Data Supplement to "Diversity and Conflict"

Data Appendix A Variable Definitions and Data Sources for the Country-Level Analyses

Migratory Distance and Population Diversity

1. **Migratory distance from East Africa:** The great circle distance from Addis Ababa, Ethiopia to a country's capital city along a land-restricted path forced through one or more of five intercontinental waypoints, including Cairo, Egypt; Istanbul, Turkey; Phnom Penh, Cambodia; Anadyr, Russia; and Prince Rupert, Canada. Distances are calculated using the Haversine formula and are measured in units of ten thousand kilometers. The methodology underlying the construction of this measure is adopted from Ramachandran et al. (2005). The geographical coordinates of the waypoints are obtained from Ramachandran et al. (2005) and those of the capital cities are obtained from the Central Intelligence Agency's (CIA) World Factbook. See Ashraf and Galor (2013a) for additional details.

2. **Population diversity (precolonial):** The expected heterozygosity (neutral genetic diversity) of a country's precolonial population as predicted by migratory distance from East Africa (i.e., Addis Ababa, Ethiopia) to the country's capital city. This measure is calculated by applying the regression coefficients obtained from regressing expected heterozygosity on migratory distance at the ethnic group level, using a worldwide sample of 53 ethnic groups from the HGDP-CEPH Human Genome Diversity Cell Line Panel. The expected heterozygosities and geographical coordinates of the ethnic groups are from Ramachandran et al. (2005). See Ashraf and Galor (2013a) for additional details.

3. **Population diversity (ancestry adjusted):** The expected heterozygosity (neutral genetic diversity) of a country's contemporary national population, as developed by Ashraf and Galor (2013a). This measure is based on migratory distances from East Africa to the year 1500 locations of the ancestral populations of the country's component ethnic groups in 2000 and on the pairwise migratory distances among these ancestral populations. The source countries of the ancestral populations are identified from the World Migration Matrix, 1500–2000 (Putterman and Weil, 2010), and the capital cities of these countries are used to compute the aforementioned migratory distances. The measure of population diversity is then computed by applying (i) the coefficients obtained from regressing expected heterozygosity on migratory distance from East Africa at the ethnic group level, using a worldwide sample of 53 ethnic groups from the HGDP-CEPH Human Genome Diversity Cell Line Panel; (ii) the coefficients obtained from regressing pairwise genetic distance on pairwise migratory distance in a sample of 1,378 HGDP-CEPH ethnic group pairs, and (iii) the ancestry weights representing the fractions of the year 2000 national population (i.e., of the country for which the measure is being computed) that can trace their ancestral origins to different source countries in the year 1500. The data at the ethnic-group (or group-pair) level on expected heterozygosities, geographical coordinates, and pairwise genetic distances are obtained from Ramachandran et al. (2005), and the country-level data on ancestry weights are obtained from the World Migration Matrix, 1500–2000. See Ashraf and Galor (2013a) for a detailed discussion of the methodology underlying the construction of this measure.

Conflict outcomes

1. **PRIO civil conflict and civil war outcomes:** Our primary measures of civil conflict are based on Version 18.1 of the UCDP/PRIO Armed Conflict Dataset (ACD), covering the 1946–2017 time period (Gleditsch et al., 2002; Pettersson and Eck, 2018). In this dataset, an armed 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 calendar year." In our study, the term *PRIO25 civil conflict* indicates an internal armed conflict between the government of a state and one or more internal opposition group(s), without any intervention from other states as independent actors or intervention from other states to support either side of the conflict. Thus, the measures of civil conflict in our study exclude internationalized internal armed conflicts. In addition, extrasytemic and interstate conflicts are also excluded from the analysis, following the standard definition of civil conflict. For further information on the data underlying our various civil conflict measures (discussed below), the interested reader is referred to the codebook for Version 18.1 of the UCDP/PRIO ACD.

   The main conflict variable examined in our cross-sectional analyses of civil conflict is the ***log number of new PRIO25 civil conflict onsets per year*** during the 1960–2017 time period. This measure is obtained by first computing the total count of *new* civil conflicts that took place on the territory of a country in our sample during this period. Then, this count is divided by the number of years over the same time period in which the territory was home to one or more entities included in the Gleditsch and Ward list of independent states, as employed by the UCDP/PRIO ACD. Finally, the resulting average annual conflict frequency is scaled up by 1 and log-transformed. Each *new* conflict is identified by a unique conflict identifier provided by the UCDP/PRIO ACD. In this definition, two or more conflict episodes involving the same actors fighting over the same incompatibility are not treated as separate (new) conflicts. Instead, they are assigned the same conflict identifier.

   The main outcome examined by our regressions using annually repeated cross-country data is ***annual PRIO25 civil conflict onset***. It is equal to 1 for each year when at least one new PRIO25 conflict broke out and zero otherwise. The date of a *new* conflict outbreak (or onset) is the starting year of the first conflict episode for a given conflict, and it reflects the first year in which the conflict reached or surpassed the annual fatality threshold of 25 battle-related deaths. Subsequent years of a given conflict episode or outbreaks of subsequent conflict episodes of the same conflict are not considered *new* conflict onsets.

   Quinquennial PRIO25 civil conflict incidence is the main outcome examined by our regressions using quinquennially repeated cross-country data over the 1960–2017 time period. It is equal to 1 for a given 5-year interval for a country if there was an active (ongoing) PRIO25 civil conflict in at least one year during that time interval and zero otherwise. A conflict is deemed active in a given calendar year if it resulted in at least 25 battle-related deaths during that year. ***Annual PRIO25 civil conflict incidence*** is defined in a similar manner except that the incidence is coded for each country-year observation instead of a 5-year time interval for a country.

   Quinquennial PRIO1000 civil war incidence is an alternative outcome examined by our robustness checks in regressions using quinquennially repeated cross-country data. This variable is constructed in a manner similar to ***quinquennial PRIO25 civil conflict incidence***. The only difference is that for civil wars, a conflict is deemed as active (ongoing) in

a given year only if a much higher fatality threshold of 1,000 (instead of 25) battle-related deaths is exceeded in that year.

2. **Intragroup (intracommunal) factional conflict:** The outcome variables employed by the analysis of intragroup conflict are based on the All Minorities At Risk (AMAR) Sample Data of the AMAR Phase I Project (Birnir et al., 2018). The AMAR sample contains longitudinal data on 365 AMAR ethnic groups. Of these groups, 291 were included in the original Minorities At Risk (MAR) Project (Phases I–V), and the remaining 74 were selected randomly from the sample frame of socially relevant groups outlined by Birnir et al. (2015), according to the new AMAR criteria summarized in the AMAR codebook.

   The measures of intragroup factional conflict we employ are constructed using the *INTRACON* variable in the AMAR Sample Data. This is a dummy variable, coded for each group in the AMAR sample, indicating the presence of an intracommunal conflict within that group in a given year. Specifically, the variable is coded for each year during the 1980–2006 time period. However, since the coverage of AMAR groups for the 1980–1984 time period is rather limited, our measures of intragroup conflict are based on information for the 1985–2006 time frame. Thus, the outcome variable in our cross-country analysis of intragroup conflict is the ***share of AMAR group-years with at least one intracommunal conflict*** within a country during this time period. Further, the outcome variable in our analysis of intragroup conflict using annually repeated cross-country data is ***annual intracommunal conflict incidence***, coded 1 for each country-year in which there was at least one AMAR sample group with an active intracommunal conflict and zero otherwise. For further information on the data underlying our measures of intragroup conflict, the reader is referred to Version 1 of the codebook for the AMAR Phase I Project.

3. **Historical conflict outcomes:** To construct historical conflict outcomes between the 15th and 19th centuries, we make use of information on the locations of violent conflicts during the 1400–1799 time period, as compiled by Brecke (1999) and georeferenced by Dincecco et al. (2015). The georeferenced conflict locations are used to map historical conflicts to territories, as defined by their contemporary national borders. It may be noted that in the catalog of conflicts from Dincecco et al. (2015), there were a small number of instances where the country assignment did not match the country implied by the georeferenced location of the conflict in ArcGIS. In such cases, supplementary information from the catalog (e.g., the actors in the conflict or the place where the conflict occurred) was consulted to first determine if the mismatch was due to an error in the original country assignment or an error in the supplied coordinates. Then, either the country assignment or the coordinates were altered to match our understanding of the true location of the conflict. In addition, for naval conflicts or for conflicts between actors that took place on lands to which neither actor was native, these specific conflicts were assigned to either one of the actors' countries (rather than the country implied by the location of the conflict) but only if the actors possessed comparable levels of diversity (e.g., if the actors were both European colonial powers engaged in a conflict on a colonized territory).

   As for the underlying conflict data, the definition of a violent conflict in Brecke's dataset is based on Cioffi-Revilla (1996): "An occurrence of purposive and lethal violence among 2+ social groups pursuing conflicting political goals that results in fatalities, with at least one belligerent group organized under the command of authoritative leadership. The state does not have to be an actor. Data can include massacres of unarmed civilians or territorial

conflicts between warlords." The list is comprised of conflicts that resulted in at least 32 fatalities. This fatality level corresponds to a magnitude of 1.5 or higher on Richardson's (1960) base-10 log conflict scale. Although the dataset does not systematically distinguish between intrastate and interstate conflicts, the latter appear to form the basis of the recorded conflicts, and while the recorded conflicts do not necessarily represent the whole universe of conflict events during the sample period, the list contains almost all major conflicts that have been documented by historians. The conflict catalog is also considered to be fairly comprehensive in terms of its broad regional coverage, including five regions of the world: Western Europe, Eastern Europe, North Africa, West & Central Africa, East & Southern Africa, as well as Central Asia & Siberia.

Based on these conflict data, our study employs two distinct categories of country-level outcome measures: (1) the number of distinct conflicts, occurring in each century of the 1400–1799 time period or across this entire time frame; and (2) the likelihood of observing one or more conflicts, either during the entire 1400–1799 time period or in each century therein.

4. **MEPV civil conflict severity:** This variable is constructed using information provided by the Major Episodes of Political Violence (MEPV) War List (1946–2017), maintained by the Center for Systemic Peace. This list is a regularly updated version of Appendix C from Marshall (1999) and further detailed in Marshall (2002).

   A major episode of political violence is defined as the systematic and sustained use of lethal violence by one or more organized groups, resulting in at least 500 directly-related deaths over the course of the episode. Episodes are coded for both time span and a general magnitude of societal-systemic impact (an eleven-point scale, 0-10). These magnitude scores are considered to be consistent and comparable across categories and cases. Further, each episode is assigned to one of seven categories of armed conflict: international violence (IV), international war (IW), international independence war (IN), civil violence (CV), civil war (CW), ethnic violence (EV), and ethnic war (EW). Episodes belonging to the last four of these categories constitute the universe of intrastate episodes that are of interest to our analysis. The magnitude scores for these episodes are aggregated into the *CIVTOT* variable in the MEPV dataset. *CIVTOT* is an annual ordinal index of civil conflict intensity at the country level that underlies the particular measure of ***quinquennial MEPV civil conflict severity*** we employ – namely, the maximum value of *CIVTOT* across all years in any given 5-year interval during the 1960–2017 time period. For further information on the data underlying our measure of civil conflict severity, the reader is referred to the codebook for the MEPV dataset.

5. **CNTS social conflict index:** This variable is based on the Domestic Conflict Event Data from the Cross-National Time Series (CNTS) Data Archive 2018 Edition (Banks and Wilson, 2018), which covers the 1815–2017 time period.

   Specifically, the basis of our CNTS social conflict index is the variable *Domestic9* from the CNTS Data Archive. *Domestic9* is an annual continuous index of the degree of social unrest, computed by first taking the weighted sum of the counts of different unrest/conflict events (given by the variables *domestic1-8*) in a country-year. As of October 2007, the weights employed were as follows: Assassinations (25), Strikes (20), Guerrilla Warfare (100), Government Crises (20), Purges (20), Riots (25), Revolutions (150), and Anti-Government Demonstrations (10). In a second step, the weighted sum is multiplied by 100/8 to obtain

Domestic9. The specific measure used in our study is a ***quinquennial CNTS social conflict index***, calculated for each country as the maximum value of *Domestic9* across all years in any given 5-year interval during the 1960–2017 time period. For further information on the source data for our social conflict index, the reader is referred to the website of the CNTS Data Archive.

6. **UCDP nonstate conflict incidence:** This measure is based on information from Version 18.1 of the UCDP Non-State Conflict Dataset, covering the 1989–2017 time period (Sundberg et al., 2012).

   A non-state conflict is defined by the Uppsala Conflict Data Program (UCDP) as "the use of armed force between two organized armed groups, neither of which is the government of a state, which results in at least 25 battle-related deaths in a year." An *organized group* can be either (i) a *formally organized group*, i.e., any non-governmental group of people having announced a name for their group and using armed force against another similarly organized group; or (ii) an *informally organized group*. The latter type of group does not have an announced name, but it uses armed force against another similarly organized group such that there is a clear pattern of violent incidents that are connected and in which both groups use armed force against the other. ***Quinquennial UCDP nonstate conflict incidence*** is coded 1 for any 5-year interval for a country if in any year during this interval there was at least one active (ongoing) non-state conflict in the country. A conflict is deemed active in a given calendar year if it resulted in at least 25 battle-related deaths during that year. For further information on the source data for our measure of non-state conflict incidence, the reader is referred to the codebook for Version 18.1 of the UCDP Non-State Conflict Dataset.

Other outcomes

1. **Number of ethnic groups:** The total number of distinct ethnic groups in a country's population, as compiled by Fearon (2003). The specific variable employed by our analysis is the natural logarithm of one plus the number of ethnic groups. See Fearon (2003) for additional details on primary data sources and methodological assumptions.

2. **Prevalence of interpersonal trust:** This variable is constructed using information from the World Values Survey (2006, 2009) (henceforth, WVS) on the prevalence of generalized interpersonal trust in a country's population. In particular, this well-known measure of social capital at the country level reflects the proportion of all respondents (from across five different waves of the WVS, conducted over the 1981–2009 time period) that opted for the answer "Most people can be trusted" (as opposed to "Can't be too careful") when responding to the survey question "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" For additional details, the reader is referred to documentation available on the WVS website.

3. **Variation in political attitudes:** The intra-country dispersion in self-reported individual political positions on a "left"–"right" categorical scale, based on data from the WVS. Specifically, this measure of heterogeneity in political attitudes at the country level is calculated as the intra-country standard deviation across all respondents (sampled over five different waves of the WVS during the 1981–2009 time period) of their self-reported positions on a categorical scale from 1 (politically "left") to 10 (politically "right") when answering the survey question "In political matters, people talk of 'the left' and 'the right.' How would you place your

views on this scale, generally speaking?" Given that this variable's unit of measurement does not possess a natural interpretation, we standardize the cross-country distribution of this variable prior to conducting our regressions. For additional details, the reader is referred to documentation available on the WVS website.

Main Control Variables

1. **Ethnic fractionalization:** This is the well-known ethnic fractionalization index of a country, reflecting the probability that two individuals, randomly selected from the country's population, will belong to different ethnic groups. Formally, a country's ethnic fractionalization index is calculated as follows:

$$FRAC = 1 - \sum_{i=1}^{n} p_i^2,$$

where p_i is the proportional representation of ethnic group i in the national population; and n is the total number of ethnic groups in the country. The specific variable we employ is based on the list of ethnic groups (and their national population shares) by country as compiled by Alesina et al. (2003). See Alesina et al. (2003) for additional details on primary data sources and methodological assumptions.

2. **Ethnolinguistic polarization:** An ethnolinguistic polarization index at the country level, calculated by applying the following definition of polarization due to Reynal-Querol (2002) and Montalvo and Reynal-Querol (2005):

$$POL = 4 \sum_{i=1}^{n} p_i^2 \left[1 - p_i\right],$$

where p_i is the proportional representation of linguistic group i in the national population; and n is the total number of linguistic groups in the country. The employed ethnolinguistic polarization index is sourced from the replication dataset of Desmet et al. (2012). The authors provide measures of several such polarization indices, constructed at different levels of aggregation of linguistic groups in a country's population (based on hierarchical linguistic trees). The specific polarization measure we use corresponds to the most disaggregated level of the linguistic tree, and it reflects the extent of polarization across subnational groups classified according to modern-day languages. See Desmet et al. (2012) for additional details on primary data sources and methodological assumptions.

3. **Absolute latitude:** The absolute value of the latitude of a country's geodesic centroid, as reported by the *At These Coordinates* resource repository, based on metadata from (i) the National Geospatial-Intelligence Agency's (NGA) GEOnet Names Server (GNS); and (ii) the United States Geological Survey's (USGS) Geographic Names Information System (GNIS).

4. **Ruggedness:** A measure of the degree of terrain ruggedness of a country's territory. Based on Riley et al. (1999), the ruggedness of a grid cell, i, is defined as

$$RIX(i) = \sqrt{\sum_{k=1}^{8} (h_i - h_{j_k})^2},$$

where h_l is the elevation (in meters above sea level) of cell $l = i, j_1, j_2, ..., j_8$, and the cells indexed by j are the eight neighboring cells of i. The country-level measure of ruggedness used by our study is the mean value of $RIX(i)$ across all 1 km $\times$ 1 km grid cells of a country. The cell-level ruggedness index is computed by Özak (2010), based on topographical data from the Global Land One-Kilometer Base Elevation (GLOBE) digital elevation model (Hastings et al., 1999).

5. **Mean and range of elevation:** The country-level mean and range of elevation (in thousands of kilometers above sea level), calculated using geospatial elevation data at a 1-degree resolution from the Geographically based Economic data (G-ECON) project (Nordhaus, 2006), based on similar data at a 10-minute resolution from New et al. (2002). The mean of elevation at the country level reflects the average value across the grid cells that are located within a country's national borders, whereas the range of elevation reflects the difference between the maximum and minimum values across the same set of grid cells. See the G-ECON project website for additional details.

6. **Mean and range of land suitability:** The country-level mean and range of a geospatial index of the suitability of land for agriculture, based on ecological indicators of climate suitability for cultivation, such as growing degree days and the ratio of actual to potential evapotranspiration, as well as on ecological indicators of soil suitability for cultivation, such as soil carbon density and soil pH. This index was initially developed at a half-degree resolution by Ramankutty et al. (2002), and it has been aggregated to the country level by Michalopoulos (2012), with the mean at the country level reflecting the average value of the index across the grid cells that are located within a country's national borders, and the range reflecting the difference between the maximum and minimum values of the index across the same set of grid cells. See Michalopoulos (2012) for additional details.

7. **Island nation:** An indicator for whether a country shares a land border with any other country, as reported by the CIA's World Factbook. Of the 147 countries in our baseline sample, the following 7 are coded as island nations: Australia, Cuba, Japan, Sri Lanka, Madagascar, New Zealand, and Philippines.

8. **Distance to nearest waterway:** The distance (in thousands of kilometers) from a grid cell to the nearest ice-free coastline or sea-navigable river, averaged across the grid cells of a country. This variable was originally constructed by Gallup et al. (1999) and is available from the Research Datasets online repository maintained by Harvard University's Center for International Development.

9. **Colonial history:** A set of three indicators reflecting a country's experience of colonial rule by (i) the U.K., (ii) France, or (iii) any other major colonizing power, respectively. Therefore, the omitted category is the absence of colonial rule. These variables are constructed based on information from various sources, including the CIA's World Factbook, the Encyclopaedia Brittanica, Country Studies of the Library of Congress, and rulers.org amongst others. Additional details are available from the authors upon request.

   In cross-sectional regressions at the country level, the relevant measures comprise time-invariant indicators for the historical presence of colonial rule – i.e., whether the country

has ever been ruled by the colonizing power in question. In regressions using repeated cross-country data, the relevant measures comprise time-varying indicators of the lagged prevalence of colonial rule – i.e., whether the country was ruled by the colonizing power in question at any point in the preceding 5-year time interval or in the preceding year, depending on the temporal dimension of the repeated cross-section.

10. **Legal origins:** A set of two time-invariant indicators for British and French legal origins, as reported by La Porta et al. (1999). Specifically, these indicators identify whether the legal origin of country's Company Law or Commercial Code is (i) the English Common Law or (ii) the French Commercial Code, respectively. The omitted category is German, Scandinavian, or Socialist legal origins, as recognized by La Porta et al. (1999).

11. **Executive constraints:** An index, reported at an annual frequency as a 7-point categorical variable (from 1 to 7) by the Polity IV Project (Version 2017), quantifying the extent of institutionalized constraints on the decision-making power of chief executives (Marshall et al., 2017). The specific version of the Polity IV Project dataset employed by our study covers the 1800–2017 time period. For further information on the index of executive constraints, the reader is referred to the codebook for Version 2017 of the Polity IV Project dataset.

    In cross-sectional regressions at the country level, the relevant measure is the temporal average of the index across all years in the 1960–2017 time period. In regressions using quinquennially repeated cross-country data, the relevant measure is the temporal average of the index across all years in the preceding 5-year time interval. Finally, in regressions based on annually repeated cross-country data, the relevant measure is the value of the index from the preceding year.

12. **Type of political regime:** Our measures of the type of political regime are based on two indicators reflecting whether a country is classified as a democracy (or not) and as an autocracy (or not) in a given year. The omitted category is anocracy, a hybrid regime that constitutes the middle range of the autocracy-democracy political spectrum. This regime classification is based on the *POLITY2* index (the Revised Combined Polity Score), as reported at an annual frequency by the Polity IV Project (Version 2017) for the 1800–2017 time period (Marshall et al., 2017). *POLITY2* is a discrete index that ranges from -10 (strongly autocratic) to +10 (strongly democratic). Following the norm in the literature, a country-year is coded as a *democracy* if the *POLITY2* score is above 5 or as an *autocracy* if the score is below -5. The prevalence of *anocracy*, occurring when the *POLITY2* score is between -5 and 5 for a country-year, therefore serves as the omitted political regime category. For further information on the *POLITY2* index, the reader is referred to the codebook for Version 2017 of the Polity IV Project dataset.

    In cross-sectional regressions at the country level, the relevant measures of regime type are the fractions of years during the 1960–2017 time period that a country spent as a democracy and as an autocracy, respectively. In regressions using quinquennially repeated cross-country data, the relevant measures are the fractions of years during the preceding 5-year time interval that a country spent as a democracy and as an autocracy, respectively. Finally, in regressions based on annually repeated cross-country data, the relevant measures are the indicators for democracy and autocracy for the preceding year.

13. **Oil or gas reserve discovery:** A time-invariant indicator of at least one petroleum (oil or gas) reserve on the land territory of a country. This variable is based on information provided in the Petroleum Dataset (Version 1.2), covering the 1946–2003 time period (Lujala et al., 2007). Therefore, the available data does not provide information about any petroleum deposit discovered after 2003. The dataset is compiled for the main purpose of investigating the relationship between armed civil conflict and natural resources. Each on-shore petroleum (oil or gas) reserve – identified as polygons in the shapefile accompanying the dataset – is assigned to a modern-day country based on the coordinates of the centroids of the deposit polygons. For additional information, the reader is referred to the codebook for Version 1.2 of the Petroleum Dataset, available from the Geographical and Resource Datasets online repository maintained by PRIO.

14. **Log population size:** The log-transformed size of a country's population, as reported by the World Bank's World Development Indicators (WDI) online data catalog.

    In cross-sectional regressions at the country level, the relevant measure is the log-transformed temporal average of annual population observations across all years in the 1960–2017 time period. In regressions using quinquennially repeated cross-country data, the relevant measure is the log-transformed temporal average of observations across all years in the preceding 5-year time interval. Finally, in regressions based on annually repeated cross-country data, the relevant measure is the log-transformed observation from the preceding year.

15. **Log GDP per capita:** The log-transformed per-capita GDP (in current US$) of a country, as reported by the World Bank's World Development Indicators (WDI) online data catalog.

    In cross-sectional regressions at the country level, the relevant measure is the log-transformed temporal average of annual per-capita GDP observations across all years in the 1960–2017 time period. In regressions using quinquennially repeated cross-country data, the relevant measure is the log-transformed temporal average of observations across all years in the preceding 5-year time interval. Finally, in regressions based on annually repeated cross-country data, the relevant measure is the log-transformed observation from the preceding year.

Other Control Variables (for Robustness Checks)

1. **Ecological fractionalization and polarization:** These measures of ecological diversity are motivated by Fenske (2014). The measure of *ecological fractionalization* is a Herfindahl index, constructed as

$$\text{Ecological fractionalization}_i = 1 - \sum_{t=1}^{t=18} (s_i^t)^2;$$

and *ecological polarization* index is given by

$$\text{Ecological polarization}_i = 1 - \sum_{t=1}^{t=18} \left(\frac{0.5 - s_i^t}{0.5} \right)^2 s_i^t,$$

where s_i^t is the share of the area of country i that is occupied by ecological type t. The polarization index measures the degree to which a country's area approximates a territory in which two ecological types each occupy half the total area. The relevant information on the spatial distribution of ecological types across the land surface of the earth is derived from

global maps of agro-ecological zones from the Food and Agriculture Organization (FAO) of the United Nations.

2. **Mean and volatility of temperature and precipitation:** These four variables are constructed using information on mean temperature (in degree Celcius) per annum and total precipitation (in mm) per annum as reported by the Climate Research Unit (CRU) (Harris et al., 2014). Specifically, we employ the country-level spatial aggregates of annual mean temperature and annual total precipitation, provided the CRU CY Version 4.01 dataset, which spans the 1901–2016 time period.

   In cross-sectional regressions at the country level, the relevant measures of mean temperature and total precipitation reflect the temporal averages of the annual observations of these variables across all years in the 1960–2017 time period, whereas the corresponding volatility measures capture their respective temporal standard deviations during the same time span. In regressions using quinquennially repeated cross-country data, the relevant mean and volatility measures are similarly defined, except that the temporal averages and standard deviations are calculated across the years of the preceding 5-year time interval (rather than the full sample period). Finally, in regressions based on annually repeated cross-country data, the relevant measures are the one-year lags of annual mean temperature and annual total precipitation as well as the interannual standard deviations of temperature and precipitation over a 5-year rolling window that ends in the preceding year.

3. **Log years since Neolithic Revolution:** The log-transformed number of thousand years elapsed (as of the year 2000) since the majority of the population residing in a territory defined by a country's modern national borders began practicing sedentary agriculture as the primary mode of subsistence. This measure, initially reported by Putterman (2008), is compiled using a host of both region- and country-specific archaeological studies as well as more general encyclopedic works on the transition from hunting and gathering to agriculture during the Neolithic Revolution. The reader is referred to Putterman's website for a detailed description of the primary and secondary data sources employed in the construction of this variable.

4. **Log index of state antiquity:** The log-transformation of an index reflecting a country's cumulative experience with institutionalized statehood since antiquity. Specifically, we employ the State Antiquity Index (version 3.1), first introduced by Bockstette et al. (2002). The underlying index quantifies the exposure of a territory – as defined by a country's modern national borders – to formal statehood (i.e., being an independent nation-state or part of a larger kingdom or an empire) since the year 1 CE and until 1950. In particular, for each 50-year time interval, information on a territory's status with respect to the following 3 questions (each with specific weights applied) is employed: (i) is there a government above the tribal level?; (ii) is this government foreign or locally based?; and (iii) how much of the territory of the modern country was ruled by this government? These information are then aggregated over time to produce an index that ranges between 0 and 1. The reader is referred to Putterman's website for a detailed description of the methodology and data sources employed in the construction of this index.

5. **Log duration of human settlement:** The natural logarithm of the maximum duration (in tens of thousands of years) of uninterrupted settlement by anatomically modern humans

across locations in a territory defined by a country's modern national borders. The underlying measure is obtained from the dataset of Ahlerup and Olsson (2012). The reader is therefore referred to that work for additional details on data sources and methodological assumptions.

6. **Log distance from regional frontier in 1500:** The great circle distance from a country's capital city to the closest regional technological frontier around the year 1500. The variable is obtained from the dataset of Ashraf and Galor (2013a). The set of regional frontiers comprises the two most populous cities, reported for the year 1500 and belonging to different civilizations or sociopolitical entities, from each of Africa, Europe, Asia, and the Americas. Distances are calculated using the Haversine formula and are measured in kilometers. The historical urban population data used to identify the frontiers are sourced from Chandler (1987) and Modelski (2003), and the geographical coordinates of ancient urban centers are sourced from online resources such as Wikipedia.

7. **Ethnic inequality in luminosity:** A measure of intra-country economic inequality as captured by the subnational spatial distribution of per-capita adjusted nighttime luminosity in the year 2000 across the georeferenced homelands of ethnic groups. This measure is sourced from the replication dataset of Alesina et al. (2016). The reader is therefore referred to that work for additional details on data sources and methodological assumptions.

8. **Spatial inequality in luminosity:** A measure of intra-country economic inequality as captured by the subnational spatial distribution of per-capita adjusted nighttime luminosity in the year 2000 across 2.5×2.5-degree geospatial grid cells. This measure is sourced from the replication dataset of Alesina et al. (2016). The reader is therefore referred to that work for additional details on data sources and methodological assumptions.

9. **Linguistic fractionalization and polarization (georeferenced):** These are the country-level counterparts of the measures of linguistic fractionalization and polarization that are used in our analysis of conflicts at the ethnic homelands level. Specifically, these measures are constructed using georeferenced information on the spatial distribution of language homelands from the World Language Mapping System (WLMS) along with gridded population data from the Gridded Population of the World (GPW) dataset.

10. **Ethnic fractionalization (Fearon, 2003):** The ethnic fractionalization index compiled by Fearon (2003). The index reflects the probability that two individuals, randomly selected from a country's population, will belong to different ethnic groups.

11. **Linguistic fractionalization (Alesina et al., 2003):** The linguistic fractionalization index compiled by Alesina et al. (2003). The index reflects the probability that two individuals, randomly selected from a country's population, will belong to different linguistic groups.

12. **Religious fractionalization (Alesina et al., 2003):** The religious fractionalization index compiled by Alesina et al. (2003). The index reflects the probability that two individuals, randomly selected from a country's population, will belong to different religions.

13. **Ethnolinguistic fractionalization (Esteban et al., 2012):** An index of ethnolinguistic fractionalization, as represented by the *frac_fear* variable in the replication dataset of Esteban

et al. (2012). The underlying ethnolinguistic population shares are sourced from Fearon (2003).

14. **Ethnolinguistic polarization (Esteban et al., 2012):** The Esteban-Ray index of ethnolinguistic polarization with $\delta = 0.05$, as represented by the *er_fear_delta005* variable in the replication dataset of Esteban et al. (2012). The underlying ethnolinguistic population shares are sourced from Fearon (2003).

15. **Gini index of ethnolinguistic diversity (Esteban et al., 2012):** The gini index of ethnolinguistic diversity per capita with $\delta = 0.05$, as represented by the variable named *gini_fear_delta005_PERCAPTA* in the replication dataset of Esteban et al. (2012). It is obtained after dividing the gini index of ethnolinguistic diversity by population size. The underlying ethnolinguistic population shares are sourced from Fearon (2003).

16. **Log percentage mountainous terrain:** The log-transformation of the proportion (in percentage) of a country's territory that is "mountainous" according to the codings of the geographer A.J. Gerard. This variable is sourced from the replication dataset of Fearon and Laitin (2003), where it is used to test the hypothesis that "rough terrain, poorly served by roads, at a distance from the centers of state power should favor insurgency and civil war."

17. **Noncontiguous state dummy:** A time-invariant indicator of whether a country possesses a territory with a population of at least 10,000 that is separated from the region containing its capital city either by land or 100 kilometers of water. This variable is sourced from the replication dataset of Fearon and Laitin (2003), where it is used to test the hypothesis that "the presence of a territory that is separated from the center of national governance by water or distance can help rebels more easily sustain insurgent activity and, thereby, make civil war more likely."

18. **Disease richness:** The total number of different types of infectious diseases in a country as reported by Fincher and Thornhill (2008), based on the Global Infectious Disease and Epidemiology Network (GIDEON; www.gideononline.com).

19. **Ethnic dominance:** A time-invariant indicator of whether the largest ethnic group in a country constitutes 45-90% of the national population. This variable is sourced from the replication dataset of Hegre and Sambanis (2006), but the primary source of the measure is Collier and Hoeffler (2004).

20. **Political instability:** A time-varying indicator at the country-year level of whether there was a change in the Polity IV regime index by 3 or more points in any of the three years prior to the country-year in question. Periods of regime transition (-88) and "interruptions" (indicating a complete collapse of central authority) are also coded as cases of political instability. Episodes of foreign occupation, however, are treated as missing observations. In robustness checks of our civil conflict onset regressions, the one-year lagged value of this variable is employed. This variable is sourced from the replication dataset of Hegre and Sambanis (2006), but the primary source is Fearon and Laitin (2003).

21. **New state dummy:** A time-varying indicator at the country-year level for whether the current year is the first year of the country's existence (e.g., as a newly independent state from colonial rule). In robustness checks of our civil conflict onset regressions, the one-year lagged value of this variable is employed. This variable is sourced from the replication dataset of Hegre and Sambanis (2006).

22. **Commodity export price shocks:** A set of four variables capturing different types of commodity export price shocks on an annual basis, sourced from the replication dataset of Bazzi and Blattman (2014). The first variable reflects *aggregate price shocks* and is computed as the annual change in a country's log commodity export price index (a geometric average of all commodity export prices weighted by lagged export shares). The remaining variables reflect three types of disaggregated price shocks. The first of these reflects *annual crop price shocks*, i.e., price shocks to annual agricultural goods, such as oilseeds, food crops, and livestock, that are more likely to accrue to households. The second reflects *perennial crop price shocks*, i.e., price shocks to perennial tree crops like cocoa, coffee, rubber, or lumber. Finally, the third type of disaggregated price shocks captures *extractive crop price shocks*, i.e., price shocks to extractive products, namely, minerals, oil, and gas, that are more likely accrue to states. By construction, the sum of the three disaggregated types of shocks yields the *aggregate price shock* variable. In robustness checks of our civil conflict onset regressions, we employ the contemporaneous as well as the one- and two-year lagged values of these various commodity export price shock variables. For additional details, the reader is referred to Bazzi and Blattman (2014).

TABLE DS.I: Summary Statistics of Main Variables in the Country-Level Analyses

			Percentile	
	Mean	SD	10th	90th
PANEL A	Old World sample ($N = 121$)			
New civil conflict onsets per year, 1960–2017	0.025	0.033	0	0.069
Population diversity (ancestry adjusted)	0.74	0.018	0.71	0.75
Migratory distance from East Africa (in 10,000 km)	0.51	0.24	0.26	0.83
Absolute latitude	0.029	0.017	0.0065	0.052
Ruggedness	0.12	0.13	0.016	0.29
Mean elevation	0.61	0.58	0.11	1.27
Range of elevation	1.55	1.32	0.28	3.04
Mean land suitability	0.36	0.23	0.035	0.67
Range of land suitability	0.70	0.26	0.35	0.97
Distance to nearest waterway	0.38	0.48	0.039	1.04
Island nation dummy	0.033	0.18	0	0
Ethnic fractionalization	0.48	0.26	0.11	0.81
Ethnolinguistic polarization	0.49	0.22	0.18	0.75
Ever a U.K. colony dummy	0.26	0.44	0	1
Ever a French colony dummy	0.21	0.41	0	1
Ever a non-U.K./non-French colony dummy	0.20	0.40	0	1
British legal origin dummy	0.26	0.44	0	1
French legal origin dummy	0.40	0.49	0	1
Executive constraints, 1960–2017 average	3.98	1.87	1.68	7
Fraction of years under democracy, 1960–2017	0.37	0.38	0	1
Fraction of years under autocracy, 1960–2017	0.39	0.33	0	0.90
Oil or gas reserve discovery	0.67	0.47	0	1
Log population, 1960–2017 average	16.1	1.46	14.4	17.9
Log GDP per capita, 1960–2017 average	7.64	1.57	5.65	9.94
PANEL B	Global sample ($N = 147$)			
New civil conflict onsets per year, 1960–2017	0.022	0.031	0	0.064
Population diversity (ancestry adjusted)	0.73	0.027	0.69	0.75
Migratory distance from East Africa (in 10,000 km)	0.81	0.68	0.30	2.09
Absolute latitude	0.027	0.017	0.0060	0.051
Ruggedness	0.13	0.13	0.018	0.28
Mean elevation	0.59	0.55	0.10	1.25
Range of elevation	1.70	1.39	0.28	3.75
Mean land suitability	0.39	0.25	0.046	0.72
Range of land suitability	0.72	0.26	0.32	0.99
Distance to nearest waterway	0.35	0.46	0.036	1.01
Island nation dummy	0.048	0.21	0	0
Ethnic fractionalization	0.47	0.25	0.11	0.79
Ethnolinguistic polarization	0.45	0.24	0.097	0.75
Ever a U.K. colony dummy	0.26	0.44	0	1
Ever a French colony dummy	0.19	0.39	0	1
Ever a non-U.K./non-French colony dummy	0.32	0.47	0	1
British legal origin dummy	0.25	0.44	0	1
French legal origin dummy	0.46	0.50	0	1
Executive constraints, 1960–2017 average	4.14	1.83	1.84	7
Fraction of years under democracy, 1960–2017	0.41	0.38	0	1
Fraction of years under autocracy, 1960–2017	0.35	0.32	0	0.88
Oil or gas reserve discovery	0.67	0.47	0	1
Log population, 1960–2017 average	16.1	1.43	14.4	17.9
Log GDP per capita, 1960–2017 average	7.70	1.49	5.70	9.94

Data Appendix B Construction of the Georeferenced Dataset at the Ethnic-Group Level

This research constructs a novel geo-referenced data set of population diversity for a large number of ethnic groups across the globe. Two measures are constructed: (i) a measure of genetic diversity for 207 ethnic homelands for all individuals covered in the Pemberton et al. (2013) dataset that can be mapped to an ethnic homeland, and (ii) a measure of predicted population diversity for 901 ethnic homelands covered in the Geo-Referencing of Ethnic Groups (GREG) map of Weidmann et al. (2010).

The geo-referenced dataset for observed genetic diversity maps all 5,193 linkable individuals in the Pemberton et al. (2013) dataset into their ethnic homelands. This mapping results in a sample of 207 ethnic homelands for which, in addition to the measure of genetic diversity, spatial characteristics (e.g., geographic, climatic, and societal attributes) are available. Furthermore, using data on the spatial distribution of language areas in conjunction with data on the spatial distribution of population sizes, the study generates measures of linguistic fractionalization and polarization for each ethnic homeland. Finally, using gridded PRIO data (PRIO-GRID version 1.01) as reported by Tollefsen et al. (2012) based on the UCDP/PRIO Armed Conflict Dataset (Gleditsch et al., 2002) as well as data on UCDP Georeferenced conflict events (Sundberg et al., 2012; Croicu and Sundberg, 2015) the study generates a range of measures of conflict within each ethnic homeland.

The mapping of the 5,193 linkable individuals in the Pemberton et al. (2013) dataset into their ethnic homelands was based on the individual's ethnic identity, location, and geographical coordinates, where the polygons for the ethnic homelands were based on (i) polygons found in Murdock (1959) and digitized by Nunn (2008); Nunn and Wantchekon (2011), (ii) the Handbook of North American Indians (Heizer, 1978), (iii) Global Mapping International's World Language Mapping System (WLMS) (see http://worldgeodatasets.com/language), (iv) the Geo-Referencing of Ethnic Groups (GREG) map of Weidmann et al. (2010), and (v) the Database of Global Administrative Areas (GADM) map version 3.6 (gadm.org).

The geo-referenced dataset for predicted predicted population diversity for 901 ethnic homelands covered in the Geo-Referencing of Ethnic Groups (GREG) map of Weidmann et al. (2010) is constructed based on the migratory distance from Addis Ababa in East Africa to the centroid of the homeland.[1]

Data Appendix C Variable Definitions and Data Sources for the Subnational-Level Analyses

Conflict measures

1. **Conflict prevalence:** The average yearly share of the area of each ethnic homeland, over the period 1989–2008, that was within the boundaries of internal armed conflict event (between the government of a state and internal opposition groups). This measure is calculated using the gridded PRIO data (PRIO-GRID version 1.01) as reported by Tollefsen et al. (2012) based on the UCDP/PRIO Armed Conflict Dataset (Gleditsch et al., 2002).

[1] One homeland spanning territories in South America and Mauritius labeled "Indians of India and Pakistan" is excluded from the sample. The qualitative results would not be affected by the inclusion of this territory.

2. **Number of conflict events:** The number of conflict events within each ethnic homeland in the UCDP Georeferenced Event Dataset covering the period 1989–2017 (Sundberg et al., 2012; Croicu and Sundberg, 2015).

3. **Number of deaths:** The best (i.e., most likely) estimate of total fatalities resulting from a conflict event within each ethnic homeland in the UCDP Georeferenced Event Dataset covering the period 1989–2017 (Sundberg et al., 2012; Croicu and Sundberg, 2015).

4. **Number of deaths per event:** The number of deaths per event within each ethnic homeland in the UCDP Georeferenced Event Dataset covering the period 1989–2017 (Sundberg et al., 2012; Croicu and Sundberg, 2015).

Trust-related measures

1. **Intra-group trust (Africa):** The measure of an individual's trust in individuals from the same ethnic group in the 2005 Afrobarometer survey (3rd wave), as linked by Nunn and Wantchekon (2011) to the ethnicity names used in the Ethnographic Atlas. The measure takes the value 0 if the response to the question "How much do you trust each of the following types of people: People from your own ethnic group?" is "not at all", 1 if the response is "just a little", 2 if the value is "I trust them somewhat" and 3 if the value is "I trust them a lot".

2. **Slave exports (Africa):** A measure of the number of slaves taken from each ethnicity in transatlantic and Indian Ocean slave trades. The measure comes from Nunn and Wantchekon (2011) and is based on data from Nunn (2008).

3. **Other control variables (Africa):** The measures come from Nunn and Wantchekon (2011) and are based on data from 2005 Afrobarometer survey (3rd wave).

4. **Trust (US):** A measure of an individual's trust in people in general based on data from the General Social Survey 1972–2014 Release 6b Smith et al. (2018). The measure takes the value 1 if the response to the question "Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?" is "cannot trust", 2 if the response is "depends", and 3 if the value is "can trust".

Migratory distance and interpersonal population diversity

1. **Observed population diversity:** The expected heterozygosity (genetic diversity) of individuals in each of the 207 ethnic homelands, as calculated using Nei's formula (Nei, 1973), based on the individual-level data from Pemberton et al. (2013).

2. **Predicted population diversity:** The predicted level of population diversity of an ethnic homeland based on the migratory distance from East Africa to the centroid of the homeland, using the linear regression fit between observed population diversity and migratory distance from Addis Ababa obtained in sample of 207 ethnic homelands for which observed genetic diversity is available. The migratory distance from Addis is defined as the shortest traversable paths from Addis Ababa to the centroid of each ethnic group was computed. Given the limited

ability of humans to travel across large bodies of water, the traversable area included bodies of water at a distance of 100km from land mass (excluding migration from Africa into Europe via Italy or Spain).[2]

Control variables

1. **Linguistic fractionalization and polarization:** The degree of fractionalization in the ethnic homeland, using the formula $1 - \sum_i s_i^2$, and the degree of polarization in the ethnic homeland, using the formula $4 \sum_i s_i^2(1 - s_i)$, where s_i is an estimate of the population share of language group i in the homeland. Using the WLMS map of the spatial distribution of language areas in conjunction with the Gridded Population of the World dataset, the study estimates the number of individuals living in each intersection between ethnic homelands and language areas, assuming that population counts in overlapping language areas are equally split between these languages.

2. **Absolute latitude:** The absolute value of the latitude of an ethnic homeland's geodesic centroid, or, when the centroid is outside of the homeland, a representative interior point.

3. **Ruggedness:** The average level of the Terrain Ruggedness Index measure of Nunn and Puga (2012) across the grid cells that are located within a homeland.

4. **Mean and range of elevation:** The mean and range of elevation above sea level of an ethnic homeland, calculated using geospatial data from the *Atlas of the Biosphere* project (nelson.wisc.edu/sage/data-and-models/atlas/), across the grid cells that are located within a homeland.[3]

5. **Mean and range of land suitability:** The mean and range of the post-1500 optimal Caloric Suitability Index, measured by Galor and Özak (2016), across the grid cells that are located within a homeland.

6. **Island location:** A dummy variable indicating if the land type of an ethnic homeland's geodesic centroid (or a representative interior point) is a "small island" or a "very small island" as reported in the *World Countries* geographical dataset provided by ESRI (arcgis.com/home/item.html?id=ac80670eb213440ea5899bbf92a04998).

7. **Distance to nearest waterway:** The mean of the geodesic distance to the nearest coast or river, across the grid cells that are located within a homeland. Coastline locations are reported in the *Global Self-consistent, Hierarchical, High-resolution Geography Database* (http://soest.hawaii.edu/pwessel/gshhg). River locations are reported in the 1:10m *Natural Earth River + Lake Centerlines* dataset version 4 (http://naturalearthdata.com/downloads/10m-physical-vectors/10m-rivers-lake-centerlines).

[2]For the computation of predicted population diversity, distances to islands, where travel on water exceeds 100kms, are ignored since the Serial Founder Effect requires the serial foundation of populations along the migratory path and this was not feasible on water.

[3]The mean elevation can be negative in some cases due to the existence of places on land with elevation below sea level or the inclusion of territories at sea in the homeland polygon, for which the elevation is negative.

8. **Temperature:** The mean of the daily average temperature (in degree Celcius), across the grid cells that are located within a homeland, based on data from the CRU TS dataset version 3.21 for the period 1901–2012, as reported by Climate Research Unit (CRU) (Harris et al., 2014).

9. **Precipitation:** The mean of the annual total precipitation (in mm), across the grid cells that are located within a homeland, based on data from the CRU TS dataset version 3.21 for the period 1901–2012, as reported by Climate Research Unit (CRU) (Harris et al., 2014).

10. **Time since settlement:** The earliest year with a positive population count estimate in the ethnic homeland. Specifically, the study employs the population count data from the *History Database of the Global Environment* dataset version 3.1 (themasites.pbl.nl/tridion/en/themasites/hyde/download/index-2.html), described in Klein Goldewijk et al. (2010, 2011).

11. **Malaria:** The mean level of plasmodium falciparum malaria endemicity in 2010, across the grid cells that are located within a homeland. Specifically, the current study employs the data on the age-standardised plasmodium falciparum Parasite Rate from Gething et al. (2011). It represents the estimated proportion of 2–10 year olds in the general population that are infected with plasmodium falciparum, averaged over the months of 2010. The estimates are based on data from parasite rate surveys and a geostatistical model that produces a range of predicted endemicities for each location. The model includes environmental covariates which improves the accuracy of the prediction. The environmental covariates include rainfall, temperature, land cover and urban/rural status. The endemicity data reports the mean value for the probability distribution at each location (approx. $1km^2$).

12. **Oil or gas reserve discovery:** A time-constant dummy for the presence of at least one petroleum (oil or gas) reserve on the territory of an ethnic homeland. The variable is based on information provided in the Petroleum Dataset (version 1.2, dated 2009) covering the period 1946–2003 (Lujala et al., 2007). The dataset is compiled for the main purpose of investigating the relationship between armed civil conflict and natural resources. Each onshore petroleum reserve (oil or gas) – indicated as polygons in the shapefile accompanying the dataset – is assigned to an ethnic homeland using the coordinates of the centerpoints of the deposit polygons.

13. **Luminosity:** The mean level of cloud-free nighttime light intensity for the years 1992–2013, accross the grid cells that are located within a homeland. Specifically, the current study employs all available data in version 4 of the Defense Meteorological Satellite Program – Operational Linescan System (DMSP-OLS) Nighttime Lights Time Series (ngdc.noaa.gov/eog/dmsp/downloadV4composites.html). Since the log of zero is undefined, log luminosity is defined as the log of the sum of 0.001 and the luminosity measure.

TABLE DS.II: Summary Statistics of Variables in the Ethnic-Group-Level Analyses

			Percentile	
	Mean	SD	10th	90th
PANEL A	Observed population diversity sample ($N = 207$)			
Population diversity (observed)	0.72	0.045	0.65	0.76
Population diversity (predicted)	0.72	0.042	0.65	0.76
Conflict prevalence	0.14	0.27	0	0.63
Number of conflicts	1.04	2.78	0	3
Number of deaths (in thousands)	3.56	39.3	0	1.49
Ethnolinguistic fractionalization	0.26	0.30	0	0.74
Ethnolinguistic polarization	0.33	0.36	0	0.85
Absolute latitude	15.2	15.1	1.85	38.0
Ruggedness	133.4	144.1	14.7	299.8
Elevation	0.75	0.75	0.066	1.67
Range of elevation	1.60	1.25	0.31	3.36
Mean land suitability	8.50	3.50	3.69	12.4
Range of land suitability	5.09	4.42	0.36	11.8
Small island dummy	0.0097	0.098	0	0
Distance to nearest waterway	56.4	61.0	0	140.5
Temperature	21.1	7.79	8.94	27.2
Precipitation	123.1	100.3	31.3	285.7
Years since settlement (centuries from present)	104.9	31.9	40.2	120.2
Malaria	0.16	0.19	0	0.49
Oil or gas discovery	0.27	0.45	0	1
Luminosity	1.20	2.95	0	3.70
PANEL B	Predicted population diversity sample ($N = 901$)			
Population diversity (predicted)	0.71	0.042	0.64	0.75
Conflict prevalence	0.19	0.32	0	0.76
Number of conflicts	1.13	4.30	0	3
Number of deaths (in thousands)	2.22	20.7	0	1.62
Ethnolinguistic fractionalization	0.49	0.28	0.023	0.83
Ethnolinguistic polarization	0.55	0.28	0.045	0.87
Absolute latitude	21.7	17.1	2.92	48.2
Ruggedness	172.2	176.7	16.3	403.9
Elevation	0.73	0.86	0.069	1.75
Range of elevation	1.84	1.37	0.34	3.69
Mean land suitability	8.24	3.61	2.09	12.2
Range of land suitability	5.56	4.64	0.55	13.2
Small island dummy	0.026	0.16	0	0
Distance to nearest waterway	43.7	56.3	0	94.9
Temperature	18.8	9.36	3.83	26.7
Precipitation	118.8	75.5	32.6	225.7
Years since settlement (centuries from present)	112.5	23.6	90.2	120.2
Malaria	0.10	0.15	0	0.37
Oil or gas discovery	0.35	0.48	0	1
Luminosity	1.47	3.69	0.0012	3.76

TABLE DS.III: Summary Statistics of Variables in the Individual-Level Trust Analyses

			Percentile		
	Mean	SD	10th	90th	*N*
PANEL A	African sample				
Intra-group trust	1.52	1.00	0	3	3,212
Population diversity (observed)	0.76	0.0039	0.76	0.77	3,212
Age	35.8	14.5	20	58	3,212
Male	0.49	0.50	0	1	3,212
Ethnic fractionalization	0.27	0.28	0	0.72	3,212
Ethnolinguistic polarization	0.53	0.13	0.30	0.62	3,212
Proportion of ethnic group in district	0.73	0.33	0.12	1	3,212
School present	0.84	0.37	0	1	3,208
Electricity present	0.65	0.48	0	1	3,210
Piped water present	0.44	0.50	0	1	3,157
Sewage present	0.23	0.42	0	1	3,054
Health clinic present	0.58	0.49	0	1	3,060
Living in an urban area	0.44	0.50	0	1	3,212
Living condition categories	2.65	1.25	1	4	3,206
Education categories	3.51	2.10	0	6	3,207
Occupation categories	18.9	92.1	1	23	3,201
Religion categories	10.5	51.4	2	12	3,204
Slave exports (Atlantic and Indian)	277.4	262.5	0.17	666.0	3,212
PANEL B	US sample				
Trust	1.88	0.97	1	3	2,294
Population diversity (predicted)	0.72	0.024	0.67	0.74	2,294
GSS year	1993.9	10.6	1980	2010	2,294
Age	54.4	19.5	27	80	2,284
Sex	1.55	0.50	1	2	2,294
Family income categories	2.73	0.89	2	4	1,803
Religion categories	2.02	1.29	1	3	2,283
Highest educational degree categories	1.30	1.20	0	3	2,290
Ethnic fractionalization (ancestral)	0.23	0.18	0.11	0.54	2,294
Ethnolinguistic polarization (ancestral	0.41	0.21	0.12	0.67	2,294
Absolute latitude (ancestral)	46.1	11.8	23	60	2,294
Ruggedness (ancestral)	131.8	94.1	30.6	237.8	2,294
Mean elevation (ancestral)	436.4	339.3	105.8	1015.3	2,294
Mean land suitability (ancestral)	0.48	0.21	0.098	0.75	2,294
Range of land suitability (ancestral)	0.92	0.12	0.82	1.00	2,294
Distance to nearest waterway (ancestral)	223.0	496.4	29.4	332.6	2,294

References

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

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

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

Ashraf, Q. and O. Galor (2013a): "The "Out of Africa" Hypothesis, Human Genetic Diversity, and Comparative Economic Development," *American Economic Review*, 103, 1–48.

Banks, A. S. and K. A. Wilson (2018): "Cross-National Time-Series Data Archive [Data file]," Databanks International, Jerusalem, Israel. https://www.cntsdata.com/.

Bazzi, S. and C. Blattman (2014): "Economic Shocks and Conflict: Evidence from Commodity Prices," *American Economic Journal: Macroeconomics*, 6, 1–38.

Birnir, J. K., D. D. Laitin, J. Wilkenfeld, D. M. Waguespack, A. S. Hultquist, and T. R. Gurr (2018): "Introducing the AMAR (All Minorities at Risk) Data," *Journal of Conflict Resolution*, 62, 203–226.

Birnir, J. K., J. Wilkenfeld, J. D. Fearon, D. D. Laitin, T. R. Gurr, D. Brancati, S. M. Saideman, A. Pate, and A. S. Hultquist (2015): "Socially Relevant Ethnic Groups, Ethnic Structure, and AMAR," *Journal of Peace Research*, 52, 110–115.

Bockstette, V., A. Chanda, and L. Putterman (2002): "States and Markets: The Advantage of an Early Start," *Journal of Economic Growth*, 7, 347–369.

Brecke, P. (1999): "Violent Conflicts 1400 A.D. to the Present in Different Regions of the World," Paper presented at the 1999 Annual Meeting of the Peace Science Society, October 8–10.

Central Intelligence Agency (2018): "The World Factbook," The Central Intelligence Agency, Washington, DC. Data retrieved at https://www.cia.gov/library/publications/the-world-factbook/.

Chandler, T. (1987): *Four Thousand Years of Urban Growth: An Historical Census*, Lewiston, NY: The Edwin Mellen Press.

Cioffi-Revilla, C. (1996): "Origins and Evolution of War and Politics," *International Studies Quarterly*, 40, 1–22.

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

Croicu, M. and R. Sundberg (2015): "UCDP Georeferenced Event Dataset Codebook version 4.0," Department of Peace and Conflict Research, Uppsala University. http://ucdp.uu.se/downloads/ged/ucdp-ged-40-codebook.pdf.

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

Dincecco, M., J. Fenske, and M. G. Onorato (2015): "Is Africa Different? Historical Conflict and State Development," IMT Lucca EIC Working Paper No. 08/2015, IMT Institute for Advance Studies Lucca.

Esteban, J., L. Mayoral, and D. Ray (2012): "Ethnicity and Conflict: An Empirical Study," *American Economic Review*, 102, 1310–1342.

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

Fearon, J. D. and D. D. Laitin (2003): "Ethnicity, Insurgency, and Civil War," *American Political Science Review*, 97, 75–90.

Fenske, J. (2014): "Ecology, Trade and States in Pre-Colonial Africa," *Journal of the European Economic Association*, 12, 612–640.

Fincher, C. L. and R. Thornhill (2008): "Assortative Sociality, Limited Dispersal, Infectious Disease and the Genesis of the Global Pattern of Religion Diversity," *Proceedings of the Royal Society B: Biological Sciences*, 275, 2587–2594.

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

Galor, O. and Ö. Özak (2016): "The Agricultural Origins of Time Preference," *American Economic Review*, 106, 3064–3103.

Gething, P. W., A. P. Patil, D. L. Smith, C. A. Guerra, I. R. Elyazar, G. L. Johnston, A. J. Tatem, and S. I. Hay (2011): "A New World Malaria Map: *Plasmodium falciparum* Endemicity in 2010," *Malaria Journal*, 10.

Gleditsch, N. P., P. Wallensteen, M. Eriksson, M. Sollenberg, and H. Strand (2002): "Armed Conflict 1946-2001: A New Dataset," *Journal of Peace Research*, 39, 615–637.

Harris, I., P. D. Jones, T. J. Osborn, and D. H. Lister (2014): "Updated High-Resolution Grids of Monthly Climatic Observations – The CRU TS3.10 Dataset," *International Journal of Climatology*, 34, 623–642.

Harris, I. C. and P. D. Jones (2013): "CRU TS3.21: Climatic Research Unit (CRU) Time-Series (TS) Version 3.21 of High Resolution Gridded Data of Month-by-Month Variation in Climate (Jan. 1901 - Dec. 2012) [Data file]," University of East Anglia Climatic Research Unit, NCAS British Atmospheric Data Centre, 24 September 2013. doi:10.5285/D0E1585D-3417-485F-87AE-4FCECF10A992.

——— (2017): "CRU CY4.01: Climatic Research Unit (CRU) Year-by-Year Variation of Selected Climate Variables by Country (CY) version 4.01 (Jan. 1901 - Dec. 2016) [Data file]," University of East Anglia Climatic Research Unit, Centre for Environmental Data Analysis, 4 December 2017. doi:10.5285/d4e823f0172947c5ae6e6b265656c273.

Hastings, D. A., P. K. Dunbar, G. M. Elphingstone, M. Bootz, H. Murakami, H. Maruyama, H. Masaharu, P. Holland, J. Payne, N. A. Bryant, et al. (1999): "The Global Land One-kilometer Base Elevation (GLOBE) Digital Elevation Model, Version 1.0," National Oceanic and Atmospheric Administration, National Geophysical Data Center, Boulder, CO. Data retrieved at https://www.ngdc.noaa.gov/mgg/topo/globe.html.

Hegre, H. and N. Sambanis (2006): "Sensitivity Analysis of Empirical Results on Civil War Onset," *Journal of Conflict Resolution*, 50, 508–535.

Heizer, R. F. (1978): *Handbook of North American Indians, Vol. 8: California*, Washington, DC: Smithsonian Institution.

Klein Goldewijk, K., A. Beusen, and P. Janssen (2010): "Long-Term Dynamic Modeling of Global Population and Built-Up Area in a Spatially Explicit Way: HYDE 3.1," *The Holocene*, 20, 565–573.

Klein Goldewijk, K., A. Beusen, G. van Drecht, and M. de Vos (2011): "The HYDE 3.1 Spatially Explicit Database of Human-Induced Global Land-Use Change Over the Past 12,000 Years," *Global Ecology and Biogeography*, 20, 73–86.

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

Lujala, P., J. Ketil Rod, and N. Thieme (2007): "Fighting over Oil: Introducing a New Dataset," *Conflict Management and Peace Science*, 24, 239–256.

Marshall, M. G. (1999): *Third World War*, Lanham, MD: Rowman & Littlefield Publishers.

——— (2002): "Measuring the Societal Impact of War," in *From Reaction to Conflict Prevention: Opportunities for the UN System*, ed. by F. O. Hampson and D. M. Malone, Boulder, CO: Lynne Rienner, 63–105.

——— (2017): "Major Episodes of Political Violence (MEPV) and Conflict Regions, 1946–2017," Center for Systemic Peace, Vienna, VA. Data retrieved at http://www.systemicpeace.org/inscrdata.html.

Marshall, M. G., T. R. Gurr, and K. Jaggers (2017): "Polity IV Project: Political Regime Characteristics and Transitions, 1800–2017," Center for Systemic Peace, Vienna, VA. Data retrieved at http://www.systemicpeace.org/inscrdata.html.

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

Modelski, G. (2003): *World Cities: -3000 to 2000*, Washington, DC: FAROS 2000.

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

Murdock, G. P. (1959): *Africa: Its Peoples and Their Culture History*, New York, NY: McGraw-Hill Book Co., Inc.

Nei, M. (1973): "Analysis of Gene Diversity in Subdivided Populations," *Proceedings of the National Academy of Sciences*, 70, 3321–3323.

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, 1–25.

Nordhaus, W. D. (2006): "Geography and Macroeconomics: New Data and New Findings," *Proceedings of the National Academy of Sciences*, 103, 3510–3517.

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

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

Nunn, N. and L. Wantchekon (2011): "The Slave Trade and the Origins of Mistrust in Africa," *American Economic Review*, 101, 3221–3252.

Özak, Ö. (2010): "The Voyage of Homo-Economicus: Some Economic Measures of Distance," Unpublished manuscript. Department of Economics, Southern Methodist University.

Pemberton, T. J., M. DeGiorgio, and N. A. Rosenberg (2013): "Population Structure in a Comprehensive Genomic Data Set on Human Microsatellite Variation," *G3: Genes, Genomes, and Genetics*, 3, 891–907.

Pettersson, T. and K. Eck (2018): "Organized Violence, 1989–2017," *Journal of Peace Research*, 55, 535–547.

Putterman, L. (2008): "Agriculture, Diffusion, and Development: Ripple Effects of the Neolithic Revolution," *Economica*, 75, 729–748.

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

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza (2005): "Support from the Relationship of Genetic and Geographic Distance in Human Populations for a Serial Founder Effect Originating in Africa," *Proceedings of the National Academy of Sciences*, 102, 15942–15947.

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, 377–392.

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

Riley, S. J., S. D. DeGloria, and R. Elliot (1999): "A Terrain Ruggedness Index that Quantifies Topographic Heterogeneity," *Intermountain Journal of Sciences*, 5, 23–27.

Smith, T. W., M. Davern, J. Freese, and S. L. Morgan (2018): "General Social Surveys, 1972–2018 [Data file]," National Data Program for the Social Sciences, Chicago, IL. Data retrieved at gss.norc.org.

Sundberg, R., K. Eck, and J. Kreutz (2012): "Introducing the UCDP Non-State Conflict Dataset," *Journal of Peace Research*, 49, 351–362.

Tollefsen, A. F., H. Strand, and H. Buhaug (2012): "PRIO-GRID: A Unified Spatial Data Structure," *Journal of Peace Research*, 49, 363–374.

Weidmann, N. B., J. K. Rød, and L.-E. Cederman (2010): "Representing Ethnic Groups in Space: A New Dataset," *Journal of Peace Research*, 47, 491–499.

World Bank (2018): "World Development Indicators," The World Bank, Washington, DC. Data retrieved at https://datacatalog.worldbank.org/dataset/world-development-indicators.

World Values Survey (2006): "European and World Values Surveys, Four-Wave Integrated Data File, 1981–2004, version 20060423," The World Values Survey Association, Stockholm, Sweden. Data retrieved at http://www.worldvaluessurvey.org.

——— (2009): "World Values Survey, 1981–2008 Official Aggregate, version 20090914," The World Values Survey Association, Stockholm, Sweden. Data retrieved at http://www.worldvaluessurvey.org.