Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Economia

Ciclo XXVII

Settore Concorsuale di afferenza: 13/A1 – ECONOMIA POLITICA

Settore Scientifico disciplinare: SECS-P/01 – ECONOMIA POLITICA

Essays on Ethnic Diversity and Development

Giorgio Chiovelli

Coordinatore Dottorato	**Relatore**
Matteo Cervellati	Matteo Cervellati

Esame finale anno 2015

Essays on Ethnic Diversity and Development

Abstract

This dissertation consists of three papers. The first paper "*Ethnicity, Migration and Conflict: Evidence from Contemporary South Africa*" exploits some of the institutional changes intervened in South Africa during the end of apartheid to investigate the relationship between ethnic diversity and conflict. I find within-ethnic polarization to be significantly related to the intensity of armed confrontations among black-dominated groups. My investigation thus gives strong and robust empirical support to the theoretical arguments which identify ethnic diversity as one of the determinants of civil conflict. The second chapter, "*Pre-Colonial Centralization, Colonial Activities and Development in Latin America*", investigates the hypothesis that pre-colonial ethnic institutions shaped contemporary regional development in Latin America. I document a strong and positive relationship between pre-colonial centralization and regional development. Results are in line with the view that highly centralized pre-colonial societies acted as a persistent force of agglomeration of economic activities and a strong predictor of colonial state capacity. The results provide a first evidence of the existence of a link between pre-colonial centralization, colonial institutional arrangements and contemporary economic development. The third paper "*Bite and Divide: Malaria and Ethnic Diversity*" investigates the role of malaria as a fundamental determinant of modern ethnic diversity. This paper explores the hypothesis, that a large exposure to malaria has fostered differential interactions that reduced contacts between groups and increased interactions within them Results document that malaria increases the number of ethnic groups at all levels of spatial disaggregation and time periods (exploiting historical and current ethnic diversity). Regressions' results show that endogamous marriages are more frequent in areas with higher geographic suitability to malaria. The results are in line with the view that malaria increases intra-ethnic interactions while decreasing inter-ethnic ones.

Keywords: Ethnic Diversity, Development, Long-Run Growth, Conflict, Colonial Institutions, Diseases, Geography, Apartheid, South Africa, Latin America, Africa.

Table of Contents

Ethnicity, Migration and Conflict: Evidence from Contemporary South Africa*

Francesco Amodio† Giorgio Chiovelli‡

August 12, 2014

Abstract

This paper explores the extent to which changes over time in ethnic distribution correlate with contemporaneous changes in conflict incidence. We focus on the history of contemporary South Africa. Migration flows following the implementation and repeal of apartheid segregation laws induced cross-sectional and time variation in districts' ethnic composition. Ethnolinguistic diversity within the black majority is shown to be strongly informative of the incidence of armed confrontations between black-dominated organized groups. In order to achieve identification, we compare the evolution of conflict across districts experiencing differential changes in ethnic composition. Results are robust when comparing neighboring localities, and to the implementation of an instrumental variable strategy where pairwise distance between districts is used to predict the location decision of internal migrants.

Keywords: conflict, ethnicity, apartheid, South Africa.
JEL Codes: D74, J15, N47, N97, O15, R23.

*We are indebted to Matteo Cervellati and Albrecht Glitz for their insightful guidance. We would like to thank the following people for helpful comments and discussion: Jean-Marie Baland, Laia Balcells, Marcelo Bérgolo, Bruno Caprettini, Francesco Caselli, Rodrigo Deiana, Giovanni Di Mauro, Ruben Enikolopov, Joan Esteban, Margherita Fort, Oded Galor, Andrea Ichino, Laura Mayoral, Stelios Michalopoulos, Massimo Morelli, Hannes Mueller, Giacomo Ponzetto, Giovanni Prarolo, Debraj Ray, Marta Reynal-Querol, Simon Roberts, Dominic Rohner, Martín Rossi, Uwe Sunde, Alessandro Tarozzi, Mathias Thoenig, John Tsoukalis, Fabrizio Zilibotti and all seminar participants at Università di Bologna, Universitat Pompeu Fabra, Toulouse School of Economics, University of Lausanne, University of Zurich, Brown University, the Barcelona GSE "Understanding Civil Conflict" Summer Forum, Aix-Marseille School of Economics, Universidad de la República in Montevideo and Universidad Torcuato Di Tella. We also thank Lynn Woolfrey from DataFirst and Ralph Sundberg from Uppsala University for their help with South Africa Census data and UCDP-GED data respectively. Errors remain our own. Amodio acknowledges financial support from the Spanish Ministry of Economy and Competitiveness. Chiovelli acknowledges financial support from the University of Bologna.

†`francesco.amodio@upf.edu`. Department of Economics and Business, Universitat Pompeu Fabra, C/ Ramon Trias Fargas 25-27, Barcelona 08005 Spain.

‡`giorgio.chiovelli2@unibo.it`. Department of Economics, Università di Bologna, Piazza Scaravilli 2, Bologna 40126 Italy.

1 Introduction

Civil war is closely related with poor economic performance. Understanding its determinants is regarded as crucial in the challenge for world development (Blattman and Miguel 2010). Conflict commonly manifests itself through ethnic markers. Ethnic traits stand out as a salient technology for either or both the generation and expression of social tensions. Indeed, the empirical evidence shows the probability of conflict outbreak and conflict incidence to be strongly correlated with measures of historical ethnic distribution (Montalvo and Reynal-Querol 2005, 2010; Desmet, Ortuño Ortín, and Wacziarg 2012; Esteban, Mayoral, and Ray 2012). However, it is still unknown whether migration flows involving ethnically diverse communities have any impact on conflict prevalence. In this paper, we study the extent to which changes over time in ethnic distribution correlate with contemporaneous changes in conflict incidence.

We focus on the history of contemporary South Africa. Starting in the 1950s, apartheid segregation laws regulated the ethnic composition of local districts. The repeal of apartheid legislation in the early 1990s restored free internal mobility of blacks within the country while democratization in 1994 ended the white minority rule in favor of the newly enfranchised black majority. Throughout this period, black-dominated parties and unions were struggling violently amongst themselves to benefit from the transition and dominate the new institutional scenario. Figure 1a plots the total number of armed confrontations between organized groups recorded in South Africa from 1989 to 2004. Non-state conflict events refer to struggles between black-dominated groups, while one-side conflict events are those in which the Government is involved.[1] Violence between black groups reached its peak in 1993, with almost 500 confrontations recorded in the country as a whole. It then decreased sharply after democratization in 1994, and died out almost completely since 2000. By the same token, Figure 1b is taken from Reed (2013) and plots the internal migration rates in South Africa from 1955 to 1999.[2] The number of moves per person-year spiked in 1991, the same year the last legal impediment to mobility (the Group Areas Act) was repealed. Migration rates remained as high as 9% in the years to follow. In this paper, we link these two facts in studying the relationship between ethnicity and conflict. In particular, we explore the extent to which the struggle between black organized groups over the period manifested itself through ethnic markers. Indeed, we relate migration-driven changes in local districts' ethnic composition to the concurrent changes in conflict incidence and find the number of conflict events to be positively correlated with contemporaneous measures of within-black ethnolinguistic polarization.[3]

[Figure 1]

[1] Source: Geo-referenced Event Dataset of the Uppsala Conflict Data Program (UCDP-GED v1.5), Department of Peace and Conflict Research, Uppsala University. The data are described in details in the Data section.

[2] Source: South Africa Migration and Health Survey (SAMHS)

[3] We use the polarization index used in Reynal-Querol (2002) and Montalvo and Reynal-Querol (2005, 2010), which falls within the class of measures proposed by Esteban and Ray (1994). The index formula is presented and discussed in Section 4.

Several theoretical contributions in the literature explore the relationship between ethnic diversity and conflict. Horowitz (1985) first noticed that conflicts seem to arise in societies where a large ethnic minority faces another ethnic majority.[4] Modeling social confrontation and its relationship with measures of heterogeneity in general, Esteban and Ray (1994, 1999) argue that polarized societies, characterized by internally homogeneous but externally very distant groups, are more prone to experience conflict. In the specific case of ethnicity, ethnic lines may become salient in conflict as they allow for separation of individuals into groups that are homogenous under one trait (ethnicity), but significantly different in their economic characteristics. Indeed, Esteban and Ray (2008, 2011a) explore the role of intra-group synergies in conflict effort that arise when ethnic groups are characterized by high within-group economic inequality. Caselli and Coleman (2013) rationalize instead the salience of ethnic trait in conflicts framing it as a technological device which prevents indiscriminate access to the expected gains of the winning group. Finally, Esteban and Ray (2011b) develop an extensive theory of conflict which highlights the role of different distribution measures. Indeed, together with Esteban, Mayoral, and Ray (2012), they show both theoretically and empirically the informativeness of the polarization measure with regard to conflict severity to be positively related to the *degree of publicness* of the disputed *prize*. We conceptualize the struggle within the black majority in South Africa through the end of apartheid as a dispute for political power in the new institutional scenario, both at the local and national level. Such prize is intrinsically public in nature. We thus refer to these last contributions as the theoretical foundation for our focus on the polarization measure.[5]

For the purpose of this paper, we combine geo-referenced information on conflict with South Africa Census data. As a preliminary analysis, we exploit cross-district variation in ethnic composition before the fall of apartheid in 1991 and find a positive correlation between ethnolinguistic polarization within the black majority and the number of recorded armed confrontations between black-dominated organized groups. As a first contribution of this paper, and despite their primarily political nature, we thus provide a qualification of these conflicts as expressed through ethnic markers. Then, we exploit the most distinguished feature of this setting: time variation in the ethnic distribution at the local district level. Mainly driven by the migration flows following the repeal of apartheid segregation laws, substantial time variation allows us to combine data from 1991 and 1996 and show the change in polarization at the district level to be positively strongly and significantly correlated with the change in the number of conflict events per year. We implement a first-difference specification which rules out unobserved differences in time-invariant determinants of conflict incidence across districts. The point estimate of the coefficient of interest is stable across specifications. Evidence suggests an increase in the within-black polarization index of one cross-district standard deviation to be associated with an increase in the number of conflict events per district of more than the 1991 national average. Moreover, conditioning on the total number of conflict events in 1989-1990 still yields highly significant results.

[4] Collier (2001) and Collier and Hoeffler (2004) further develop and empirically test this hypothesis. Horowitz himself extensively studied South Africa as a case study of a society divided along ethnic lines.

[5] For an overview of the debate over the use of alternative ethnolinguistic distribution measures in the empirical analysis see also Alesina, Devleeschauwer, Easterly, Kurlat, and Wacziarg (2003) and Desmet, Ortuño Ortín, and Weber (2009).

The relationship we find using a first-difference approach can potentially be driven by unobserved factors which affect the evolution of both ethnic composition and conflict in the same way, generating a spurious correlation between the two. For example, unobserved geographic characteristics may be correlated with both changes in ethnic diversity and the evolution of conflict (Michalopoulos 2012). We examine the severity of this issue by looking at the relationship of interest within clusters of neighboring districts. In the spirit of matching methods, we control for common trends at the cluster level. We thus exploit only residual variability in both within-black ethnolinguistic polarization and conflict incidence and still find evidence of a positive relationship between the two. Finally, we specifically focus on internal migration as the main source of variability for the change in districts' ethnic composition. Individuals' migration choice and location decision may potentially be endogenous to the evolution of conflict. In the presence of more than two groups, mapping the outcomes of individual decision problems into systematic changes in the polarization index is theoretically challenging.[6] We thus implement an instrumental variable strategy where we use pairwise distance between districts to predict the location decision of internal migrants. This because, once the migration decision is taken, distance from the origin is a strong determinant of ultimate location choice (Kok, O'Donovan, Bouare, and Van Zyl 2003). In a simulation exercise, we use distance-based predicted location probabilities to re-allocate an exogenously given fraction of migrants from each ethnolinguistic group from each district to the rest of South Africa, independently from conflict levels and changes. The resulting predicted change in the district-level polarization index is thus used as instrument for the actual one, ruling out identification threats from both endogenous location decisions of internal migrants and other sources of changes in ethnic distribution, such as disease prevalence. 2SLS point estimates are higher than previous ones, highlighting the role played by internal migration and confirming the existence of a strong link between within-black ethnolinguistic polarization and number of violent confrontations between black-dominated organized groups.

This paper contributes to the empirical literature studying relationship between ethnic polarization and conflict incidence. Such exploration was first carried out by Montalvo and Reynal-Querol (2005). More recently, Desmet, Ortuño Ortín, and Wacziarg (2012) investigate the explanatory power of diversity measures as computed at different levels of the ethnolinguistic world tree. Most of the existing studies exploit variation at the cross-country level and make use of time-invariant distribution indices, drawn from historical and encyclopedic sources of ethnic diversity.[7] One contribution of this paper is that we employ contemporaneous ethnolinguistic group populations from Census data in the computation of our time-varying index of ethnic polarization at the local district level. In particular, we draw information from South Africa Censuses on languages spoken and validate them using the *Ethnologue* linguistic database (Lewis 2009).[8] We also use information on districts' socio-economic characteristics, drawn from

[6] Esteban and Ray (1994) provide a series of examples to intuitively show the absence of a partial order for increasing polarization. The same arguments prevent establishing even a partial map between migration flows and changes in the polarization measure.

[7] One exception is Novta (2013), who studies theoretically and empirically the relationship between municipal-level ethnic composition and the spread of civil conflict in Bosnia.

[8] Detailed information about this procedure is available in the data section. Notice that, with this approach, we trade off the possible drawbacks from the use of potentially endogenous administrative boundaries with

Census data, in order to show the robustness of results in reduced form.

Finally, we contribute to a growing body of the literature that studies conflict at a more disaggregated level, such as local communities and individuals themselves. Within these, Michalopoulos and Papaioannou (2011) use information on spatial distribution of African ethnicities before colonization and find contemporary civil conflict incidence to be concentrated in the historical homeland of partitioned ethnicities. Rohner, Thoenig, and Zilibotti (2013) use individual, county and district-level data from Uganda to investigate the social and economic consequences of ethnic conflicts. Dube and Vargas (2013) exploit variation in the international price of several agricultural commodities and look at the differential effect on violent conflict across municipalities in Colombia. Besley and Reynal-Querol (2014) look instead into conflict persistence in Africa having fine geographical grids as units of observation. Finally, Harari and La Ferrara (2013) study instead the impact of negative climate shocks on conflict incidence using within-year variation at the local level.

The paper is organized as follows. In Section 2, we provide a short overview of the history of contemporary South Africa. Section 3 introduces the relevant concepts, variables and measures, coupled with the data we use. The empirical strategy and results are presented in Section 4. Section 5 concludes.

2 Historical Background: South Africa and the End of Apartheid

Since the end of World War II until 1994, South Africa was ruled under *apartheid* regime. Apartheid - meaning *apartness* in Afrikaans - was a form of government based on physical separation of blacks and whites achieved through racial discrimination and political disenfranchisement of the black majority. Divisions along racial lines were thought to be the fundamental organizing principle for the allocation of all resources and opportunities, the basis of all spatial demarcation, planning and development and the boundary for all social interactions (Posel 2001). Under apartheid, economic activity was based on the exploitation of black cheap labor force in order to ensure high returns to white-owned capital investments (Clark and Worger 2011). By the same token, white workers and farmers were given protection from the competition of their black counterparts.

An extensive legislation implementing racial segregation was put in place after the election of the National Party (NP) into government in 1948. The Population Registration Act (1950) formalized the racial classification system through identifying four different races living in the South Africa. Blacks were further classified into native ethnicities on the basis of the first language they spoke. The Bantu Authorities Act (1951) and Bantu Resettlement Act (1954) established ten black ethnicity-based *homeland* reserve areas, known as *Bantustan*: Transkei and Ciskei (Xhosa ethnicity), Bophuthatswana (Tswana), Venda (Venda), Gazankulu (Tsonga), Lebowa

the advantages of deriving a time-variant measure of ethnic polarization to employ in our analysis at the subnational level. The *Ethnologue* database has been already used as source of ethnolinguistic information in Alesina, Devleeschauwer, Easterly, Kurlat, and Wacziarg (2003), Desmet, Ortuño Ortín, and Weber (2009), Desmet, Ortuño Ortín, and Wacziarg (2012) and Esteban, Mayoral, and Ray (2012).

(Sotho), Qwaqwa (Sotho), KwaZulu (Zulu), KaNgwane (Swazi) and KwaNdebele (Sotho). The long-term objective of the law was to let the homelands become independent territories.[9] Political separation was achieved with the Promotion of Bantu Self-Government Act (1959), which implied the *de iure* disenfranchisement of blacks from white South Africa. In less than thirty years, approximately 3.5 million blacks were obliged to move to the homelands the government assigned to the different ethnolinguistic groups (Clark and Worger 2011). A number of legislative acts, such as the Group Areas Act, implemented and guaranteed re-settlement of blacks and influx controls. The Bantustans became densely populated areas with low levels of service delivery, infrastructure and employment (Kok, O'Donovan, Bouare, and Van Zyl 2003). Figure 2 presents the map of Bantustans in South Africa as defined by the apartheid legislation as of 1986.

[Figure 2]

The apartheid system began to collapse in the mid-1980s mainly due to the internal contradictions and conflicting effects of the same implemented policies. Exploitation and control of black labor force revealed itself as costly and hardly compoundable with the enforcement of physical separation of races and ethnicities. Moreover, twenty-five countries (USA and UK among them) set a trade embargo in the late 1980s, following the 1977 UN Resolution 418 on a mandatory arms embargo. At the same time, opposition parties started to organize. Together with the black Africa National Congress (ANC) - a multiethnic party established in 1912 - and the Pan-Africanist Congress (PAC) founded in 1953, the Zulu-based Inkatha Freedom Party (IFP) was formed in 1980s and the multiracial United Democratic Front (UDF) was created in 1983. Opposition to apartheid was not homogeneous and parties were not cohesive. Black political parties engaged in conflicts for two main reasons. First, there was no agreement on how to put an end to the apartheid experience. Second, there was no consensus on how a new post-apartheid South Africa should be ruled.

Among the others, confrontation between the ANC/UDF and IFP was the harshest. The IFP was the KwaZulu governing party, drawing support from the traditional Zulu chiefly structures in the province, and running on a strongly conservative political platform closely aligned with large business interests in the province (Carver 1996). On the other hand, the basis of support for both the UDF and ANC was mainly the urban youth and working class, with their political goal being the establishment of a non-racial unitary state in contrast with traditional chief power. As a result, the ANC and IFP had divergent views on the future role of Bantustans, the consequences of international embargoes, and the eradication of local traditional power. Moreover, the apartheid government took advantage of divisions within the black opposition in its struggle to retain power. Indeed, the independentist goals of the IFP and its opposition to ANC were seconded if not supported by the apartheid establishment (Clark and Worger

[9]Over time, the government granted various degrees of self-government to the Bantustans, ranging from independent state-nation - as in Transkei, Bophuthatswana, Venda, and Ciskei between 1977 and 1981 - to limited self-government - as in KwaZulu, Lebowa, Gazankulu, Qwaqwa, KaNgwane and KwaNdebele.

2011).[10]

In the late 1980s, the government decided to repeal the Pass Law and to free ANC leaders like Nelson Mandela and other black leaders with the aim of freezing the protests and starting to negotiate. Between 1990 and 1991, the Natives' Land Act, the Population Registration Act and the Group Areas Act were repealed and free mobility of blacks within South Africa was restored. As previously shown in Figure 1b, internal (especially inter-provincial) migration rates spiked. The relative economic deprivation of Bantustans acted as a centrifugal force. Indeed, Kok, O'Donovan, Bouare, and Van Zyl (2003) use data on migrants from the 1996 Census to show the substantial contribution to migration of former homelands. Moreover, Bouare (2002) identifies relative GDP as one of the most important determinants of inter-provincial moves, together with the relative number of reported crimes.[11]

Negotiations taking place in 1991-1992 were everything but smooth: while Mandela and de Klerk (NP) were signing official agreements, fightings kept going on. Free elections were held in 1994, with Mandela becoming the first black president of the Republic of South Africa. The struggles between black-dominated groups weakened but continued in the following years. A number of competing explanations have been put forward among both academics and activists: political competition, ethnic mobilization, underlying social and economic deprivation, and the role of powerful local leaders who relied on the use of violence in their struggle to retain power. Carver (1996) argues that the violence after 1994 was motivated by the willingness to eliminate pockets of support for the minority party in any given area. One example of this would be the 1995 Christmas Day massacre at Shobashobane, in the south east of KwaZulu-Natal. On that occasion, 19 ANC members were murdered by a group of about 600 Inkatha Freedom Party members in an attack to an ANC enclave within an almost exclusively pro-IFP territory.

3 The Data: Concepts and Measurement

The database for the analysis is built through combining several different data sources. The fundamental geographical unit for the within-country empirical analysis is the Magisterial District (MD), a sub-provincial territorial unit defined by the judicial system under the administration of the Department of Justice and Constitutional Development.[12] The map of South Africa in Figure 3 shows the district boundaries. These are still informed by the pre-1994 demarcations of the self-governing states and the Republic of South Africa territory. This makes them particularly valuable as units of comparative analysis on a small-scale geographical basis, as all other

[10] As shown in the Online Appendix, the fundamental results from our empirical analysis are qualitatively the same when we control for measures of the intensity of governmental repression.

[11] It is here worth noticing that existing contributions, like Kok, O'Donovan, Bouare, and Van Zyl (2003), use post-migration data from the 1996 Census to derive predictors of former migration decisions between 1991 and 1996. This constitutes a serious limitation to their study. Indeed, the authors themselves argue that their analysis could not find entirely satisfactory explanations for the observed internal migration trends.

[12] South Africa is currently divided into 9 provinces and 354 MDs. 298 of these were surveyed in the 1991 Census of the Republic of South Africa, as the remaining ones were part of independent Bantustans.

administrative divisions have been subject to frequent re-demarcations after democratization in 1994. Partly because of this, MDs have been used as unit of analysis in the economics and natural science literature (Case and Deaton 1999; Hoffman and Todd 2000).

[Figure 3]

3.1 Ethnic Polarization

Ethnolinguistic information is drawn from the 1991 and 1996 Census of South Africa[13] (Statistics South Africa 1991, 1998). These allow for separate identification of individuals belonging to different ethnolinguistic groups according to the first language they speak. The native African groups in the database are Swazi, Xhosa, Zulu, Sotho, Tswana, Tsonga and Venda.[14] For each MD, we count the total number of individuals in each group and compute the relative share of each group within the black majority.

Throughout the analysis, we employ the *binary* version of the polarization index implemented in Reynal-Querol (2002) and Montalvo and Reynal-Querol (2005, 2010), which falls within the class of measures proposed by Esteban and Ray (1994). The ethnolinguistic polarization within the black majority is computed as

$$ELP_{Within-Black} = 1 - \sum_{i=1}^{N} \left(\frac{1/2 - \pi_i}{1/2} \right)^2 \pi_i \tag{1}$$

where π_i is the within-black share of group i and N is the number of groups. The index value ranges between 0 and 1, with the maximum value being reached in presence of two groups of the same size. The index combines information on both the number of groups and their relative size, returning a distribution measure linked to the generation of social tension between equally distant groups (Esteban and Ray 1994). In the presence of more than two groups, the index value decreases monotonically with the number of groups if they are all of the same size. It is worth here noticing that, when groups have different sizes, changes in the population size of a given group do not map systematically into changes in the polarization index. Esteban and Ray (1994) provide a series of examples in this respect.

[13] Access to the 10% sample Census data for both years was kindly provided by DataFirst Research Unit at the University of Cape Town.

[14] Desmet, Ortuño Ortín, and Wacziarg (2012) show how to make use of the genealogical relationship between world languages in the construction of distribution indexes. In line with their approach, we use information contained in the *Ethnologue* linguistic database (Lewis 2009) assuming all languages at the same level to be as equally distant from the proto-languages of their respective families. We build our measure considering black ethnolinguistc groups in South Africa which correspond to level 11 in the world language tree as reported in Figure B.1 in the Online Appendix.

3.2 Conflict

The measure of conflict incidence is derived from the Geo-referenced Event Dataset of the Uppsala Conflict Data Program (UCDP-GED v1.5).[15] Assembled by the Department of Peace and Conflict Research at Uppsala University, it provides geo-referenced information on organized violence in Africa between 1989 and 2010, detailing different categories - state-based conflict, non-state conflict and one-sided violence.[16] The data are disaggregated spatially and temporally down to the level of individual events of fatal violence. For each conflict event, information is given on date of the event, place of the event (with coordinates), actors participating and estimates of fatalities. A conflict event is recorded in the database if it caused at least 1 death and it involved actors engaged in a nationwide conflict which caused at least 25 deaths in a year in the period (1989-2010).

We measure the yearly incidence of armed confrontations between black-dominated organized groups in South Africa at the MD level by counting the number of related geo-referenced conflict events in each MD per year. These amount to all non-state conflict events.[17]

3.3 Socio-economic and Geographical Controls

We aggregate further information for the surveyed territories at the MD level from Census 1991 and 1996, checking for consistency across waves. Moreover, we use data on population, rural population, number of individuals reporting no education, number of unemployed individuals, number of individuals out of the labor force and number of South African citizens.[18]

In addition to conflict and Census data, we use three additional data sources. Following Michalopoulos and Papaioannou (2013, 2014), we use NOOA (2012) night-time light satellite images data for 1992 (the first available year) and 1996 as a proxy for economic conditions in South Africa at the MD level.[19]. Consistently with their approach, we average night-time light

[15] Department of Peace and Conflict Research, Uppsala University. The dataset is available at http://www.ucdp.uu.se/ged/data.php. See Melander and Sundberg (2011) for the last data presentation. See Eck (2012) for a complete discussion of the UCDP-GED database and its comparison with the Armed Conflict Location Events Dataset (ACLED). The UCPD-GED geocoding and precision is there concluded to be far superior to ACLED's and found particularly suitable for the study of conflict at the sub-national level.

[16] *State-based* conflict is defined as a contested incompatibility that concerns government and/or territory where the use of armed force between two parties, of which at least one is the government of a state, results in at least 25 battle-related deaths in a year in the period (1989-2010). *Non-state* conflict refers to the use of armed force between two organised armed groups, neither of which is the government of a state, which results in at least 25 battle-related deaths in a year in the period. *One-sided violence* refers to the use of armed force by the government of a state or by a formally organised group against civilians which results in at least 25 deaths in a year in the period (Sundberg, Lindgren, and Padskocimaite 2010).

[17] In the study of the relationship between ethnic polarization and conflict, the literature has used conflict data from the Peace Research Institute Oslo (PRIO) Montalvo and Reynal-Querol 2005, 2010; Esteban, Mayoral, and Ray 2012). This is mainly due to the cross-countries analysis and large time span (1945 to 2010) these studies usually consider. Non-state and one-sided conflicts are there considered as outcomes to test robustness of results.

[18] In our regression specification we use the logarithm of these variables, augmenting all values by 0.01 when some of them are equal to zero. Results go through other variables specification and level shift.

[19] For an extensive discussion of these data and their validity as a proxy for economic conditions in the african territories see Michalopoulos and Papaioannou (2014, 2013) See also Doll, Muller, and Morley (2006) and Sutton, Elvidge, and Ghosh (2007). In our regression specifications, consistently with Michalopoulos and Papaioannou (2014, 2013), we augment the night-time light satellite measure by 0.01 before taking its logarithm. Results are

density across 30-second grid areas (approximately 1 square kilometer) within the same MD. In line with the existing literature, we also make use of geographical variables as controls. In our cross-sectional specification we include use a MD-level measure of terrain ruggedness(data from Nunn and Puga (2012)). The measure is computed by averaging 30-second grid area observations belonging to the same MD.[20] The terrain slope index from Global Aero-ecological Zones (GAEZ) data (IIASA/FAO 2012) is created by averaging 5-minute by 5-minute (approximately 9 km by 9 km) grid-cells observations within the same MD. Finally, area accessibility from GAEZ is computed as estimated travel time to nearest city with 50,000 or more inhabitants in year 2000.

4 Empirical Strategy and Results

4.1 Preliminary Analysis

More than 2,000 non-state conflict events are geo-referenced in MDs of South Africa in the UCDP-GED dataset in the period 1989-1996.[21] ANC and IFP confronted each other in more than 85% of events. As for the others, within the most numerous are those where the United Democratic Front (UDF) is involved against the IFP, and the ANC Greens faction against the ANC Reds faction. An average number of 1.2 non-state conflict events per MD is recorded in 1991, with cross-district variation being larger than four times the national mean. Overall, conflict prevalence decreases after democratization in 1994. The average is 0.1 in 1996, but with a cross-district variation of 0.7.[22]

[Table 1]

Table 1 shows the population sizes of the African native ethnolinguistic groups in South Africa according to the 1991 and 1996 Census. Together with nationwide stocks, the table reports the differences in the population of each ethnolinguistic group between 1991 and 1996. MDs in those Bantustans which were already granted independence are not covered by the 1991 Census of the Republic of South Africa. The inclusion of former independent Bantustan territories in the 1996 Census generates a dramatic increase in the population stocks of the corresponding ethnolinguistic groups. In the fourth column, the same overall differences are computed looking only at those 294 districts which are part of both the 1991 and 1996 Census. We find evidence of a substantial inflow into these territories of individuals belonging to ethnicities previously segregated in the independent Bantustans. At the extreme, the Tswana group population

unaltered with respect to other or no level shift.

[20] Data are available at http://diegopuga.org/data/rugged/.

[21] Table A.1 in the Appendix reports the total number of one-sided conflict events involving the Government and non-state conflict events per year in MDs in South Africa recorded in the UCDP-GED dataset, together with the estimated total number of deaths.

[22] Summary statistics for the derived sample are reported in Table A.2 of the appendix.

increases by more than 30%.

On top of migration inflows from independent Bantustans, we find evidence of substantial internal mobility.[23] According to 1996 Census data, 2.5 millions blacks moved from one MD to another in between 1991 and 1996. Figure 4 shows the change in the population share of the three biggest ethnolinguistic groups at the district level in between 1991 and 1996, plotted against the share of black population in 1991. It is worth noticing that the most relevant changes do not seem to be concentrated in those districts where the share of blacks in 1991 was either negligible or close to one.[24]

[Figure 4]

As suggested by the above figures, the negligible change in average within-black polarization between 1991 and 1996 hides substantial variability over time across districts. The time difference goes from a minimum of -0.89 to a maximum of one, spanning almost the entire support of admissible values. Remarkably, the standard deviation of the overtime chenge in the polarization index is as high as half of its 1991 cross sectional standard deviation.

About 32% of conflict events recorded in 1991 takes place in the 25% of districts with the highest level of within-black polarization. More importantly, the relationship appears to be stronger when within-province variability in conflict incidence is considered. The 25% of districts with the highest within-black polarization have on average 0.66 conflicts more than the province average, as opposed to the 0.48 less recorded in the 25% of districts with the lowest within-black polarization.

4.2 Cross-sectional Estimates, 1991

We start the regression analysis by exploiting cross-district variation in both within-black polarization and conflict in 1991. We thus compare the incidence of conflict in 1991 across districts characterized in the same year by different levels of within-black polarization, in the search of a systematic relationship between the two variables. We adopt the following linear regression specification

$$conf_{ip91} = \gamma_p + \beta\ ELP_{WB\ ip91} + \mathbf{Z}'_{ip}\omega + \mathbf{X}'_{ip91}\varphi + u_{ip91} \tag{2}$$

[23] Using data from the South Africa Migration and Health Survey (SAMHS), Reed (2013) studies internal migration patterns amongst the black population of South Africa in the second half of the twentieth century. He reports non-negligible migration rates even before the repeal of the Pass Law in 1986. Nonetheless, migration rates spike in 1991 and after, and estimates are very similar to the one we obtain using data from Statistics South Africa (1998). Indeed, SAMHS data are recognized by Reed (2013) to compare favorably to Census data.

[24] Figure A.1 in the Appendix carries further exploration of migration patterns.

where $conf_{ip91}$ is the number of non-state conflict events recorded in district i in province p in 1991. $ELP_{WB\ ip91}$, computed as discussed in Section 4, is the within-black polarization index in the same district in 1991, while Z_{ip} is a set of time-invariant district geographical characteristics (ruggedness, slope index and accessibility). X_{ip91} is the vector of time-varying demographic and economic controls, capturing time-variant district characteristics in 1991 (log of population, black population, rural population, number of individuals reporting no education, number of unemployed, number of individuals out of the labor force and number of South Africa citizens). γ_p captures province fixed effects, netting out average differences across provinces. The residual u_{ip91} captures instead those unobserved factors which affect conflict incidence.

[Table 2]

Table 2 provides the corresponding results. Throughout all specifications, an Ordinary Least Squares (OLS) estimation with province fixed effects is run on the available 1991 sample. Given the possible endogeneity of within-black polarization to conflict intensity, we start by providing results from a regression with no additional controls in the first column. The results show a positive and significant relationship between the within-blacks polarization index and number of conflict events at the district level.[25] An increase of one cross-district standard deviation of the within-black polarization index is associated with 0.8 more conflict events per year in 1991, more than half of the national average. Columns (2) and (3) include additional controls such as population, black population, and the Night-time Lights measure as proxy for local economic activity (all in logs).[26] Both the magnitude and significance of the coefficient of interest turn out to be almost unaffected. Column (4) reports estimation results after further including geographic controls such as indexes of ruggedness, terrain slope and accessibility. The coefficient of interest remains unchanged in both magnitude and significance with respect to columns (2) and (3). Results in column (5) show the results from the full specification including as controls a number of economic covariates such as the log of rural population, number of unemployed individuals and number of individuals reporting no education. The coefficient of the within-blacks polarization is now lower in magnitude and significant only at the 11% level. Nonetheless, we find results to be consistent with the hypothesis of the observed non-state conflict events to be qualifiable as expressed along ethnic lines. These findings are in line with the cross-country literature relating ethnic polarization to conflict as discussed in Section 2.

[25] We performed the same estimations using a Poisson model specification which takes into account the count nature of our dependent variable. Results are substantially unchanged. These are available in the Online Appendix. Finally, as a further check, we use the logarithm of the number of conflict events (augmented by 0.01) as outcome in all specifications and find highly consistent results.

[26] Notice that the inclusion of population stocks allows to look at the relationship of interest keeping population constant, thus being analogous to the use of the number of conflict events per capita as dependent variable. Indeed, replacing the latter as outcome yields equally significant results.

4.3 First-difference Estimates, 1991-1996

The restoration of free internal mobility of blacks after 1991 largely accounts for time variation in ethnolinguistic group population shares per district. Together with initial cross-district variation, it opens the way for the implementation of a first-difference strategy which looks at the evolution of both within-black ethnolinguistic polarization and non-state conflict, net of district-specific time-invariant characteristics. In other words, it is possible to compare the evolution of conflict across districts experiencing differential changes in ethnic composition, and test whether the observed change in within-black polarization systematically correlates with the change in non-state conflict incidence.

We combine the available information from both 1991 and 1996 and adopt the following specification

$$\Delta conf_{i96-91} = \delta + \beta\ \Delta ELP_{WB\ i96-91} + \Delta \mathbf{X}'_{i96-91}\varphi + \Delta u_{i96-91} \tag{3}$$

where $\Delta conf_{i96-91}$ is the change in the number of recorded non-state conflict events in district i in between year 1991 and 1996, while $\Delta ELP_{WB\ i96-91}$ is the corresponding change in the within-blacks polarization index. The proposed first-difference specification allows to cancel out both observable and unobservable time-invariant characteristics at the district level. The effect of nationwide events (such as democratization in 1994) and general time trends are instead captured by the constant term δ. As before, X_{it} is the vector of time-varying demographic and economic controls in year t (population, blacks, night-time light, rural, etc.). The difference residual Δu_{i96-91} captures those unobserved changes and factors which affect the change in conflict incidence.

Results are reported in Table 3. All specifications are implemented over the sample of districts for which Census data are available for both 1991 and 1996. The first column provides the results from the simple regression specification. Notice that the negative estimate of the constant term is consistent with the general decrease in conflict prevalence with the first democratic elections in 1994. More importantly, the estimated coefficient of the ethnolinguistic polarization is highly significant. The point estimate more than doubles the obtained from the 1991 cross-sectional analysis.[27] An increase in one cross-district standard deviation in the within-black ethnolinguistic polarization measure is now associated with 2 more conflict per district, more than the 1991 national average. Changes in ethnic composition at the district level are thus found to be informative of the evolution of conflict. Column (2) and (3) show that the results are robust to the inclusion of time-variant economic controls. The coefficient of interest remains significant and relatively stable.[28]

[27] We produced estimates from a non linear specification. Results are still positive and significant using a fixed-effects Poisson model. These are available in the Online Appendix. As before, we also use the logarithm of the number of conflict events (augmented by 0.01) as outcome in all specifications and still find consistent results.

[28] Table A.3 in the Appendix shows how estimates of the coefficient of the within-black ethnolinguistic polar-

[Table 3]

Column (4) provides estimation results assuming heteroskedastic difference residuals and estimating Eicker-Huber-White robust standard errors (White 1980). This allows to take into account heterogeneity in the variability of the first-difference residuals in the computation of the standard errors used for inference. The coefficient of interest is still significant at the 10% level. In column (5), we take into account cross-sectional dependence of first-difference residuals and follow Conley (1999) in allowing for non-zero correlation when coordinate distance between districts' centroids is less than 1 degree latitude and/or 1 degree longitude (approximately 110 km). The estimate of our coefficient of interest retains significance at the 10% level. The same holds when spatial spillovers are directly taken into account by including as regressor the average change in within-black ethnolinguistic polarization in neighboring districts $\overline{\Delta ELP}_{WB\ j96-91}$. Assuming heteroskedastic difference residuals, the point estimate of the coefficient of interest is somewhat lower than before, but still significant at the 10% level. The average change in within-black ethnolinguistic polarization in neighboring district is also found to be highly correlated with the change in conflict incidence, with the corresponding coefficient being significant at the 5% level.

Finally, column (7) reports instead the results from a specification which augments the first-difference one by including the number of non-state conflict events in the 1989 and 1990. Given the general decrease in the number of conflict events over the period, we want to test whether the systematic relationship we find between within-black polarization and conflict incidence is robust to conditioning on a measure of pre-1991 conflict incidence.[29] The coefficient of the latter is negative and highly significant. This means that the observed decrease in the number of conflict events in between 1991 and 1996 was higher in those districts with a high number of conflicts in 1989 and 1990. More importantly, we still find evidence of a systematic relationship between within-black polarization and conflict incidence. The estimated coefficient is highly significant. Together with the negative estimate of the pre-1991 conflict measure coefficient, the decrease in the point estimate of the change in within-black polarization suggests the existence of a negative relationship between the latter and conflict incidence in 1989-1990.

ization measure in the first-difference specification remains highly significant when controlling separately for the changes of each within-black group shares. None of the latter individually thus seem to be responsible for the relationship we find in the baseline specification.

[29] Given the adopted linear regression model specification, the estimator of the parameter in column (7) is inconsistent and not directly comparable with the ones from the simple first-difference specification. The inclusion of a lag as a regressor in the first-difference model does not allow to net out district-level time-invariant unobservable characteristics. We intentionally compromise on the consistency of our estimator in order to check whether our results are robust to a partial solution to the problem of mean-reversion and convergence.

4.4 First-difference with Cluster-specific Trends

One possible concern with the above results is that these can be driven by the presence of unobserved factors which affect the evolution of both ethnic composition and conflict in the same way, generating a spurious correlation between the two. For example, those districts where polarization increased in 1991-1996 might be systematically different from others in terms of geographic characteristics (Michalopoulos 2012). The same factors might have been on their own responsible for the change in the incidence of conflict within the black majority, generating a spurious correlation between changes in ethnic composition and conflict.

We study the extent to which these issues are likely to affect the results by looking at the evolution of within-black polarization and conflict within clusters of neighboring districts. Similarly to the neighbors-pair fixed-effects analysis in Acemoglu, Garca-Jimeno, and Robinson (2012), our argument is that districts located next to each other are highly comparable in terms of both observable and unobservable characteristics. As a result, when looking at the relationship of interest within clusters of neighboring districts, it is possible to net out those common unobserved sources of heterogeneity and time-varying omitted variables, possibly correlated with the evolution of both polarization and conflict incidence. Only residual variability in the variables of interest is therefore exploited for identification.

We define as *treated* those districts $g \in M$ where our polarization measure *decreased more than average* in 1991-1996. We then keep their neighboring *non-treated* districts $f \in N(g)$, and drop treated districts with no non-treated neighboring districts. We obtain a sub-sample of 227 out of the initial 294 districts.[30] For each treated district $g \in M$ and its non-treated neighbors $f \in N(g)$, we focus on 1991-1996 differences and consider the following model

$$\begin{aligned} \Delta conf_{g96-91} &= \delta + \gamma_g + \beta\ \Delta ELP_{WB\ g96-91} + \Delta \mathbf{X}'_{g96-91}\varphi + \Delta\varepsilon_{g96-91} \qquad g \in M \\ \Delta conf_{f96-91} &= \delta + \gamma_g + \beta\ \Delta ELP_{WB\ f96-91} + \Delta \mathbf{X}'_{f96-91}\varphi + \Delta\varepsilon_{f96-91} \quad f \in N(g) \end{aligned} \tag{4}$$

where, as before, $\Delta conf_{i96-91}$ is the change in the number of recorded non-state conflict events in district i in between year 1991 and 1996, while $\Delta ELP_{WB\ i96-91}$ is the change in the within-blacks polarization index (with $i = g, f$). Time trends are still controlled for by δ. γ_g captures cluster-specific trends, controlling for unobservable determinants of evolution of conflict in 1991-1996, possibly related with change in polarization over the period. In other words, we include a dummy for each cluster of districts, taking value one for all treated and non-treated observation in the cluster. X_{it} is the vector of time-variant district characteristics in year t (population, blacks, night-time light, rural, etc.). The difference residual $\Delta\varepsilon_{i96-91}$ captures those idiosyncratic unobserved changes and factors which affect the change in conflict incidence, net of cluster-specific trends.

[30] Following the proposed definition, the final subsample contains 105 treated and 122 non-treated districts.

In the final restricted sample, each non-treated district is possibly neighbor of more than one treated district. Thus, there exist several different ways to group districts into clusters. We implement a bootstrap-type procedure where we run a series of regressions, matching in each repetition each non-treated district to a single treated district. The results deliver an empirical distribution of parameter estimates $\hat{\beta}$ which can be used for inference.[31].

[Table 4]

Results are shown in Table 4. Column (1) reports the estimate of the coefficient of interest when only the logarithm of total black population, total population and the night-time satellite light variables are included as controls. Netting out cluster specific trends and exploiting only within-cluster residual variability, we still find a significant relationship between the change in within-black ethnolinguistic polarization and the evolution of conflict. The corresponding point estimate is highly significant and somewhat larger than the one obtained before. The full set of other time-variant economic controls is included in column (2), with results being substantially unchanged. A full set of third-degree polynomials of covariates is included in column (3). We do this in order to control for non-linear discontinuities at the border, which may themselves be correlated with the outcome variable (Acemoglu, Garca-Jimeno, and Robinson 2012). The point estimate is still highly significant, even if lower in magnitude.

Given the restricted sample, we attach to these estimates only a local interpretation. Nonetheless, results are largely consistent with the ones derived in the previous section. If anything, the presence of unobserved omitted factors which are systematically correlated with both the change in within-black polarization and conflict incidence seem to downward bias our initial first-difference estimate of the parameter of interest.

However, exploiting differential changes in ethnic composition across neighboring districts can still be problematic. Internal migration is the primary source of variation for the change in ethnic composition at the district level in between 1991 and 1996. According to 1996 Census data, migration across contiguous districts represent a high fraction of total district-level moves in the period. Focusing on the districts in our sample, Table 5 shows, for each ethnic group, the average number of moves towards neighboring districts as a fraction of total outmigration in 1991-1996. On average, 35% of Xhosa movers from a given district are estimated to relocate in neighboring districts. The same percentage is 28% for movers from the Zulu ethnic group. The relocation of internal migrants in neighboring districts and the differential changes in ethnic composition that follow could possibly be endogenous to the evolution of conflict. The purpose of the next section is to address this issue in a systematic way.

[31]The focus on clusters instead of pairs of neighboring districts and the implementation of the suggested bootstrap-type procedure differs from Acemoglu, Garca-Jimeno, and Robinson (2012) Our approach avoids duplicating observations as required by their neighbors-pair fixed-effects strategy, and delivers standard errors for inference without building up and estimating the parameters of the variance-covariance matrix of residuals as required by their random-effects strategy.

[Table 5]

4.5 Internal Migration and Instrumental Variable Strategy

With the implementation of a first-difference identification strategy we compare districts which experienced different changes in the within-black polarization measure in between 1991 and 1996, and find heterogeneity along this dimension to be significantly correlated with the observed changes in the incidence of conflict within the black majority at the district level. Changes over time in the polarization measure occur when group population sizes change at different rates: if all ethnolinguistic group populations were changing at the same rate in all districts, the polarization measure would have not changed anywhere. Therefore, one concern for the validity of our empirical exercise is that the disproportional changes in the size of ethnolinguistic groups may themselves be driven by the change of conflict incidence.

We focus on internal migration as the main source of variability for the change in districts' ethnic composition. Conflict causes displacement of individuals and households. High non-state conflict incidence in 1991 could be framed as a *push* factor that positively affected individuals' propensity to migrate out of the district after 1991. By the same token, expectations about low non-state conflict incidence in 1996 can be conceptualized as a *pull* factor in the same framework. If agents' expectations were fulfilled, migration decisions would be endogenous to conflict incidence in both 1991 and 1996. Still, this does not necessarily imply that migration decisions are endogenous to the *changes* in conflict incidence over the period. Moreover, even if this was the case, our exercise would be invalidated only if the following two conditions are met. First, endogenous migration decisions need to map into endogenous disproportional changes in ethnolinguistic group population sizes. Second, the resulting changes in the ethnolinguistic polarization index need to be positively correlated with changes in conflict incidence. Both requirements are far from being immediately framed in a coherent narrative. Suppose that one given ethnic group has comparative advantages in armed confrontation. Individuals belonging to this group may thus migrate disproportionally more towards districts where conflict incidence is expected to increase. If expectations are fulfilled, we would observe positive changes in the size of the given group in those districts which experienced higher changes in conflict incidence. However, the same percentage change in the size of a given group does not map systematically into changes in the polarization index if the latter is heterogeneous across districts to begin with. This is even more the case in the presence of more than two groups as in our case. In this respect, Esteban and Ray (1994) show with a series of examples the intuition for the absence of a partial order for increasing polarization. Furthermore, Table A.3 in the Appendix shows how changes in each within-black group shares cannot individually account for all the variability of the change in within-black ethnolinguistic polarization, whose coefficient estimates remain highly significant in the first-difference specification.

To investigate the issue, we implement an instrumental variable (IV) strategy where pairwise

distance between districts is used to predict the location decision of internal migrants. We use information from the 1996 Census on surveyed individuals who declare to have moved in between 1991 and 1996 and to have been resident in a different district in 1991. Using the same data, Kok, O'Donovan, Bouare, and Van Zyl (2003) provide evidence of a negative relationship between the number of migrants moving between two districts and pairwise distance between them. We exploit this feature and estimate a conditional logit model (Cameron and Trivedi 2005) for the location decision of migrants in the sample of the form

$$p_{ij} = \frac{e^{\beta \ distance_{ij}}}{\sum_{j \in S} e^{\beta \ distance_{ij}}} \tag{5}$$

where $distance_{ij}$ is the distance (in km) between district i and j in the set of South Africa districts S. We thus estimate the probability of each individual leaving district i after 1991 to be observed in district j in 1996 as predicted by the values of pairwise distance only.[32]

In a simulation exercise, we next assume that a fraction x of individuals from each ethnolinguistic group e leave each district $i \in S$ after 1991, and allocate them to districts $j \neq i$ using the estimated probabilities $\hat{p}_{ij}$ derived as above. The value of x is chosen by matching it with the total district-level outflow rate of blacks who located in any other district over the relevant period, equal to 8.66%.[33] We thus compute predicted population stocks for each black ethnolinguistic group in each district i in 1996 as

$$\widehat{N}_{e,i,1996} = N_{e,i,1991}(1-x) + \sum_{j \in S} \hat{p}_{ji} \ xN_{e,j,1991} \tag{6}$$

The predicted population share are next used to construct a predicted within-black ethnolinguistic polarization index for 1996 using the same formula as in section 4. We then take the difference between our predicted ethnolinguistic polarization in 1996 and the actual one in 1991 in order to obtain the predicted change in polarization between 1991 and 1996, $\Delta\widetilde{ELP}_{WB,i}$. Notice that, by forcing x to be the same for all ethnic groups in all districts, we rely on $\hat{p}_{ji}$ only in the generation of predicted time variability in the change in the polarization index. Our predicted value for the change in within-black ethnolinguistic polarization is the one which would be observed in case a given exogenous fraction of individuals in each ethnic group in each district were to leave after 1991 and locate in another district according to what predicted by distance only. Indeed, individual reallocation probabilities are computed as independent from conflict levels and their changes at both origin and destination, and assigned to each fictitiously displaced individual. We use this prediction as a source of exogenous variation for the actual change in polarization observed in the data, ruling out the potential endogeneity of individuals'

[32] Conditional Logit estimation results are provided in the Online Appendix to the paper.

[33] Focusing on the 294 districts in our sample, we calculate the total number of individuals moving from one MD to another in between 1991 and 1996 as recorded in the 1996 Census. We then divide this number by the total number of individuals belonging to any black ethnolinguistic group living in any of the 294 districts surveyed in the 1991 Census of the Republic of South Africa.

migration decisions. Furthermore, focusing on predicted internal migration allows to validate the latter as the main mechanism for the change in districts' ethnic composition and rule out other potentially endogenous sources of disproportional changes in ethnic group population sizes, such as disease (specifically HIV) prevalence.

[Table 6]

Results from both the first and the second-stage regressions are reported in Table 6. Conditional on the other variables included as controls, the *F-statistics* for the significance test of the instrument in the first-stage regression are safely above 10 in all specifications. The instrument appears strong enough in producing a relevant shift in the actual value of the change in the within-black polarization index. For consistency with the previous analysis, we start by including only the total number of blacks, total population and night-time satellite light value in logs as controls in the first column. The point estimate of the coefficient of interest is significant at the 10% level and more then doubles the one obtained in Table 3, suggesting the presence of a downward bias in previous estimates from endogenous migration decisions. An increase in within-black polarization of one 1991 cross-district standard deviation is now associated with 4.5 more conflict events, almost four times the 1991 national average. However, results from a Hausman test do not allow to reject the hypothesis of both the first-difference and the IV estimators being consistent (Hausman 1978). In the second column, we allow first-difference residuals to be heteroskedastic and estimate Eicker-Huber-White robust standard errors (White 1980). The estimate of the within-black polarization coefficient are significant at the 5% level.[34] The same pattern of significance is observed when conditioning on the full set of economic controls in column (3) and (4), with the point estimate being somewhat lower in magnitude.

Results from the implemented instrumental variable strategy confirm the existence of a positive relationship between within-black ethnolinguistic polarization and non-state conflict incidence in the data. Ruling out potential threats from both the endogeneity of individual migration decisions and other sources of differential changes in ethnic group populations sizes yields a much bigger point estimate of the effect of interest. The downward bias in the first-difference specification might indeed be caused by heterogeneous exposure and/or resistance of ethnic groups to diseases, which could be possibly negatively correlated with the change in within-black polarization, but positively correlated with the change in conflict incidence. However, given the spatial heterogeneity of internal migration flows we observe in the data, we cannot exclude differences in magnitude to be attributable to the local interpretation of the instrumental variable estimate of the parameter we wish to identify (Angrist, Imbens, and Rubin 1996).

[34] Eicker-Huber-White robust standard errors may be lower than conventional ones. Angrist and Pischke (2010) show that this can be the case whenever lower variance of the residual is associated with covariate values far from the mean of the covariate distribution. In our case, given that the mean of the predicted change in polarization is close to zero, this would mean that those districts which are assigned by our instrument to experience the smallest change in polarization also have higher variability in the first-difference residuals of the change in conflict.

4.5.1 Exclusion Restriction and Falsification Tests

One possible concern with the proposed instrumental variable approach is that, despite the role played by internal migration patterns as estimated using pairwise distances, the predicted change in polarization is still a function of local ethnic distribution in 1991. The latter may not be uncorrelated with unobserved district-level characteristics, which may themselves affect the evolution of conflict incidence in the following year. Additionally, the ethnic distribution in 1991 could have been strategically manipulated by the apartheid rulers or the very same ethnic groups in order to shape the evolution of conflict in the years to follow. In both cases, if the variability in the change of polarization induced by the proposed instrument were to come from the variation in ethnic composition at the MD level in 1991, the exclusion restriction would be violated.

In order to address this concern, we plot in Figure 5 the relationship between the change in conflict incidence 1991-1996 and the initial level of polarization across districts in 1991, together with the line fitting the relationship between the two. The figure shows the absence of any relationship between the two variables. Nonetheless, our instrument is also a function of ethnic composition in other districts, with this relationship being stronger for closer districts. In Figure 6, we explore the relationship between the change in conflict incidence 1991-1996 and the average level of polarization in neighboring districts in 1991. As before, we find no significant relationship between ethnolinguistic polarization in neighboring districts and the evolution of conflict in the following years. Given the highly non-linear nature of our instrument, we also plot the change in conflict incidence between 1991 and 1996 over the initial within-black share of each ethnic group in 1991, and the average share of each ethnic group in neighboring districts. All plots are depicted in Figures A.2 and A.3 in the Appendix. No meaningful relationships are detectable. We take evidence from these results altogether as reassuring: the absence of any relationship between ethnic composition in 1991 and the evolution of conflict in the following years speaks in favor of our exclusion restriction.

[Figures 5 and 6]

Another way to address the same concern is to show that the variation induced by the instrument relies on predicted internal migration patterns, and does not overlap with the one due to 1991 polarization values. We thus implement a falsification test where we use the within-black polarization index in the MD in 1991 as instrument for its actual change in between 1991 and 1996. The idea of the test is straightforward: if the variability we exploit in our original strategy were to be driven by the initial distribution of ethnic groups, results from this falsification test - where predicted migration is not playing any role - would be similar to the previous ones. Results are reported in Table 7. Conditional on the other variables included as controls, within-black polarization index in 1991 is found to be a strong negative predictor of the change in polarization. This is due to the mechanical negative correlation that arises when regressing

the predicted change in polarization over the level of polarization in the previous period, as the latter appears in the former with a negative sign. Nonetheless, second-stage regression results for the coefficient of interest are insignificant in all specifications.

[Table 7]

We repeat the same exercise in Table 8, but using now as instrument the average level of polarization in neighboring districts in 1991. No significant relationship is here found in the first stage. Both the initial level of polarization in the district and the one in neighboring districts are used as instruments in Table 9. Despite the strong first stage, we again find no significant relationship in the second stage.

[Tables 8 and 9]

Again, we interpret these results altogether as validating our original instrumental variable strategy. We there rely on a non-trivial source of variation which stems out of internal migration moves, as the variability in the endogenous regressor induced by measures of ethnic distribution in 1991 does not relate systematically with the evolution of conflict in the following years. Internal migration flows are thus found to be the relevant mechanism for the generation of changes in within-black polarization index which mattered for the evolution of conflict in the period.[35]

As a final check, we investigate whether the change in ethnolinguistic polarization is systematically related with any pre-existing trend in conflict incidence. We argue this is unlikely to be the case, as the years between 1991 and 1996 exhibit an overall decrease in conflict incidence following democratization in 1994. Nonetheless, given that our conflict data are available since 1989, we can replace as outcome the change in conflict incidence between 1989 and 1991 in our first-difference specification. Results are reported in Table 10. By construction, the first stage regression results are identical to those belonging to our main IV specification. Second stage estimates are instead non-significant. This allows us to conclude that the relationship we found between the change in ethnic polarization between 1991 and 1996 does not capture a general trend in the evolution of conflict, and its relationship with the latter is indeed specific of the period under investigation.

[Table 10]

[35]We also use 1980 Census data and replicate the falsification test using the within-black polarization at the district level in 1980. Apartheid legislation was still fully in place in the period, as the Pass Law was only repealed in 1986. Using observations for the 279 districts available in the sample, we find within-black polarization in 1980 to have no predictive power for its change in between 1991 and 1996.

5 Conclusions

This paper studies the extent to which conflict incidence correlates over time with contemporaneous measures of ethnic distribution. The history of contemporary South Africa in 1991 through 1996 carries with it substantial variation in the variables of interest. Combining Census data with geo-referenced information on conflict, we show the incidence of violent struggles amongst black-dominated organized groups during the fall of apartheid to be positively correlated with within-black ethnolinguistic polarization. We thus provide a qualification of these conflicts as expressed through ethnic markers, despite their primarily political nature. Time variability along both dimensions allows to investigate the relationship of interest after clearing out the impact of unobserved time-invariant characteristics at the local level. A one cross-district standard deviation increase in within-black polarization is found to be associated with an increase in the number of conflict events of more than the 1991 national average. Findings are robust to several additional checks. Comparing the evolution of within-black polarization and conflict within clusters of neighboring districts still yields highly significant results. The same holds when focusing specifically on the internal migration channel. We simulate migration patterns and the resulting changes in ethnic distribution using location probabilities as predicted by pairwise distance between districts. Instrumental variable point estimates are still positive and bigger than first-difference ones, validating internal migration as the main source of variability for the changes in ethnic distribution which mattered for the evolution of conflict.[36]

The approach and empirical results in this paper contribute to disclose the potential of the use of micro-level data in the study of social conflict and its determinants. On one hand, migration-driven changes in the ethnic distribution at the local level are here revealed to be informative of conflict prevalence. The exploration of how this interacts with displacement caused by conflict itself constitutes a fruitful avenue for future research. On the other hand, the specificities of the South Africa setting can be further investigated in order to study the effect of democratization on conflict. Evidence from this paper suggests the heterogeneity and divisions within the newly enfranchised majority to interact with nationwide institutional changes and still inform conflict incidence at the local level even after democratization in 1994. The need for a theoretical and empirical investigation of this argument motivates our future research agenda.

References

Acemoglu, D., C. Garca-Jimeno, and J. A. Robinson (2012): "Finding Eldorado: Slavery and long-run development in Colombia," *Journal of Comparative Economics*, 40(4), 534–564.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003): "Fractionalization," *Journal of Economic Growth*, 8(2), 155–94.

[36] Results are confirmed when the estimated number of deaths in non-state conflicts per MD is used as measure of conflict incidence, and a measure of the intensity of governmental repression is controlled for. These findings are not for publication and can be found in the Online Appendix to the paper.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91(434), pp. 444–455.

Angrist, J. D., and J. S. Pischke (2010): "A Note on Bias in Conventional Standard Errors under Heteroskedasticity," Unpublished.

Besley, T., and M. Reynal-Querol (2014): "The Legacy of Historical Conflict: Evidence from Africa," *American Political Science Review*, 108, 319–336.

Blattman, C., and E. Miguel (2010): "Civil War," *Journal of Economic Literature*, 48(1), 3–57.

Bouare, O. (2002): "Determinants of Internal Migration in South Africa," *Southern African Journal of Demography*, 8(1), 23, 2001–2002.

Cameron, A. C., and P. K. Trivedi (2005): *Microeconometrics: Methods and Applications*. Cambridge University Press, New York.

Carver, R. (1996): *Kwazulu-Natal - Continued Violence and Displacement*. WRITENET.

Case, A., and A. Deaton (1999): "School Inputs and Educational Outcomes in South Africa," *The Quarterly Journal of Economics*, 114(3), pp. 1047–1084.

Caselli, F., and W. J. Coleman (2013): "On the Theory of Ethnic Conflict," *Journal of the European Economic Association*, 11, 161–192.

Clark, N. L., and W. H. Worger (2011): *South Africa. The Rise and Fall of Apartheid*. Pearson, 2nd edn.

Collier, P. (2001): "Implications of Ethnic Diversity," *Economic Policy*, 16(32), 127–166.

Collier, P., and A. Hoeffler (2004): "Greed and Grievance in Civil War," *Oxford Economic Papers*, 56(4), 563–595.

Conley, T. G. (1999): "GMM Estimation with Cross Sectional Dependence," *Journal of Econometrics*, 92(1), 1–45.

Desmet, K., I. Ortuño Ortín, and R. Wacziarg (2012): "The Political Economy of Linguistic Cleavages," *Journal of Development Economics*, 97(2), 322–338.

Desmet, K., I. Ortuño Ortín, and S. Weber (2009): "Linguistic Diversity and Redistribution," *Journal of the European Economic Association*, 7(6), 1291–1318.

Doll, C. N., J.-P. Muller, and J. G. Morley (2006): "Mapping Regional Economic Activity from Night-time Light Satellite Imagery," *Ecological Economics*, 57(1), 75 – 92.

Dube, O., and J. F. Vargas (2013): "Commodity Price Shocks and Civil Conflict: Evidence from Colombia," *Review of Economic Studies*, 80(4), 1384–1421.

Eck, K. (2012): "In Data We Trust? A Comparison of UCDP GED and ACLED Conflict Events Datasets," *Cooperation and Conflict*, 47(1), 124–141.

Esteban, J., L. Mayoral, and D. Ray (2012): "Ethnicity and Conflict: An Empirical Study," *American Economic Review*, 102(4), 1310–42.

Esteban, J., and D. Ray (1994): "On the Measurement of Polarization," *Econometrica*, 62(4), 819–51.

——— (1999): "Conflict and Distribution," *Journal of Economic Theory*, 87(2), 379–415.

——— (2008): "On the Salience of Ethnic Conflict," *American Economic Review*, 98(5), 2185–2202.

——— (2011a): "A Model Of Ethnic Conflict," *Journal of the European Economic Association*, 9(3), 496–521.

——— (2011b): "Linking Conflict to Inequality and Polarization," *American Economic Review*, 101(4), 1345–74.

Harari, M., and E. La Ferrara (2013): "Conflict, Climate and Cells: A Disaggregated Analysis," CEPR Discussion Papers 9277, C.E.P.R. Discussion Papers.

Hausman, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46(6), 1251–71.

Hoffman, T. M., and S. W. Todd (2000): "A National Review of Land Degradation in South Africa: the Influence of Biophysical and Socioeconomic Factors," *Journal of Southern African Studies*, 26(4), 743–758.

Horowitz, D. L. (1985): *Ethnic Groups in Conflict*. Berkeley: University of California Press.

IIASA/FAO (2012): *Global Agro-ecological Zones (GAEZ v3.0)*. available at http://www.gaez.iiasa.ac.at/ (accessed April 20, 2012), Laxenburg, Austria and FAO, Rome, Italy.

Kok, P., M. O'Donovan, O. Bouare, and J. Van Zyl (2003): *Post Apartheid Patterns of Internal Migration in South Africa*. Human Sciences Research Council.

Lewis, M. P. (2009): *Ethnologue: Languages of the World, Sixteenth edition*. SIL international Dallas, TX, available at http://www.ethnologue.com/ (accessed March 11, 2012).

Melander, E., and R. Sundberg (2011): "Climate Change, Environmental Stress, and Violent Conflict: Tests Introducing the UCDP Georeferenced Event Dataset," Paper presented at the International Studies Association, March 16-19, Montreal, Canada.

Michalopoulos, S. (2012): "The Origins of Ethnolinguistic Diversity," *American Economic Review*, 102(4), 1508–39.

Michalopoulos, S., and E. Papaioannou (2011): "The Long-Run Effects of the Scramble for Africa," NBER Working Papers 17620, National Bureau of Economic Research, Inc.

——— (2013): "Pre-colonial Ethnic Institutions and Contemporary African Development," *Econometrica*, 81(1), 113–152.

——— (2014): "National Institutions and Subnational Development in Africa," *The Quarterly Journal of Economics*, 129(1), 151–213.

Montalvo, J., and M. Reynal-Querol (2005): "Ethnic Polarization, Potential Conflict and Civil War," *American Economic Review*, 95(3), 796–816.

——— (2010): "Ethnic Polarization and the Duration of Civil Wars," *Economics of Governance*, 11(2), 123–143.

NOOA (2012): *Version 4 DMSP-OLS Nighttime Lights Time Series*. available at http://www.ngdc.noaa.gov/dmsp/downloadV4composites.html (accessed April 20, 2012).

Novta, N. (2013): "Ethnic Diversity and the Spread of Civil War," Unpublished.

Nunn, N., and D. Puga (2012): "Ruggedness: The Blessing of Bad Geography in Africa," *The Review of Economics and Statistics*, 94(1), 20–36.

Posel, D. (2001): "What's in a Name? Racial Categorisations Under Apartheid and Their Afterlife," *Transformation*, 47, 50–74.

Reed, H. E. (2013): "Moving Across Boundaries: Migration in South Africa, 1950-2000," *Demography*, 50(1), 71–95.

Reynal-Querol, M. (2002): "Ethnicity, Political Systems, and Civil Wars," *Journal of Conflict Resolution*, 46(1), 29–54.

Rohner, D., M. Thoenig, and F. Zilibotti (2013): "Seeds of Distrust: Conflict in Uganda," *Journal of Economic Growth*, 18(3), 217–252.

Statistics South Africa (1991): *South African Census 1991 [dataset]. Version 1.1.* Pretoria: Statistics South Africa [producer]. Cape Town: DataFirst [distributor]. http://www.datafirst.uct.ac.za/ (accessed Februrary 3, 2012).

——— (1998): *South African Census 1996 [dataset]. Version 1.2.* Pretoria: Statistics South Africa [producer]. Cape Town: DataFirst [distributor]. http://www.datafirst.uct.ac.za/ (accessed Februrary 3, 2012).

——— (2007): *Using the 2001 Census: Approaches to Analysing Data*. Statistics South Africa, Pretoria.

Sundberg, R., M. Lindgren, and A. Padskocimaite (2010): *UCDP GED Codebook Version 1.0-2011*. Department of Peace and Conflict Research, Uppsala University.

Sutton, P. C., C. D. Elvidge, and T. Ghosh (2007): "Estimation of Gross Domestic Product at Sub-national Scales Using Nighttime Satellite Imagery," *International Journal of Ecological Economics & Statistics*, 8(S07), 5–21.

University of Texas at Austin (1986): *South Africa: Black Homelands 1986*. Perry-Castañeda Library Map Collection, http://en.wikipedia.org/wiki/File:Southafricanhomelandsmap.png (accessed June 8, 2012).

White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48(4), 817–38.

Tables and Figures

Table 1: Ethnic Group Population Sizes

	Total Population		Difference 1996-1991	
	1991 Census 298 MDs	1996 Census 354 MDs	Overall	Study Sample 294 MDs
Xhosa	2,493,382	7,206,005	4,712,623	77,370
Zulu	8,416,125	9,216,413	800,288	163,555
Sotho	6,414,684	7,378,473	963,789	234,038
Swazi	943,989	1,017,233	73,244	45,365
Tswana	1,437,660	3,299,902	1,862,242	463,044
Tsonga	1,450,874	1,757,589	306,715	82,791
Venda	116,533	876,546	760,013	2,296

Notes. The Table shows the total ethnic group population sizes in our sample in 1991 and 1996, together with the overall difference and the difference using only all districts for which information is available in both 1991 and 1996 samples (Sources: Statistics South Africa 1991, 1998).

Table 2: 1991 Cross-sectional Estimation

	Total Number of Non-state Conflict Events 1991				
	(1)	(2)	(3)	(4)	(5)
ELP_{WB}	2.399	2.464	2.444	2.381	1.888
	(1.07)	(1.09)	(1.10)	(1.10)	(1.17)
Blacks (log)		-0.046	-0.029	-0.257	-0.348
		(0.42)	(0.43)	(0.47)	(0.52)
Population (log)		0.367	0.312	0.518	0.226
		(0.52)	(0.55)	(0.59)	(3.89)
Night-time Lights (log)			0.038	0.018	-0.014
			(0.12)	(0.13)	(0.13)
Rural Population (log)					0.001
					(0.13)
No Education (log)					0.325
					(1.64)
Unemployed (log)					1.203
					(0.74)
Constant	-0.425	-3.973	-3.414	-1.032	0.607
	(0.74)	(3.19)	(3.66)	(4.57)	(5.28)
Province Fixed Effects	Y	Y	Y	Y	Y
Geographic Controls	N	N	N	Y	Y
Other Economic Controls	N	N	N	N	Y
Observations	294	294	294	291	291
R^2	0.180	0.185	0.185	0.199	0.214

Notes. Standard errors in parenthesis. The table reports Ordinary Least Squares coefficients estimates from the 1991 cross-sectional specification. The unit of observation is a MD in South Africa for which information is available in 1991. The dependent variable is the total number of non-state conflict events coded in the MD in 1991 in the UCDP-GED dataset. ELP_{WB} is the district-level within-black polarization measure. Other controls are defined as in the data section (Sources: Statistics South Africa 1991, 1998; Nunn and Puga 2012; IIASA/FAO 2012; NOOA 2012).

Table 3: First-difference Estimation

	Change in Total No. of Non-state Conflict Events 1991-1996						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ΔELP_{WB}	5.924	5.562	5.178	5.178	5.178	4.758	4.107
	(1.59)	(1.63)	(1.66)	(2.96)	(3.04)	(2.79)	(1.41)
Non-state Conf 89-90							-0.273
							(0.03)
$\overline{\Delta ELP}_{WB\ j}$						7.692	
						(3.58)	
Δ Blacks (log)		-0.276	-0.325	-0.325	-0.325	-0.354	-0.228
		(0.43)	(0.45)	(0.27)	(0.26)	(0.34)	(0.38)
Δ Population (log)		0.019	5.366	5.366	5.366	2.885	-0.796
		(0.75)	(4.80)	(6.70)	(5.10)	(6.34)	(4.11)
Δ Night-time Lights (log)		-0.277	-0.239	-0.239	-0.239	-0.246	-0.129
		(0.16)	(0.16)	(0.25)	(0.29)	(0.25)	(0.13)
Δ Rural Population (log)			-0.682	-0.682	-0.682	-0.677	-0.378
			(0.22)	(0.26)	(0.29)	(0.26)	(0.19)
Δ No Education (log)			-1.011	-1.011	-1.011	0.142	0.540
			(1.75)	(1.94)	(1.43)	(1.78)	(1.49)
Δ Unemployed (log)			-0.113	-0.113	-0.113	-0.037	-0.112
			(0.83)	(0.61)	(0.68)	(0.61)	(0.70)
Constant	-1.068	-0.882	-3.718	-3.718	-3.718	-3.589	-1.290
	(0.28)	(0.30)	(1.91)	(1.80)	(1.89)	(1.75)	(1.63)
Other Economic Controls	N	N	Y	Y	Y	Y	Y
Observations	294	294	294	294	294	294	294
R^2	0.045	0.058	0.096	0.096	0.096	0.112	0.354

Notes. Standard errors in parenthesis. The table reports first-difference coefficients estimates. The unit of observation is a MD in South Africa for which information is available in both periods. The dependent variable is the change in total number of non-state conflict events coded in the MD in between 1991 and 1996 in the UCDP-GED dataset. ΔELP_{WB} is the district-level change in within-black polarization measure. $\overline{\Delta ELP}_{WB\ j}$ is the average change in neighboring districts. *Non-state Conf 89-90* is the total number of conflict events coded in the MD in 1989-1990. Other controls are defined as in the data section (Sources: Statistics South Africa 1991, 1998; NOOA 2012). (1), (2), (3) are first-difference estimates assuming homoskedastic difference residuals; those in (4) and (6) are first-difference estimates assuming heteroskedastic difference residuals; in (5) we allow for cross-sectional dependence in the structure of difference residuals (Conley 1999) allowing for non-zero correlation when coordinate distance between districts' centroids is less than 1 degree latitude and/or 1 degree longitude (approximately 110 km). Those in (7) are estimates from the first-difference model augmented with the measure of pre-1991 conflict incidence, assuming homoskedastic difference residuals.

Table 4: First-difference with Cluster-specific Trends

	Change in Total No. of Non-state Conflict Events		
	(1)	(2)	(3)
ΔELP_{WB}	6.08	6.36	5.26
	(1.83)	(1.98)	(1.51)
Other Economic Controls	N	Y	Y
Polynomial of Controls	N	N	Y
Observations	227	227	227
Repetitions	200	200	200

Notes. Empirical standard errors in parenthesis. The table reports coefficients estimates and standard errors from a first-difference specification with cluster-specific trends. A restricted sample of *treated* districts are kept from the initial sample, together with their neighboring *control* districts (see text for details). A bootstrap-type procedure is implemented, where at each repetition every control district is randomly matched to a single treated district and coefficients from the first-difference specification with cluster-specific trends are estimated. The total number of blacks and total population, and night-time satellite light value in logs are used as controls in (1), while all economic controls are included in (2). Specification (3) is augmented with a 3rd order polynomial of controls (Sources: UCDP-GED v1.5; Statistics South Africa 1991, 1998; NOOA 2012).

Table 5: Migration Across District Boundaries

	% of Moves towards Neighboring Districts		
	Districts of Origin	Mean	St. Dev.
Xhosa	266	34.7	33.5
Zulu	229	28.3	30
Sotho	238	33.2	28.3
Swazi	116	29	34.5
Tswana	162	32.7	34.4
Tsonga	128	31.1	32.1
Venda	77	31.9	37.5

Notes. The Table shows separately for each ethnic group the percentage of movers in between 1991 and 1996 which are estimated to relocate in neighboring districts with respect to the one of origin. (Sources: Statistics South Africa 1998).

Table 6: Instrumental Variable Estimation

	(1)	(2)	(3)	(4)
1st Stage	Change in Polarization Measure ΔELP_{WB}			
$\Delta\widetilde{ELP}_{WB}$	0.244	0.244	0.249	0.25
	(0.06)	(0.06)	(0.06)	(0.06)
F-stat	17.13	17.13	17.60	17.60
robust *F-stat*		14.82		16.07
2nd Stage	Change in Total Number of Non-state Conflict Events			
$\Delta\widehat{ELP}_{WB}$	13.438	13.438	12.074	12.074
	(7.10)	(5.85)	(6.95)	(5.28)
Δ Blacks (log)	-0.121	-0.121	-0.117	-0.117
	(0.46)	(0.26)	(0.50)	(0.28)
Δ Population (log)	0.361	0.361	2.640	2.640
	(0.83)	(0.76)	(5.54)	(6.15)
Δ Night-time Lights (log)	-0.252	-0.252	-0.243	-0.243
	(0.16)	(0.24)	(0.16)	(0.24)
Constant	-0.969	-0.969	-2.822	-2.822
	(0.32)	(0.29)	(2.12)	(1.75)
Economic Controls	N	N	Y	Y
Observations	294	294	294	294

Notes. Standard errors in parenthesis. The table reports first-stage and second-stage instrumental variable estimates of the first-difference baseline model. The unit of observation is a MD in South Africa for which information is available in both periods. The dependent variable is the change in the total number of non-state conflict events coded in the MD in between 1991 and 1996 in the UCDP-GED dataset. ELP_{WB} is the district-level within-black polarization measure. The total number of blacks and total population, and night-time satellite light value in logs are used as controls in both the first and second stage in (1) and (2), while all other economic controls are included in (3) and (4). First-difference residuals are assumed to be homoskedastic in columns (1) and (3), and heteroskedastic in (2) and (4), where Eicker-Huber-White robust standard errors (White 1980) are estimated (Sources: UCDP-GED v1.5; Statistics South Africa 1991, 1998; NOOA 2012).

Table 7: Instrumental Variable Falsification (I)

Instrument: Within-black Polarization in 1991

	(1)	(2)	(3)	(4)
1st Stage	Change in Polarization Measure ΔELP_{WB}			
$ELP_{1991,WB}$	-0.159	-0.159	-0.156	-0.156
	(0.03)	(0.03)	(0.03)	(0.03)
F-stat	30.52	30.52	28	28
2nd Stage	Change in Total Number of Non-state Conflict Events			
$\Delta \widehat{ELP}_{WB}$	2.862	2.862	3.461	3.461
	(5.25)	(5.72)	(5.45)	(5.16)
Δ Blacks (log)	-0.329	-0.329	-0.377	-0.377
	(0.44)	(0.22)	(0.47)	(0.30)
Δ Population (log)	-0.098	-0.098	6.045	6.045
	(0.78)	(0.65)	(5.16)	(6.51)
Δ Night-time Lights (log)	-0.286	-0.286	-0.238	-0.238
	(0.16)	(0.27)	(0.16)	(0.25)
Constant	-0.853	-0.853	-3.942	-3.942
	(0.30)	(0.25)	(2.00)	(1.87)
Economic Controls	N	N	Y	Y
Observations	294	294	294	294

Notes. Standard errors in parenthesis. The table reports first-stage and second-stage instrumental variable estimates of the first-difference baseline model. The unit of observation is a MD in South Africa for which information is available in both periods. The dependent variable is the change in the total number of non-state conflict events coded in the MD between 1991 and 1996 in the UCDP-GED dataset. ELP_{WB} is the district-level within-black polarization measure. The total number of blacks and total population, and night-time satellite light value in logs are used as controls in both the first and second stage in (1) and (2), while all other economic controls are included in (3) and (4). First-difference residuals are assumed to be homoskedastic in columns (1) and (3), and heteroskedastic in (2) and (4), where Eicker-Huber-White robust standard errors (White, 1980) are estimated (Sources: UCDP-GED v1.5; Statistics South Africa, 1991, 1998; NOOA, 2012).

Table 8: Instrumental Variable Falsification (II)

Instrument: Average Within-black Polarization in Neighboring Districts in 1991

	(1)	(2)	(3)	(4)
1st Stage	Change in Polarization Measure ΔELP_{WB}			
$\overline{ELP}_{-i\ 1991,WB}$	-0.042	-0.042	-0.034	-0.033
	(0.04)	(0.03)	(0.04)	(0.03)
F-stat	1.21	2.05	0.73	1.35
2nd Stage	Change in Total Number of Non-state Conflict Events			
$\Delta\widehat{ELP}_{WB}$	-18.736	-18.736	-18.581	-18.581
	(33.28)	(28.43)	(42.36)	(36.50)
Δ Blacks (log)	-0.752	-0.752	-1.044	-1.044
	(0.86)	(0.86)	(1.40)	(1.36)
Δ Population (log)	-1.037	-1.037	14.758	14.758
	(1.75)	(1.75)	(17.83)	(16.95)
Δ Night-time Lights (log)	-0.355	-0.355	-0.227	-0.227
	(0.23)	(0.38)	(0.21)	(0.30)
Constant	-0.615	-0.615	-6.806	-6.806
	(0.54)	(0.42)	(6.02)	(5.58)
Economic Controls	N	N	Y	Y
Observations	294	294	294	294

Notes. Standard errors in parenthesis. The table reports first-stage and second-stage instrumental variable estimates of the first-difference baseline model. The unit of observation is a MD in South Africa for which information is available in both periods. The dependent variable is the change in the total number of non-state conflict events coded in the MD between 1991 and 1996 in the UCDP-GED dataset. ELP_{WB} is the district-level within-black polarization measure. The total number of blacks and total population, and night-time satellite light value in logs are used as controls in both the first and second stage in (1) and (2), while all other economic controls are included in (3) and (4). First-difference residuals are assumed to be homoskedastic in columns (1) and (3), and heteroskedastic in (2) and (4), where Eicker-Huber-White robust standard errors (White, 1980) are estimated (Sources: UCDP-GED v1.5; Statistics South Africa, 1991, 1998; NOOA, 2012).

Table 9: Instrumental Variable Falsification (III)

Instruments: Within-black Polarization in 1991 and
Average Within-black Polarization in Neighboring Districts in 1991

	(1)	(2)	(3)	(4)
1st Stage	Change in Polarization Measure ΔELP_{WB}			
$ELP_{1991,WB}$	-0.324	-0.324	-0.316	-0.316
	(0.04)	(0.06)	(0.04)	(0.06)
$\overline{ELP}_{-i\ 1991,WB}$	0.270	0.270	0.268	0.269
	(0.05)	(0.07)	(0.05)	(0.07)
F-stat	28.76	15.72	27.03	15.38
2nd Stage	Change in Total Number of Non-state Conflict Events			
$\Delta\widehat{ELP}_{WB}$	6.323	6.323	6.264	6.264
	(3.96)	(5.72)	(4.08)	(5.52)
Δ Blacks (log)	-0.261	-0.261	-0.293	-0.293
	(0.43)	(0.20)	(0.45)	(0.25)
Δ Population (log)	0.052	0.052	4.937	4.937
	(0.76)	(0.63)	(4.95)	(6.41)
Δ Night-time Lights (log)	-0.275	-0.275	-0.240	-0.240
	(0.16)	(0.25)	(0.16)	(0.25)
Constant	-0.891	-0.891	-3.577	-3.577
	(0.30)	(0.25)	(1.94)	(1.83)
Economic Controls	N	N	Y	Y
Observations	294	294	294	294

Notes. Standard errors in parenthesis. The table reports first-stage and second-stage instrumental variable estimates of the first-difference baseline model. The unit of observation is a MD in South Africa for which information is available in both periods. The dependent variable is the change in the total number of non-state conflict events coded in the MD between 1991 and 1996 in the UCDP-GED dataset. ELP_{WB} is the district-level within-black polarization measure. The total number of blacks and total population, and night-time satellite light value in logs are used as controls in both the first and second stage in (1) and (2), while all other economic controls are included in (3) and (4). First-difference residuals are assumed to be homoskedastic in columns (1) and (3), and heteroskedastic in (2) and (4), where Eicker-Huber-White robust standard errors (White, 1980) are estimated (Sources: UCDP-GED v1.5; Statistics South Africa, 1991, 1998; NOOA, 2012).

Table 10: Instrumental Variable Falsification (IV)

Dependent Variable: Change in Total Number of Non-state Conflict Events 1989-1991

	(1)	(2)	(3)	(4)
1st Stage	Change in Polarization Measure ΔELP_{WB}			
$\Delta\widetilde{ELP}_{WB}$	0.244	0.244	0.249	0.25
	(0.06)	(0.06)	(0.06)	(0.06)
F-stat	17.13	17.13	17.60	17.60
2nd Stage	Change in Total Number of Non-state Conflict Events 1989-1991			
$\Delta\widehat{ELP}_{WB}$	-3.982	-4.506	-4.080	-4.080
	(6.12)	(4.54)	(4.96)	(3.84)
Δ Blacks (log)	0.202	0.191	0.273	0.273
	(0.40)	(0.21)	(0.43)	(0.25)
Δ Population (log)	0.026	0.003	-2.223	-2.223
	(0.71)	(0.89)	(4.69)	(4.73)
Δ Night-time Lights (log)	0.046	0.044	0.032	0.032
	(0.14)	(0.26)	(0.14)	(0.26)
Constant	0.100	0.106	0.671	0.671
	(0.27)	(0.17)	(1.82)	(1.86)
Economic Controls	N	N	Y	Y
Observations	294	294	294	294

Notes. Standard errors in parenthesis. The table reports first-stage and second-stage instrumental variable estimates of the first-difference baseline model. The unit of observation is a MD in South Africa for which information is available in both periods. The dependent variable is the change in the total number of non-state conflict events coded in the MD between 1989 and 1991 in the UCDP-GED dataset. ΔELP_{WB} is the district-level change in within-black polarization measure between 1991 and 1996. The total number of blacks and total population, and night-time satellite light value in logs are used as controls in both the first and second stage in (1) and (2), while all other economic controls are included in (3) and (4). First-difference residuals are assumed to be homoskedastic in columns (1) and (3), and heteroskedastic in (2) and (4), where Eicker-Huber-White robust standard errors (White, 1980) are estimated (Sources: UCDP-GED v1.5; Statistics South Africa, 1991, 1998; NOOA, 2012).

FIGURE 1: CONFLICT AND MIGRATION IN CONTEMPORARY SOUTH AFRICA

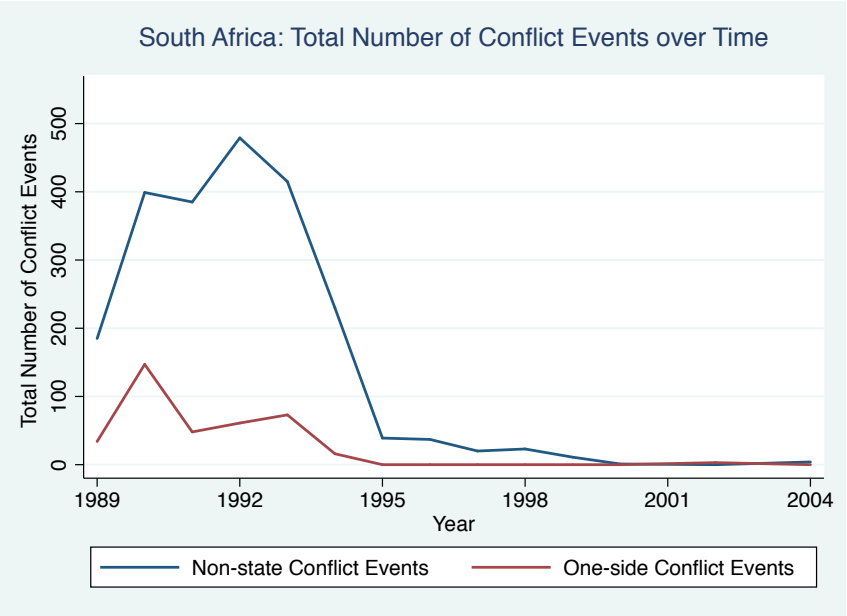


Figure 1a. The figure plots the total number of conflict events in South Africa from 1989 to 2004. Non-state conflict events refer to struggles between black-dominated groups, while one-side conflict events are those in which the Government is involved. The data are described in details in the Data section (Source: UCDP-GED).

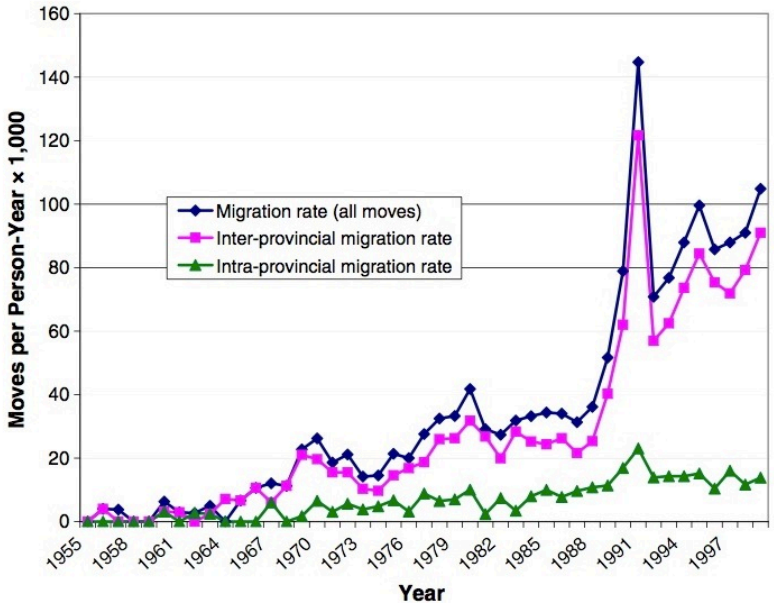
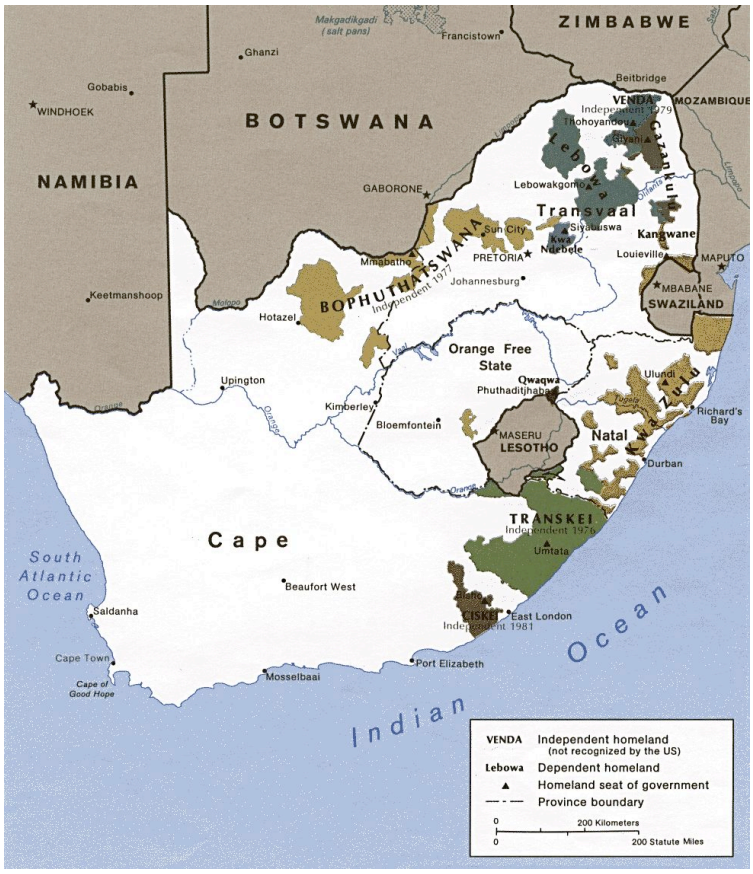


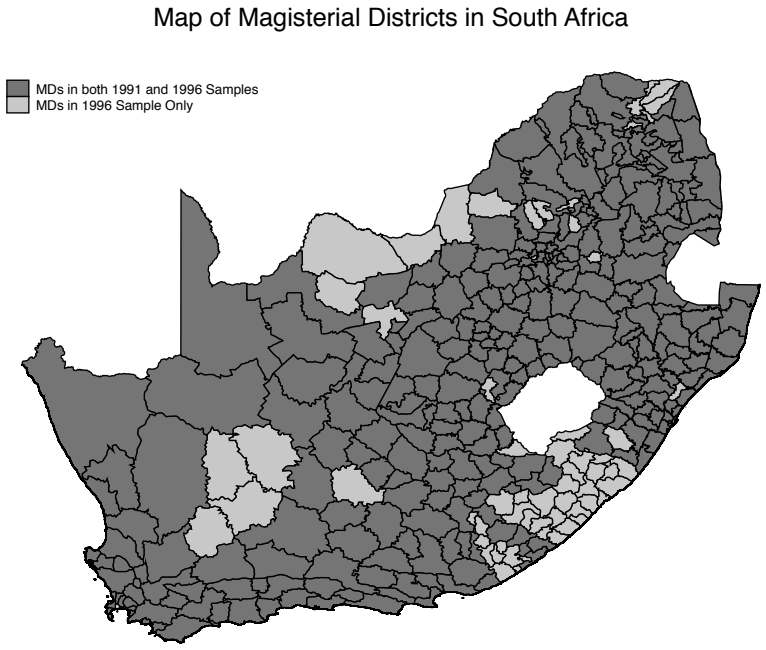
Figure 1b. The figure plots the migration rates in South Africa from 1955 to 1999. Source: South Africa Migration and Health Survey (SAMHS), Reed 2013.

Figure 2: Map of Bantustans



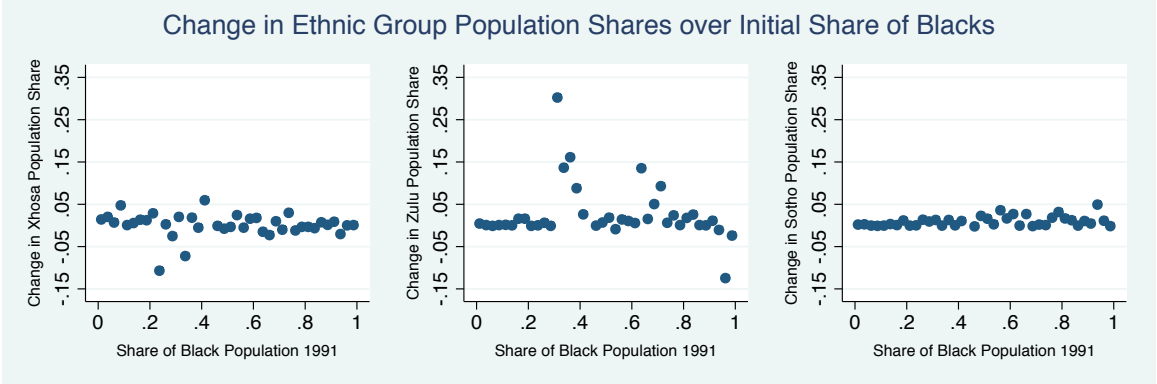
Notes. Map of Bantustans in South Africa as of 1986 (produced by the U.S. CIA. Source: University of Texas at Austin 1986).

FIGURE 3: MAP OF MAGISTERIAL DISTRICTS IN SOUTH AFRICA



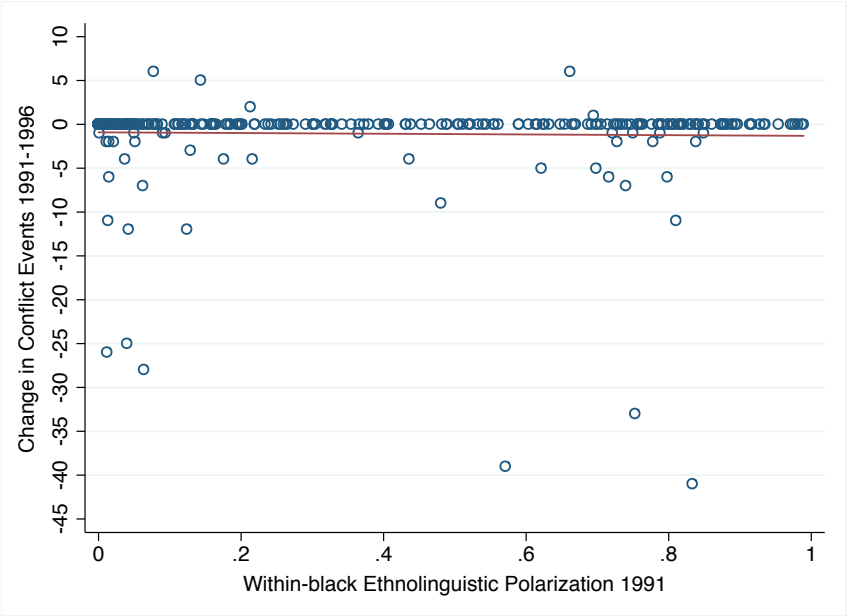
Notes. Map of Magisterial Districts in South Africa, indicating in dark grey those we for which information can be retreived from both the 1991 and 1996 Census (Statistics South Africa 1991, 1998). Those Bantustans which were already granted independence are not covered by the 1991 Census of the Republic of South Africa (Source: authors' elaboration using Stata).

FIGURE 4: ETHNIC GROUP POPULATIONS PER DISTRICT: CHANGES 1991-1996



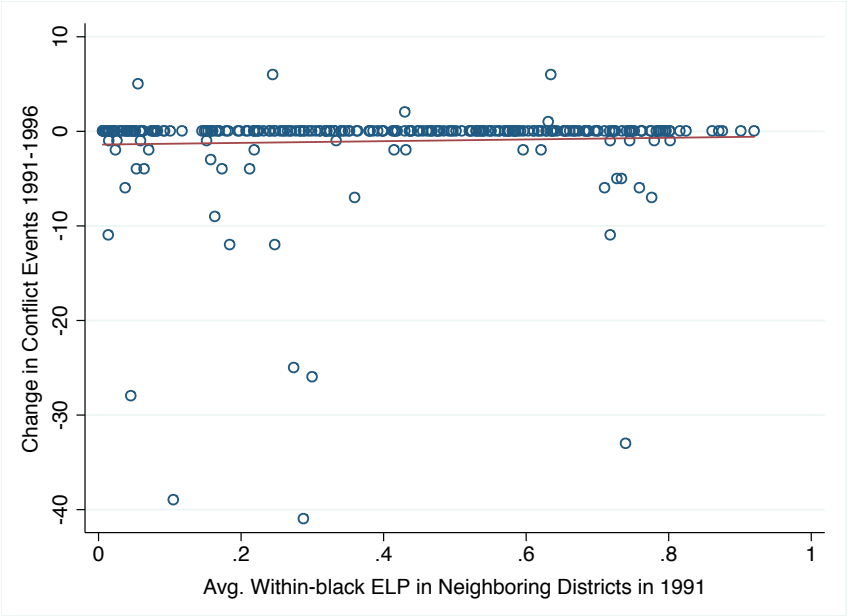
Notes. The figure shows the change in population share of each ethnolinguistic group at the district level plotted over the initial share of blacks. Observations are districts for which we are able to retrieve information consistently from both 1991 and 1996 Census (Source: Statistics South Africa 1991, 1998). This excludes districts in the apartheid homelands which were granted independence and were no longer part of the Republic of South Africa. Observations are averaged per bins of share of blacks of size 2.5%.

Figure 5: Change in Conflict and Initial Polarization



The figure plots the change in the total number of non-state conflict events between 1991 and 1996 over the levels of within-black ethnolinguistic polarization in the same district in 1991 (Source: UCDP-GED; Statistics South Africa 1991, 1998).

Figure 6: Change in Conflict and Initial Polarization in Neighboring Districts



The figure plots the change in the total number of non-state conflict events between 1991 and 1996 over the average levels of within-black ethnolinguistic polarization in neighboring district in 1991 (Source: UCDP-GED; Statistics South Africa 1991, 1998).

A Appendix

Table A.1: One-sided and Non-state Conflict events: 1989-1998

	Gov. Repression		Non-state Conflicts	
Year	Conflict Events	Est. Deaths	Conflict Events	Est. Deaths
1989	34	54	185	226
1990	147	195	399	1243
1991	48	49	385	657
1992	61	61	479	665
1993	73	67	415	643
1994	16	14	230	444
1995	0	0	39	143
1996	0	0	37	156
1997	0	0	20	30
1998	0	0	23	44
1999	0	0	11	19
2000	0	0	1	0
2002	3	2	0	0
2004	0	0	4	0
2010	1	1	0	0

Notes. The Table shows the total number of One-sided and Non-state conflict events per year in South Africa recorded in the geo-referenced Event Dataset of the Uppsala Conflict Data Program (UCDP-GED v1.5) which we are able to map into MDs, together with the estimated total number of deaths.

Table A.2: Summary Statistics

Panel A: Year 1991

Variable	Mean	St. Dev.	Min	Max	N
Non-state Conflict Events	1.271	5.192	0	41	303
ELP_{WB}	0.368	0.336	0	0.989	296
Blacks	72.418	110.795	0	972.838	299
Population	103.67	156.61	3.04	1546.067	299
Rural Population	44.876	76.325	0	419.321	299
No Education	29.879	37.919	1.136	253.145	299
Unemployed	7.194	15.502	0.034	189.362	299
Night-time Lights	4.133	12.79	0	63	354
Accessibility	2.04	0.762	1	4	351
Ruggedness	2.249	1.448	0.237	6.538	354
Slope Index	73.009	22.623	13	99	351

Panel B: Year 1996

Variable	Mean	St. Dev.	Min	Max	N
Non-state Conflict Events	0.105	0.692	0	6	354
ELP_{WB}	0.34	0.332	0	0.997	350
Blacks	87.97	109.609	0	896.042	354
Population	114.63	139.528	3.557	902.861	354
Rural Population	53.123	73.148	0	404.352	354
No Education	21.674	23.123	0.929	137.231	354
Unemployed	13.403	20.166	0.214	189.748	354
Night-time Lights	4.816	13.26	0	63	354

Panel C: Difference 1991-1996

Variable	Mean	St. Dev.	Min	Max	N
Non-state Conflict Events	-1.044	4.9	-41	6	298
ELP_{WB}	0.002	0.177	-0.889	0.997	294
Blacks	3.541	69.902	-665.71	338.236	298
Population	2.752	78.429	-788.692	347.736	298
Rural Population	-2.697	40.793	-224.642	175.135	298
No Education	-10.822	23.11	-195.972	52.845	298
Unemployed	4.695	12.893	-118.249	78.861	298
Night-time Lights	0.785	3.15	-15	29	298

Notes. Data for Blacks, Population, Rural Population, No Education, Unemployed, Not Economically Active and Citizens of South Africa are in thousands. All variables are discussed in Section 3 (Sources: UCDP-GED v1.5, Statistics South Africa 1991, 1998; Nunn and Puga 2012; IIASA/FAO 2012; NOOA 2012).

Table A.3: First-difference Estimation: Within-blach Group Shares as Controls

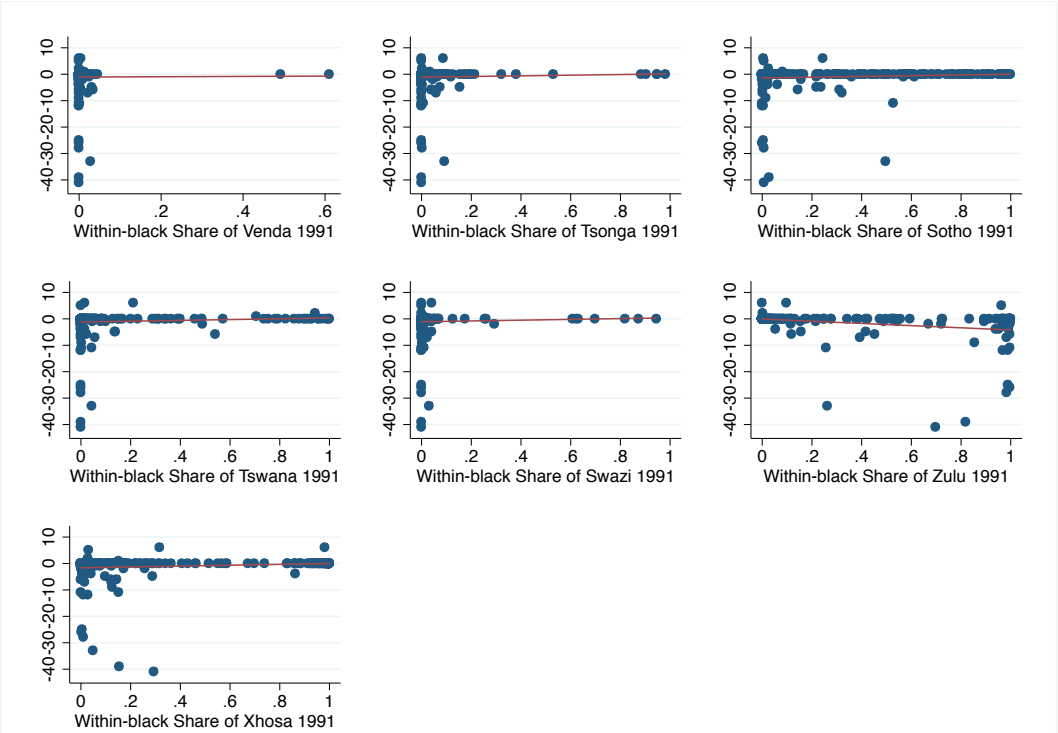
	Change in Total No. of Non-state Conflict Events 1991-1996						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ΔELP_{WB}	6.499	4.961	5.403	5.362	5.174	5.215	5.234
	(1.68)	(1.64)	(1.78)	(1.68)	(1.66)	(1.68)	(1.66)
Δ Sh. Xhosa$_{WB}$	10.865						
	(3.40)						
Δ Sh. Zulu$_{WB}$		-13.569					
		(4.31)					
Δ Sh. Sotho$_{WB}$			-1.503				
			(4.22)				
Δ Sh. Swazi$_{WB}$				-7.346			
				(9.80)			
Δ Sh. Tswana$_{WB}$					-1.648		
					(4.64)		
Δ Sh. Tsonga$_{WB}$						1.114	
						(7.38)	
Δ Sh. Venda$_{WB}$							32.563
							(46.68)
Δ Blacks (log)	-1.299	-0.323	-0.441	-0.342	-0.342	-0.325	-0.328
	(0.54)	(0.44)	(0.55)	(0.45)	(0.45)	(0.45)	(0.45)
Δ Population (log)	4.485	5.612	5.312	5.213	5.265	5.347	5.301
	(4.73)	(4.73)	(4.81)	(4.81)	(4.82)	(4.81)	(4.81)
Δ Night-time Lights (log)	-0.240	-0.229	-0.242	-0.237	-0.237	-0.239	-0.242
	(0.16)	(0.16)	(0.16)	(0.16)	(0.16)	(0.16)	(0.16)
Δ Rural Population (log)	-0.682	-0.704	-0.684	-0.681	-0.679	-0.683	-0.685
	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)	(0.22)
Δ No Education (log)	-1.267	-1.285	-1.018	-0.937	-1.037	-1.022	-0.904
	(1.73)	(1.73)	(1.76)	(1.76)	(1.76)	(1.76)	(1.76)
Δ Unemployed (log)	0.077	0.056	-0.090	-0.165	-0.120	-0.114	-0.111
	(0.82)	(0.82)	(0.83)	(0.83)	(0.83)	(0.83)	(0.83)
Constant	-3.314	-4.042	-3.631	-3.696	-3.731	-3.722	-3.839
	(1.88)	(1.88)	(1.93)	(1.91)	(1.91)	(1.91)	(1.92)
Other Economic Controls	Y	Y	Y	Y	Y	Y	Y
Observations	294	294	294	294	294	294	294
R^2	0.127	0.126	0.096	0.098	0.096	0.096	0.097

Notes. Standard errors in parenthesis. The table reports first-difference coefficients estimates. The unit of observation is a MD in South Africa for which information is available in both periods. The dependent variable is the change in total number of non-state conflict events coded in the MD in between 1991 and 1996 in the UCDP-GED dataset. ΔELP_{WB} is the district-level change in within-black polarization measure. For each black ethnolinguistic group, ΔSh. $Group_{WB}$ is the change in the within-black share of that group. Other controls are defined as in the data section (Sources: Statistics South Africa 1991, 1998; NOOA 2012). All estimates are derived assuming homoskedastic difference residuals.

FIGURE A.1: ETHNIC GROUP POPULATIONS PER DISTRICT: CHANGES 1991-1996

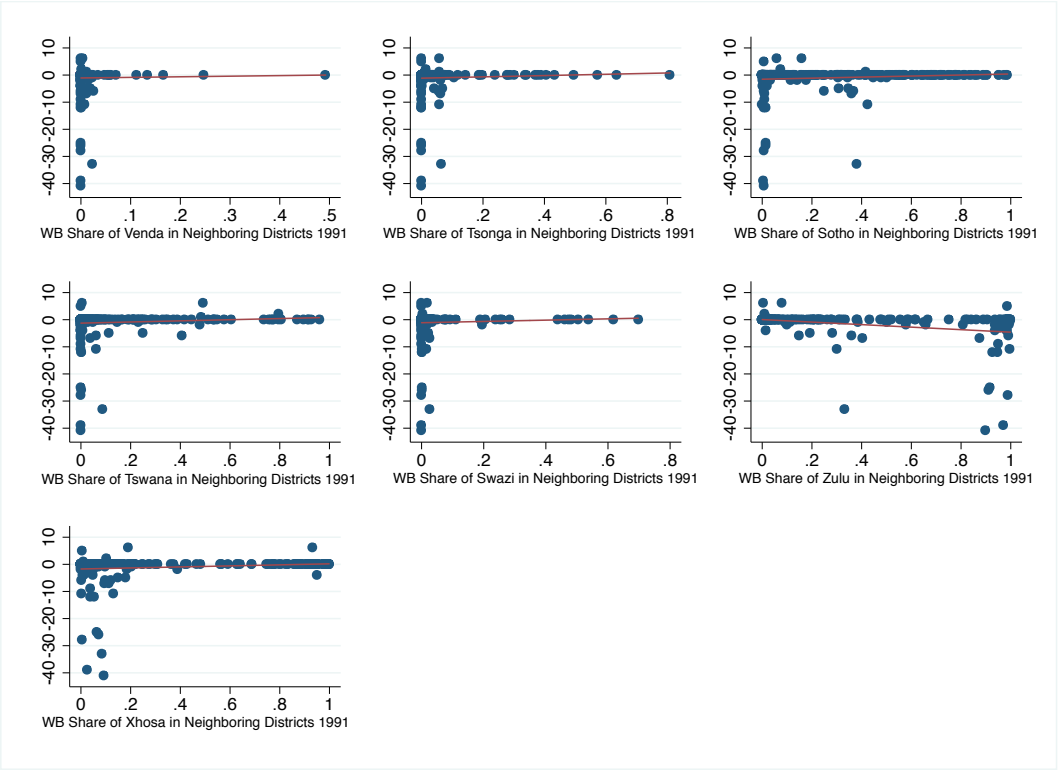
Notes. The top graphs in the figure show the change in population share of each of the three biggest ethnolinguistic groups at the district level plotted over the initial share of each group. The bottom graphs plot instead the changes for the two biggest groups, Xhosa and Zulu, over the population share of the other in 1991. Observations are districts for which we are able to retrieve information consistently from both 1991 and 1996 Census (Source: Statistics South Africa 1991, 1998). This excludes districts in the apartheid homelands which were granted independence and were no longer part of the Republic of South Africa. Observations are averaged per bins of share of blacks of size 2.5%. Taking these figures and Figure 4 together, a few patterns emerge. Districts where the population share of the Zulu ethnolinguistic group grew the most between 1991 and 1996 are those were both the Zulu population share and the share of blacks in 1991 were large but not close to one. According to the bottom graphs, the Xhosa population share was very low in 1991 in these same districts. The Zulu group population seems thus to increase more than proportionally in those districts with a relevant presence of non-black population, and where the Zulu themselves were already a large share of the population. Similarly, looking at changes in the Xhosa population shares, these seem to be disproportionally more positive in those districts where the Zulu group population share was low in 1991, and more negative otherwise. Again, the most important changes in absolute terms seem to be observed where non-black population shares were positive in 1991.

Figure A.2: Change in Conflict and Initial Within-black Shares



The figure plots the change in the total number of non-state conflict events between 1991 and 1996 over the within-black share of each ethnolinguistic group in 1991 (Source: UCDP-GED; Statistics South Africa 1991, 1998).

Figure A.3: Change in Conflict and Initial Within-black Shares in Neighboring Districts



The figure plots the change in the total number of non-state conflict events between 1991 and 1996 over the average within-black share of each ethnolinguistic group in neighboring districts in 1991 (Source: UCDP-GED; Statistics South Africa 1991, 1998).

B Appendix for Online Publication

B.1 Poisson Model Estimation Results

Table B.1: Poisson Model, 1991 Cross-sectional Estimation

	Total Number of Non-state Conflict Events 1991				
	(1)	(2)	(3)	(4)	(5)
ELP_{WB}	2.542	3.747	3.749	3.863	3.947
	(0.20)	(0.28)	(0.28)	(0.32)	(0.37)
Blacks (log)		0.512	0.517	0.631	0.151
		(0.13)	(0.13)	(0.18)	(0.35)
Population (log)		-0.083	-0.095	-0.167	-2.921
		(0.12)	(0.13)	(0.17)	(1.20)
Night-time Lights (log)			0.005	-0.054	-0.115
			(0.02)	(0.03)	(0.03)
Rural Population (log)					0.050
					(0.02)
No Education (log)					0.695
					(0.39)
Unemployed (log)					1.459
					(0.31)
Constant	-3.105	-6.736	-6.631	-20.579	-19.571
	(0.51)	(0.83)	(0.93)	(834.81)	(1443.28)
Province Fixed Effects	Y	Y	Y	Y	Y
Geographic Controls	N	N	N	Y	Y
Other Economic Controls	N	N	N	N	Y
Observations	294	294	294	291	291

Notes. Standard errors in parenthesis. The table reports Poisson model coefficients estimates from the 1991 cross-sectional specification. The unit of observation is a MD in South Africa for which information is available in 1991. The dependent variable is the total number of non-state conflict events coded in the MD in 1991 in the UCDP-GED dataset. ELP_{WB} is the district-level within-black polarization measure. Other controls are defined as in the data section (Sources: Statistics South Africa 1991, 1998; Nunn and Puga 2012; IIASA/FAO 2012; NOOA 2012).

Table B.2: Poisson Model, Fixed-Effect Estimation

	Total Number of Non-state Conflict Events		
	(1)	(2)	(3)
ELP_{WB}	5.814	17.129	24.092
	(2.35)	(5.38)	(8.14)
Blacks (log)		-4.149	-5.645
		(1.19)	(1.60)
Population (log)		6.182	4.175
		(1.45)	(4.44)
Night-time Lights (log)		-0.458	-0.456
		(0.27)	(0.29)
Rural Population (log)			-0.073
			(0.08)
No Education (log)			-2.426
			(1.84)
Unemployed (log)			0.097
			(1.21)
Constant	-2.076	-2.108	-4.958
	(0.19)	(0.24)	(2.48)
Other Economic Controls	N	N	Y
Observations	94	94	94

Notes. Standard errors in parenthesis. The table reports coefficients estimates from a fixed-effects Poisson model. With two time periods only (1991 and 1996), estimates from a first-difference linear regression model are equivalent to the ones of a fixed-effects linear regression model. Estimation is performed over those MDs for which a positive number of non-state conflict events is observed in at least one of the two considered year. The dependent variable is the total number of non-state conflict events coded in the MD in the year in the UCDP-GED dataset. ELP_{WB} is the district-level within-black polarization measure. Other controls are defined as in the data section (Sources: Statistics South Africa 1991, 1998; NOOA 2012).

B.2 Conditional Logit Estimation Results

Table B.3: Conditional Logit Estimation

	Migrant from i is observed in j
Distance between i and j (km)	-0.009
	(5.31×10^{-6})
Observations	894'758'631
Pseudo R^2	0.21

Notes. Standard errors in parenthesis. The table reports the estimated coefficient of the distance variable for the Conditional Logit Model in equation (5) in the text. The number of observations is given by multiplying the number of migrants as recorded in the 1996 Census (Statistics South Africa 1998) for the total number of MDs they had the option to locate in (353).

B.3 Governmental Repression as Control

Table B.4: 1991 Cross-sectional Estimation

	Total Number of Non-state Conflict Events 1991				
	(1)	(2)	(3)	(4)	(5)
ELP_{WB}	2.120	2.145	2.155	2.107	2.073
	(0.99)	(1.01)	(1.01)	(1.02)	(1.08)
Gov. Repression	3.335	3.305	3.311	3.280	3.369
	(0.46)	(0.46)	(0.47)	(0.47)	(0.50)
Blacks (log)		-0.032	-0.041	-0.296	-0.098
		(0.39)	(0.39)	(0.44)	(0.48)
Population (log)		0.171	0.199	0.453	-3.744
		(0.48)	(0.51)	(0.55)	(3.64)
Night-time Lights (log)			-0.020	-0.038	-0.052
			(0.11)	(0.12)	(0.12)
Rural Population (log)					-0.027
					(0.12)
No Education (log)					-1.044
					(1.53)
Unemployed (log)					0.355
					(0.69)
Constant	-0.446	-2.015	-2.303	-0.258	0.198
	(0.68)	(2.96)	(3.38)	(4.22)	(4.89)
Province Fixed Effects	Y	Y	Y	Y	Y
Geographic Controls	N	N	N	Y	Y
Other Economic Controls	N	N	N	N	Y
Observations	294	294	294	291	291
R^2	0.309	0.309	0.309	0.319	0.329

Notes. Standard errors in parenthesis. The table reports Ordinary Least Squares coefficients estimates from the 1991 cross-sectional specification. The unit of observation is a MD in South Africa for which information is available in 1991. The dependent variable is the total number of non-state conflict events coded in the MD in 1991 in the UCDP-GED dataset. ELP_{WB} is the district-level within-black polarization measure. Governmental repression is measured counting in each MD the total number of one-sided conflict events recorded in the UGDP-GED database where the Government is involved against civilians. Other controls are defined as in the data section (Sources: Statistics South Africa 1991, 1998; Nunn and Puga 2012; IIASA/FAO 2012; NOOA 2012).

Table B.5: First-difference Estimation

	Change in Total No. of Non-state Conflict Events 1991-1996				
	(1)	(2)	(3)	(4)	(5)
ΔELP_{WB}	5.175	4.787	4.309	4.309	3.642
	(1.47)	(1.50)	(1.51)	(2.79)	(1.33)
Δ Gov. Repression	3.285	3.296	3.575	3.575	2.560
	(0.45)	(0.45)	(0.46)	(1.58)	(0.42)
Non-state Conf 89-90					-0.233
					(0.03)
Δ Blacks (log)		-0.171	-0.286	-0.286	-0.213
		(0.40)	(0.41)	(0.26)	(0.36)
Δ Population (log)		-0.184	-1.386	-1.386	-4.725
		(0.69)	(4.45)	(7.63)	(3.92)
Δ Night-time Lights (log)		-0.289	-0.254	-0.254	-0.156
		(0.15)	(0.14)	(0.22)	(0.13)
Δ Rural Population (log)			-0.718	-0.718	-0.448
			(0.20)	(0.24)	(0.18)
Δ No Education (log)			-2.143	-2.143	-0.498
			(1.60)	(1.99)	(1.41)
Δ Unemployed (log)			-0.646	-0.646	-0.493
			(0.76)	(0.52)	(0.67)
Constant	-0.586	-0.391	-2.647	-2.647	-0.879
	(0.27)	(0.28)	(1.74)	(1.75)	(1.54)
Other Economic Controls	N	N	Y	Y	Y
Observations	294	294	294	294	294
R^2	0.191	0.204	0.255	0.255	0.430

Notes. Standard errors in parenthesis. The table reports first-difference coefficients estimates. The unit of observation is a MD in South Africa for which information is available in both periods. The dependent variable is the change in total number of non-state conflict events coded in the MD in between 1991 and 1996 in the UCDP-GED dataset. ELP_{WB} is the district-level within-black polarization measure. *Non-state Conf 89-90* is the total number of conflict events coded in the MD in 1989-1990. Governmental repression is measured counting in each MD the total number of one-sided conflict events recorded in the UGDP-GED database where the Government is involved against civilians. Other controls are defined as in the data section (Sources: Statistics South Africa 1991, 1998; NOOA 2012). (1), (2), (3) and (5) are first-difference estimates assuming homoskedastic difference residuals; those in (4) are first-difference estimates assuming heteroskedastic difference residuals.

B.4 Conflict-related Deaths as Alternative Measure of Conflict Incidence

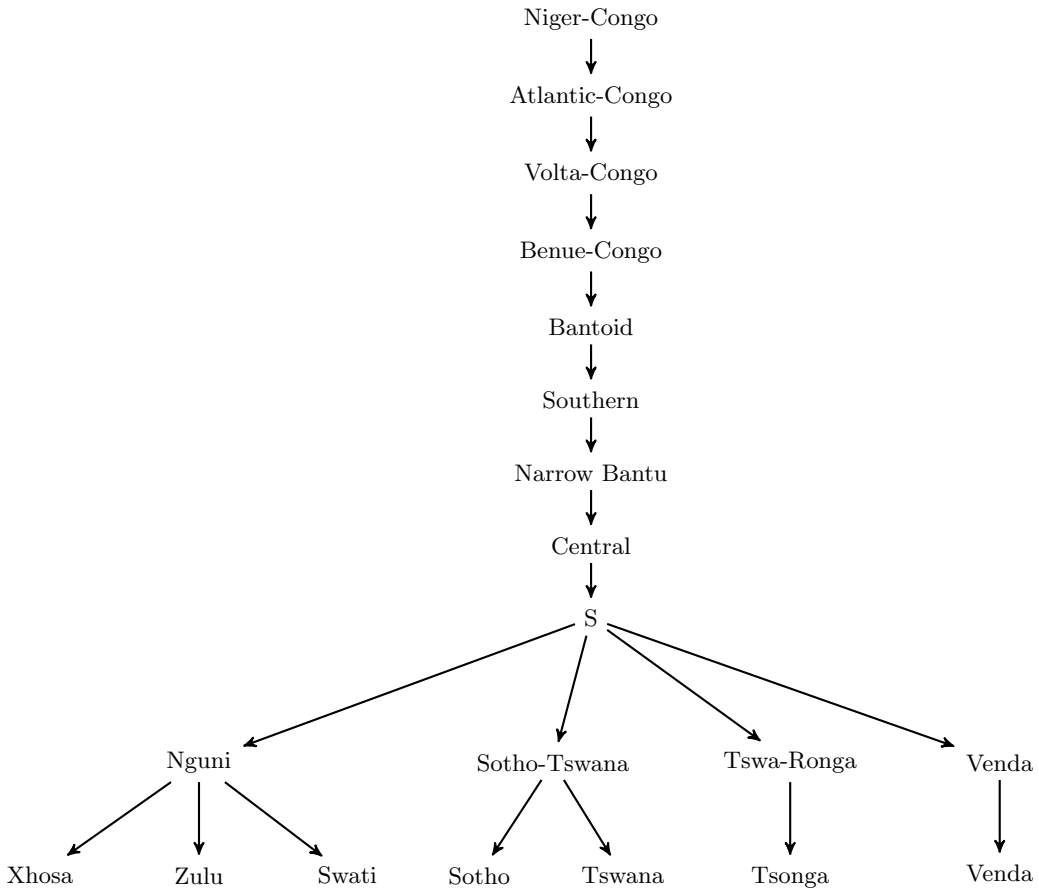
Table B.6: First-difference Estimation

	Change in Total No. of Non-state Conflict Deaths 1991-1996				
	(1)	(2)	(3)	(4)	(5)
ΔELP_{WB}	8.660	8.660	8.259	8.259	5.900
	(3.71)	(4.80)	(3.83)	(4.77)	(3.03)
Est. Deaths 89-90					-0.246
					(0.02)
Δ Blacks (log)	-0.632	-0.632	-0.494	-0.494	-0.490
	(0.98)	(0.45)	(1.03)	(0.50)	(0.81)
Δ Population (log)	0.332	0.332	7.448	7.448	-4.767
	(1.71)	(1.42)	(11.07)	(9.63)	(8.79)
Δ Night-time Lights (log)	-0.834	-0.834	-0.800	-0.800	-0.413
	(0.36)	(0.63)	(0.37)	(0.61)	(0.29)
Δ Rural Population (log)			-0.571	-0.571	0.295
			(0.51)	(0.27)	(0.41)
Δ No Education (log)			-0.238	-0.238	1.314
			(4.04)	(4.29)	(3.19)
Δ Unemployed (log)			1.138	1.138	0.047
			(1.92)	(1.18)	(1.52)
Constant	-1.012	-1.012	-8.046	-8.046	-2.699
	(0.68)	(0.57)	(4.40)	(3.88)	(3.50)
Other Economic Controls	N	N	Y	Y	Y
Observations	294	294	294	294	294
R^2	0.042	0.042	0.056	0.056	0.414

Notes. Standard errors in parenthesis. The table reports first-difference coefficients estimates. The unit of observation is a MD in South Africa for which information is available in both periods. The dependent variable is the change in total number of deaths in non-state conflict events coded in the MD in 1991 and 1996 in the UCDP-GED dataset. ELP_{WB} is the district-level within-black polarization measure. *Est. Deaths 89-90* is the total number of deaths in non-state conflict events coded in the MD in 1989-1990. Other controls are defined as in the data section (Sources: Statistics South Africa 1991, 1998; NOOA 2012). (1), (2), (3) and (5) are first-difference estimates assuming homoskedastic difference residuals; those in (4) are first-difference estimates assuming heteroskedastic difference residuals.

B.5 Ethnolinguistic Information

FIGURE B.1: THE *Ethnologue* TREE OF NATIVE ETHNOLINGUISTIC GROUPS IN SOUTH AFRICA



Source: Lewis (2009).

Pre-Colonial Centralization, Colonial Activities and Development in Latin America

Giorgio Chiovelli*

March, 2015

Abstract

This paper investigates the empirical relationship between pre-colonial ethnic centralization and contemporary development in Latin America at sub-national level. Using historical ethnic homelands as units of observation, we find a strong positive correlation between pre-colonial institutions and levels of development today, as measured by light per capita. This finding is robust to the inclusion of local geographical factors and other pre-colonial observable characteristics (e.g. pre-colonial population density). We argue that this long-term effect relies on the influence played by pre-colonial institutions in shaping early colonial activities in Latin America. To test this hypothesis, we employ a new georeferenced dataset on Spanish American colonial treasuries - *Cajas Reales* - as a proxy for colonial state capacity. We document a strong positive effect of pre-colonial institutions on sub-national colonial institutions. We interpret this as evidence of a direct link between pre-colonial institutions, colonial institutional arrangement and contemporary economic development.

Keywords: Latin America, Ethnicities, Development, Colonial Institutions.
JEL Codes: O10, O40, O43, N16, Z10

*giorgio.chiovelli2@unibo.it. Department of Economics, Università di Bologna, Piazza Scaravilli 2, Bologna 40126 Italy.

1 Introduction

Several contributions had shed lights on the role of early development on contemporary development. At the same time, a growing body of the economic literature has started to focus on the positive effect of state capacity for economic perfermance. This paper studies the long-term effect of pre-colonial state capacity at local level in Latin America.

This paper provides a new empirical framework to study the historical deep rooted factors of regional development in Latin America. The starting point for the analysis is the digitalization of the series of maps contained in Murdock 1951. Using the ethnic homelands as unit of observations, we match to each ethnic group the corresponding entry in the Ethnographic Atlas (EA from now on) (Murdock 1967) containing socio-economic and cultural information on selected ethnic groups. In order to increase the sample numerosity we rely on narratives from Murdock 1951 containing comparable information to those of the EA. Following Gennaioli and Rainer 2007 and Michalopoulos and Papaioannou 2013, we relates pre-colonial centralization and levels of development today, as measured by light per capita. As preliminary analysis, we document the existence of a strong and positive effect of pre-colonial centralization on light per capita. The effect is sizable. Results are robust to the inclusion of several geographical and location characteristics. We exploit within country variabily in order to take into account country-specific unobservable characteristics. At the same time, we digitize the so called *Intendecias* - sub-national colonial administrative units - and we assign them to each corresponding ethnic group. To achive identification, we thus rely on the within-country within-intendency variation in both pre-colonial centralization and light per capita.
The relationship we find might be biased by unobservable factors at pre-colonial time. In order to address this concern we show that the effect of pre-colonial centralization on light per capita holds when controlling for a set of pre-colonial observable characteristics (e.g. pre-colonial population density). We then turn the analysis at more disaggregated pixel level of analysis in order to show that our findings are stable to a finer set of geographic controls. Using a 0,125x0,125 degree pixel level resolution we show that our results are confirmed.

We then turn to investigate a possibile channel operating between pre-colonial centralization and regional development today. The impact of colonial activities in Latin America is undeniable Dell, 2010. Recent contributions started to investigate the determinants of sub-national colonial activieties (Arias and Girod 2011). We suggest that the long-term effect of pre-colonial centraliziation on regional development relies on the influence played by the former in shaping early colonial activities in Latin America. The ideal institution to study the determinants and the consequences of local colonial activities should be highly decentralized and operating on a well-defined terroritorial basis. We georeference a novel dataset on Spanish American colonial treasuries - *Cajas Reales*. Given its tax colletcion function, we present this institution as a proxy for subnational colonial state capacity. We show that pre-colonial centralization is strongly correlated with the presence of *Cajas Reales* in a given location. We interpret this as evidence of a direct link between pre-colonial institutions, colonial institutional arrangement and contemporary economic development.

2 Data

2.1 Spatial Distribution of Pre-Colonial Ethnic Homelands

The first step for the creation of the dataset is the ditigitalization of the series of Ethnolinguistic Maps contained in George Peter Murdock's (1951, 1967). In his work, Murdock 1951 maps the spatial distribution of pre-colonial ethnic homelands in Central and South America as well in Mexico. Murdock's work represents the best attempt to map the spatial distribution of the pre-colonial ethnic homeland at the eve of European colonization[1]. The ethnolingustic maps are presented following the modern country divisions. For each country in Central and South America we have a map depicting the geographical distribution of its respective pre-colonial ethnic groups. Geographic coordinates are assigned to each single map facilitating the spatial aggregation of different maps. In Murdock and O'Leary 1975, the whole continental North America is mapped. I digitize this map and clip it, keeping only the ethnic homelands belonging to the actual borders of Mexico, with the one of Central and South America's map. Finally we are able to retrieve 330 ethnic homelands for Latin America.

[Figure 1]

2.2 Ethnic Characteristics

Ethnographic Atlas (Murdock 1967) provides several information on ethnic-level cultural, geographic, and socio-economic characteristics of 1270 ethnicities around the world. Using the updated version of the Ethnographic Atlas (Gray 1999), I match 121 ethnic groups to 95 ethnic homelands in Latin America.

Latin America is characterized by a paucity of ethnic information retrievable from the Ethnographic Atlas (Nunn, Alesina, and Giuliano 2013; Michalopoulos 2012). To solve this limitation, I digitize ethnic narratives from Murdock (1951). reports a short description of each ethnic homelands' ethnographic information. Using these narratives,I follow the coding procedure of the Ethnographic Atlas (Gray 1999). This procedure allows me to increase the sample dimension for Central and South America from 40 to 50 %[2].

As a measure of pre-colonial state capacity I use the "Jurisdictional Hierarchy beyond the Local Community" index as derived from the procedure described above. This variable identifies the political jurisdictional structure above the local community level for each society. In order words, this index

[1] "[..] data always pertain to the period of earliest European contact unless otherwise explicitly stated" pag 3

[2] In Appendix are reported all the information regardin this procedure

measures the degree of political complexity of pre-colonial villages' network. The degree of political centralization ranges from 0 to 4. At the bottom of the classification are the statless societies *lacking any form of political organization*. A score of 1 is assigned to petty chiefdoms; a score of 2 identifies paramount chiefdoms; and 3 and 4 are the scores assigned to ethnic homelands characterized by pre-modern states as political ogranization[3]. Figure 2 shows the variation in the distribution of pre-colonial centralization across ethnic homelands in Latin America.

[Figure 2]

2.3 Satellite Night Light

We are interested in measuring regional development at the highest resolution as possible. A growing body of the literature on comparative development uses luminosity data from the Defense and Metereological Satellite Program's Operational Linescan System. (DMPS-OLS). DMPS-OLS records satellite images of the earth at night. The satellite detects lights from human settlements, fires, gas flares, lightning, and the aurora. The measure is a six-digit number (rangin from 0 to 63). The high resolution of the satellite images (1km x 1km) makes them particularly valuable for the porpuse of this study. We construct average light density for the period 2008-2012 averaging across pixels at the level of our unit of observation.

We then construct additional measures of regional development as captured by luminosity at night. In order to cope with the blooming problem, we use all the available data on luminosity from the DMPS-OLS to construct an avarage light density measure for the period 1992-2012 (Ashraf, Galor, and Klemp 2014). If blooming problem was due to a particular yearly climatic condition, this procedure should mitigate the issue. As an additional measure of regional development, we construct a measure of light per capita for the period 2008-2012 (Pinkovskiy 2013; Ashraf, Galor, and Klemp 2014).

3 Empirical Strategy and Results

To study the relationship between pre-colonial centralization and contemporary regional development across ethnic homeland, we estimate the following specification:

$$y_{i,c,d} = a_0 + a_c + a_d + \beta Centr_i + X_{i,c,d}\Psi + \lambda PD_{i,c,d} + \epsilon_{i,c,d}. \quad (1)$$

[3]It is worth to notice that a main difference between my classification and the classification of the Ethnographic Atlas is that in Murdock (1957) the ethnographic information are usually reported at the ethnic homeland level; while in the Ethnographic Atlas, ethnographic information are provided at the ethnic group level. Several ethnic groups may be present in the same ethnic homeland. This may in part explain the differences in the coding. In the Appendix the reader will find a clear exposition of this issue.

The dependent variable $y_{i,c,d}$ captures the level of regional development in country i in intendency d, as proxied by light intensity. $Centr_i$ is the local ethnic centralization following the degree of the jurisdictional hierarchy beyond the local level. If the ethnicity is split by a country or intendency border, each partition is assigned to the corresponding country c or intedency d. Almost all specification includes country fixed effect (a_c) and intendency fixed effect (a_d) in order to exploit the within-country within-intendency variation. This procedure may magnify the measurument error problem absorbing a sizable part of the overall variation. Nevertheless we are willing to trade this problem with the chance of ruling out potential unobservable characteristic both at the country level (e.g. quality of national institutions) and at the intendency level (local-specific colonial policies - *Mita system*).

Results in Table 1 unconver a positive and significant relation between pre-colonial institutions and regional development today. The effect is stable across several specification in which we control for several geographic and location characteristics. Results hold when we control for pre-colonial proxy of development (population density in 1492).

3.1 Inference

There are at least two important dimensions on which serial correlation should be accounted for. The first one is the ethnolinguistic family to which an ethnic group belong to. It can be the case that inside the same ethnolinguistic family the pre-colonial institution is not randomly distributed. In all specification we cluster the standard error at the ethnolinguistic family level as derived by the level 1 of the Ethnologue family tree. The number of cluster on this dimension is 67. Another source of potential serial correlation is the country dimension. Notice that the number of country is 19 (15 in the Spanish American sample). This makes inference based on the cluster dimension unreliable. In order to try to mitigate this issue, we offer three different solution. First, we first clusterize standard errors at the country-ethnolinguistic level. Secondly, we allow our standard error to be spatially serially correlated and we implement a Conley's correction procedure [4]. Finally, in order to be as conservative as possibile, we estimate a linear regressions with multi-way clustered standard errors (both at the country and ethnolinguistic level) bootstrapped for one cluster dimension (country level) [5].

4 Pre-Colonial Omitted Variable

The relationship between pre-colonial centralization and contemporary regional development today might be affected by other characteristics at pre-colonial time. If one of these characteristics is correlated with both the pre-colonial centralization and luminosity in 2008-2012 this may lead to an omitted variable problem for our estimates. In order to mitigate this issue we try to show that other pre-colonial

[4] michalopoulos2012 show that Conley correction for spatial correlation is similar to a double clustering at the country and ethnolinguistic level.

[5] We use the STATA command **cgmwildboot** as written by **?**. Note that this last procedure does not provide standard errors. It returns a p-value to test the null hypothesis that the coefficient of interest is not different from 0.

characteristics are not correlated with contemporary regional development today. When such correlation is present, we them show that, conditioning on these variables, pre-colonial centralization is still an important predictor of regional development.

[Table]

In columns A we test light density on the ethnic-level variables controlling for country and intendency fixed effects. The general trend is that almost all pre-colonial characteristics are insignificant. In columns B we add the our pre-colonial centralization and we want to see whether the correlation with regional development conditional is driven by other ethnic traits. Reassuringly, The coefficient of the pre-colonial centralization enters positive and significant in all specifications.

5 Channels of Persistence

Previous findings unconver a positive correlation between pre-colonial centralization and contemporary level of regional development in Latin America. What can be the mechanism behind this long-term effect? In this section we argue that this empirical regularity may rely on the influence played by pre-colonial centralization in shaping early colonial activities in Latin America. The idea is that pre-colonial structure and political complexity altered the incentives of colonizers in their settling decisions. The existence of complexed societies should lower the cost of setting an institution in a particular place if some elements of that institutions (e.g. tributes) were already present in that place. Acting as a cost-reducing factor, highly centralized pre-colonial societies facilitates the formation of colonial state presence. Along with its extractive nature, colonial state presence is related to provision of local public good - like roads - and to higher fiscal capacity. This directly links pre-colonial centralization, colonial institutional arrangement and contemporary regional economic development.

5.1 *Caja Reales*

In order to test this conjecture, we employ a novel dataset on sub-national colonial institutions, namely *Cajas Reales*. *Cajas Reales* were sub-national treasury offices with the task of tax collection in a unique, well-defined geographic area. Anytime a new political unit was established a *Caja Real* was created too. The "system consisted of a network of quasi autonomous interdependent fiscal districts and authorities" (Irigoin and Grafe 2008). Though organized in a hierarchical system [6], local officials were responsible for "gathering and spending royal revenues". This generates regional differences in both fiscal capacity and local investement in public goods. In fact, the main task of the centralized authority survising the regional cajas was to eliminate regional differences and to try to homogenize administritive practices.

[6]

We take adavantage of the seminal contributon TePaske 1982, 1990 collected data on *Caja Reales* for the period 1520-1800. We focus on the sub-sample of the *Cajas Reales* as digitized by . Moreover, we include cajas for Colombia (Jaramillo, Meisel, and Urrutia 2006) and Venezuela (García 1991). We match the name of each caja real to the corresponding contemporary city in order to obtain latitude and longitude information. With this procedure, we are able to retrieve information on 90 local treasuries location. We overalap the ethnic-country-intendencia with the caja reales in our database. We then count the number of caja reales that are contained in each ethnic-country-intendencia. This variable will be our proxy for colonial state presenceat subnational level. Figure shows the spatial destribution of cajas reales in Latin America.

[Figure]

5.2 Pre-Colonial Centralization and Colonial State Presence

We want to test whether the degree of pre-colonial centralization correlates with colonial state presence at sub-national level, as proxied by the presence of at least one caja real. For this reason, we adopt the following specification:

$$Colonial_{e,i,c} = \alpha + \beta Jurisdictional_e + \gamma Z_{e,i,c} + \phi E_{e,i,c} + \mu_c + \mu_i + \epsilon_{e,i,c} \tag{2}$$

The dependent variable $Colonial_{e,i,c}$ captures the number of $CajaReal$ in ethnic homeland e in intendency i in country e. $Jurisdictional_e$ is the local pre-colonial ethnic centralization following the degree of jurisdictional hierarchy beyond the local level in ethnic group e. As before, if the ethnicity is split by a country or intendency border, each partition is assigned to the corresponding country c or intedency d. In almost all specification, we include country fixed effect (a_c) and intendenc fixed effect (a_d) in order to exploit the within-country within-intendency variation. In several specification we control for land endowments (elevation and area under water), ecological features (a malaria stability index, land suitability for agriculture, precipitation, temperature), and natural resources (diamond mines and petroleum fields). We then include a set of location control of each ethnic homeland in each country in each intendency. We include for each ethnic group distance from its centroids to the corresponding country capital, country border, closest major river, and closest coast.

Results in Table 3 show a positive and significant effect of pre-colonial institutions on subnational colonial institutions. The effect is stable across several specification controlling for a rich set of geographical and location characteristics. This evidence is consistent with the view that pre-colonial early institutions shaped colonial strategy and affecting current development today.

[Table 4]

References

ACEMOGLU, D., S. JOHNSON, AND J. A. ROBINSON (2001): "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 91(5), 1369–1401.

ARIAS, L. M., AND D. GIROD (2011): "Indigenous origins of colonial institutions," Discussion paper, Working Paper, University of California San Diego.

ASHRAF, Q., O. GALOR, AND M. KLEMP (2014): "The Out of Africa Hypothesis of Comparative Development Reflected by Nighttime Light Intensity," Working Papers 2014-4, Brown University, Department of Economics.

GARCÍA, J. A. (1991): *La intendencia en Venezuela: don Esteban Fernández de León, intendente de Caracas, 1791-1803*, vol. 25. Editum.

GENNAIOLI, N., AND I. RAINER (2007): "The modern impact of precolonial centralization in Africa," *Journal of Economic Growth*, 12(3), 185–234.

GRAY, J. P. (1999): "A corrected ethnographic atlas," *World Cultures*, 10(1), 24–136.

IRIGOIN, A., AND R. GRAFE (2008): "Bargaining for absolutism: a Spanish path to nation-state and Empire Building," *Hispanic American Historical Review*, 88(2), 173–209.

JARAMILLO, J., A. MEISEL, AND M. URRUTIA (2006): "Continuities and Discontinuities in the Fiscal and Monetary Institutions of New Granada 1783-1850," *Transferring wealth and power from the Old to the New World: Monetary and fiscal institutions in the 17th through the 19th Centuries*, p. 414.

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.

MICHALOPOULOS, S. (2012): "The Origins of Ethnolinguistic Diversity," *American Economic Review*, 102(4), 1508–39.

MICHALOPOULOS, S., AND E. PAPAIOANNOU (2013): "Pre-Colonial Ethnic Institutions and Contemporary African Development," *Econometrica*, 81(1), 113–152.

MURDOCK, G. P. (1951): *Outline of South American Cultures*, vol. 2. Human Relations Area Files.

——— (1967): "Ethnographic atlas," .

MURDOCK, G. P., AND T. J. O'LEARY (1975): *Ethnographic Bibliography of North America: General North America*, vol. 1. Human Relations Area Files Press.

NEW, M., D. LISTER, M. HULME, AND I. MAKIN (2002): "A high-resolution data set of surface climate over global land areas," *Climate research*, 21(1), 1–25.

NUNN, N., A. ALESINA, AND P. GIULIANO (2013): "On the Origins of Gender Roles: Women and the Plough," *Quarterly Journal of Economics*, 128(2).

PINKOVSKIY, M. L. (2013): "Economic discontinuities at borders: Evidence from satellite data on lights at night," Discussion paper, Working Paper.

RAMANKUTTY, N., J. A. FOLEY, J. NORMAN, AND K. MCSWEENEY (2002): "The global distribution of cultivable lands: current patterns and sensitivity to possible climate change," *Global Ecology and Biogeography*, 11(5), 377–392.

TEPASKE, JOHN J., K. H. S. E. A. (1982, 1990): *The Royal Treasuries of the Spanish Empire in America*, vol. 1-4. Duke University Press.

6 Tables and Figures

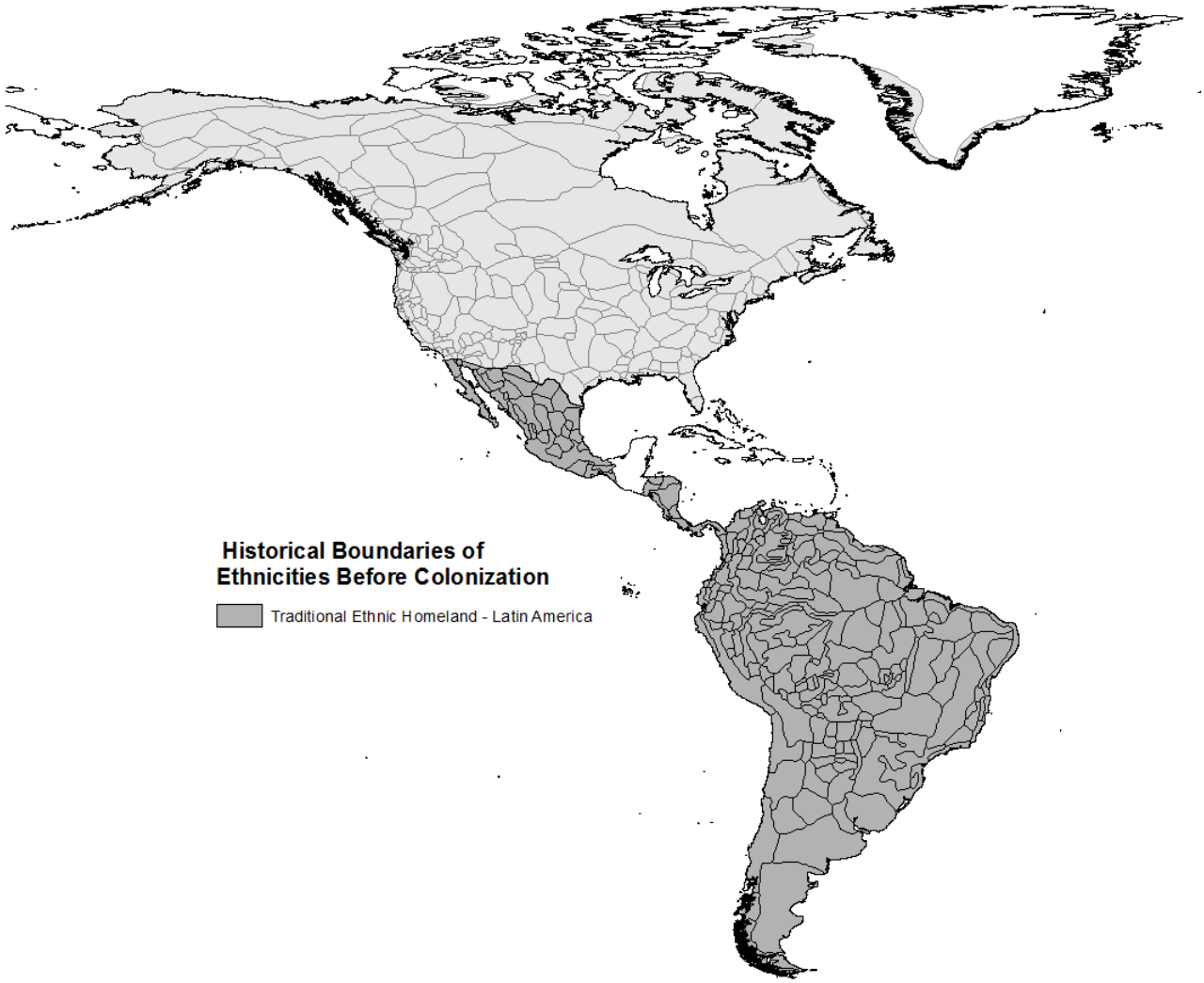


Figure 1. Spatial Distribution of Ethnic Homelands in Latin America

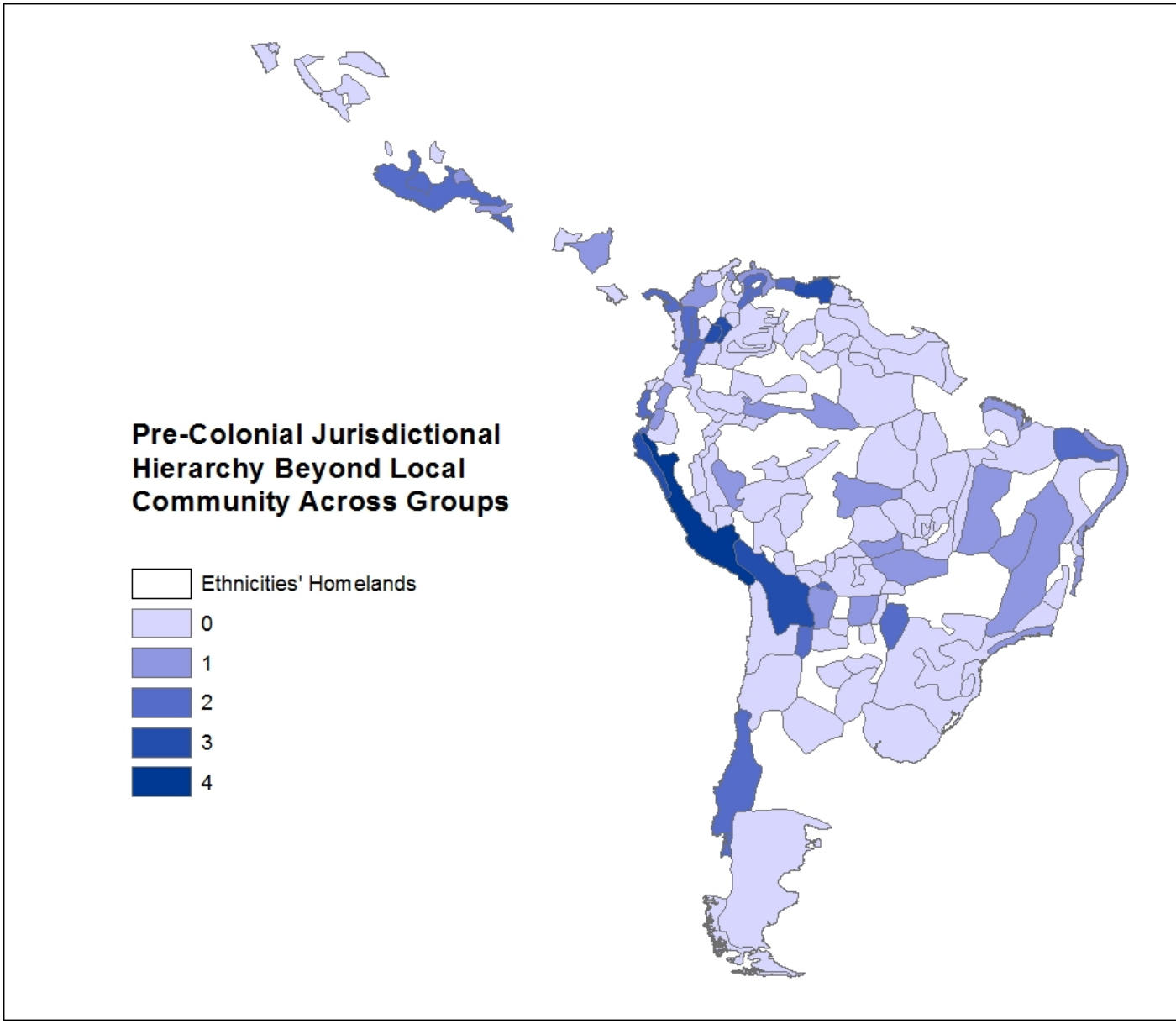


Figure 2. Ethnic Pre-Colonial Centralization in Latin America

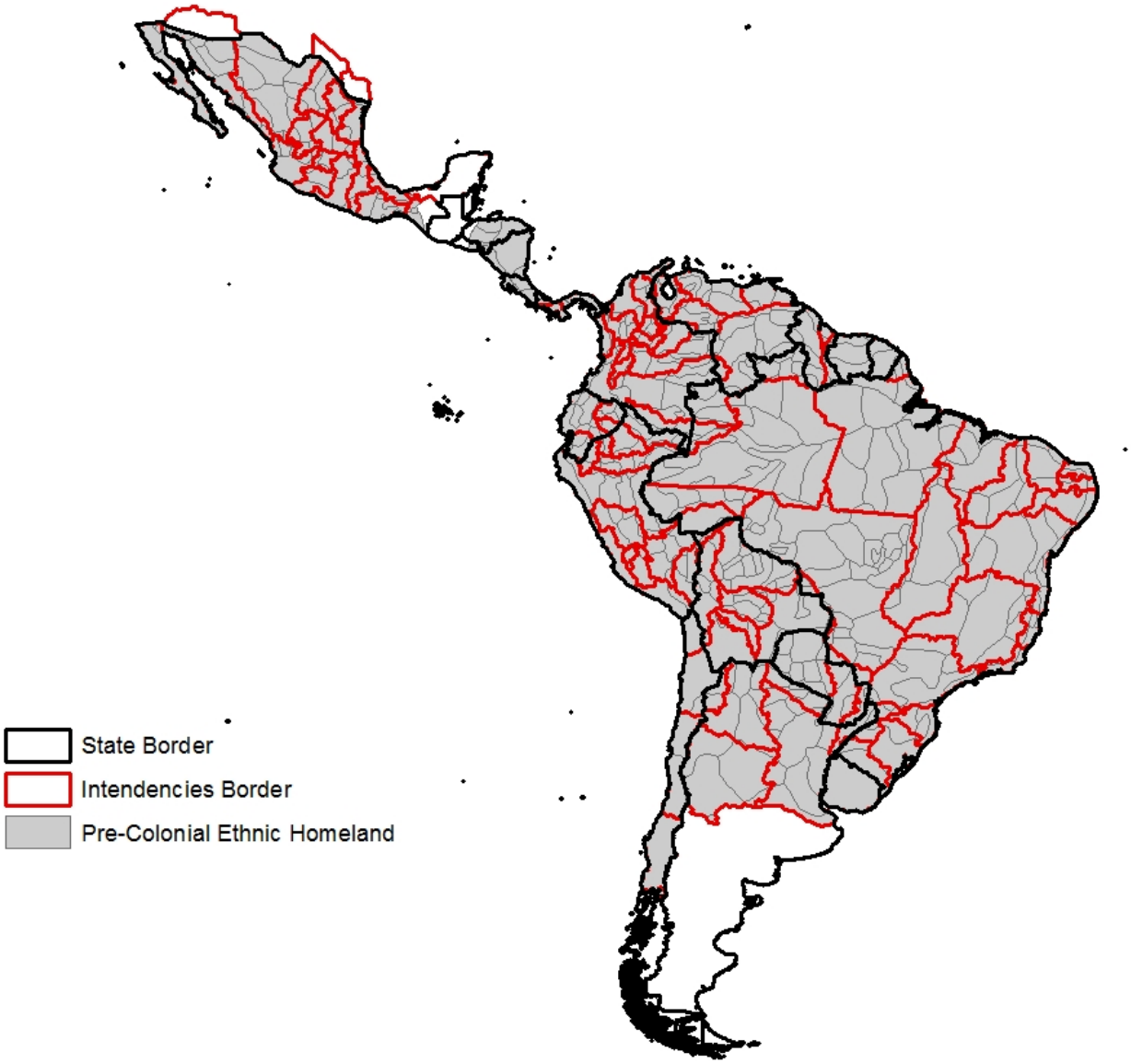


Figure 2. Ethnic Pre-Colonial Centralization in Latin America

Table 1: Pre-Colonial Ethnic Centralization and Regional Development

	Light per capita in 2012 (log)					
	Latin America Full Sample					
Jurisdictional Hierarchy	0.253	0.298	0.247	0.240	0.192	0.144
Double Cluster S.E.	(0.079)	(0.112)	(0.071)	(0.073)	(0.067)	(0.064)
Conley S.E.	[0.112]	[0.079]	[0.063]	[0.061]	[0.053]	[0.050]
R^2	0.058	0.230	0.542	0.565	0.629	0.679
N	486	486	486	486	486	486
	Spanish America Sample					
Jurisdictional Hierarchy	0.249	0.292	0.264	0.262	0.214	0.174
Double Cluster S.E.	(0.090)	(0.088)	(0.075)	(0.077)	(0.073)	(0.064)
Conley S.E.	[0.098]	[0.077]	[0.065]	[0.066]	[0.060]	[0.049]
R^2	0.071	0.248	0.473	0.502	0.579	0.648
N	336	336	336	336	336	336
Country FE	N	Y	N	Y	Y	Y
Intendency FE	N	N	Y	Y	Y	Y
Location Controls	N	N	N	N	Y	Y
Geographic Controls	N	N	N	N	N	Y
PopDens1492	N	N	N	N	N	Y

Notes. The Table reports OLS estimates relating regional development with pre-colonial jurisdictional hierarchy beyond the local community index as derived from Murdock 1951 and Murdock 1967. The unit of observation ethnic-country-intendency. The dependent variable is log(0.01+light per capita) at the ethnicity-country-intendency level. In columns 5-6 we control for location and geographic controls. The set of location controls include: log (1 + area under water), log (surface area), distance of the centroid of each ethnicity-country-intendency area from the respective capital city, the distance from the sea coast, distance from the major river, the distance from the national border. The set of geographic controls include: land suitability for agriculture, elevation, ruggedness, a malaria stability index, temperature, precipitation a diamond mine indicator, and an oil field indicator. In column 6 we control for log (pre-colonial population density).

The Data Appendix contains detailed variable definitions and data sources. Below the estimates we report in parentheses clustered standard errors at the country dimension. We report in brackets Conley?s (1999) standard errors that account for 2-dimensional spatial auto-correlation.

Table 2: The Long-Run Effects of Pre-colonial Political Complexity Conditioning on Other Pre-colonial Ethnic Features

	Dependent: Dependent: Light per capita in 2008-2012 (log)				
	Specification A		Specification B		
	Variable	Obs.	Variable	Jurisd Hierarchy	Obs.
Gathering	-0.242	311	0.102	0.458***	311
	(0.23)		(0.24)	(0.10)	
Hunting	0.0847	310	0.558	0.470***	310
	(0.49)		(0.56)	(0.08)	
Fishing	-0.769	310	-0.186	0.422***	310
	(0.62)		(0.53)	(0.12)	
Animal Husbandry	0.330*	312	0.0475	0.431***	312
	(0.16)		(0.14)	(0.11)	
Agricultural Dependence	0.114*	312	0.0471	0.418***	312
	(0.05)		(0.04)	(0.10)	
Polygyny	-0.161	308	0.0650	0.444***	308
	(0.31)		(0.31)	(0.10)	
Clan Communities (dummy)	-2.079***	266	-2.035***	0.310*	266
	(0.23)		(0.22)	(0.13)	
Agricultural Type	0.440***	292	0.276**	0.315**	292
	(0.10)		(0.10)	(0.12)	
Settlements	0.0700	292	0.0167	0.424***	292
	0.06		0.06	(0.11)	
Complex Settlement	0.162	291	-0.100	0.437***	291
	(0.32)		0.30	(0.11)	
Local Community	0.140	292	0.0569	0.431***	292
	0.25		0.26	0.12	
Local Community (dummy)	0.222	290	0.134	0.418***	290
	(0.24)		(0.26)	(0.12)	
Milking	-0.0939	292	-0.568	0.469***	292
	(0.76)		(0.76)	0.12	
Slavery	1.005***	263	0.757**	0.474***	263
	(0.24)		(0.24)	(0.14)	
Election (dummy)	-0.832	177	-0.386	0.765**	177
	(0.51)		(0.59)	(0.24)	

Notes. The Table shows the within-country within-intendency OLS estimates linking regional development with precolonial ethnic characteristics in Murdock 1967 Ethnographic Atlas. The dependent variable is the log (0.01 + light per capita) at the ethnic-intendency-country level. All specification include a set of country fixed effect(constants not reported) and a set of intendency fixed effect (constants not reported). In all specification A (in columns (1)-(2)) We regress the log(0.01 + light per capita) on various ethnic traits from Murdock 1967. In specification B (in columns (3)-(5)) We regress the log(0.01 + light per capita) on each Murdock's additional variables and the jurisdictional hierarchy beyond the local community index as derived from Murdock (1951). The Data Appendix contains detailed variable description and construction. Below the estimates we report in parenthesis Conley's corrected standard errors.

Table 3: Pre-Colonial Centralization and Sub-National Colonial Institutions

	Number of Cajas Reales				
	Spanish America Sample				
Jurisdictional Hierarchy	0.176***	0.176***	0.174**	0.174**	0.142***
Double Cluster S.E.	0.050	0.056	0.070	0.069	0.045
Conley S.E.	0.047	0.047	0.053	0.053	0.040
R^2	0.173	0.248	0.331	0.368	0.528
N	328	328	328	328	328
	Spanish America Sample w/ PopDens1492				
Jurisdictional Hierarchy	0.171***	0.160***	0.163**	0.159**	0.133***
Double Cluster S.E.	0.044	0.053	0.067	0.063	0.042
Conley S.E.	0.047	0.050	0.052	0.054	0.041
PopDens1492	0.013	0.043*	0.041	0.064	0.054
Double Cluster S.E.	0.020	0.024	0.054	0.062	0.075
Conley S.E.	0.024	0.028	0.042	0.045	0.047
R^2	0.174	0.255	0.334	0.375	0.531
N	328	328	328	328	328
Country FE	N	Y	N	Y	Y
Intendency FE	N	N	Y	Y	Y
PopDens1492	N	N	N	Y	Y
Location Controls	N	N	N	N	Y
Geographic Controls	N	N	N	N	Y

Channel: Effect of Caja Real on Contemporary Development					
	Light Density in 2012 (Log)				
Number of Cajas	0.502***	0.608***	0.647***	0.698***	0.507***
Double Cluster S.E.	0.117	0.120	0.151	0.160	0.161
Conley S.E.	0.110	0.088	0.116	0.110	0.092
Jurisdictional Hierarchy	0.257*	0.278**	0.235**	0.225*	0.174*
Double Cluster S.E.	0.144	0.138	0.113	0.127	0.094
Conley S.E.	0.130	0.114	0.100	0.106	0.063
R^2	0.129	0.296	0.502	0.531	0.555
N	328	328	328	328	328
Country FE	N	Y	N	Y	Y
Intendency FE	N	N	Y	Y	Y
Location Controls	N	N	N	N	Y
Geographic Controls	N	N	N	N	Y

7 Appendix A

	Population Density in 2000 (log)					
	Latin America Full Sample					
Jurisdictional Hierarchy	0.723***	0.664***	0.506***	0.494***	0.324***	0.270**
Double Cluster S.E.	(0.16)	(0.10)	(0.11)	(0.12)	(0.10)	(0.11)
Conley S.E.	(0.15)	(0.08)	(0.09)	(0.11)	(0.09)	(0.10)
R^2	0.139	0.299	0.599	0.619	0.659	0.758
N	478	478	478	478	478	478
	Spanish America Sample					
Jurisdictional Hierarchy	0.670***	0.643***	0.518***	0.520***	0.333***	0.284**
Double Cluster S.E.	(0.15)	(0.11)	(0.12)	(0.14)	(0.12)	(0.12)
Conley S.E.	(0.14)	(0.09)	(0.10)	(0.13)	(0.11)	(0.11)
R^2	0.156	0.352	0.570	0.599	0.658	0.764
N	328	328	328	328	328	328
Country FE	N	Y	N	Y	Y	Y
Intendency FE	N	N	Y	Y	Y	Y
PopDens1492	N	N	N	N	Y	Y
Location Controls	N	N	N	N	N	Y
Geographic Controls	N	N	N	N	N	Y

7.1 Data Construction and Sources

Variables at the ethnicity-country-intendency / Pixel level

Night Lights
Average population in 1x1 degree cell. Source: constructed as the 1x1 degree cell average of population across raster grids, computed using ArcGIS with data in EASE GLOBAL GRID projection, with data from the Center for International Earth Science Information Network - CIESIN - Columbia University, United Nations Food and Agriculture Programme - FAO, and Centro Internacional de Agricultura Tropical - CIAT. 2005. Gridded Population of the World, Version 3 (GPWv3): Population Count Grid. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://sedac.ciesin.columbia.edu/data/set/gpw-v3-population-count.

Ln Population Density
Natural logarithm of the average night lights intesity. Source: NOOA National Geophysical Data Centre for the year 2000.

Malaria Ecology
Average Malaria Ecology Index. Source: Malaria Ecology index from Kiszewski, Mellinger, Spielman, Malaney, Sachs, and Sachs 2004.

Average Temperature
Mean annual temperature (baseline period 1961-1990). Source: FAO/IIASA, 2011-2012. Global Agro-ecological Zones (GAEZ v3.0). FAO Rome, Italy and IIASA, Laxenburg, Austria.

Average Precipitation
Average monthly precipitation mm/month (baseline period 1961-1990). Source: CRU CL 2.0 data from New, Lister, Hulme, and Makin 2002.

Land Suitability
Average land suitability. Source: land suitability index from Ramankutty, Foley, Norman, and McSweeney 2002.

Mean Elevation
Average elevation. Source: mean elevation is constructed with data from National Oceanic and Atmospheric Administration (NOAA) and U.S. National Geophysical Data Center, TerrainBase, release 1.0 (CD-ROM), Boulder, Colo.

Ruggedness
Average ruggedness (Terrain Ruggedness Index, 100 m). Source: mean ruggedness is constructed with data from Terrain Ruggedness Index originally devised by Riley, DeGloria, and Elliot (1999), obtained through $http$: $//diegopuga.org/data/rugged/\#grid.$

Total Water Area
Total area occupied by water within the 1x1 degree cell. Source: constructed with ArcGIS by intersecting the 1x1 degree cell grid and the Digital Chart of the World inwater shapefile, by intersecting the 1x1 degree cell grid and the Digital Chart of the World oceans and sea shapefile. We sum up total in-cell water area and the areas of the cell occupied by seas and oceans, areas computed with data in EASE GLOBAL GRID projection.

Total Area
Total area of the 1x1 degree cell. Source: constructed with ArcGIS by intersecting the 1x1 degree cell grid and the World Language Mapping System shapefile from the Digital Chart of the World. We exclude cell parts not covered by World Language Mapping System data (and by the Africa Murdock Map, and the Murdock map for North and South America), areas computed with data in EASE GLOBAL GRID projection.

Ln Distance Coast
Distance to closest coast. Source: constructed with ArcGIS the digital Chart of the World coastline shapefile.

Ln Distance Border
Distance to closest border. Source: constructed with ArcGIS the digital Chart of the World boundaries shapefile.

Ln Distance Capital

Distance to the capital of the country where lies the centroid of the 1x1 degree cell. Source: constructed with ArcGIS the digital World Capital shapefile.

Ln Distance River
Distance to closest river. Source: constructed with ArcGIS using $Major_Rivers_World$ from www.naturalearth.com

8 Omline Appendix - Not intended for Publication

Table 5: Recoding of Jurisdictional Hierarchy Beyond Local Community Level

Ethnic Homeland	Jurisdictional Hierarchy	Narrative
Abipon	No Level	The tribe is organized into politically autonomous migratory bands
Aburra	Missing	No information
Acaxee		
Achagua	No Level	there is no political integration above the level of the local headman or clan
Acroa	Missing	No information
Aguano	Missing	No information
Alacaluf	Missing	No information
Amahuaca	No Level	Communities are politically autonomous
Amayane	Missing	No information
Apalai	Missing	No information
Apiaca	No Level	Villages are politically indipendent
Apinaye	No Level	Each village is politically autonomous
Arara	Missing	No information
Araucanians	Two Level	Communities, each with a headman, are politically organized into districts
Arawak	Missing	No information
Arawine	Missing	No information (like Yaruma)
Arikem	Missing	No information
Arua	Missing	No information
Ashluslay	Missing	No information
Atacama	No Level	They are clustered in villages under presumably autonomous headmen
Atsahuaca	Missing	No information
Auake	Missing	No information
Aueto	No Level	A headman with limited authority presides over an entire village, which is
Aweikoma	No Level	Autonomous migratory bands
Aymara	Three Level	Nation was divided into a number of states
Aztec		
Bacairi	No Level	No political integration beyond local level
Baniwa	Missing	No information
Barauna	Missing	No information
Barbacoa	Missing	No information
Betoi	No Level	villages are small, politically independent villages
Bororo	One Level	Paramount chief in at least one district (not in all)
Botucudo	No Level	Autonomous migratory bands
Cagaba	Missing	No information
Cahita		
Cahuapana	Missing	No information
Caingang	No Level	Local groups are politically autonomous

Table 6: Recoding of Jurisdictional Hierarchy Beyond Local Community Level

Ethnic Homeland	Jurisdictional Hierarchy	Narrative
Camacan	Missing	No information
Camaracoto	Missing	No information
Campa	No Level	Villages are politically autonomous
Canari	One Level	Local headmen are subject to the rulers of petty states with a complex offic
Canelo	Missing	No information
Canichana	Missing	No information
Caquetio	One Level	The communities of district are politically organized as petty state under a
Caraca	Two Level	Villages are politically organized into districts under paramount chiefs
Caraja	No Level	Villages are politically autonomous
Carib	No Level	Each community is politically autonomous under a headman
Carijona	Missing	No information
Cariri	No Level	Villages are politically autonomous
Cashinawa	One Level	Socially aggregated villages
Catio	Two Level	Political authority is exercized by local headmen and paramount chiefs of d
Catukina	Missing	No information
Cawahib	One Level	Paramount chiefs are present in some region (not all)
Cayamo	No Level	like Coroa
Cayapa	Missing	No information
Cayapo	Missing	No information
Cayuvava	Missing	No information
Cazcan		
Cenu	One Level	Large villages ruled by local chiefs and sub-chiefs, who commonly pay trib
Chake	Missing	No information
Chama	No Level	Each settlement is politically autonomous
Chandule	Missing	No information
Chane	Two Level	like Chiriguano
Chapacura	Missing	No information
Charrua	No Level	Politically autonomous migratory bands
Chibcha	Three Level	Chibcha are organized into five warring feudal states, each with a king who
Chinantec		
Chinato	No Level	lack political integration beyond local headmen with limited authority
Chinipa		
Chipanec		
Chiquito	No Level	Community is governed by an headman of great prestige assisted by lesser
Chiricaua		
Chiriguano	One Level	Communities are often aggregated under paramount chiefs (not always)
Choco	No Level	Communities are politically automnomous
Chono	Missing	No information

Table 7: Recoding of Jurisdictional Hierarchy Beyond Local Community Level

Ethnic Homeland	Jurisdictional Hierarchy	Narrative
Chorotega		
Choroti	No Level	There are nominal chiefs over larger groups, but they do not have real autho
Coahuilteco		
Cocama	No Level	People live in large politically autonomous villages
Cochimi		
Cofan	Missing	No information
Colorado	Missing	No information
Comechingon	No Level	Villages are politically autonomous
Concho		
Copopa		
Cora		
Coroa	No Level	The village is governed by two headman and a council, and is politically au
Cumana	Two Level	In addition to local headmen, there are paramount chiefs over small states
Cuna	Two Level	Endogamous villages, each under an elective headman, are organized into c
Diaguita	No Level	Politically autonomous head man
Diegueno		
Emerillon	Missing	No information
Encabellado	No Level	The people live in politically autonomous villages
Esmeralda	No Level	They live in politically autonomous villages
Fulnio	Missing	No information
Gauyupe	No Level	People live in politically autonomous villages
Gayon	Missing	No information
Goajira	No Level	politically autonomous migratory bands
Guachi	Missing	No information
Guachichil		
Guahibo	Missing	No information
Guaitaca	Missing	No information
Guaja	Missing	No information
Guamo	Missing	No information
Guamontey	Missing	No information
Guana	No Level	Each village is politically autonomous
Guarani	No Level	Politically autonomous communities
Guarayu	Missing	No information
Guasave		
Guato	Missing	No information
Guayaki	No Level	Autonomous migratory bands
Guaymi		
Gueren	No Level	like Botucudo

Table 8: Recoding of Jurisdictional Hierarchy Beyond Local Community Level

Ethnic Homeland	Jurisdictional Hierarchy	Narrative
Huarpe	Missing	No information
Huave		
Huaxtec		
Huichol		
Inca	Four Level	Political integration is the most complex in native America
Ipurina	No Level	Communities are politically autonomous
Itonama	Missing	No information
Janambre		
Jeico	Missing	No information
Jicaque		
Jirajara	Two Level	Paramount chiefs rule over districts or even, at least nominally, over entire t
Jivaro	No Level	community with an indipendent family head
Jumano		
Kamia		
Kitemoca	No Level	like Chiquito
Lagunero		
Lama	Missing	No information
Leco	Missing	No information
Lenca		
Lipan		
Maca	Missing	No information
Macu	Missing	No information
Macurap	Missing	No information
Macusi	No Level	Each settlement has a headman
Manao	One Level	The authority of chiefs apparently extends in some instances beyond the loc
Manta	Two Level	Paramount chiefs rule over a number of [..] settlements
Masco	Missing	No information
Mascoi	No Level	Local groups are politically autonomous
Mashacali	Missing	No information
Mataco	No Level	Each band is politically autonomous under a local headman
Maue	Missing	No information
Mayoruna	Missing	No information
Mbaya	Two Level	Politically aggregated under a paramount chief
Mixtec		
Mocovi	No Level	No information
Mojo	Missing	No information
Mosetene	No Level	Politically autonomous settlements
Mosquito		

Table 9: Recoding of Jurisdictional Hierarchy Beyond Local Community Level

Ethnic Homeland	Jurisdictional Hierarchy	Narrative
Motilon	Missing	No information
Movima	Missing	No information
Mundurucu	Missing	No information
Mura	Missing	No information
Nambicuara	No Level	Each village is politically autonomous
Nevome		
Omagua	Missing	No information
Omaguaca	Two Level	They (villages) are aggregated under paramount chiefs
Ona	No Level	true chiefship is absent
Opata		
Opaye	Missing	No information
Otomac	No Level	Extreme local autonomy prevails
Otomi		
Otuke	No Level	like Chiquito
Pacaguara	Missing	No information
Paez	Two Level	Paramount chiefs rule over districts composed of a number of such disperse
Palikur	Missing	No information
Pame		
Pantangoro	No Level	Political integration does not extend beyond the local community
Papago		
Paressi	One Level	Instances of Paramount chief (not always)
Pasto	No Level	Dwellings aggregated into compact villages, each politically autonomous u
Patagon	Missing	No information
Patasho	Missing	No information
Paumary	No Level	Settlements are politically autonomous
Paya		
Payagua	Missing	No information
Pilaga	Missing	No information
Pima		
Piro	No Level	Politically independent settlements
Puelche	Missing	No information
Puinave	Missing	No information
Puri	No Level	Autonomous migratory bands
Quimbaya	Two Level	Villages are organized into districts under paramount chiefs
Quito	One Level	Paramount chiefs rule over a number of such settlements (not all)
Rama		
Rukuyen	No Level	Each community is politically autonomous under a headman
Saliva	Missing	No information

Table 10: Recoding of Jurisdictional Hierarchy Beyond Local Community Level

Ethnic Homeland	Jurisdictional Hierarchy	Narrative
Saraveca	No Level	like Chiquito
Seri		
Shacriaba	One Level	like Sherente
Sherente	One Level	Neighborhood villages are loosely organized under a council composed of
Shiriana	No Level	Local headmen exercise considerable authority
Siriono	Missing	No information
Suya	No Level	Each village is politically autonomous
Tacana	No Level	Each [community] is ruled by an headman
Tahue		
Talamanca		
Tamanac	Missing	No information
Tamaulipeco		
Tapiete	Missing	No information
Tarahumara		
Tarairu	Two Level	Villages are apparently politically integrated under a paramount chiefs
Tarasco		
Taripare	Missing	No information
Taulipang	No Level	Villages are political autonomous
Tehuelche	Missing	No information
Tempqui	Missing	No information
Tenetehara	No Level	Villages are politically autonomous
Tepehuan		
Tequistlateco		
Terembe	Missing	No information
Timba	Two Level	Paramount chiefs rule over a number of villages
Timbira	No Level	Each village enjoy political autonomy
Timbu	Missing	No information
Timote	Two Level	Villages are politically organized into districts under paramount chiefs
Toba	No Level	Politically autonomous migratory bands
Tonocote	Missing	No information
Totonac		
Totorame		
Trumai	No Level	Villages are political autonomous
Tucano	No Level	People live in politically autonomous settlements
Tucuna	No Level	Each community is politically independent
Tunebo	Three Level	similar to Chibcha
Tupi	One Level	apparently share culture of Tupinamba
Tupinamba	One Level	One chief often (not always) exercises paramount authority over a number

Table 11: Recoding of Jurisdictional Hierarchy Beyond Local Community Level

Ethnic Homeland	Jurisdictional Hierarchy	Narrative
Uru	Missing	No information
Vilela	Missing	No information
Waicuri		
Waiwai	Missing	No information
Wapishana	No Level	Each community is politically autonomous under a headman
Waraicu	Missing	No information
Warrau	No Level	Each village has a headman
Waura	No Level	Villages are political autonomous
Witoto	Missing	No information
Yabuti	Missing	No information
Yagua	No Level	People live in politically autonomous settlements
Yahgan	Missing	No information
Yamamdi	No Level	Each village has a headman and is politically autonomous
Yaruma	Missing	No information
Yaruro	Missing	No information
Yecuana	No Level	Politically autonomous under a headman
Yuma	Missing	No information
Yuma		
Yunca	Three Level	[..] Large towns. Yunca were organized in a poweful feudal state
Yuracare	No Level	politically autonomous settlements
Yurimagua	Missing	No information
Yuruna	No Level	Each community is politically independent
Zacalteco		
Zamuco	One Level	Chamoco (at least one tribe not all) keeps slaves and are politically organiz
Zaparo	Missing	No information
Zapotec		
Zoque		

Notes. The Table shows the coding procedure for the Jurisdictional Hierarchy Beyond the Local Community Level for Central and South America (excluding Mexi
the codebook from Gray(1999).

Table 12: Caja Reales

Caja Real	Virreinato	Latitude	Longitude	Source
Acapulco	Nueva Espana	16,85000038	-99,8666687	
Antioquia	Nueva Granada	6,559721947	-75,82805634	
Arequipa	Peru	-16,39882278	-71,53688049	
Arica	Upper Peru	-18,48333359	-70,33333588	
Arispe	Nueva Espana	30,33083344	-110,1691666	
Bolanos	Nueva Espana	21,83055687	-103,7805557	
Buenos Aires	Rio de la Plata	-34,60833359	-58,37194443	
Cailloma	Peru	-15,63669395	-71,60227203	
Campeche	Nueva Espana	18,83638954	-90,40333557	
Carabaya	Peru	-14,06950092	-70,4312439	
Caracas	Nueva Granada	10,5	-66,91666412	
Caragnas	Upper Peru	-18,45000076	-67,44999695	
Cartagena	Nueva Granada	10,39999962	-75,5	
Cartago	Nueva Granada	4,814278126	-75,69455719	
Castrovirreyna	Peru	-12,78583336	-74,97277832	
Catamarca	Rio de la Plata	-28,46666718	-65,78333282	
Chachapoyas	Peru	-6,216667175	-77,84999847	
Charcas	Upper Peru	-19,04999924	-65,25	
Chiloe	Chile	-42,60610962	-73,80502319	
Chiuahua	Nueva Espana	28,6352787	-106,0888901	
Chucuito	Chile	-16,21439552	-69,4573822	
Citara	Nueva Granada	5,692276955	-76,6581955	
Cochabamba	Upper Peru	-17,39361191	-66,15694427	
Concepcion	Peru	-11,91847801	-75,3128891	
Cordoba de Tucuman	Rio de la Plata	-31,41666794	-64,18333435	
Coro		11,4	-69,683333	
Corrientes	Rio de la Plata	-27,46666718	-58,81666565	
Cuenca(Loja)	Ecuador	-2,883300066	-78,98329926	
Cumana	Nueva Granada	10,44999981	-64,16666412	
Cuzco	Peru	-13,52614117	-71,97130585	
Durango	Nueva Espana	24,01666641	-104,6666641	
Giron	Nueva Granada	7,070833206	-73,17305756	
Guadalajara	Nueva Espana	20,66666794	-103,3499985	
Guanajuato	Nueva Espana	21,0177784	-101,2566681	
Guayana	Nueva Granada	8,373100281	-62,64360046	
Guayaquil	Ecuador	-2,18333292	-79,8833313	
Hacha	Nueva Granada	11,54416656	-72,90694427	
Honda	Nueva Granada	5,204166889	-74,7416687	
Huancavelica	Peru	-12,78583336	-74,97277832	
Humanga	Peru	-13,17000008	-74,22000122	
Jaen de Bracamoras	Ecuador	-5,699999809	-78,80000305	
Jauja	Peru	-11,77499962	-75,5	
Jujui	Rio de la Plata	-24,20000076	-65,30000305	
La Guaira		10,6	-66,933056	

Notes. The Table shows the coding procedure for the Jurisdictional Hierarchy Beyond the Local Community Level for Central and South America (excluding Mexico). Narratives are extracted from Murdock (1951). The coding is taken from the codebook from Gray(1999).

Caja Real	Virreinato	Latitude	Longitude	Source
La Paz	Upper Peru	-16,50222206	-68,16555786	
La Rioja	Rio de la Plata	-29,43333244	-66,84999847	
Lima	Peru	-12,03499985	-77,01860809	
Maldonado	Rio de la Plata	-34,90000153	-54,95000076	
Mantucana	Peru	-11,84472179	-76,3861084	
Maracaibo	Nueva Granada	10,64999962	-71,6166687	
Mendoza	Chile	-32,8833313	-68,81666565	
Merida	Nueva Espana	20,96999931	-89,62000275	
Mexico	Nueva Espana	19,43333244	-99,1333313	
Michoacan	Nueva Espana	19,16861153	-101,8997192	
Mompox	Nueva Granada	9,241944313	-74,42666626	
Montevideo	Rio de la Plata	-34,88360977	-56,1819458	
Neiva	Nueva Granada	2,92750001	-75,28749847	
Novita	Nueva Granada	4,956110954	-76,60610962	
Oaxaca	Nueva Espana	17,08333206	-96,75	
Ocana	Nueva Granada	8,233332634	-73,34999847	
Oruro	Upper Peru	-17,96666718	-67,1166687	
Pachuca	Nueva Espana	20,12166786	-98,73583221	
Pamplona	Nueva Granada	7,378056049	-72,65249634	
Panama	Nueva Granada	8,966667175	-79,53333282	
Paraguay	Rio de la Plata	-25,29638863	-57,64138794	
Piura y Paita	Peru	-5,092735767	-81,10197449	
Popayan	Nueva Granada	2,444166899	-75,6213913	
Portobelo	Nueva Granada	9,553889275	-79,65583038	
Potosi	Upper Peru	-19,58333206	-65,75	
Presidio del Carmen	Nueva Espana	20,62750053	-87,08110809	
Puebla de los Angeles	Nueva Espana	19,04527855	-98,19750214	
Puerto Cabello		10,466667	-68,016667	
Puno	Peru	-15,84333324	-70,02361298	
Quito	Ecuador	-0,216700003	-78,5	
Remedios	Nueva Granada	7,030832767	-74,53333282	
Rosario	Nueva Espana	30,06027794	-115,726944	
Salta	Rio de la Plata	-24,78333282	-65,41666412	
Saltillo	Nueva Espana	25,43333244	-101	
San Juan	Rio de la Plata	-31,53722191	-68,52527618	
San Luis de Potosi	Nueva Espana	22,14972115	-100,9749985	
Sana	Peru	-7,085278034	-79,71611023	
Santa Cruz de la Sierra	Upper Peru	-17,79999924	-63,16666794	
Santa Fe	Nueva Granada	4,611700058	-74,0759964	
Santa Fe de Veracruz	Rio de la Plata	-31,63333321	-60,70000076	
Santa Marta	Nueva Granada	11,24722195	-74,20166779	
Santiago de Chile	Chile	-33,43783188	-70,65032959	
Santiago del Estero	Rio de la Plata	-27,79999924	-64,26667023	

Notes. The Table shows the coding procedure for the Jurisdictional Hierarchy Beyond the Local Community Level for Central and South America (excluding Mexico). Narratives are extracted from Murdock (1951). The coding is taken from the codebook from Gray(1999).

Caja Real	Virreinato	Latitude	Longitude	Source
Sombrerete	Nueva Espana	23,63299942	-103,6500015	
Tabasco	Nueva Espana	17,97222137	-92,58889008	
Trujillo	Peru	-8,016667366	-79	
Tucuman	Rio de la Plata	-26,83333206	-65,19999695	
Valdivia	Chile	-39,80833435	-73,2416687	
Veracruz	Nueva Espana	19,1902771	-96,15333557	
Vico y Pasco	Peru	-10,68360043	-76,26519775	
Zacatecas	Nueva Espana	22,77166748	-102,5752792	
Zimapan	Nueva Espana	20,73333359	-99,3833313	

Notes. The Table shows the coding procedure for the Jurisdictional Hierarchy Beyond the Local Community Level for Central and South America (excluding Mexico). Narratives are extracted from Murdock (1951). The coding is taken from the codebook from Gray(1999).

Bite and Divide: Malaria and Ethnic Diversity

Matteo Cervellati*, Giorgio Chiovelli†, Elena Esposito‡§

June 5, 2014

Abstract

In this paper we argue that malaria exposure represented a fundamental determinant of modern ethno-linguistic diversity. We conjecture that in highly malarial areas the necessity to adapt and develop immunities specific to the local disease environment historically reduced mobility and increased isolation, thus leading to the formation of a higher number of different ethno-linguistic groups. We conduct the first part of the analysis at a disaggregated level by creating a grid of artificial countries of 1x1 degree of size (around 110 square km at the equator) to employ as units of observation. We use a rich set of new data on ethno-linguistic diversity constructed from geolocalized maps of ethno-linguistic groups around the world at several points in history, as well as new data on historical malaria endemicity. Results point to a strong positive correlation between historical malaria endemicity and the number of ethno-linguistic groups at all levels of spatial disaggregation. In the second part of the exercise, we explore the micro-channels behind the above empirical regularity. We hypothesize that the increased isolation caused by malaria and the need to exploit and preserve location-specific immunities (i) strengthened ethnic identity and (ii) increased the propensity to marry within the group. Therefore, we employ georeferenced indivual data from the third wave of the Afrobarometer survey and show that malaria exposure is positively correlated with the strenght of ethnic identity. Moreover, we exploit modern data on marriage patterns in 22 African countries retrieved from the Demographic and Health Survey. Regressions' results show that endogamous marriages are more frequent in areas with higher geographic suitability to malaria.

Keywords: Ethno-linguistic Diversity, Diseases, Geography, Long-Run Growth.
JEL Classification: J12, J15, O43, Z13.

*University of Bologna and IZA
†University of Bologna
‡University of Bologna
§*Contact details:* University of Bologna, Department of Economics, Piazza Scaravilli 2, Bologna, ITALY.

1 Introduction

A vast and consolidated stream of economic literature has been documenting the economic consequences of ethno-linguistic diversity [among many others] (Easterly and Levine, 1997; Alesina and Ferrara, 1999, 2004; Fearon and Laitin, 2003). More recent contributions, and in particular Michalopoulos (2008) and Ashraf and Galor (2013), make the point that ethno-linguistic diversity is in itself a consequence of more fundamental features of the geographical environment, showing that geo-physical characteristics, such as the heterogeneity in land suitability and the migratory distance from the cradle of humankind in East Africa, are associated with higher contemporary ethnic diversity today.

In this paper we document that an additional important geographical feature - geographic suitability to malaria - explains part of the currently observed ethno-linguistic diversity across and within modern countries. We hypothesize that in highly malarial areas the necessity to adapt and preserve immunities specific to the local disease environment increased isolation, reduced mobility and population admixture with neighboring groups, with the consequence of producing and preserving a larger number of different ethno-linguistic groups. From one hand, as we know from historical narratives, the danger associated with moving into diseased environments has been historically an obstacle to trade and commerce. Ramen (2002) mentions the practice of "silent trade" - i.e. trade at a distance between caravans from North Africa and population at the south of the Sahara - devised to prevent the traders from exchanging dangerous germs on top of the goods. It is well known that malaria represented one of the most lethal diseases in history and, for this reason, it likely constituted a major obstacle to trade and, more generally, to population movements. As a matter or fact, the Boer trekkers faced high morbidity and mortality while adventuring into northern tropical areas, having in the end to resettle in more temperate regions (Becker, 1985). Malaria represented a barrier between places where malaria was endemic and places where it was absent, but also between highly malarial areas and equally malarial neighboring areas. Diamond and Ford (2000) argued that "tropical Africans were combating malaria with more than just antibodies... by living in relatively small communities, spread out over vast areas...[to] limit the level of malaria transmission". On top of this, it is important to mention that malaria exists in various strains and that a whole set of location-specific strains were discovered, against which only strain-specific resistance can offer some form of protection. Curtin (1968, 1998) showed that African troops had higher mortality rates in foreign African countries than in their home-countries, possibly a consequence of different strains of malaria encountered in foreign African countries. Historian William McNeill (1976) argued that in the Indian sub-continent the heterogeneity in epidemiological endowments between intrusive Aryans and local "forest folks" prevented the local primitive communities from being "digested" into the invaders' civilization, remaining semi-autonomous and separated. Malaria, for instance, protected for centuries the isolation of Tharu people of Nepal (Brower and Johnston, 2007). After centuries of residence in malaria-infested regions, Tharu people developed significant genetic resistance to malaria, to the point that they faced an about sevenfold lower

malaria incidence than synpatric non-Tharu people (Modiano, Morpurgo, Terrenato, Novelletto, Di Rienzo, Colombo, Purpura, Mariani, Santachiara-Benerecetti, Brega, et al., 1991). Malaria immunities allowed Tharu people to live undisturbed from neighboring powerful civilization, as they were the only group able to survive in those infested lands. Strict endogamy practiced by the Tharu confined these traits to this indigenous group and preserved this location-specific advantage.

In this work, we provide empirical evidence of the relationship linking malaria incidence with the historical and contemporaneous number of ethno-linguistic groups observed across and within countries. In order to do so we exploit the framework devised by Michalopoulos (2008) and conduct the analysis at a disaggregated level, by superimposing over modern countries a grid of artificial squares, measuring 1x1 degree of size. By doing so, we are able to concentrate on within-country variation of malaria incidence and diversity, as cross-country variation in diversity could hide key historical state-formation processes as unobservable confounders. One particularly challenging obstacle faced by the literature exploring the consequences of disease exposure lies in the search for exogenous measures of disease incidence, as it is well known that disease prevalence is highly influenced by living standards (diets, housing condition, public health, agricultural practices) and by patterns of population density and isolation. For these reasons, our attempt is to exploit only the exogenous component of historical malaria incidence, exploiting two reasonably exogenous disaggregated indexes. The first index we use is the malaria stability index devised by Kiszewski, Mellinger, Spielman, Malaney, Sachs, and Sachs (2004), which aims at measuring the force and stability of malaria transmission, based on the biological characteristics of the regionally dominant vector mosquitoes, such as their propensity to feed on humans and their daily survival, and how these features interact with the climatic environment. The second index is a recently available historical measure of malaria endemicity measured at the beginning of the twentieth century, produced by Lysenko (1968) and recently digitalized by (Hay S.I., 2004). In our view, the index has to be preferred to modern measures of clinical malaria incidence, insofar as it measures the degree of malaria endemicity in a time where the main massive malaria eradication campaigns had still to be conceived and realized.

We depart from previous work exploring the relationship between disease and diversity along several dimensions[1]. First, we concentrate on the role played by malaria alone, based on the fact that malaria has likely been the most destructive disease in human history. Malaria is known as the "strongest known selective pressure in the recent history of the human genome" Kwiatkowski (2005), and just as much as it affected the evolution of the human genome, we argue it molded the social structure and habits of societies residing in the affected areas. We do not exclude that malaria acted in combination with other relevant human diseases, however,

[1]A previous empirical work by (Cashdan, 2001) highlighted a correlation between ethnic diversity and pathogens load, she concentrated on six diseases: leishmanias (three species), trypanosomes (two species), malaria (four species), schistosomes (three species), filariae (two species), spirochetes (two species and one genus), and leprosy. Birchenall (2010) exploits the same of set of diseases and a similar framework to investigate the long term effect of diseases on development

we maintain that malaria was the major driving force, even if its effects may have been worsened by the interaction with other harmful co-existing diseases. Secondly, we highlight that the relation between malaria incidence and diversity exists at different points in history, today and before colonization. Exploiting the fact that malaria was not present in the New World before colonization, we show that the relationship between geographic suitability to malaria and ethno-linguistic diversity is absent in pre-colonial Americas. This exercise, which we perform as a sort of "placebo test", mitigates the concern that our index of geographic suitability to malaria is capturing the effect on diversity of climatic and geographic characteristics spuriously related with malaria suitability. In order to test this, we exploit Murdock (1959) data on the distribution of ethnic groups in pre-colonial North and South America, recently digitized by Chiovelli (2014).

Our main contribution to the literature exploring the fundamental determinants of ethno-linguistic diversity is to propose and explicitly test two of the possible channels that link malaria and ethno-linguistic diversity. Firstly, we investigate the link between malaria exposure and the saliency of ethnic identity. Using data from the third wave of the Afrobarometer, a large-scale survey assessing prevalent cultural and political individual attitudes, from 16 countries in Sub-Saharan Africa. Respondents are asked to declare how strongly they identify with their national country versus their ethnic group. Our results show that in places more suitable for malaria respondents tend to identify more with their ethnic group rather than with their nationality, the relation holds when controlling for a large set of geographical and individual controls. In the second part of the exercise, we hypothesize that in highly malarial environment the need to exploit and preserve location-specific immunities reduced the practice of marrying away from home and from former kinship ties, thus increasing the propensity to arrange endogamous marriages. We test this hypothesis using contemporary data on marriage patterns in Africa from the Demographic and Health Survey. We exploit several waves of the survey conducted in 1815 clusters across 22 African countries, whenever information on both the ethic identity of the wife and the husband was available. Since information on ethnicity varies along waves and countries - as for some waves very detailed information on ethnic identity is provided whereas for other waves more aggregated ethnic families are reported - we associate all reported ethnic identities to the various branches of the Ethnologue tree. In this way, we are able to compute the propensity to marry within the ethnic family at various level of ethnic family aggregation. We show that, at intermediate levels of ethnic disaggregation, individuals living in areas with a higher suitability to malaria have higher propensity to marry somebody from their same group. Results hold when controlling for individual level of education, age and whether the individual is a urban or rural dweller.

2 Malaria Exposure and Ethno-Linguistic Diversity

Our first attempt is to document the existence of a macro correlation strong and sizable in magnitude between ethnic diversity and malaria exposure.

2.1 Data and Empirical Strategy

Empirical Strategy We follow the framework devised by Michalopoulos (2008) and create a grid of cells of 1x1 degree of size, corresponding to about 110 km at the equator. By doing so, we create a set of squared "artificial countries" which we exploit as unit of observation. This approach has two main advantages: from one hand, the small size of cells allow us to measure more accurately geographical and environmental features, variables such as elevation, average precipitation and temperature are more accurately measured this way than in the aggregated form of a standard cross-country regressions; ii) from another hand, by looking at small artificial cells we can address an additional source of variability, i.e. within country variability. In fact, following this framework, we can easily account for common country effects and get rid the potential confounding factors which act at the country level, such as institutional or cultural features. This approach has also a set of shortcomings, the most compelling one being that many variables for which we would like to control for, such as GDP or average human capital, are not available at such a finer spatial resolution. However, it is also true that all these controls would be endogenous to our variable of interests and could not be included into the regression without fears of biased estimates.

We estimate the relationship between the natural logarithm of the number of linguistic groups in the 1x1 degree cell and the incidence of malaria. Our baseline specification follows:

$$Ln(NumberofGroups)_{i,c} = \beta_0 + \beta_1 Malaria_{i,c} + \beta_2 \mathbf{X_{i,c}} + \mu_{\mathbf{c}} + \epsilon_{\mathbf{i,c}}$$

where i indicates the cell, and c the country. $\mathbf{X_{i,c}}$ includes a vast set of climatic, geographical and location controls, μ_c stands for country fixed effects. Based on our conceptual framework, we hypothesize further that the effect of malaria on linguistic diversity could be higher in places with a high heterogeneity in the stock of acquired and innate immunities. We expect this heterogeneity to be higher in places with a higher variation in elevation since, because of its reliance on mosquitoes for transmission, malaria cannot be transmitted in highlands even whenever present in the neighboring lowlands. Because of this precise reason, we might find neighboring areas where people living in the lowlands developed strong immunities to malaria and people in the highland did not. This heterogeneity in immunity endowment could in principle represent a barrier to the admixing of these groups. We test this prediction by looking at the interaction between malaria incidence with a measure of the standard deviation of the elevation in the cell $Mal_{i,c} * STDElev_{i,c}$:

$$Ln(NofGroups)_{i,c} = \beta_0 + \beta_1 Mal_{i,c} + \beta_2 STDElev_{i,c} + \beta_3 Mal_{i,c} * STDElev_{i,c} + \beta_2 \mathbf{X_{i,c}} + \mu_{\mathbf{c}} + \epsilon_{\mathbf{i,c}}$$

The Data The World Language Mapping System database offers, to our knowledge, the most comprehensive mapping of the world's known living languages[2]. We compute the number of languages spoken in all cells of our dataset, excluding languages spoken by less than 10000 people overall. Figure 1 shows how the number of languages per cell varies across the world. Note that in about half of the cells composing our sample only one language is spoken, while in one fourth of the cell 3 or more languages are currently in use.

In order to measure malaria exposure, we rely on the work of Kiszewski, Mellinger, Spielman, Malaney, Sachs, and Sachs (2004), which constructed their index as follows: they associated to each country a dominant vector of *Anopheles* mosquitoes (for countries with different dominant mosquitoes, mosquitoes were association to sub-regions). A monthly index of stability was then computed as a parametric function of the share of blood meal taken by the mosquito, the daily survival rate and the extrinsic incubation period. Once this regional monthly index (constructed for about 260 regions in the world) was created, in order to obtain a finer data resolution, a mimimum lagged threshold of precipitation (10 mm) was exploited as a pre-condition for malaria transmission. The yearly aggregation of such an index is the malaria stability index which we exploit in our analysis.

The second measure of malaria incidence that we use is a historical one. It was produced by Lysenko (1968) and recently digitized by (Hay S.I., 2004), it aims at measuring the level of malaria endemicity at the beginning of the 20th century. Endemicity is defined as the parasite rate (PR) in the 2-10 year age cohort [3]. The index takes value 0 wherever malaria is absent, 1 for epidemic areas, 2 where malaria is hypoendemic, 3 for mesoendemic areas, 4 for hyperendemic and 5 for holoendemic areas. As previously mentioned, the level of malaria endemicity could be endogenous to several factors such as agricultural activity and population density, however, insofar as it is measured at the begininning of the 20th it precedes the timing of the massive malaria eradication campaings which took place after the IIWW. It is reasonable to consider than before the era of massive eradication campaing, the incidence of malaria was more tightly related to climatic factors than it is today.

As a robustness, we exploit what we consider three additional proxies of historical malaria incidence: three blood related genetic variants which have been shown to be strictly associated with historical malaria incidence. HbS stands for sickle haemoglobin allele frequency in 2010, G6PD for allele frequency for G6PD deficiency in 2010, and Duffy for Duffy negative phenotype again measured in 2010.

[2]Note that we exploit a definition of ethnicity based on language, this choice follows both theoretical and empirical consideration.

[3]Hypoendemic with PR lower than 0.1; mesoendemic with PR between 0.11-0.5; hyperendemic for 0.51-0.75 for the holoendemic class (PR higher than 0.75), the PR refers to the 1-year age group

2.2 Results

Table 1 reports baseline results. An increase in malaria incidence is associated with a sizable and robust increase in the number of languages spoken in a cell, see Column 1-5 in Table 1. In terms of magnitude, going from a cell with no malaria to a cell which was historically holoendemic increases the average number of language spoken in the cell by almost one fourth. We obtain a similar estimate when we look at the results of the regression with the malaria ecology index of stability in transmission, in this case going from a place with 0 malaria stability to a place with an average stability of 34, increases the number of groups by 0.23%, see Column 6-10 in Table 1. Results are confirmed when looking at the effect of genetic immunities, Table 2 and when looking at the number of pre-colonial ethnic group as reported by Murdock (1959), see Table 4. Consistent with our predictions, the correlation between the indexes of malaria and the number of ethno-linguistc groups is not present when looking at the distribution of ethnic groups of pre-colonial Americas, since the disease was not present in the New World before colonization. Finally, there is an extra effect of malaria incidence on ethnic diversity in areas with a larger than average heterogeneity in elevation (Table 3).

3 Looking for the Channel

In the previous section we documented that areas more suitable for malaria present higher levels of ethno-linguistic diversity. In this section we explore two micro-channels linking malaria exposure to contemporaneous levels of ethno-linguistic diversity.

3.1 Malaria Exposure and Ethnic Saliency

We first focus on the saliency of ethnic lines in defining individual identity. Ethnic traits are crucial technlogy in aggregating preferences and beliefs. Our hypothesis is that the degree to which an individual identify himself as belonging to a particular ethnic group should be positive related to the exposure of malaria.

Data In order to test this hypothesis we employ data from the third wave of Afrobarometer. Afrobarometer is an indivual-level survey covering 16 Sub-Saharan African in 2005. The surveys are conducted on a random sample ranging from 1200 to 2400 individuals in each district in each country. The survey records information on the ethnic identity of each respondent, georeferenced information on the location of the respondent, individual information and a vast set of questions exploring cultural and political attitudes. Thanks to this setting, for each individual in the survey we know the exact location in which the interview was performed. This allows us to merge the GPS information in the Afrobarometer with all the geographic and ecological variables employed in the previous analysis.

Table 1: Malaria and Ethno-Linguistic Diversity

Dependent Variable	Ln(Number of Languages in the Cell)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Malaria Endemicity	0.200***	0.162***	0.090***	0.052***	0.051***					
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)					
Malaria Ecology						0.045***	0.039***	0.023***	0.007*	0.008**
						(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Variation Suitability		0.784**	0.433*	0.651***	0.626***		0.611*	0.315	0.636***	0.600**
		(0.33)	(0.23)	(0.23)	(0.22)		(0.33)	(0.22)	(0.24)	(0.23)
Variation Elevation		0.000***	0.000***	0.000***	0.000***		0.000***	0.001***	0.000***	0.000***
		(0.00)	(0.00)	(0.00)	(0.00)		(0.00)	(0.00)	(0.00)	(0.00)
Geographic Controls	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Cell Area and Water Area	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Location and Distances	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Population and Night Lights	No	No	No	No	Yes	No	No	No	No	Yes
Country FE	No	No	No	Yes	Yes	No	No	No	Yes	Yes
N	9450	9450	9450	9450	9450	9450	9450	9450	9450	9450
R-Squared	0.187	0.289	0.391	0.549	0.549	0.163	0.295	0.393	0.546	0.547

The dependent variable is the natural logarithm of the number of Ethnologue linguistic groups in the cell. The "Geographic Controls" include absolute latitude, mean elevation, average soil suitability, average terrain ruggedness, average precipitation, average temperature. The "Location and Distances" controls includes the natural logarithm of the distance to the country capital, to the coast, to the country border, to the closest river and to Adis Ababa, the number of countries in the cell and a dummy variable taking value 1 for cells that completely belong to a single country. OLS estimates. The unit of observation is a 1 x 1 degree cell. Standard errors are clustered at the country level. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

Table 2: Genetic Malaria Immunities and Ethno-Linguistic Diversity

Dependent Variable	Ln(Number of Languages in the Cell)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Hbs	7.785***	-0.254	-0.292						
	(1.26)	(0.99)	(0.99)						
G6PD				3.750***	1.675**	1.633**			
				(1.10)	(0.77)	(0.77)			
Duffy							0.484***	0.618***	0.590***
							(0.13)	(0.18)	(0.19)
Geographic Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Cell Area and Water Area	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Location and Distances	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Population and Night Lights	No	No	Yes	No	No	Yes	No	Yes	Yes
Country FE	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
N	9368	9368	9368	5599	5599	5599	5643	5643	5643
R-Squared	0.093	0.546	0.546	0.071	0.532	0.533	0.068	0.530	0.531

The dependent variable is natural logarithm of the number of Ethnologue linguistic groups in the cell. The "Geographic Controls" include absolute latitude, mean elevation, standard deviation of elevation, average soil suitability and standard deviation of soil suitability, average terrain ruggedness, average precipitation, average temperature. The "Location and Distances" controls includes the natural logarithm of the distance to the country capital, to the coast, to the country border, to the closest river and to Adis Ababa, the number of countries in the cell and a dummy variable taking value 1 for cells that completely belong to a single country. OLS estimates. The unit of observation is a 1 x 1 degree cell. Standard errors are clustered at the country level. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

Table 3: Malaria, Variation in Elevation and Ethno-Linguistic Diversity

Dependent Variable	Ln(Number of Languages in the Cell)			
	(1)	(2)	(3)	(4)
Malaria Endemicity	0.016		0.017	
	(0.01)		(0.01)	
Malaria Endemicity x Variation Elevation	0.000***		0.000***	
	(0.00)		(0.00)	
Malaria Ecology		0.004		0.005
		(0.00)		(0.00)
Malaria Ecology x Variation Elevation		0.000**		0.000*
		(0.00)		(0.00)
Variation Elevation	0.000	0.000***	0.000	0.000***
	(0.00)	(0.00)	(0.00)	(0.00)
Geographic Controls	Yes	Yes	Yes	Yes
Cell Area and Water Area	Yes	Yes	Yes	Yes
Location and Distances	Yes	Yes	Yes	Yes
Population and Night Lights	Yes	Yes	Yes	Yes
Country FE	No	No	Yes	Yes
N	9450	9450	9450	9450
R-Squared	0.552	0.548	0.553	0.548

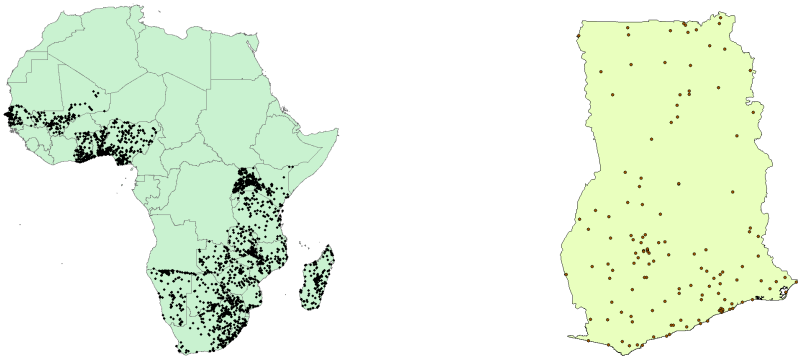
The dependent variable is natural logarithm of the number of Ethnologue linguistic groups in the cell. The "Geographic Controls" include absolute latitude, mean elevation, average soil suitability, average terrain ruggedness, average precipitation, average temperature. The "Location and Distances" controls includes the natural logarithm of the distance to the country capital, to the coast, to the country border, to the closest river and to Adis Ababa, the number of countries in the cell and a dummy variable taking value 1 for cells that completely belong to a single country. OLS estimates. The unit of observation is a 1 x 1 degree cell. Standard errors are clustered at the country level. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

Table 4: Malaria and Pre-Colonial Ethno-Linguistic Diversity

Dependent Variable	PANEL A Ln(Number of Languages in Pre-Colonial Africa)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Malaria Endemicity	0.115***	0.061***	0.043**	0.055**	0.049**					
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)					
Malaria Ecology						0.017***	0.017***	0.010**	0.006**	0.006*
						(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dependent Variable	PANEL B Ln(Number of Languages in Pre-Colonial America)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Malaria Endemicity	0.051	0.005	-0.015	-0.010	-0.012					
	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)					
Malaria Ecology						0.034*	0.022*	0.014	0.012	0.007
						(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Geographic Controls	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Cell Area and Water Area	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Location and Distances	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Location and Modern Country	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Country FE	No	No	No	Yes	Yes	No	No	No	Yes	Yes
N	1972	1972	1972	1972	1972	1972	1972	1972	1972	1972
R-Squared	0.103	0.166	0.223	0.322	0.326	0.091	0.186	0.227	0.318	0.323

The dependent variable in Panel A is the natural logarithm of the number of linguistic groups in the cell in Pre-Colonial Africa, while in Panel B is the natural logarithm of the number of linguistic groups in the cell in Pre-Colonial America. The "Geographic Controls" include absolute latitude, mean elevation, standard deviatin in elevation, average soil suitability, standard deviation in soil suitability, average terrain ruggedness, average precipitation, average temperature. The "Location and Distances" controls includes the natural logarithm of the distance to the country capital, to the coast, to the closest river and to Adis Ababa. The "Location and Modern Countries" controls include the number of countries in the cell, a dummy variable taking value 1 for cells that completely belong to a single country and distance to the country border. OLS estimates. The unit of observation is a 1 x 1 degree cell. Standard errors are clustered at the country level. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

In order to proxy for the strength and salience of ethnic identity, we focus on the question *Let us suppose that you had to choose betweeb being a [Ghanaian/Kenyan/etc.] and being a [respondent's identity group] and (respondent's group). Which of these two do you feel most strongly attached to?* (q82). The variable takes value from one (begin only identified with your ethnic group) to five (begin only identified with your country nationality).



(a) Across Africa (b) Within Ghana

Figure 1: Afrobarometer Cluters

Empirical Strategy Our unit of observation is the individual respondent in location z, district d and country c. In order to test whether ethnic identity is more salient in place with higher malaria exposure, we exploit within-country - across Afrobarometer clusters - variability using the following specification:

$$EthnicSaliency_{i,z,d,c} = \beta_0 + \beta_1 Malaria_{z,d,c} + \beta_2 \mathbf{X}_{z,d,c} + \mathbf{Z}_{i,z,d,c} + \mu_c + \epsilon_{z,d,c}$$

where $EthnicSaliency_{i,z,d,c}$ is an ordered variable taking value from one (ethnic identity only) to five (country identity only). $Mal_{z,d,c}$ is a measure of malaria suitability in respondent location z in district d in country c. $X_{z,d,c}$ is a vector of time invariant geographical controls (elevation (level and standard deviation), suitability of agriculture, temperature, precipitation. $Z_{i,z,d,c}$ is a vector of individual characteristics (urban resident, education, age), μ_c is a country dummy.

Results Table 5 summarizes main results. The coefficient of malaria ecology is significantly negative: going from a cluster with no malaria to a cluster with the maximum level of malaria increases the propensity to identify with her own ethnic group by a bit less than 1 point. Importantly, the coefficient estimate does not change when including a vast set of geographic controls. Interestingly, the correlation holds when exploiting variation within ethnic group. In other words, conditional on the ethnic group the respondent belong to, he/she identifies more

Table 5: Malaria and Ethnic Saliency

Dependent Variable	Identifies with Country vs Ethnicity				
	(1)	(2)	(3)	(4)	(5)
Malaria Ecology	-0.019***	-0.017***	-0.017***	-0.016***	-0.016***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Geographic Controls	No	Yes	Yes	Yes	Yes
Location and Distances	No	Yes	Yes	Yes	Yes
Humidity and TseTse Suitability	No	No	Yes	Yes	Yes
Individual Controls	No	No	No	Yes	Yes
Share of own group in District	No	No	No	Yes	Yes
Size of own group in District	No	No	No	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Ethnicity fixed effects	No	No	No	Yes	Yes
N	20137	19901	18329	18329	18329
R-Squared	0.113	0.112	0.123	0.145	0.147

The dependent variable measures the strength of national identity vs. ethnic identity, it goes from 1 (only ethnic identity) to 5 (only national identity). Malaria Ecology measures the force and transmission of malaria in the location of the Afrobarometer respondent. The "Geographic Controls" include mean elevation, average soil suitability, average terrain ruggedness, average precipitation and average temperature. The "Location and Distances" controls includes the natural logarithm of the distance to the country capital and to the coast. OLS estimates. The unit of observation is the Afrobarometer respondent. Standard errors are clustered at the country level. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

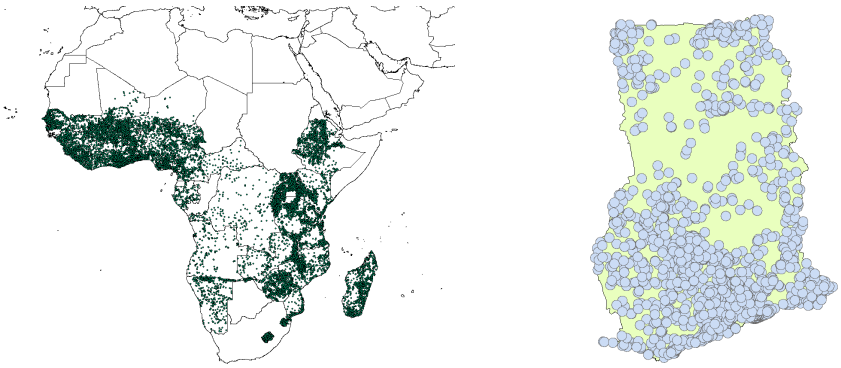
with the group (and less with the country) in clusters with more malaria.

3.2 Malaria Exposure and Endogamous Marriages

The fact the malaria exposure increases ethno-lingustic heterogeneity is consistent with the idea of ethno-linguistic diversity being affected by human interactions. Along with fertility and mortality rates, human behaviors - such as migration and ethnic mixing - may alter the distribution of ethnic groups in a given area. In a partial equilibrium world, endogamous ethnic marriage is a crucial factor for the persistence of a given ethnic group in a given location. Moreover, historical narratives (McNeill, 1975; Browner and Johnston, 2007) and genetic studies (Kwaiatkowski, 2005) point out that heterogeneity in epidemiological endowments is conducive to the development of location-specific immunities. These two facts coupled together imply that ethnic groups historically exposed to malaria may have a comparative advantage in terms of differential mortality with respect to groups that were not exposed to malaria, or to the same indigenous strains. Our hypothesis is that ethnic endogamicity may be an optimal individual behavioral response to malaria exposure. If marriage decisions are driven by the observed differential mortality among ethnic groups, marrying within the group could maximize the survival probability of the couple and of their descendants. This response is optimal in the sense that it reduces the risk of losing the ethnic-specific immunities of the couple.

Data In order to test this hypothesis we use contemporary georeferenced data on marriage patterns in Africa provided by the Demographic and Health Survey (DHS). DHS surveys are

constructed to provide detailed information on a sub-national representative sample of women of age ranging from 15 to 49. We exploit several waves of the survey conducted in 1815 cluster across 22 African countries, see Figure 2. In most of the waves, data on married couples are provided. In this exercise we use those waves for which information on both the ethnic identity of the wife and the husband is available.



(a) Across Africa (b) Within Ghana

Figure 2: DHS Clusters

The exercise of identification of endogamous marriages is complicated by the fact that the definition of ethnicity is not homogenous and varies along waves and countries. More precisely, for the same country we might have a wave where it is specified the ethnicity x of individual W_i, while in the following wave individual W_j is reported to belong to ethnicity $x1$, which actually represents a sub-family of ethnicity x. In fact, in some cases the ethno-lingustic level reported is a dialect, while in other cases we have only official national languages. This is problematic to the extent to which the variability of ethno-linguistic diversity across waves and across country is just artificially generated by the different levels of details and precision that characterized the process of DHS data collection. To mitigate this issue, our strategy is to associate all ethnic groups reported in the DHS to the various branches of the Ethnologue trees (Gordon and Grimes, 2005) (see 3). Following this procedure, we are able to assign each ethno-linguistic entry in the DHS to the respective ethno-linguistc family in the Ethnologue. For every spouse in the sample, we are therefore able to see whether he/she formed a couple with an individual belonging to her ethnic group at various level of ethno-linguistic family aggregation.

From the DHS data, we extract individual - both for female and male - socio-economic (religion, urban residence, education) and democraphic (age) characteristics and the size of ethnic group population in the region. Moreover, we exploit the spatial disaggregation of the DHS data in order to construct our geographical controls. Notice that, in order to preserve respondents confindenciality, DHS uses random displacement in the georeferencing process of different clusters in each country. Urban clusters are displaced with a minimum of 0 and a maximum of 2 kilometers of error. Rural clusters are instead displaced with an error of a minimum of 0

and a maximum of 5 kilometers of positional error with a further 1% of the rural clusters displaced a minimum of 0 and a maximum of 10 kilometers (DHS website) [4]. In order to minimize this displacement issue, we construct in ArcGIS a buffer of 10 km radius around the coordinates of each cluster. In such a way, we are able to use the area inside this circle to construct cluster-level measure of malaria suitability, elevation, suitability of agriculture, temperature and precipitation. Since the area of the buffer contains the "true" location of each cluster, we think this strategy represents a reasonable solution to the random displacement adopted in DHS data.

Empirical Strategy Our unit of observation is the individual respondent in cluster c in country z for each avalaible wave t. In order to test whether there is a behavioral endogamous respone to malaria exposure, we are going to exploit the within-country across cluster variability using the following specification:

$$Endogamy_{i,c,z,t} = \beta_0 + \beta_1 Mal_{c,z} + \beta_2 \mathbf{X}_{c,z} + \mathbf{Z}_{i,c,z} + \mu_z + \mu_t + \epsilon_{i,c,z,t}$$

where $Endogamy_{i,c,z,t}$ is a dummy variable taking value one if the indivual is married with somebody belonging to the same ethnic group in cluster c in country z in wave t. $Mal_{c,z}$ is a measure of malaria suitability in cluster c in country z. $X_{c,z}$ is a vector of time invariant geographical controls (elevation (level and standard deviation), suitability of agriculture, temperature, precipitation. $Z_{i,c,z}$ is a vector of individual characteristics (urban resident, education, age). μ_z is a country dummy, μ_e is a female ethnic dummy, and μ_t is a wave dummy.

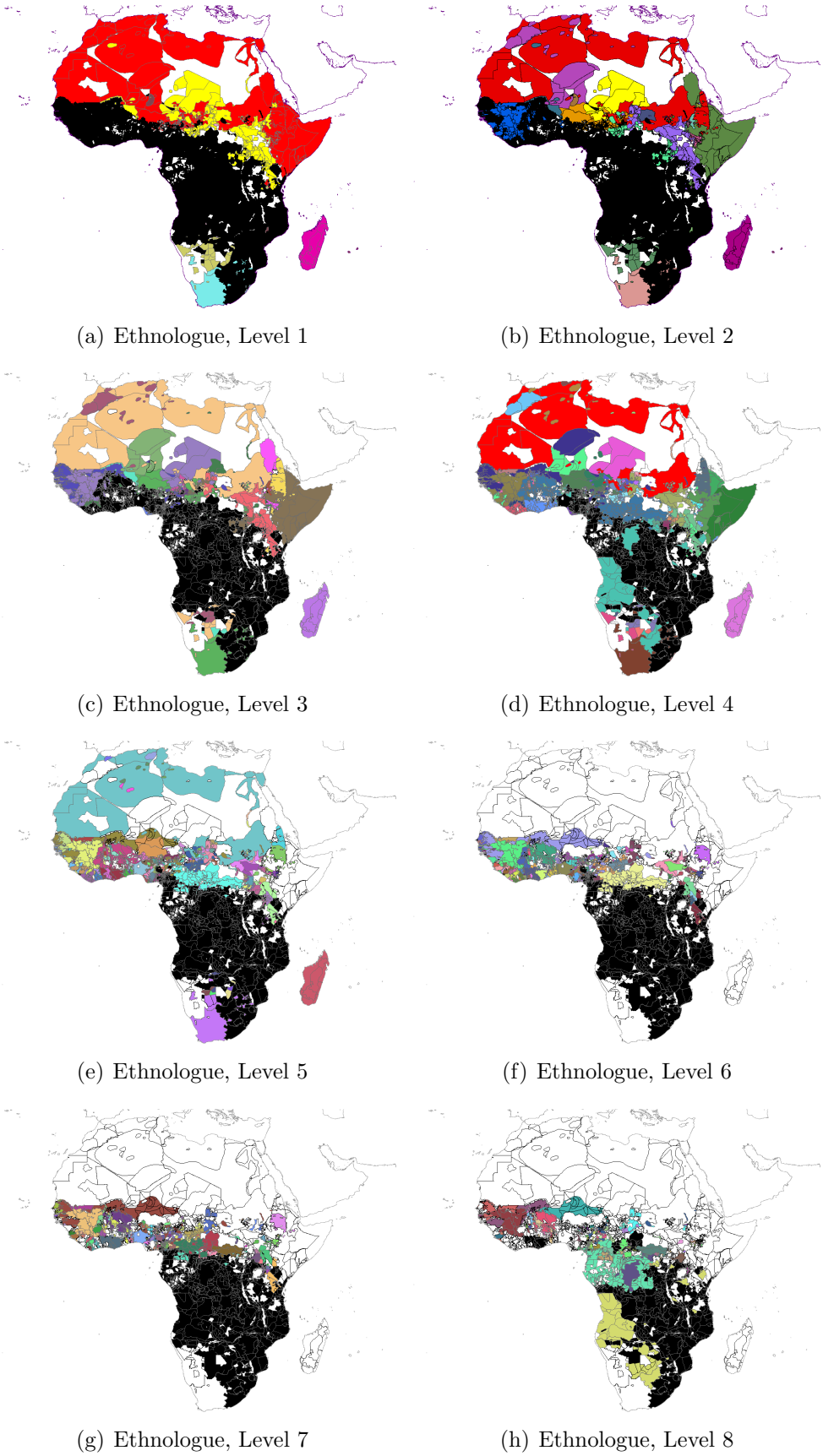
3.3 Results

Results are presented in different tables for levels of ethnolinguist familiy aggregation of Ethnologue Level 2 to 7. Results for each Ethnologue Level are reported in Table 6 to 7. Our baseline specification is the one exploiting within-country across cluster variation in the relationship between endogamous marriages and malaria exposure. Results show a robust correlation between endogamous marriages and malaria exposure when we look at at intermediate levels of ethnic disaggregation. Within a country, individuals living in areas with a higher suitability to malaria have higher propensity to marry somebody of the same group. The result holds when controlling for individual level of education, age and whether the individual is a urban or rural dwellers.

[4]Notice that there is no issue in the assignment of cluster near to country borders. In fact, "displacement is restricted so that the points stay within the country and within the DHS survey region (DHS website). Morevoer, "In surveys released since 2009 the displacement is restricted to the country's second administrative level where possible." (DHS website)

4 Conclusion

In this paper we hypothesize that malaria is a historical determinant of contemporary ethnolinguistic diversity. In order to test this hypothesis, we show that higher historical malaria incidence and higher geographic suitability to malaria are both associated to higher linguistic diversity. We propose a channel linking malaria incidence and diversity, we conjecture that malaria increased the propensity to marry within the groups. We test this hypothesis by looking at marriage patterns within 22 African countries, we show that geographic suitability to malaria is associated with a higher rate of endogamous marriages.



(a) Ethnologue, Level 1

(b) Ethnologue, Level 2

(c) Ethnologue, Level 3

(d) Ethnologue, Level 4

(e) Ethnologue, Level 5

(f) Ethnologue, Level 6

(g) Ethnologue, Level 7

(h) Ethnologue, Level 8

Figure 3: Ethnologue Language Groups, by Level of Disaggregation.

Table 6: Malaria and Endogamous Marriages

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	Endogamous Marriage - Ethnologue Level 2							
Malaria Endemicity Buf 10km	0.003	0.005*	0.000	-0.002				
	(0.00)	(0.00)	(0.00)	(0.00)				
Malaria Ecology Buf 10km					0.001	0.002**	0.002	0.001
					(0.00)	(0.00)	(0.00)	(0.00)
N	76888	76871	73887	73887	76888	76871	73887	73887
R-Squared	0.037	0.041	0.064	0.077	0.038	0.042	0.064	0.077
Dependent Variable	Endogamous Marriage - Ethnologue Level 3							
Malaria Endemicity Buf 10km	0.002	0.003	-0.001	-0.004				
	(0.00)	(0.00)	(0.00)	(0.00)				
Malaria Ecology Buf 10km					0.002***	0.003***	0.002*	0.001
					(0.00)	(0.00)	(0.00)	(0.00)
N	75049	75032	72050	72050	75049	75032	72050	72050
R-Squared	0.044	0.046	0.070	0.091	0.045	0.048	0.071	0.091
Dependent Variable	Endogamous Marriage - Ethnologue Level 4							
Malaria Endemicity Buf 10km	0.003	0.005*	0.001	-0.004				
	(0.00)	(0.00)	(0.00)	(0.00)				
Malaria Ecology Buf 10km					0.002*	0.003**	0.002	0.001
					(0.00)	(0.00)	(0.00)	(0.00)
N	73125	73108	70163	70163	73125	73108	70163	70163
R-Squared	0.044	0.046	0.069	0.092	0.045	0.048	0.070	0.092
Geographic Controls	No	Yes	Yes	Yes	Yes	No	Yes	Yes
Individual Control	No	Yes	Yes	Yes	Yes	No	Yes	Yes
Group Size in the Region	No	No	Yes	Yes	Yes	No	No	Yes
Ethnic Group fixed effects	No	No	No	Yes	No	No	No	Yes
Religion fixed effects	No	No	Yes	Yes	No	No	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

The dependent variable is a dummy taking value 1 for women who married somebody from the same linguistic group defined at various Ethnologue levels. Malaria Ecology and Malaria Endemicity are the average malaria incidence measured in the 10 km radius from the centroid of the cluster. The "Geographic Controls" include average suitability, average temperature and precipitation, average and standard deviation of elevation in a radius of 10 km from the cluster centroid. The "Individual Controls" include a dummy for urban residence, age and highest education of the wife and of the husband. OLS estimates. The unit of observation is a 1 x 1 degree cell. Standard errors are clustered at the country level. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

Table 7: Malaria and Endogamous Marriages

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	Endogamous Marriage - Ethnologue Level 5							
Malaria Endemicity Buf 10km	0.000	0.003	-0.002	-0.003				
	(0.00)	(0.00)	(0.00)	(0.00)				
Malaria Ecology Buf 10km					0.003***	0.004***	0.004***	0.003**
					(0.00)	(0.00)	(0.00)	(0.00)
N	66107	66090	63164	63164	66107	66090	63164	63164
R-Squared	0.055	0.057	0.076	0.103	0.058	0.060	0.079	0.104
Dependent Variable	Endogamous Marriage - Ethnologue Level 6							
Malaria Endemicity Buf 10km	0.004	0.006	0.000	-0.001				
	(0.00)	(0.00)	(0.00)	(0.00)				
Malaria Ecology Buf 10km					0.004***	0.005***	0.003**	0.002
					(0.00)	(0.00)	(0.00)	(0.00)
N	58790	58773	55886	55886	58790	58773	55886	55886
r2	0.078	0.082	0.114	0.141	0.083	0.086	0.116	0.142
Dependent Variable	Endogamous Marriage - Ethnologue Level 7							
Malaria Endemicity Buf 10km	0.001	0.003	-0.002	-0.001				
	(0.01)	(0.01)	(0.00)	(0.00)				
Malaria Ecology Buf 10km					0.003*	0.003**	0.002	0.002
					(0.00)	(0.00)	(0.00)	(0.00)
N	51647	51630	49325	49325	51647	51630	49325	49325
R-squared	0.064	0.066	0.092	0.113	0.066	0.068	0.093	0.114
Geographic Controls	No	Yes	Yes	Yes	Yes	No	Yes	Yes
Individual Control	No	Yes	Yes	Yes	Yes	No	Yes	Yes
Group Size in the Region	No	No	Yes	Yes	Yes	No	No	Yes
Ethnic Group fixed effects	No	No	No	Yes	No	No	No	Yes
Religion fixed effects	No	No	Yes	Yes	No	No	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

The dependent variable is a dummy taking value 1 for women who married somebody from the same linguistic group defined at various Ethnologue levels. Malaria Ecology and Malaria Endemicity are the average malaria incidence measured in the 10 km radius from the centroid of the cluster. The "Geographic Controls" include average suitability, average temperature and precipitation, average and standard deviation of elevation in a radius of 10 km from the cluster centroid. The "Individual Controls" include a dummy for urban residence, age and highest education of the wife and of the husband. OLS estimates. The unit of observation is a 1 x 1 degree cell. Standard errors are clustered at the country level. ***, **, * indicate significance at 1-, 5-, and 10-% level, respectively.

References

Alesina, A., and E. Ferrara (1999): "Participation in heterogeneous communities," Discussion paper, National Bureau of Economic Research.

——— (2004): "Ethnic diversity and economic performance," Discussion paper, National Bureau of Economic Research.

Ashraf, Q., and O. Galor (2011): "The" Out of Africa" Hypothesis, Human Genetic Diversity, and Comparative Economic Development," Discussion paper, National Bureau of Economic Research.

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

Becker, P. (1985): *The pathfinders: the saga of exploration in southern Africa*. Viking.

Birchenall, J. (2010): "Disease and diversity in Africa's Long-term Economic Development," .

Brower, B., and B. R. Johnston (2007): *Disappearing peoples?: indigenous groups and ethnic minorities in South and Central Asia*. Left Coast Press.

Cashdan, E. (2001): "Ethnic diversity and its environmental determinants: effects of climate, pathogens, and habitat diversity," *American Anthropologist*, 103(4), 968–991.

Curtin, P. (1968): "Epidemiology and the slave trade," *Political Science Quarterly*, pp. 190–216.

Curtin, P. D. (1998): *Disease and empire: The health of European Troops in the Conquest of Africa*. Cambridge University Press.

Diamond, J., and L. Ford (2000): "Guns, germs, and steel: the fates of human societies," *Perspectives in Biology and Medicine*, 43(4), 609.

Easterly, W., and R. Levine (1997): "Africa's growth tragedy: policies and ethnic divisions," *The Quarterly Journal of Economics*, 112(4), 1203.

Fearon, J., and D. Laitin (2003): "Ethnicity, insurgency, and civil war," *American Political Science Review*, 97(1), 75–90.

Gordon, R. G., and B. F. Grimes (2005): *Ethnologue: Languages of the world*, vol. 15. SIL international Dallas, TX.

Hay S.I., Guerra C.A., T. A. N. A. . S. R. (2004): "The global distribution and population at risk of malaria: past, present and future," *Lancet Infectious Diseases*, 4(6), 327–336.

Kiszewski, A., A. Mellinger, A. Spielman, P. Malaney, S. E. Sachs, and J. Sachs (2004): "A global index representing the stability of malaria transmission," *American Journal of Tropical Medicine and Hygiene*, 70(5), 486–498.

Kwiatkowski, D. P. (2005): "How malaria has affected the human genome and what human genetics can teach us about malaria," *The American Journal of Human Genetics*, 77(2), 171–192.

Lysenko, A.J., S. I. (1968): "Geography of malaria. Amedico-geographic profile of an ancient disease," *Medicinskaja Geografija*, p. 25146.

McNeill, W. (1976): "Plagues and Peoples. 1976," *Garden City, NY: Anchor P.*

Michalopoulos, S. (2008): "The Origins of Ethnolinguistic Diversity: Theory and Evidence," MPRA Paper 11531, University Library of Munich, Germany.

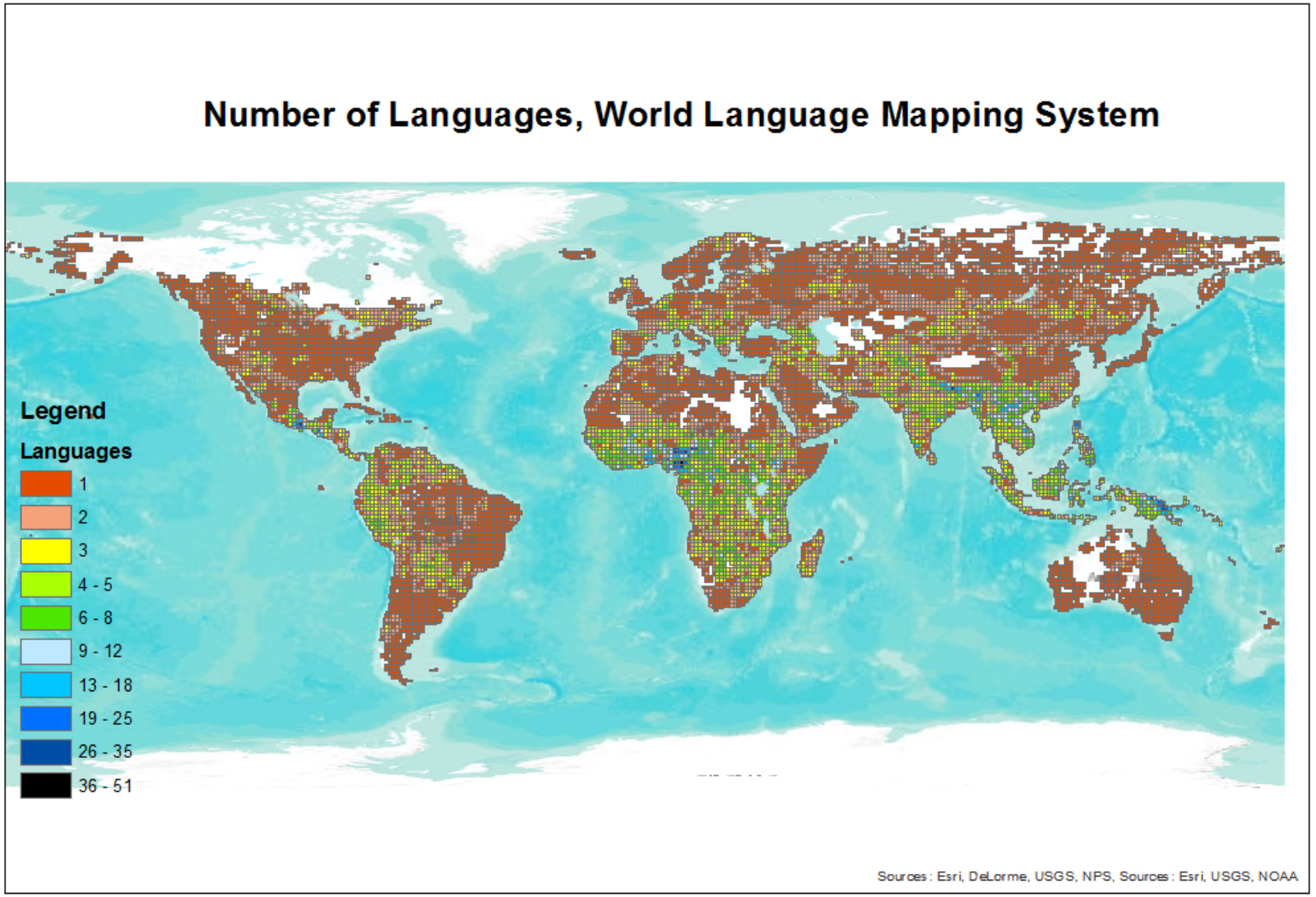
Modiano, G., G. Morpurgo, L. Terrenato, A. Novelletto, A. Di Rienzo, B. Colombo, M. Purpura, M. Mariani, S. Santachiara-Benerecetti, A. Brega, et al. (1991): "Protection against malaria morbidity: near-fixation of the α-thalassemia gene in a Nepalese population," *American journal of human genetics*, 48(2), 390.

Ramen, F. (2002): *Sleeping Sickness and Other Parasitic Tropical Diseases*. The Rosen Publishing Group.

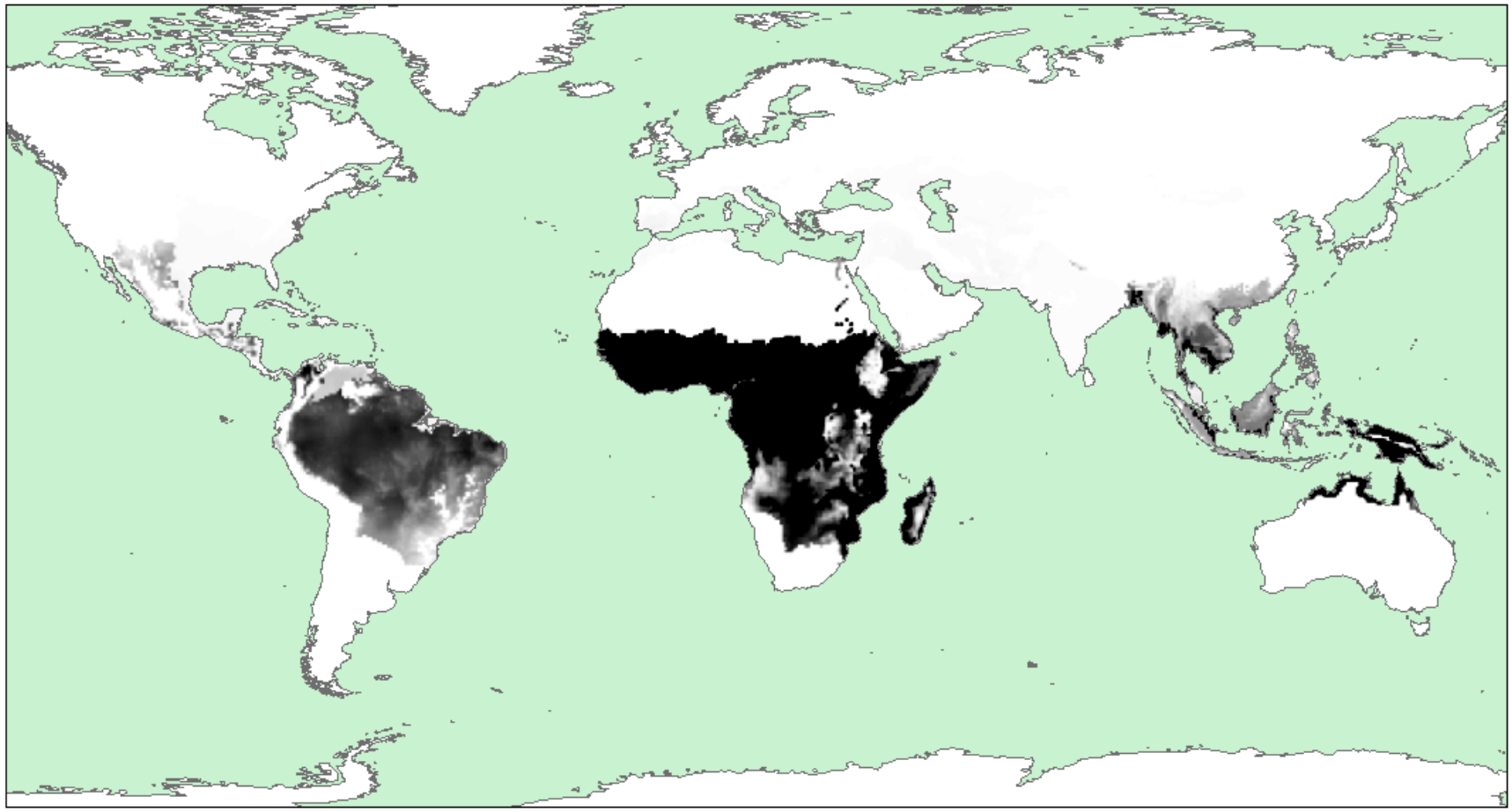
5 Appendix

Table 8: Summary statistics

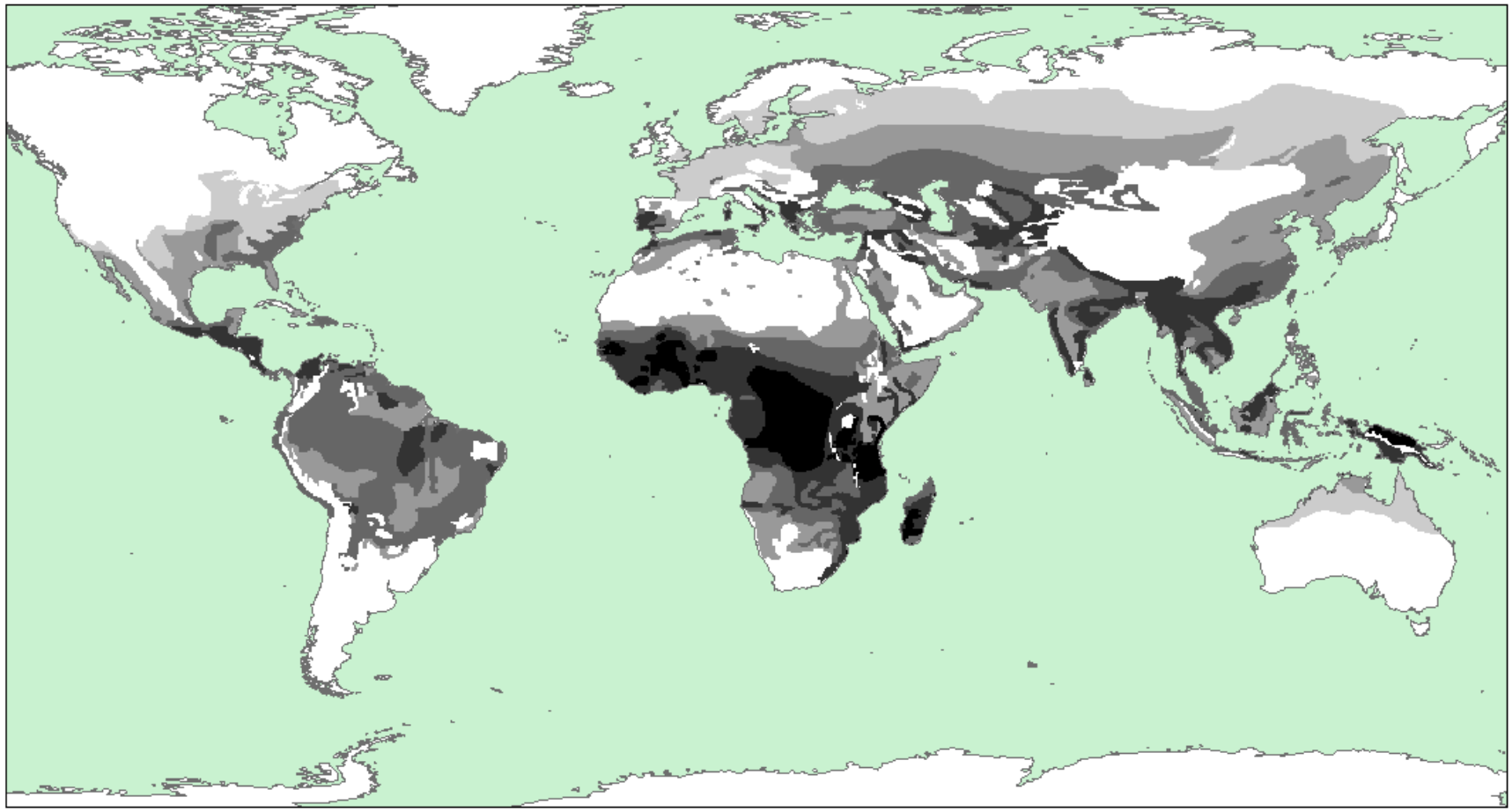
Variable	Mean	Std. Dev.	Min.	Max.	N
Ln(Number of Language)	0.533	0.65	0	3.932	12762
Ln(Number of Language, Africa)	0.735	0.554	0	2.565	2662
Ln(Number of Language, Americas)	0.427	0.465	0	2.197	4637
Malaria Endemicity	1.31	1.477	0	5	15683
Malaria Ecology	1.809	4.943	0	34.728	16522
Average Temperature	8.081	15.203	-30.331	30.386	17263
Average Precipitation	60.697	59.003	0	640.889	17272
Mean Suitability	0.27	0.308	0	0.999	16498
Suitability	0.032	0.047	0	0.409	16498
Mean Elevation	624.722	792.77	-720.643	5725.512	17213
Variation Elevation	137.56	163.493	0	1868.89	17213
Ruggedness	85098.69	112044.254	30.148	1016771.313	17272
Total Water	0.513	0.877	0	9.146	17272
Total Area	8.072	3.31	1.001	12.308	12762
Number of Country	1.148	0.402	1	5	17272
Within Country	0.869	0.337	0	1	17272
Ln(Migratory Distance)	9.15	0.79	4.358	10.247	17263
Ln(Distance Coast)	11.746	2.055	0.357	14.546	17272
Ln(Distance Border)	11.035	1.83	-1.409	13.854	17272
Ln(Distance River)	12.653	1.456	3.875	15.878	17272
Ln(Distance Adis Ababa)	14.004	1.231	8.247	17.001	17170
Absolute Latitude	40.81	22.502	0.25	83.417	17272
Night Lights	3.604	3.169	0	49.196	16385
Ln(Population)	2.534	1.959	-2.197	9.037	10297

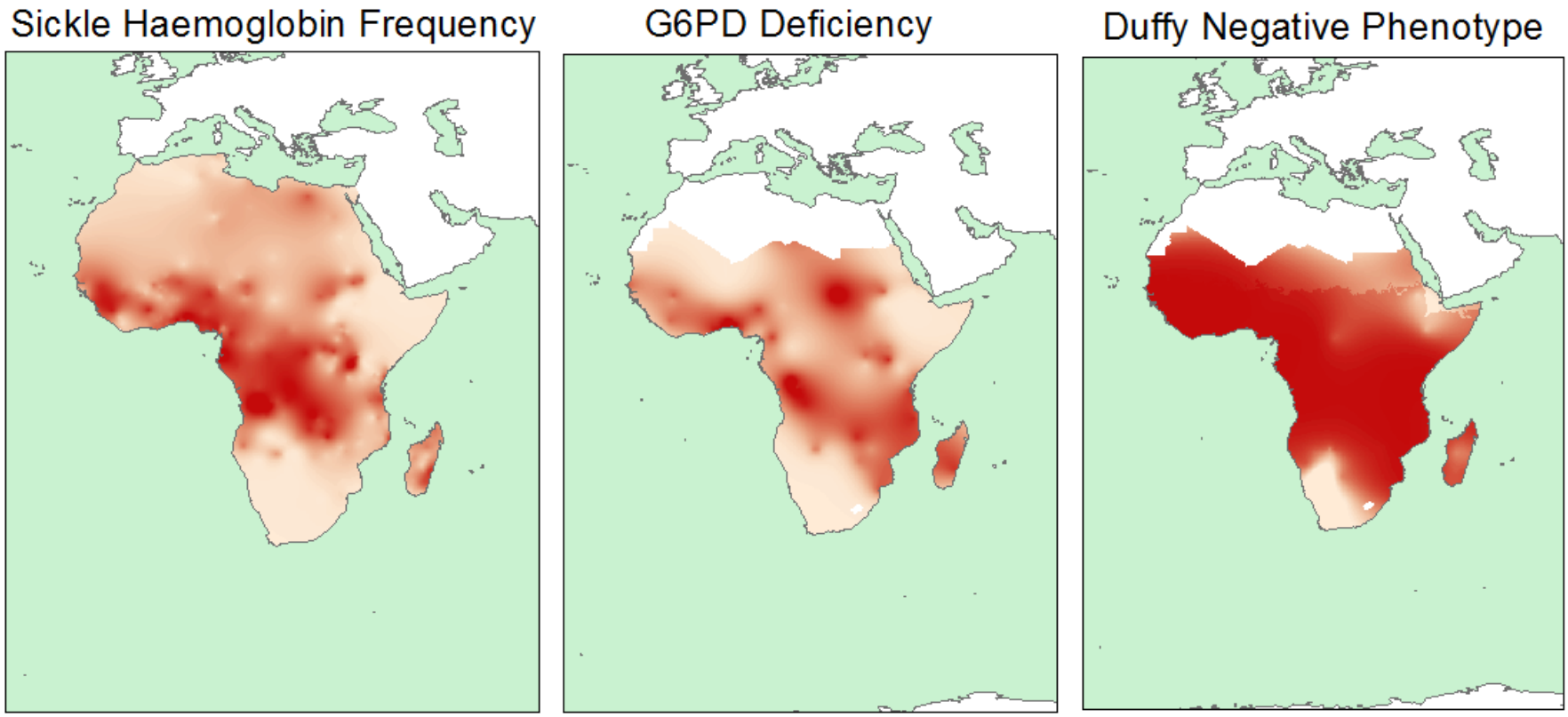

Number of Languages, World Language Mapping System
Legend
Languages
1
2
3
4 - 5
6 - 8
9 - 12
13 - 18
19 - 25
26 - 35
36 - 51
Sources: Esri, DeLorme, USGS, NPS, Sources: Esri, USGS, NOAA

Malaria Ecology Index



Malaria Endemicity




Sickle Haemoglobin Frequency
G6PD Deficiency
Duffy Negative Phenotype

Number of Languages, World Language Mapping System
Number of languages in the 1x1 degree cell. Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the World Language Mapping System language shapefile. We keep only languages that are spoken by more than 1000 people overall. We exclude all intersections that measures less than 10 km squared.

Number of Ethnic Group, Murdock (1959)

Number of ethnic groups in the 1x1 degree cell. Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the digitalized map of Murdock Africa map, from Nunn (2009). We exclude all intersections that measures less than 10 km squared.

Number of Ethnic Group, Murdock (1959)

Number of ethnic groups in the 1x1 degree cell. Source: constructed with a (ArcGIS) spatial join between a 1x1 degree grid and the digitalized map of Murdock map for North and South America, digitalize by Chiovelli (2013). We exclude all intersections that measures less than 10 km squared.

Malaria Ecology

Average Malaria Ecology Index in the 1x1 degree grid cell. Source: average Malaria Ecology is constructed as the 1x1 degree cell average of the Malaria Ecology index from Kiszewski (2004) across grids, computed using ArcGIS with data in EASE GLOBAL GRID projection.

Malaria Endemicity

Average Historical Malaria Endemicity in the 1x1 degree grid cell. Source: average Historical Malaria Endemicity is constructed as the 1x1 degree cell average of the Malaria Endemicity level, devised by Lysenko (1968) and digitalized by Hay 2004, computed using ArcGIS with data in EASE GLOBAL GRID projection.

Average Temperature

Mean annual 1x1 degree cell temperature (baseline period 1961-1990). Source: average temperature is constructed as the 1x1 degree cell of the mean annual temperature across raster grids, computed using ArcGIS with data in EASE GLOBAL GRID projection, from FAO/IIASA, 2011-2012. Global Agro-ecological Zones (GAEZ v3.0). FAO Rome, Italy and IIASA, Laxenburg, Austria.

Average Precipitation

Average 1x1 degree cell monthly precipitation mm/month (baseline period 1961-1990). Source: average monthly precipitation is constructed as the 1x1 degree cell average of the mean monthly precipitation across 10 minute grids, computed using ArcGIS with data in EASE GLOBAL GRID projection, with CRU CL 2.0 data from New (2002).

Land Suitability

Average land suitability in the 1x1 degree cell. Source: average Land Suitability is constructed as the 1x1 degree cell average of the land suitability index from Ramankutty (2002) across 0.5 degree raster grids, computed using ArcGIS with data in EASE GLOBAL GRID projection.

Variation in Land Suitability

Standard deviation of land suitability in the 1x1 degree cell. Source: standard deviation for Land Suitability is constructed as the 1x1 degree cell standard deviation of the land suitability index from Ramankutty (2002) across 0.5 degree raster grids, computed using ArcGIS with data in EASE GLOBAL GRID projection.

Mean Elevation

Average 1x1 degree cell elevation. Source: mean elevation is constructed as the 1x1 degree cell average of elevation across grids, computed using ArcGIS with data in Africa EASE GLOBAL GRID projection, with data from National Oceanic and Atmospheric Administration (NOAA) and U.S. National Geophysical Data Center, TerrainBase, release 1.0 (CD-ROM), Boulder, Colo.

Variation in Elevation

Standard deviation of elevation in the 1x1 degree cell. Source: constructed as the 1x1 degree cell standard deviation of elevation across grids, computed using ArcGIS with data in Africa EASE GLOBAL GRID projection, with data from National Oceanic and Atmospheric Administration (NOAA) and U.S. National Geophysical Data Center, TerrainBase, release 1.0 (CD-ROM), Boulder, Colo.

Ruggedness

Average 1x1 degree cell ruggedness (Terrain Ruggedness Index, 100 m). Source: mean ruggedness is constructed as the 1x1 degree cell average of elevation across grids, computed using ArcGIS with data in EASE GLOBAL GRID projection, with data from Terrain Ruggedness Index originally devised by Riley, DeGloria, and Elliot (1999), obtained through http://diegopuga.org/data/rugged/grid.

Total Water Area

Total area occupied by water within the 1x1 degree cell. Source: constructed with ArcGIS by intersecting the 1x1 degree cell grid and the Digital Chart of the World inwater shapefile, by intersecting the 1x1 degree cell grid and the Digital Chart of the World oceans and sea shapefile. We sum up total in-cell water area and the areas of the cell occupied by seas and oceans, areas computed with data in EASE GLOBAL GRID projection.

Total Area

Total area of the 1x1 degree cell. Source: constructed with ArcGIS by intersecting the 1x1 degree cell grid and the World Language Mapping System shapefile from the Digital Chart of the World. We exclude cell parts not covered by World Language Mapping System data (and by the Africa Murdock Map, and the Murdock map for North and South America), areas computed with data in EASE GLOBAL GRID projection.

Number of Countries

Total number of countries in the 1x1 degree cell. Source: constructed with ArcGIS by intersecting the 1x1 degree cell grid and the country boundaries shapefile from the Digital Chart of the World, areas computed with data in EASE GLOBAL GRID projection.

Within Country

Dummy variable taking value one if the 1x1 degree cell belong to one single country, 0 otherwise. Source: constructed with ArcGIS by intersecting the 1x1 degree cell grid and the country boundaries shapefile from the Digital Chart of the World.

Ln Migratory Distance

Migratory distance, on a land path, from Adis Ababa. Source: computed following **?**. The distance of the centroid of 1x1 degree cell from from Adis Abeba is computed using the Haversine formula. In order to replicate the most likely migration pattern followed by early men, we calculated the distance from Adis Ababa of the path that connect several obligatory intermediate points, and namely: Cairo, Istambul, Phnom Phen, Anadyr and Prince Rupert.

Ln Distance Coast

Distance to closest coast. Source: constructed with ArcGIS using the digital Chart of the World coastline shapefile.

Ln Distance Border

Distance to closest border. Source: constructed with ArcGIS using the digital Chart of the World boundaries shapefile.

Ln Distance Capital

Distance to the capital of the country where lies the centroid of the 1x1 degree cell. Source: constructed with ArcGIS using the digital World Capital shapefile.

Ln Distance River

Distance to closest river. Source: constructed with ArcGIS using Major Rivers World selected p3w shapefile, retrived from www.naturalearth.com.

Absolute Latitude

Absolute latitudinal distance from the equator in decimal degrees of the 1x1 degree cell. Source: computed using ArcGIS with data in WSG1984.

Night Lights

Data Construction and Sources: Malaria and Diversity

Average population in 1x1 degree cell. Source: constructed as the 1x1 degree cell average of population across raster grids, computed using ArcGIS with data in EASE GLOBAL GRID projection, with data from the Center for International Earth Science Information Network - CIESIN - Columbia University, United Nations Food and Agriculture Programme - FAO, and Centro Internacional de Agricultura Tropical - CIAT. 2005. Gridded Population of the World, Version 3 (GPWv3): Population Count Grid. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://sedac.ciesin.columbia.edu/data/set/gpw-v3-population-count.

Ln Population Density

Average night lights intesity in 1x1 degree cell. Source: constructed as the 1x1 degree cell average of night lights intesity grids, computed using ArcGIS with data in EASE GLOBAL GRID projection, with data from NOOA National Geophysical Data Centre for the year 2000.

Data Construction and Sources: Malaria Exposure and and endogamous Marriages

Intra-Ethnic Marriages
Variable equals to one for a marriage between two individuals belonging to the same ethnic group. Zero otherwise. Source: Demographic and Health Survey (DHS). We create this variable starting from ethnic variables ($v131$ for female and $mv131$ for male for most of the waves) in the DHS. We validate the ethnic groups using Ethnologue.

Malaria Ecology
Average Malaria Ecology Index in the 10km radius around the coordinates of each cluster. Source: Malaria Ecology index from Kiszewski (2004).

Malaria Endemicity
Average Historical Malaria Endemicity in the 10km radius around the coordinates of each cluster. Source: devised by Lysenko (1968) and digitalized by Hay 2004.

Average Temperature
Mean annual temperature in the 10km radius around the coordinates of each cluster (baseline period 1961-1990). Source: FAO/IIASA, 2011-2012. Global Agro-ecological Zones (GAEZ v3.0). FAO Rome, Italy and IIASA, Laxenburg, Austria.

Average Precipitation
Average monthly precipitation mm/month in the 10km radius around the coordinates of each cluster (baseline period 1961-1990). Source: CRU CL 2.0 data from New (2002).

Land Suitability
Average land suitability in the 10km radius around the coordinates of each cluster. Source: Ramankutty (2002).

Mean Elevation
Average elevation in the 10km radius around the coordinates of each cluster. Source: National Oceanic and Atmospheric Administration (NOAA) and U.S. National Geophysical Data Center, TerrainBase, release 1.0 (CD-ROM), Boulder, Colo.

Variation in Elevation
Standard deviation of elevation in the 10km radius around the coordinates of each cluster. Source: National Oceanic and Atmospheric Administration (NOAA) and U.S. National Geophysical Data Center, TerrainBase, release 1.0 (CD-ROM), Boulder, Colo.

Urban Residence Female
Dummy variable taking value one the individual is living in urban areas. Zero otherwise. Source: Demographic and Health Survey (DHS) ($v025$ for female and $mv025$ for male).

Education
Highest year of education of the respondent individual. Source: Demographic and Health Survey (DHS) ($v133$ for female and $mv133$ for male).

Age
Age of the respondent individual. Source: Demographic and Health Survey (DHS) ($vv012$ for female and $mv012$ for male).

Population in the Relevant Ethnic Aggregation
Absolute number of DHS survey respondent belonging to any given ethnolinguistic family in the region. Source: Demographic and Health Survey (DHS) ($v012$ for female and $mv012$ for male).