(1)	(2)
Linguistic Distance	-46.09***	-41.06**
	(15.10)	(15.95)
Observations	440	379
R^2	0.464	0.333

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. All variables are based on survey data from Ugandan villages collected by Larson and Lewis (2017). They use an experimental setting that seeds identical information in two different villages of Uganda. "The seeded information was that in three days an event would be held at which all adults in attendance would receive a valuable block of soap in exchange for taking a survey" Larson and Lewis (2017). The dependent variable is a 0-1 binary variable which takes the value of 1 if individuals have heard about the event that was announced by the experimenters and 0 otherwise. A circle of radius 50 km has been considered for calculating the linguistic distance variable. All columns include controls for education, material of walls of house, type of job, religion, gender, if respondent was a seed in the information spread intervention, occupation, and age. Column 1 also includes a dummy identifying whether the individual attended the soap event. Column 2 restricts the sample to individuals who did not attend the soap event. Please refer to Larson and Lewis (2017) for further details.

K.2 Selective Fertility

Table K2: Test for Selective Fertility

	(1) Education	(2) Partner's Education	(3) Currently Working	(4) Worked Past	(5) Wealth Index
Linguistic Distance	0.0341	0.142**	0.0361	0.0383	0.0716
	(0.0626)	(0.0556)	(0.0253)	(0.0255)	(0.0826)
ELF	0.0340	0.0415	0.0122	0.0305*	0.0937
	(0.0436)	(0.0507)	(0.0159)	(0.0165)	(0.0770)
Observations	205986	193735	205704	196289	205988
R^2	0.413	0.415	0.189	0.185	0.441
	Female child	Height	Weight	HAZ respondent	No. of children
Linguistic Distance	-0.00190	-0.221	0.482	-5.043*	-0.0272
	(0.00354)	(0.162)	(0.485)	(2.763)	(0.0631)
ELF	-0.000480	-0.215*	0.157	-3.010	0.0363
	(0.00364)	(0.129)	(0.256)	(3.057)	(0.0572)
Observations	861386	205988	205988	144728	205988
R^2	0.012	0.331	0.378	0.162	0.545

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The column headings indicate the individual mother-level dependent variable for each specification, except for column 1 of panel 2 where the regression is at the child level and the dependent variable is at the child level. The mother-level dependent variables are: educational attainment, partner0s educational attainment, current work status, past work status, height in cms, weight in cms, height-for-age z-score, and total no. of children. A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns (except column 1 of panel 2) include controls for survey-wave FE, region FE, ethnicity FE, religion FE, year of birth FE, dummies for wealth index, and geographic isolation controls described in the notes of Table B1. Column 1 of panel 2 includes controls for survey-wave FE, region × year FE, ethnicity × year FE, religion FE, individual controls and geographic isolation controls described in the notes of Table B1.

K.3 Additional Information Variables

Table K3: Additional Information Variables

	(1) Hand washing	(2) Seeking Medical	(3) Where to Go	(4) TB Curable
Panel 1. Full Sample				
Linguistic Distance	-0.0288	-0.00729	-0.0189	-0.00911
	(0.0200)	(0.0290)	(0.0177)	(0.0262)
Observations	71108	65301	63870	43030
R^2	0.140	0.184	0.066	0.096
Panel 2. Migrant Sample				
Linguistic Distance	0.0229**	0.0441	-0.0276	-0.0520**
	(0.00881)	(0.0278)	(0.0255)	(0.0234)
Observations	35951	33253	33750	19766
R^2	0.091	0.145	0.057	0.104
Panel 3. Non-Migrant Sample				
Linguistic Distance	-0.0718***	-0.0562**	-0.00221	0.0147
	(0.0239)	(0.0213)	(0.0186)	(0.0556)
Observations	34967	31886	29932	14065
R^2	0.185	0.220	0.088	0.096

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses are clustered at the region level. The column headings indicate the individual mother-level dependent variable for each specification. The dependent variable in column 1 is a binary 0-1 variable indicating whether the individual washed her hands before preparing last meal. In column 2 it indicates whether the respondent can decide by herself whether or not the child should be taken for medical treatment when child is seriously ill. It takes the value of 1 if they answer either "yes" or "depends" to the question and 0 if they answer "no". In column 3 the dependent variable asks if knowing where to go when they are sick is a major problem preventing the respondent from getting a medical advice or treatment. It takes the value of 0 if they answer "no problem" or "small problem" to the question and 1 if they answer "big problem". The dependent variable in column 4 is based on a question that asks respondents if TB can be cured. It take the value 0 if they answer "No" or "don't know" and 1 if they answer "yes". A circle of radius 50 km has been considered for calculating the linguistic distance and ELF variables. All columns include controls for ELF, survey-wave FE, region FE, ethnicity FE, religion FE, year of birth FE, dummies for wealth index, and geographic isolation controls described in the notes of Table B1.

L Adjustments for Multiple Comparisons

Table L1: Child-Level Regressions: Full Sample

	p values							
variable	unadjusted	bonferroni	sidak	holm	holland	hochberg	simes	yekutieli
Child	0.00	0.02	0.01	0.02	0.01	0.01	0.01	0.02
infant	0.00	0.02	0.01	0.02	0.01	0.01	0.01	0.02
neonatal	0.00	0.05	0.04	0.04	0.04	0.04	0.02	0.05
HAZ	0.02	0.36	0.31	0.26	0.23	0.26	0.07	0.24
stunted	0.02	0.30	0.26	0.24	0.22	0.24	0.07	0.24
WAZ	0.12	1.00	0.85	0.96	0.64	0.74	0.22	0.75
Tetanus	0.63	1.00	1.00	1.00	0.93	0.74	0.68	1.00
Measles	0.41	1.00	1.00	1.00	0.93	0.74	0.55	1.00
polio	0.46	1.00	1.00	1.00	0.93	0.74	0.56	1.00
dpt	0.18	1.00	0.95	1.00	0.76	0.74	0.31	1.00
iron tablets	0.07	1.00	0.68	0.73	0.53	0.68	0.16	0.53
bcg	0.36	1.00	1.00	1.00	0.93	0.74	0.53	1.00
antenatal visits	0.49	1.00	1.00	1.00	0.93	0.74	0.56	1.00
full immunization	0.74	1.00	1.00	1.00	0.93	0.74	0.74	1.00
skilled birth attendance	0.08	1.00	0.69	0.73	0.53	0.68	0.16	0.53

Notes: The p-values have been adjusted for multiple comparisons using seven different alternatives based on Newson (2010).

Table L2: Child-Level Regressions: Migrant Sample

	p values (Migrant sample)							
variable	unadjusted	bonferroni	sidak	holm	holland	hochberg	simes	yekutieli
Child	0.04	0.54	0.42	0.50	0.40	0.50	0.27	0.90
infant	0.26	1.00	0.99	1.00	0.93	0.79	0.48	1.00
neonatal	0.34	1.00	1.00	1.00	0.95	0.79	0.48	1.00
HAZ	0.54	1.00	1.00	1.00	0.95	0.79	0.58	1.00
stunted	0.53	1.00	1.00	1.00	0.95	0.79	0.58	1.00
WAZ	0.35	1.00	1.00	1.00	0.95	0.79	0.48	1.00
Tetanus	0.02	0.30	0.26	0.30	0.26	0.30	0.27	0.90
Measles	0.79	1.00	1.00	1.00	0.95	0.79	0.79	1.00
polio	0.21	1.00	0.97	1.00	0.92	0.79	0.48	1.00
dpt	0.09	1.00	0.75	1.00	0.70	0.79	0.45	1.00
iron tablets	0.29	1.00	0.99	1.00	0.94	0.79	0.48	1.00
bcg	0.39	1.00	1.00	1.00	0.95	0.79	0.48	1.00
antenatal visits	0.19	1.00	0.95	1.00	0.92	0.79	0.48	1.00
full immunization	0.35	1.00	1.00	1.00	0.95	0.79	0.48	1.00
skilled birth attendance	0.22	1.00	0.98	1.00	0.92	0.79	0.48	1.00

Notes: The p-values have been adjusted for multiple comparisons using seven different alternatives based on Newson (2010).

Table L3: Child-Level Regressions: Non-Migrant Sample

	p values (Non-Migrant sample)							
variable	unadjusted	bonferroni	sidak	holm	holland	hochberg	simes	yekutieli
Child	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
infant	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
neonatal	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HAZ	0.04	0.59	0.45	0.35	0.30	0.35	0.08	0.28
stunted	0.00	0.03	0.03	0.02	0.02	0.02	0.01	0.02
WAZ	0.29	1.00	0.99	1.00	0.88	0.87	0.44	1.00
Tetanus	0.05	0.76	0.54	0.41	0.34	0.41	0.10	0.32
Measles	0.57	1.00	1.00	1.00	0.96	0.87	0.71	1.00
polio	0.71	1.00	1.00	1.00	0.97	0.87	0.81	1.00
dpt	0.82	1.00	1.00	1.00	0.97	0.87	0.87	1.00
iron tablets	0.00	0.03	0.03	0.02	0.02	0.02	0.01	0.02
bcg	0.87	1.00	1.00	1.00	0.97	0.87	0.87	1.00
antenatal visits	0.00	0.05	0.04	0.03	0.03	0.03	0.01	0.02
full immunization	0.36	1.00	1.00	1.00	0.89	0.87	0.49	1.00
skilled birth attendance	0.07	1.00	0.67	0.50	0.40	0.50	0.12	0.39

Notes: The p-values have been adjusted for multiple comparisons using seven different alternatives based on Newson (2010).

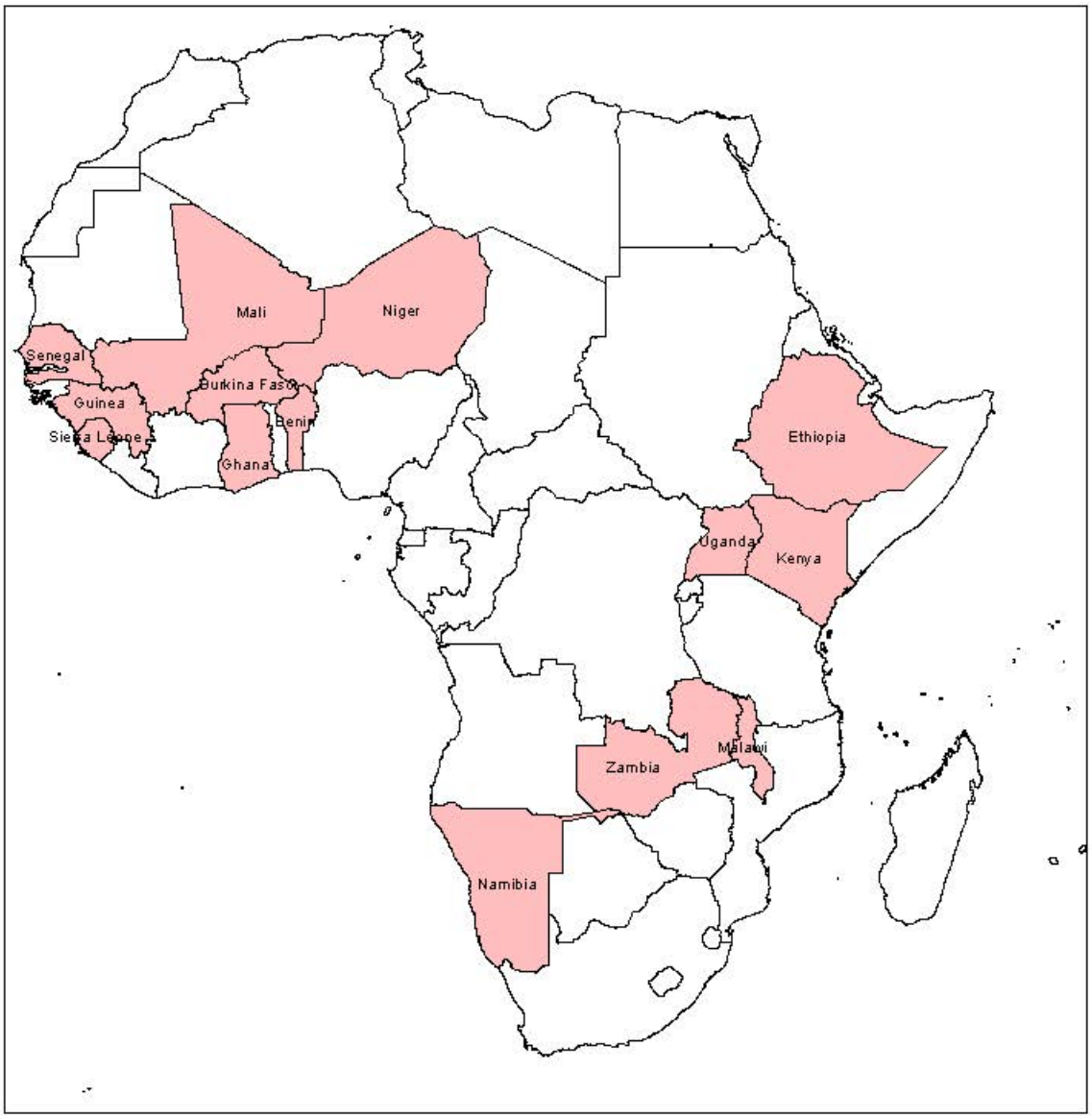
Table L4: Mother-Level Regressions: : Non-Migrant Sample

	p values							
variable	unadjusted	bonferroni	sidak	holm	holland	hochberg	simes	yekutieli
Education	0.99	1.00	1.00	1.00	0.99	0.99	0.99	1.00
Literacy	0.65	1.00	0.99	1.00	0.96	0.99	0.87	1.00
Water	0.20	0.98	0.67	0.79	0.58	0.79	0.49	1.00
Electricity	0.70	1.00	1.00	1.00	0.96	0.99	0.87	1.00
ORS	0.03	0.14	0.13	0.14	0.13	0.14	0.14	0.31

Notes: The p-values have been adjusted for multiple comparisons using seven different alternatives based on Newson (2010).

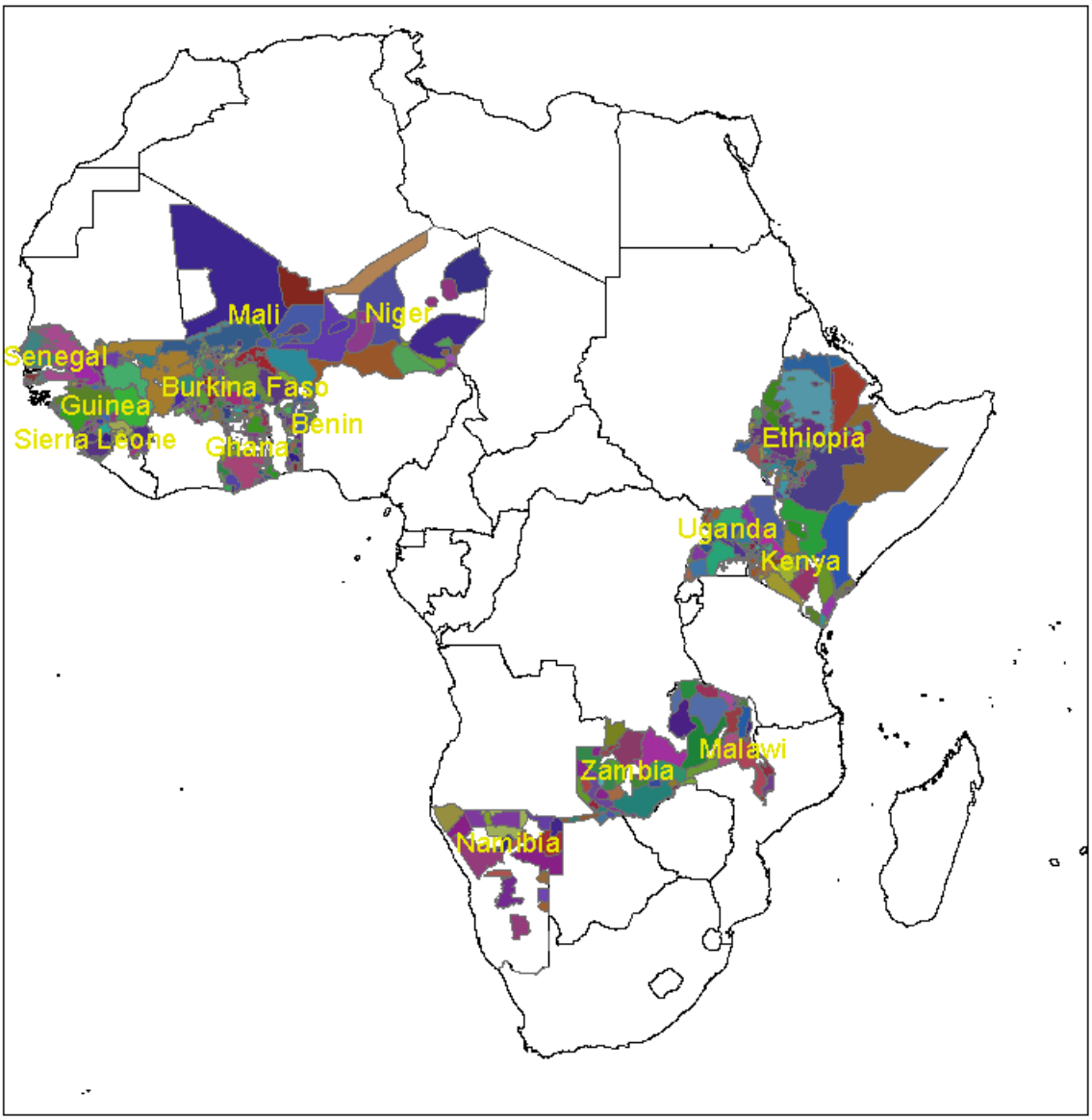
M Appendix Figures

Figure M1: Countries used



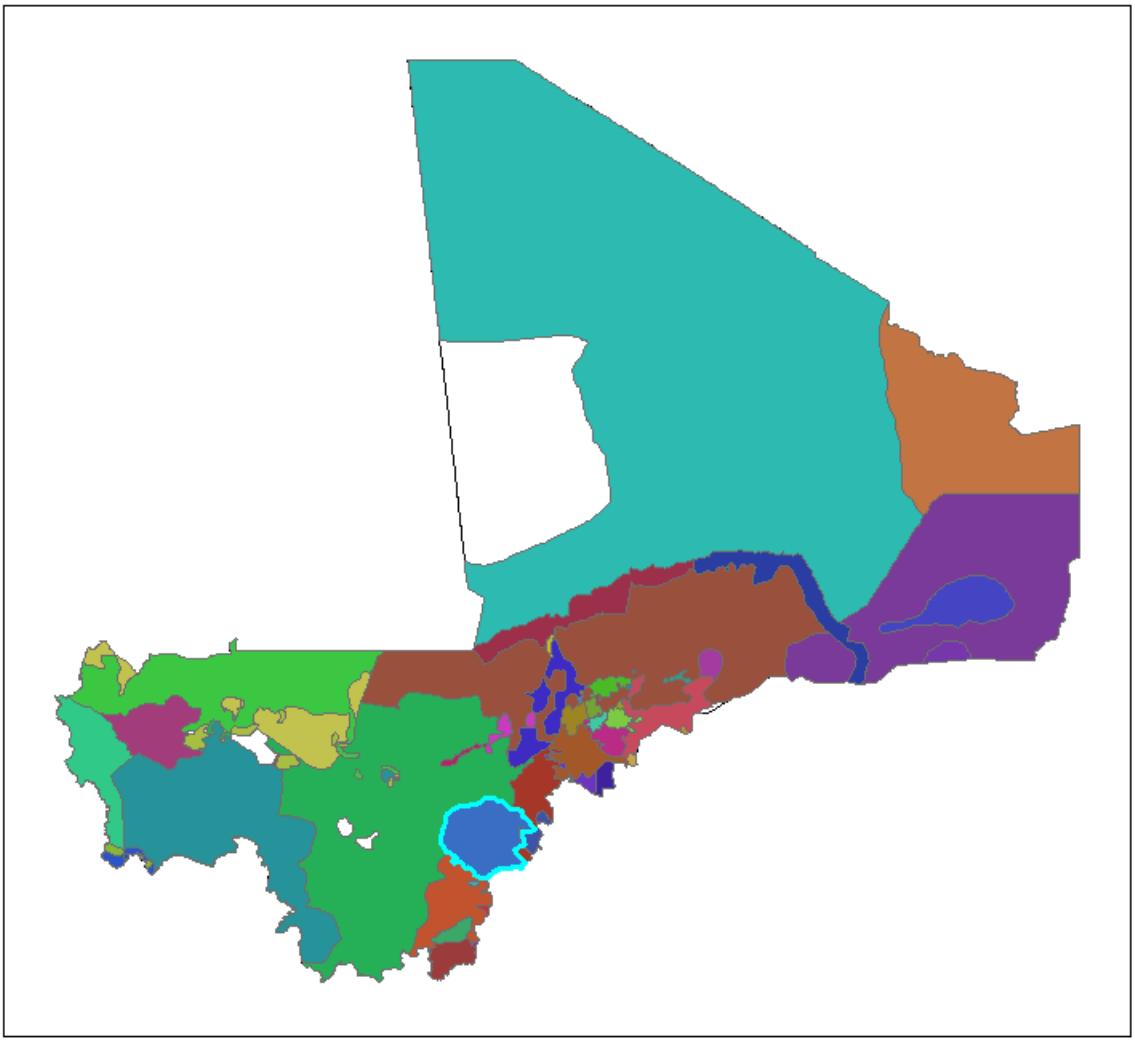
NOTES: This map plots the fourteen countries used in the study: Benin, Burkina Faso, Ethiopia, Ghana, Guinea, Kenya, Malawi, Mali, Namibia, Niger, Senegal, Sierra Leone, Uganda, and Zambia.

Figure M2: Languages used



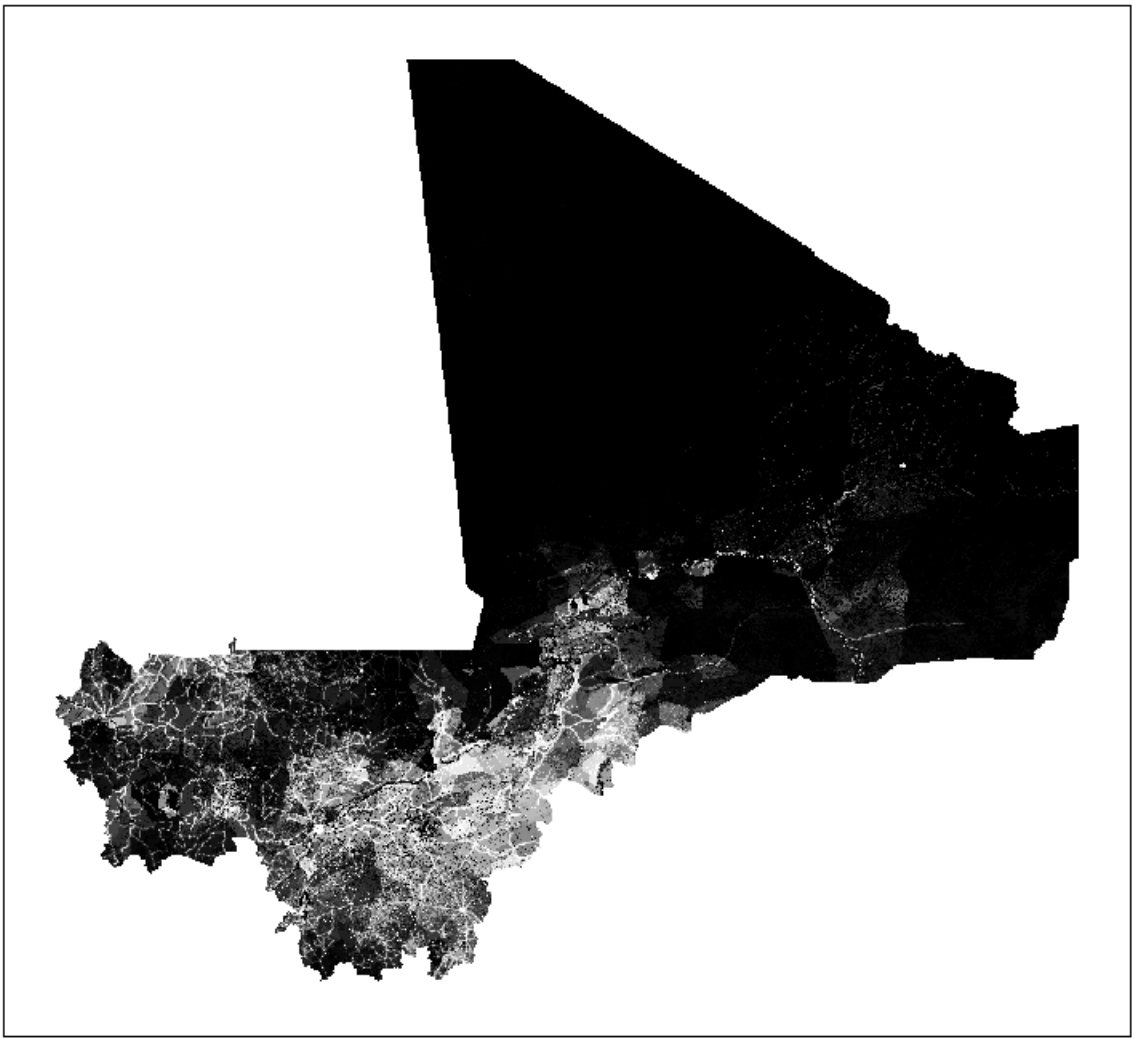
NOTES: This map plots the linguistic groups for the fourteen countries used in the study from the Ethnologue database. Polygons of different colours represent the different language groups. Areas where multiple languages are spoken are represented by overlapping polygons, which are not distinguishable in this map.

Figure M3: The Languages of Mali



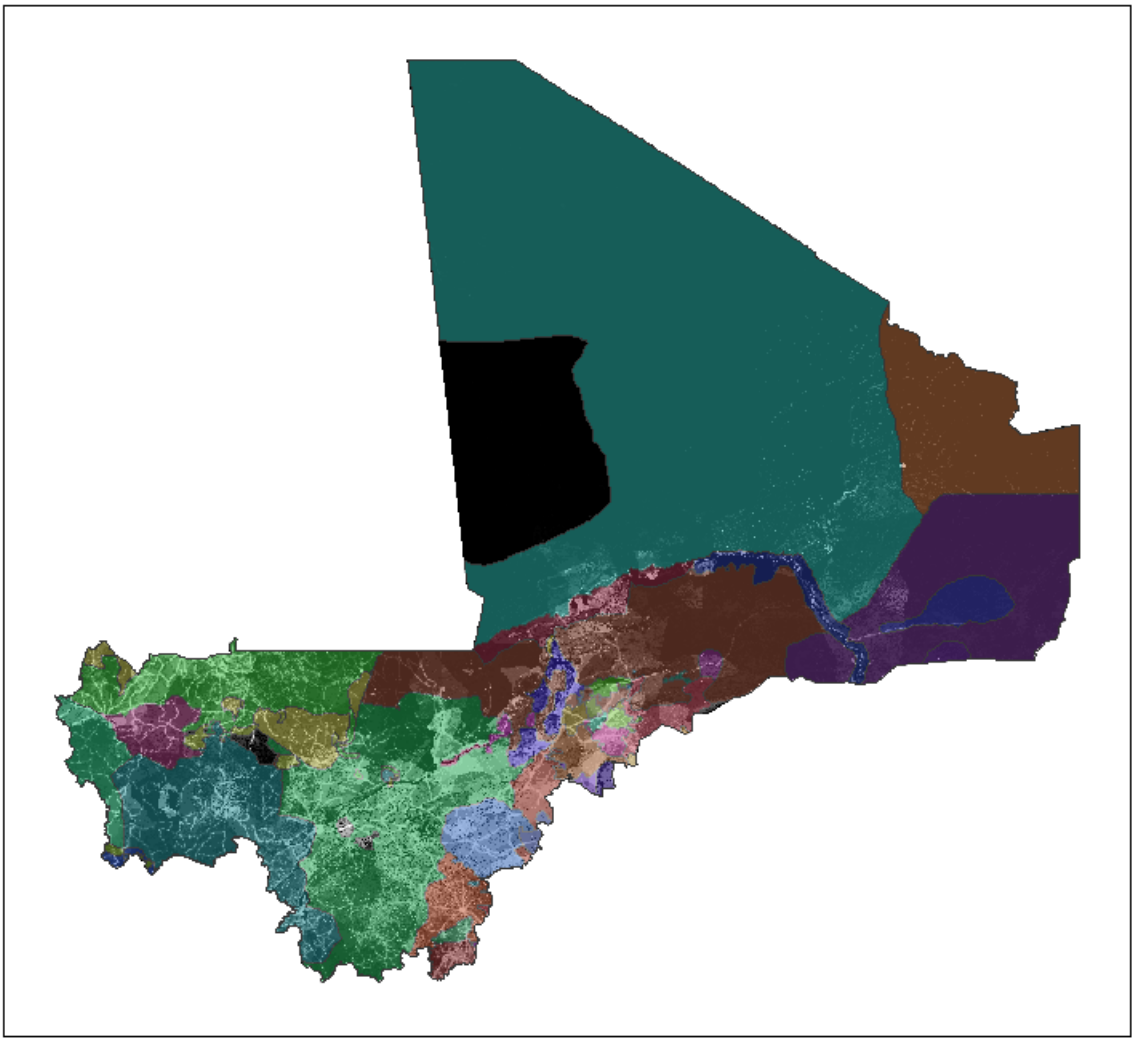
NOTES: This map plots the linguistic groups of Mali from the Ethnologue database. Polygons of different colours represent the different language groups. Areas where multiple languages are spoken are represented by overlapping polygons, which are not distinguishable in this map. The polygon highlighted in blue in the south-eastern corner of the map demarcates the linguistic homeland of the Mamara Senoufo language speakers. Figure 1 from the main paper zooms into this region.

Figure M4: The Population of Mali



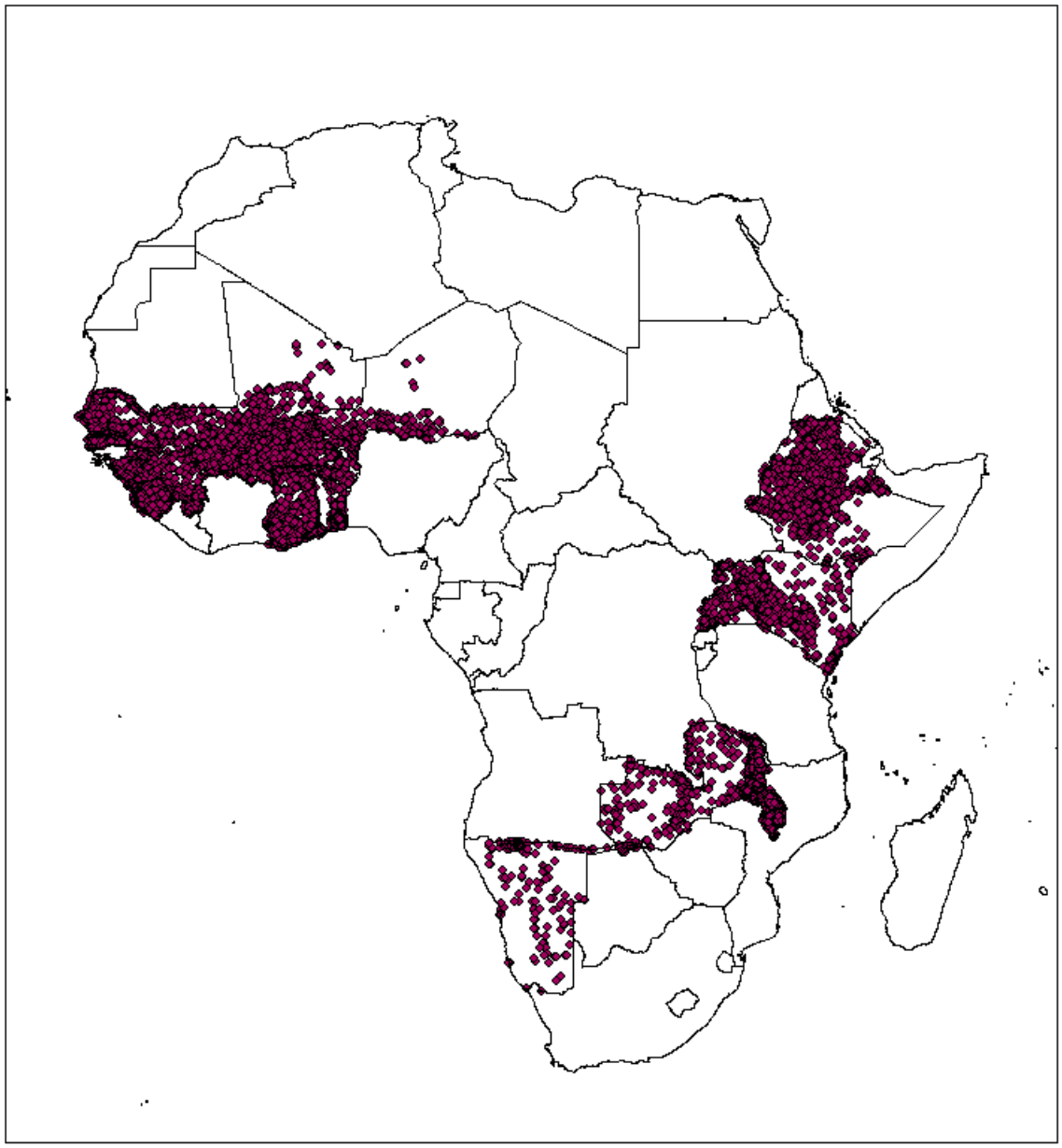
NOTES: This map plots the population distribution of Mali from the LandScan database at the 30 arc seconds x 30 arc seconds (approximately 1 km $\times$ 1 km at the equator) resolution. The brighter (darker) pixels within the geographic boundaries of Mali represent more (less) populated areas.

Figure M5: Mali Overlay



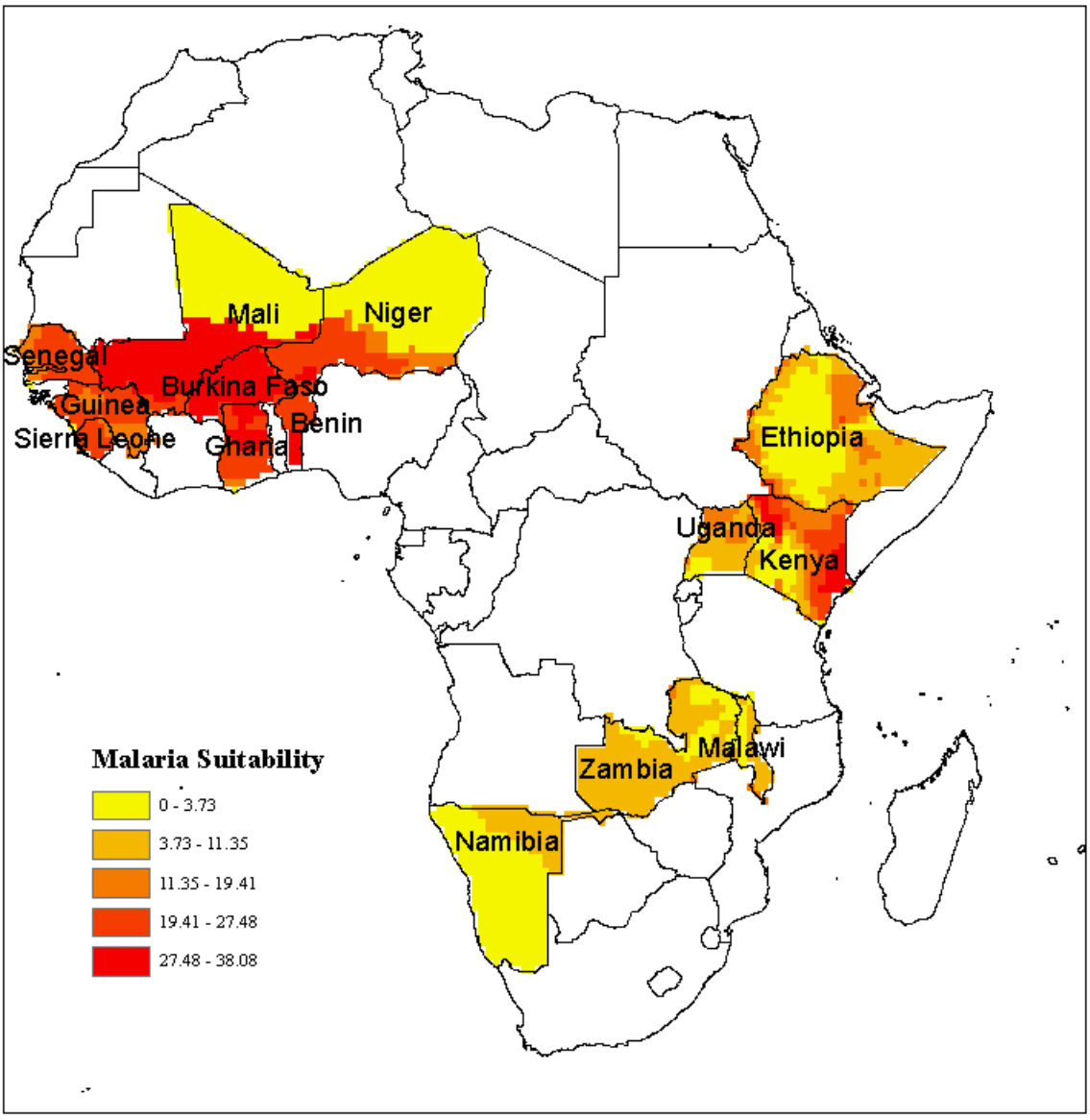
NOTES: This map overlays the language polygons for Mali from the Ethnologue database (see Figure M3) on the population distribution of Mali from the LandScan database (See Figure M4). Polygons of different colours represent the different language groups in the language group map. Areas where multiple languages are spoken are represented by overlapping polygons, which are not distinguishable in this map. The brighter (darker) pixels within the geographic boundaries of Mali in the population map represent more (less) populated areas.

Figure M6: Mothers' Locations



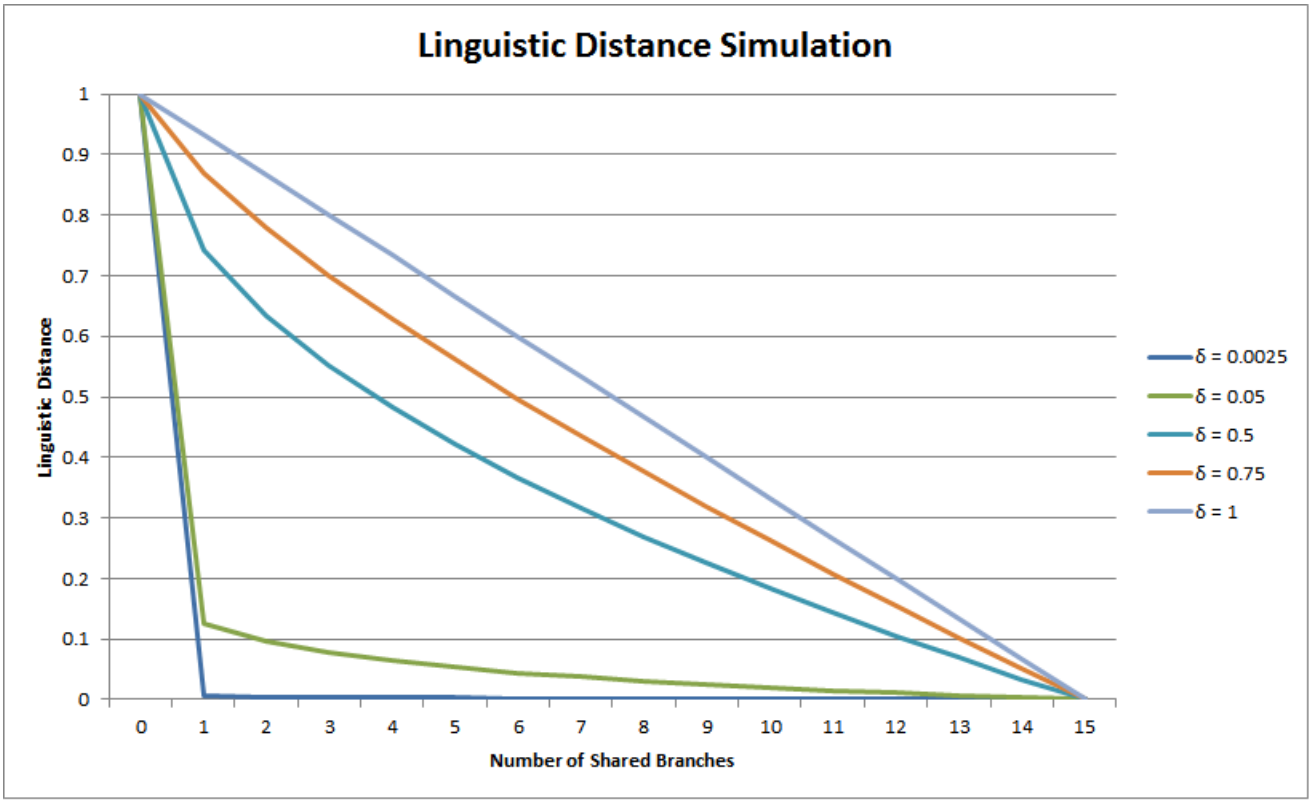
NOTES: This map plots the locations of the 28,839 DHS clusters where the 208,898 individual mothers that comprise the sample used in this paper are located.

Figure M7: Malaria Stability Index



NOTES: This map plots the malaria stability index / malaria suitability index originally generated by Kiszewski et al. (2004) and made available in a 5 km $\times$ 5 km resolution raster format by McCord and Anttila-Hughes (2017).

Figure M8: Linguistic Distance for alternative values of the decay factor δ



NOTES: This graph simulates how linguistic distance changes for alternative values of the decay factor δ. The x-axis gives how many branches any two languages share and the y-axis gives the corresponding values of linguistic distance for different values of δ.

INSTITUT DE RECHERCHE
ÉCONOMIQUES ET SOCIALES

Place Montesquieu 3
1348 Louvain-la-Neuve

ISSN 1379-244X D/2020/3082/05

UCLouvain

IRES | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics