Diversity, Conflict and Agglomeration in African Cities *

Andre Gray,

Faculty Advisor: Sam Bazzi, Sara Lowes

October 11, 2024

Abstract

As African countries urbanize, migrants from disparate cultures are meeting and working in close proximity in burgeoning cities. A literature in political economy suggests ethnic diversity might limit economic growth and increase conflict, while a long tradition in urban economics highlights the benefits of density and agglomeration. How does ethnic diversity impact the returns to urbanization in Africa? Using historical information on ethnic diversity and plausibly exogenous shocks to regional productivity, I study how the ethnic mix of regions affects city growth and development. I model ethnic diversity as a labor supply problem for growing cities, where regions must draw labor from surrounding areas, trading off agglomeration benefits with the congestion force of potential conflict arising from a diverse mix of workers. I show that cities that emerge in more ethnically homogeneous regions benefit from higher light density and lower conflict today. I then explore the implications of these labor supply constraints for contemporary climate-induced migration.

*Working Draft, Please Do Not Circulate

Contents

1 Introduction **4**

2 City Growth, Ethnic Diversity and Labor Supply **5**

3 Literature Review **7**

4 Data **8**
- 4.1 Ethnic and Linguistic Diversity . . . 8
- 4.2 Population and City Locations . . . 8
- 4.3 Development Outcomes . . . 9
- 4.4 DHS Level Outcomes . . . 9
- 4.5 Geographic Variables . . . 9

5 Empirical Strategy **10**
- 5.1 Measures of Ethnic Diversity . . . 11
- 5.2 Correlations of Diversity and Development . . . 12
- 5.3 Correlations of Diversity and Geographic Covariates . . . 14
- 5.4 IV Strategy 1: Railroad Towns and Least-Cost Path . . . 16
- 5.5 IV Strategy 2: Portage Sites . . . 20

6 Results **22**
- 6.1 Rail IV Results . . . 23
  - 6.1.1 First Stage . . . 23
  - 6.1.2 Second Stage . . . 23
- 6.2 Portage Score IV Results . . . 27
  - 6.2.1 First Stage . . . 27
  - 6.2.2 Second Stage . . . 28

7 Climate and Diversity **30**
- 7.1 Measuring Moments of Drought Shocks . . . 31
- 7.2 Cross-Sectional Evidence with Drought Instruments . . . 31
- 7.3 Panel Evidence with Drought Instruments . . . 34
- 7.4 2SLS Results with Drought Instruments . . . 34

8 City Size and Diversity **36**
- 8.1 Heterogeneity by City Size . . . 36
- 8.2 Occupational Segregation and City Size . . . 37
- 8.3 Residential Segregation and City Size . . . 41

9 Conclusion **41**

A Model **48**
A.1 Labor Demand 48
A.2 Labor Supply 48
A.3 Production 49
A.4 Identifying Parameters with Productivity Shocks 50
A.5 Computing Spatial Equilibrium 50

B Portage Score Validation **51**
B.1 Calculating Portage Score 51
B.2 Validating Portage Score with Hydrological Data 51

C Additional Figures & Tables **52**

1 Introduction

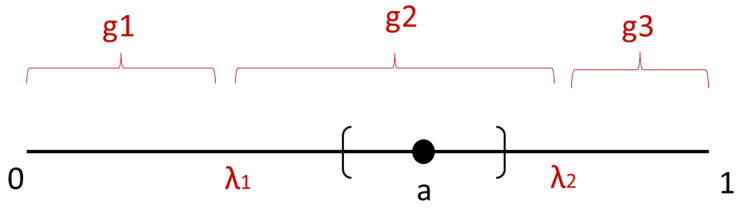
Africa is marked by its diversity. Nigeria alone contains upwards of 300 ethnic groups and languages. In turn urbanization in Africa has brought together an array of cultures in large, dense cities like Lagos, Addis Ababa and Kinshasa. Work in the political economy of development has pointed to Africa's ethnic diversity as a source of conflict, where groups compete for resources and political power within colonial borders. At the same time, work in urban economics stresses the many benefits of dense cities, and developed states often boast large cosmopolitan cities as growth centers.

Given these competing narratives across fields about diversity, density and growth, a natural question is: does diversity help or hinder development in African cities? We may think that the positive benefits of ethnic diversity outweigh the negatives. Diversity may drive city growth through the love of variety or ethnicity-specific knowledge and ideas that boost technological change (Montalvo and Reynal-Querol, 2021; Mueller et al., 2022; Ashraf and Galor, 2013). On the other hand, ethnic diversity may work against agglomeration as a congestion force, dampening the returns to density by increasing the probability of conflict or limiting the productivity of team-based labor (Hjort, 2014). Lastly there may be size effects. Perhaps large cities are able to manage their ethnic diversity through workers' segregated sorting or firm-ethnicity specialization, producing an inverted U-shaped relationship between diversity and growth.

To explore the effect of diversity on city development, I compare the trajectories of different urban centers that emerged in regions with more or less ethnic heterogeneity. I leverage regional productivity shocks from the colonial era to create exogenous variation in a region's propensity to become a city, unrelated to its underlying ethnic make-up. In particular I leverage the fact that cities were more likely to emerge near colonial rail lines built along the least-cost path between a coastal town and an inland resource. As a second identification strategy, I use portage sites along inland rivers as an instrument for contemporary city location. In a second stage I measure the impact of regional diversity in a city on nighttime light density and conflict intensity.

I find that cities that emerged in more diverse locations have lower nighttime light density and more conflict today, suggesting that city growth is constrained by the ethnic mix of the labor supply they draw from. This finding motivates a model of urban growth in which urban centers must draw workers from neighboring regions. Due to migration costs, cities are constrained to drawing labor supply from nearby areas, which may be more or less diverse ex-ante. Cities trade-off demand for new workers and the agglomeration benefits of density with a congestion force of ethnic diversity, measured by the ethnic fractionalization of the urban center.

Figure 1: Example of City and Worker Location with 3 Ethnic Groups



Note: Workers are uniformly distributed, but total ethnic group sizes may vary. The size of the population that move to a is governed by the offered wage p which determines the catchment area for the city.

2 City Growth, Ethnic Diversity and Labor Supply

As a motivation for the empirical approach and the spatial model, consider a region on a line with a continuum of potential migrant workers distributed uniformly from 0 to 1. Workers differ only by their ethnicity $g \in (g1, g2, g3)$. The ethnic groups are spatially segregated along the line, and may be of different relative sizes. The relative proportions of each worker type are parametrized by λ_1, λ_2.

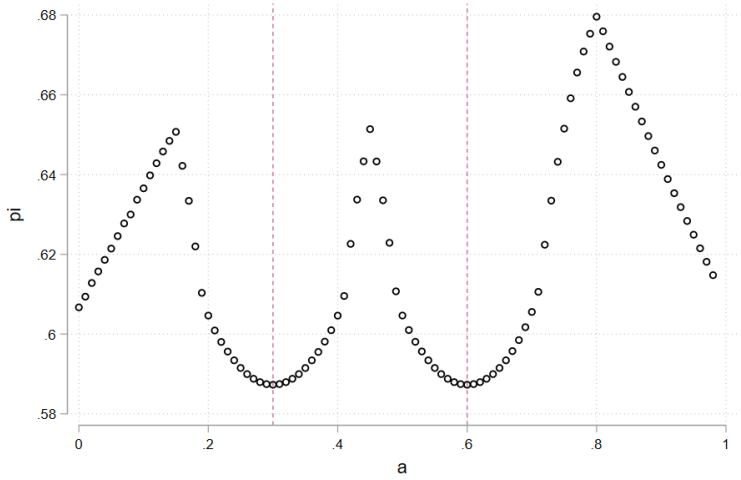
Consider a city a that is placed randomly along this line, and must draw workers from nearby regions. The city produces a normal good according to AL^{α}, and pays wage p. Workers must choose whether to migrate to the city at wage p, or take the reservation wage in their home region. A worker located at x that chooses to move to the city pays a migration cost proportional to their distance, $t(|a-x|)$.

Ethnic diversity can enter either in the city's production function or in the worker's migration decision. We can think of a fractionalization index F that measures ethnic diversity as a function of the proportion of workers that migrate to the city form each group w_1, w_2, w_3. Examples of this kind of function include the standard fractionalization index $\text{F} = \sum_1^3 \frac{w_i}{L}(1-\frac{w_i}{L})$, or the Herfindahl-Hirschman Index (HHI) $\text{F} = \sum_1^3 (\frac{w_i}{L})^2$.

Diversity index F can be incorporated into production by setting $A = \bar{A}L^{\gamma}F^{-}\upsilon$. The diversity can also be incorporated into the worker's decision. Suppose a worker at location x from group i has gains from migration $p - t(|a-x|) - \upsilon F$. Note that the effect of F on production or migration may be positive or negative. It may be that ethnic diversity increases production through the introduction of new ideas, independent of the total agglomeration benefits from density L. Ethnic diversity may also play a negative role by increasing ethnic conflict and urban violence in the city, dampening production or reducing the amenity value for residents.

If ethnic diversity is a negative force, total city production is constrained by the L that the city is able to recruit for a given wage p. The city's ability to recruit works at a given

Figure 2: Simulation of City Profits and Location in Linear Region



Note: Simulation with migration cost $t = 1$, fractionalization index weighted by $\beta = 0.2$, production $y = L^{0.3}$, and three ethnic groups parsed at 0.3 and 0.6.

wage is is a function of the underlying ethnic diversity in the vicinity of the city, ie. the placement of a relative to λ_1, λ_2. A city placed in the middle of the territory of ethnic group g_2 can recruit more workers at a given p due to reduced diversity congestion costs, relative to a city placed at the border between g_2, g_3. If city a experiences a shock to total productivity A, the resulting increase in L is also a function of where the city is located along the distribution of ethnicities. Figure 2 shows simulations from an example of the model where a cost of ethnic fractionalization F is imposed on the production function of city. The figure shows that as we vary the location of city a along the line, total profit of the city varies. Cities located at the center of ethnic territories benefit from a homogeneous worker population and are therefore larger and more profitable.

My empirical strategy leverages shocks to regional productivity that are exogenous to the underlying ethnic distribution of workers in the surrounding area. I use these historical shocks to instrument for where cities are located across space, which amounts to varying a and comparing development outcomes across cities with different underlying levels of diversity in their potential labor supply. The example provided above can be extended into a more general framework, as a spatial equilibrium model with many worker types, many regions and endgenous amenities. I describe this model in detail in Appendix Section A.

3 Literature Review

This paper contributes to a literature on the role of diversity in economic development (Arbatli et al., 2020; Alesina and Ferrara, 2005; Mueller et al., 2022; Esteban et al., 2012; Gisselquist et al., 2016; Gören, 2014; Adhvaryu et al., 2021). Work in this area has generally focused on the impact of ethnic diversity at the state or district level, leveraging variation from colonial borders or redrawn political boundaries to study the effect of diversity on either conflict, as measured by number of armed conflicts, or economic growth. Papers on Africa in particular find that ethnic diversity, as measured by fractionalization or polarization indexes, is associated with higher levels of conflict and lower economic growth.

Montalvo and Reynal-Querol (2021) reconsider this conclusion by focusing their analysis at the city level rather than the state, finding a positive correlation between ethnic diversity and growth. Using geolocated trading markets from Porteous (2019), they suggest that the positive returns at the local level are driven by cross-ethnic trade. The role of ethnic borderlands in the facilitation of trade is complex. A fairly large literature in trade suggests that cultural and ethnic similarity helps to facilitate trade across regions and national borders (Aker et al., 2014; Melitz and Toubal, 2014). However, if particular ethnic groups specialize in different goods, we might expect that markets and eventually cities emerged at these intersections between territories. One of the goals of this paper is to bring causal inference to this discussion. I add to this previous work by considering the role diversity plays in dampening benefits to agglomeration. By exploiting exogenous variation in city location, I show how ex-ante diversity may hinder city growth and development.

This paper also contributes to a literature in urban economics that considers the role of labor supply constraints and labor heterogeneity on city growth and wages (Almagro and Dominguez-Iino, 2022; Diamond, 2016; Duranton and Puga, 2020; Monte et al., 2018). The paper by Diamond (2016) considers the effect of worker preferences for high-skilled neighbors on the endogenous amenity value of regions. In this paper, I consider how the ethnicity mix of workers may produce a kind of endogenous amenity value that impacts migration decisions. Monte et al. (2018) study the role of transport infrastructure in relieving local labor supply constraints, which in turn affects the benefits of productivity shocks. In my paper I study the role of ethnicity-driven labor supply constraints on a city's ability to benefit from regional productivity shocks.

Lastly this paper contributes to a literature on the location of cities, and the role that path-dependence and geography play in determining where cities emerge (Bleakley and Lin, 2015; Ullman, 1970; Michaels et al., 2012; Bleakley and Lin, 2012; Harari, 2020). In my paper I leverage the placement of colonial rail lines and geographically determined portage sites to identify regions that became cities independent of the particular ethnic mix of the region.

4 Data

4.1 Ethnic and Linguistic Diversity

To understand the historical distribution of ethnic groups across space, I use a variety of maps that record the approximate boundaries of ethnic tribes historically across Africa. The first is the Ethnographic Atlas, an anthropological database that charts polygons of historic ethnic groups across Africa (Murdock, 1967). The second is the "Geo-referencing of ethnic groups" (GREG) dataset, which was assembled using the Soviet Atlas Narodov Mira (ANM) (Weidmann et al., 2010). Both of these sources were created by anthropologists in the 1960s, and are meant to be representative of the precolonial arrangement of ethnic groups across Africa. These maps have been used by economists to study the impact of cultural traits on long-term economic outcomes (McGuirk and Nunn, 2024; Lowes, 2017).

I also use the Ethnologue, a database of world languages that includes a map of the distribution of commonly spoken languages across Africa (Paolillo and Das, 2006; Gershman and Rivera, 2018). This is a more contemporary source of linguistic variation, and therefore reflects sorting of groups overtime. However to the extent that indigenous languages are persistent and local, we might expect this map to at least partly cohere with the historic maps. As a check against these sources I create a measure of contemporary ethnic diversity using georeferenced surveys from the Demographic and Health Surveys, which often ask questions about the respondent's ethnic background. I also create a county level measure of diversity using available 10% census data for a subset of African countries which report ethnicity.

4.2 Population and City Locations

I use three sources of population data at granular spatial levels. Historical data on population in Africa is limited, and relies on a combination of colonial censuses and interpolation. My main source is Africapolis, a database of African cities that tracks the spatial distribution of human settlements greater than 10k people across Africa. Africapolis has a record of city and town locations and estimated population counts that goes back to 1950 (Heinrigs, 2020). Grids that overlap with the Africapolis layer are marked as cities in my dataset. I supplement this source with Worldpop data, which provides population counts at 1km resolution, and the History Database of the Global Environment (HYDE). HYDE provides spatial data on land use change over time, and includes estimated population counts in precolonial periods.

4.3 Development Outcomes

Beyond population counts, my outcomes of interest are levels of development and levels of conflict across cities and time. As a granular measure of regional development I use a harmonized dataset of nighttime lights from DMSP and VIIRS, which has light density across Africa from 1992 to 2013 (Li et al., 2020). To measure conflict intensity I use the UCDP event dataset, which contains georeferenced information on conflict events from 1975 to 2021 (Sundberg and Melander, 2013). Conflict events in this data consist of battles between two organized groups, typically involving the state military and an insurgent group. Of course, not all conflicts in this dataset are ethnically motivated. While the UCDP data does not explicitly code conflict events as ethnically motivated, each battle event is linked to a source article. In a robustness check I code conflict events by whether the source article mentions words such as "ethnic", "race" or "tribe".

For each grid, I aggregate an average light density measure for each year between 1992 and 2013, as well as cumulative measures for 1992-1999, 2000-2009 and 2010-2013. For conflict events, I create a probability of conflict measure that represents the probability that a conflict event falls within the grid across the period 1975 to 2021. This is simply the average number of years in which that grid observes a conflict event. I also calculate the average number of conflict deaths across this time period for each grid.

The last development outcome I leverage is colonial railroads. The universe of colonial rail projects was collected by Jedwab and Moradi (2016).

4.4 DHS Level Outcomes

As another measure of contemporary regional wealth, I take advantage of the geolocated Demographic and Health Survey (DHS) data from IPUMS. Leveraging all available surveys from African states, I create a measure of durables consumption at the household level by taking the principle component of a variety of household assets. This measure provides a proxy for consumption in a given region, following Gollin et al. (2021). Analysis using this durables consumption measure is done at the DHS cluster level, rather than using all grid observations.

4.5 Geographic Variables

In order to create an index of portage site propensity, I use data on major African river networks from the HydroSHEDS database (Lehner and Grill, 2013). I also use data on elevation variation across the continent, which is measured with the ruggedness index from Nunn and Puga (2012). Supplementary geographic information at the grid level includes malaria ecology (Kiszewski et al., 2004) and soil suitability (Ramankutty et al., 2002).

5 Empirical Strategy

The empirical strategy takes inspiration from the literature on estimating labor demand curves from shocks to labor demand (Diamond, 2016; Notowidigdo, 2020) and housing supply elasticities (Saiz, 2010; Guedes et al., 2023). In this work inverse demand and supply elasticities are estimated using an interaction between a labor demand shock and a housing supply constraint.

Initially we are interested in an equation like the following. For each grid, what is the long-term relationship between diversity, population and GDP?

$$Log(Y)_i = \beta_0 + \beta_1 L_i + \beta_2 D_i + \beta_3 L_i * D_i + X_i + \epsilon_i \tag{1}$$

Where D_i captures a historic measure of a region's potential exposure to diverse migrants, and L_i captures a measure of employment density. The interaction of fractionalization and labor is what we're interested in – how does historic fractionalization affect the labor demand elasticity, and in turn output and productivity?

There are a few things to note about this equation.

1. A historic fractionalization measure will be related to a variety of geographic fundamentals that governed the migration of groups over space and may also affect productivity or GDP
2. If we're using night lights to measure GDP, then this equation is mainly capturing some facts about urbanization –the relationship between rural productivity and night lights is more tenuous. (Pérez-Sindín et al., 2021)
3. D_i may have direct effects on GDP, β_2, but it will also move total labor through its effects on labor demand elasticity. Regressing an outcome like $GDP/capita$ obscures this endogenous relationship that in part governs the selection effects taking place between the initial distribution of groups and today's economic output.

Of course population density is an endogenous variable. Our IV strategy will predict population density and its interaction by exploiting temporary shocks to regional productivity ΔA_i that drove labor demand historically, but that are no longer correlated with unobserved productivity fundamentals today.

In our first pass analysis, we'll use a dummy that marks a "city" as our indicator for employment density. A city is defined as a town with more than 10k in population, as recorded by the Africapolis dataset. This helps us to avoid having to use interpolated population measures that are in part backed out from light density estimates. Our IV instruments will predict city locations, allowing us to estimate our parameter of interest β_3:

$$Log(Y)_i = \beta_0 + \beta_1 \hat{C}_i + \beta_2 D_i + \beta_3 \widehat{C_i * D_i} + X_i + \epsilon_i \tag{2}$$

We need our productivity shock ΔA_i to be uncorrelated with unobserved other factors in ϵ_i that drive our outcomes of light density, wealth, lights/capita,and conflict. Note we do not need that D_i is uncorrelated with ΔA_i in this estimation. Our concern will be if we are missing an interaction term of an unobserved fundamental that is correlated with $\widehat{C_i * D_i}$. For example, if $Ruggedness_i * \Delta A_i$ is correlated with $\widehat{C_i * D_i}$ and the outcome, we need to include this term in the estimation.

I split the African continent into equally sized hexagonal grids of approximately $1200km^2$, which I use as my regions i. For each grid I aggregate data on conflict, light density, population across years and the geographic variables including malaria suitability, ruggedness and soil suitability. Rather than limit the aggregation to data falling within the grid region itself, I create a buffer zone with a radius of 50km from each grid, and aggregate population, lights and geographic data within this area. In the Appendix I do the analysis at other buffer distances, including 20km and 100km. A grid is coded as urban if it overlaps with a city identified in the Africapolis data set. Because the data includes small towns and settlements, I define a settlement as a city if it hosts greater than 10k people for any year in the dataset. In some of the analysis that follows, I will restrict to the sample to "new cities". These are cities that specifically emerge in the dataset after 1960, so we can think of them as colonial or post-colonial settlements, which are the sites we expect to be affected by colonial era shocks.

5.1 Measures of Ethnic Diversity

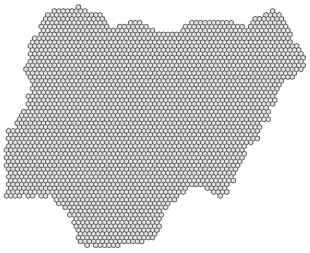
Using the various ethnic maps, I first calculate the density of nearby ethnic groups for each grid as the number of different groups that intersect within a radius of the grid centroid. I set the radius at 20, 50 and 100km. Using the same radii, I also calculate a fractionalization index that measures the share of land area taken up by different ethnicities. In particular it measures the probability that two randomly placed points in the region land in the territory of two different ethnic tribes.

In the main results, I will use two diversity measures: the 50km specification of the Murdock fractionalization index, and the first principal component of the fractionalization measures across data sources including the Murdock map, GREG atlas and Ethnologue (at the 50km boundary level). Results for other specifications are provided in the Appendix. Both of these measures are standardized in the regression analysis. The correlation between the various data sources is shown in Table B2. While the measures are indeed correlated as expected, it's important to note that correlation across data sources only ranges between 0.3 and 0.6, suggesting that at least at a sub-country levels the marked ethnic boundaries are different across sources.

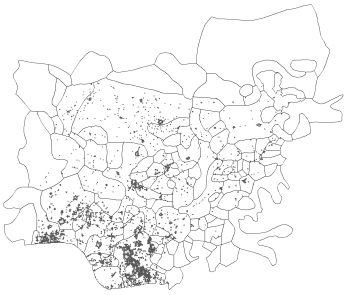
For the subsample of grids that overlap with available census data from IPUMS International I include them in the correlation matrix in Table B3. Across measures we see the

Figure 3: Examples of Data Visualizations, Nigeria

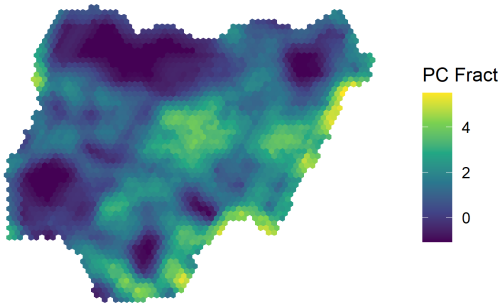
(a) Murdock Map, 1967



(b) Grid View



(c) Towns/Settlements > 10k, Africapolis



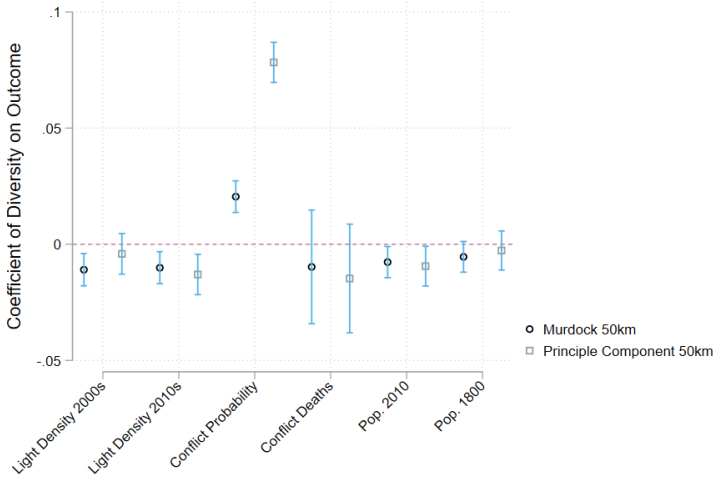
(d) Principle Component Fractionalization

higher diversity is correlated with lower county-level ethnic concentration (see equation 16 for how this ethnic concentration measure is calculated in the census data). The relationship between the contemporary census-measured ethnic diversity and the historical measures of diversity is suggestive of persistence in ethnic group locations over time, as suggested by Gershman and Rivera (2018).

5.2 Correlations of Diversity and Development

Figure 4 shows the aggregate relationship between the main measures of ethnic diversity, light density and population size at the grid level. In particular the graph shows beta coefficients from the regression of outcome y on diversity measure D at the grid level, with country fixed effects and controls for malaria, soil suitability, whether the region is a city, and ruggedness. There's some suggestive evidence of a relationship between light density and historic ethnic diversity, and a strong relationship between historic diversity and conflict.

Figure 4: Association of Diversity with Lights and Conflict



Notes: These coefficients are estimated from the regression $y = \alpha + \beta D + X + \upsilon_s$, where D is a standardized measure of diversity either Murdock fractionalization or the principle component. Covariates X include malaria suitability, land suitability and ruggedness. υ_s are state fixed effects. Light Density 2000s is the average light density for a grid cell from 2000-2009, and Light Density 2010s is the same measure for 2010-2013. Conflict probability is the proportion of years where a conflict event is observed in that grid from 1975-2021. Average deaths is the average death toll across conflicts occurring in a grid cell across that time period (this is only defined for grids that have a positive conflict probability). All outcomes are standardized.

It also appears the grid's that are more historically diverse have lower population in 2010, while this effect is not seen on the historic population measure for 1800. Figure B3 in the Appendix shows regressions of each independent diversity measure on the average light density for the years 2000-2013.

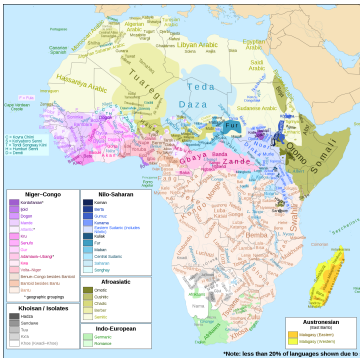
These broad patterns seem to match past associations in the literature, which typically find a strong relationship between regional diversity and conflict, while the relationship between diversity and growth is mixed, and seems to depend on the size of the analysis unit (Arbatli et al., 2020; Gören, 2014; Montalvo and Reynal-Querol, 2021). But these associations don't address the key question about the role of ethnic diversity in city growth. For this we need to compare regions that have high productivity (potential for city growth), but different levels of heterogeneity in the prospective local workforce. To do this I use two identification strategies that attempt to isolate geographic variation in city location, such that the interaction of the instrument and underlying historic ethnic diversity is unrelated to unobserved productivity fundamentals. In the first strategy, I consider regions that benefitted from the placement of a nearby colonial railroad, boosting market access. For the second strategy I consider regions located near portage sites, which became important areas for trade in the colonial era.

5.3 Correlations of Diversity and Geographic Covariates

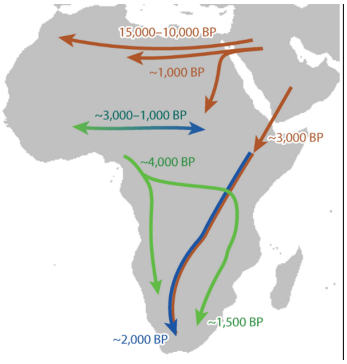
The extreme diversity of ethnic groups and languages across the African continent was set by a series of hypothesized major population migrations, including the Bantu expansion from West Africa (3000 BCE to 500 AD) and the Eurasian backflow into the Horn of Africa (circa 1000 BCE). Demographers believe these migrations followed a semi-founder pattern, where environmental factors determined how long groups stayed in particular places, before continuing south through the continent (Fortes-Lima et al., 2024; Michalopoulos, 2012; Semo et al., 2020). Geographical factors largely determined the routes these migrations, and where groups may have splintered off.

This means that ethnic diversity will also necessarily be a function of geographical factors such as elevation, ruggedness, land suitability, temperature and distance to rivers. This should be true at various spatial resolutions, and it's not obious if the correlation is more or less strong for state, subdistrict, or more granular measures of ethnic diversity. Table 1 shows regressions of our various historic measures of ethnic diversity on geographic covariates. We see that at our grid-cell level, higher ethnic diversity is associated with higher ruggedness, closeness to rivers, malaria suitability and agricultural suitability. We control for all these factors and their interactions with the instrument throughout the analysis.

Figure 5: African Diversity and Historical Migrations



(a) Language Families



(b) (Schlebusch and Jakobsson, 2018)

Table 1: Geography and Ethnic Diversity

	PC Fract	Murd Fract	Murd Count	Greg Count	Lang.
Dist River	-0.229	-0.157	-0.175	-0.426	0.080
	[0.005]***	[0.006]***	[0.006]***	[0.014]***	[0.002]***
Dist Coast	0.079	0.001	0.020	0.234	-0.029
	[0.005]***	[0.006]	[0.006]***	[0.016]***	[0.002]***
Malaria Suit.	0.135	0.083	0.121	0.242	-0.036
	[0.005]***	[0.005]***	[0.006]***	[0.014]***	[0.002]***
Pastoral Suit	-0.055	-0.021	-0.044	-0.022	-0.013
	[0.004]***	[0.004]***	[0.005]***	[0.012]*	[0.002]***
Agricultural Suit	0.186	0.175	0.219	0.247	-0.069
	[0.006]***	[0.006]***	[0.007]***	[0.018]***	[0.002]***
Past*Agr Suit	0.027	0.033	0.019	0.136	-0.016
	[0.004]***	[0.004]***	[0.004]***	[0.011]***	[0.001]***
Elevation	-0.082	-0.061	-0.056	-0.266	0.021
	[0.005]***	[0.005]***	[0.006]***	[0.012]***	[0.002]***
Ruggedness	0.086	0.034	0.057	0.266	-0.010
	[0.004]***	[0.004]***	[0.004]***	[0.011]***	[0.001]***
Mean Dep.	-0.018	-0.013	2.056	2.789	0.704
Observations	81,073	88,714	88,714	85,412	88,714

Notes: The fractionalization measures are standardized. The regressions include country fixed effects.

5.4 IV Strategy 1: Railroad Towns and Least-Cost Path

To study the effect of regional ethnic diversity on city growth and urban conflict, an ideal experiment might randomly place cities in different locations with different ex-ante levels of ethnic diversity. To approximate this idealized setting I leverage regional productivity shocks which increase the probability of a city forming in a given region, independent of the underlying distribution of ethnic groups. In particular, I consider the construction of colonial railroads, which boosted the fundamental productivity for all regions along the the railroad path by suddenly granting access to distant markets. Colonial railroads were often built to connect coastlines to a particular resource in the interior of the country. An example is the British Uganda railway, which connected Mombasa on the coast to Lake Victoria for geopolitical reasons. This railway incidentally also increased the productivity of regions that lay along the least-cost path between these points. Indeed human settlements grew everywhere along the railway, and the railway's path within Kenya predicts the location of contemporary Kenyan cities (Jedwab et al., 2017).

Depending on their location along the railway path, these emerging cities have different exposure to an ethnically diverse population of potential migrants. By using the placement of railways as an instrument for contemporary city location, I can compare the development of cities with different levels of regional ethnic diversity along the same rail line. Figure 6 shows a map of constructed colonial railways collected by Jedwab and Moradi (2016). Taking each grid's distance from the nearest rail line, I look at the association between distance to the nearest line and whether or not that grid is a city, as well as the grid's population. Figure 7 shows that being within 50km of a colonial rail line is highly predictive of city location and size. The relationship is also nonlinear. Figure 8 shows a regression discontinuity design by distance to nearest rail. Locations 20km or nearer to a rail are at the right side of the cut-off, while regions further away are on the left. We see that the probability of being a new city (that is, a city that emerged after 1950) spikes once a region is within 20km of a rail.

I use a two stage least squares strategy in which the distance of a grid to a colonial rail-line is used as an instrument for city location. Grids that are nearby colonial rail lines are more likely to become cities over time. Different grids have different levels of ethnic diversity. I additionally instrument for the interaction between my city dummy and ethnic diversity using the interaction of ethnic diversity with the distance to the rail line. This gives me 2 first-stage equations:

$$C_i = \alpha + \beta_1 Dist_i + \upsilon_r + \omega_s + \epsilon_i \tag{3}$$

$$C_i * D_i = \alpha + \beta_1 Dist_i + \beta_2 Dist_i * D_i + \upsilon_r + \omega_s + \epsilon_i \tag{4}$$

Here C_i represents a dummy for whether a grid is a city, D_i represents the grid's ethnic diversity, $Dist_i$ is the grid's distance to the nearest colonial railway, and υ_r, ω_s control for

Figure 6: Colonial rail lines (Jedwab et al., 2017)

railway and state fixed effects respectively.

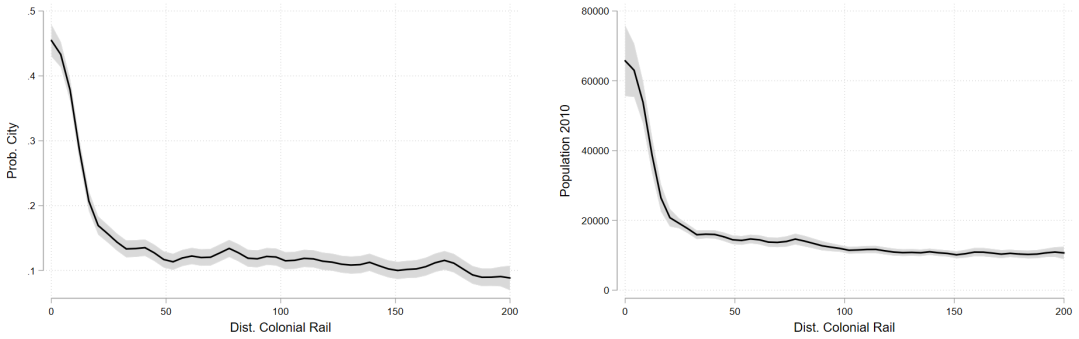
I can then estimate a second stage equation to examine the effect of a city placed in a more or less diverse area on an outcome Y_i. In the second stage we are interested in the coefficient β_3 from the equation:

$$Y_i = \alpha + \beta_1 \hat{C}_i + \beta_2 D_i + \beta_3 \widehat{C_i * D_i} + X_i + \epsilon_i \tag{5}$$

Where $\hat{C}_i$ represents the instrumented value from equation 3 and $\widehat{C_i * D_i}$ represents the instrumented value from equation 4. X_i includes the grid level controls malaria suitability, soil suitability, elevation and ruggedness.

What is the relationship between the rail line placement and a region's ethnic diversity? It may be the case for example that a rail path intentionally moved through more homogeneous ethnic areas specifically to avoid destabilizing conflict. There is indeed some significant association between distance to the rail line and a grid's fractionalization index. Figure 9 shows linear polynomials of the association between rail line distance and fractionalization for all grids within 200km of a rail line. Table 2 shows a regression of the distance to the rail line on Murdock fractionalization, controlling for geographic fundamentals, state and rail fixed effects.

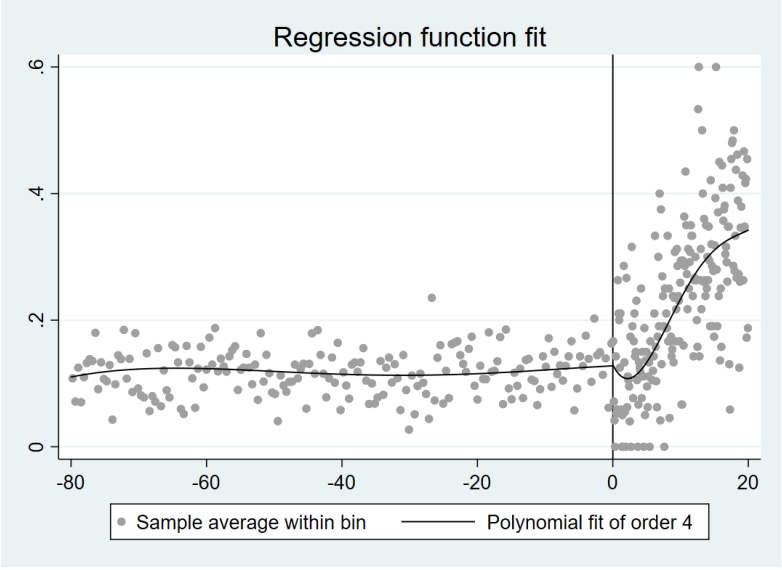
Figure 7: Local Polynomial of City Formation by Dist to Rail (km)



(a) Prob. of City Location

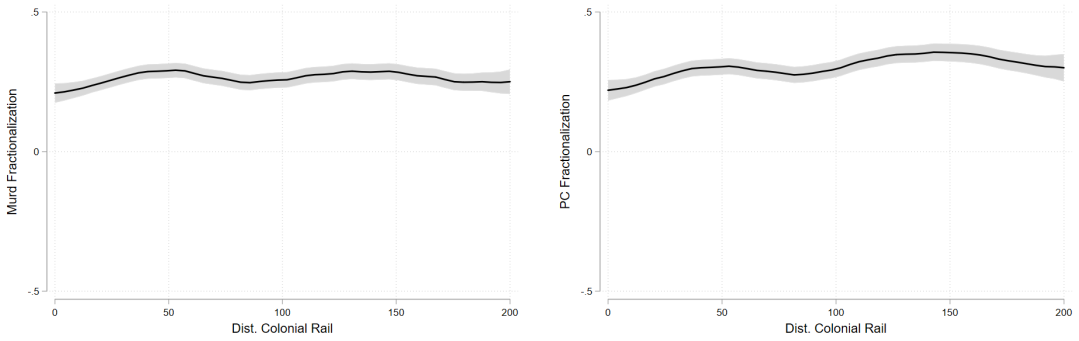
(b) City Pop. 2010

Figure 8: RD Plot at 20km Rail Distance and New City Probability



Note: Probability that grid is new city by rail distance (km), where 0 is 20km. Controls include river/coast distance, elevation/ruggedness, suitability (land/malaria/tsetse/pastoral) and historical population.

Figure 9: Local Polynomial of Ethnic Fractionalization by Distance to Rail (km)



(a) Murdock Fractionalization

(b) PC Fractionalization

Table 2: Fractionalization and Distance to Rail

	Rail Dist	Rail Dist	Rail Dist	Rail Dist
Murd Fract	2.509	2.157	3.004	2.635
	[2.054]	[2.234]	[2.078]	[2.256]
Coast Distance	4.996	4.300	5.829	5.265
	[0.639]***	[0.926]***	[0.732]***	[1.062]***
River Distance	2.055	2.987	2.338	3.151
	[0.488]***	[0.594]***	[0.499]***	[0.605]***
Elevation	-0.008	-0.007	-0.010	-0.009
	[0.002]***	[0.002]***	[0.002]***	[0.003]***
Ruggedness	2.152	1.950	2.448	2.279
	[0.567]***	[0.603]***	[0.576]***	[0.604]***
Land Suit			4.763	2.947
			[2.704]*	[3.264]
Malaria Suit			-0.133	-0.317
			[0.096]	[0.132]**
Tsetse Suit			0.178	3.534
			[5.127]	[6.165]
Agricultural Suit			-0.185	-0.702
			[0.303]	[0.393]*
Animal Suit			-0.364	-0.383
			[0.232]	[0.291]
Pastoral Suit			0.271	0.986
			[0.259]	[0.316]***
HunterGatherer Suit			0.310	0.324
			[0.278]	[0.391]
Constant	-0.000	0.000	0.000	0.000
	[0.490]	[0.483]	[0.492]	[0.484]
Rail FE	N	Y	N	Y
Dist to Rail km	<100km	<100km	<100km	<100km
Observations	18,517	18,517	18,263	18,263

Notes: Fractionalization measures are standardized. The regressions include a variety of geographic and ecological controls, as well as country and rail fixed effects. Conley standard errors with spatial correlation at 20km. Sample is cut at 100km within a rail line.

5.5 IV Strategy 2: Portage Sites

A second strategy leverages the geographic placement of portage sites (Bleakley and Lin, 2012). Maritime trade often requires ships to move inland from the coast along navigable rivers. Sharp changes in elevation along rivers create rapids and waterfalls, preventing large ships from traveling further. It becomes necessary to create infrastructure at the point at which a river is no longer navigable to transfer goods from ships to land transport. Prior work in the US has shown that many US cities developed along the Atlantic Seaboard Fall Line, which creates a point of elevation change at which inland rivers are no longer navigable on the east coast (Bleakley and Lin, 2012).

Using the same logic, I use data on land ruggedness and the river network to identify points along African river systems at which large elevation changes occur. If portage sites predict town locations, then sites that are rugged and near rivers are more likely to have a town develop over time, as maritime travel became more prevalent in colonial Africa. I create a portage site score as the interaction between a region's distance from a river and its ruggedness level. A high score represents a higher degree of ruggedness and a region closer to a river. Figure 10 shows a visual example for the Democratic Republic of Congo. Portage scores are higher along the river network, and importantly, also vary along a particular river. The mouth of the river in the West shows high portage scores near where the Congo River has large rapids that precede the city of Kinshasa.

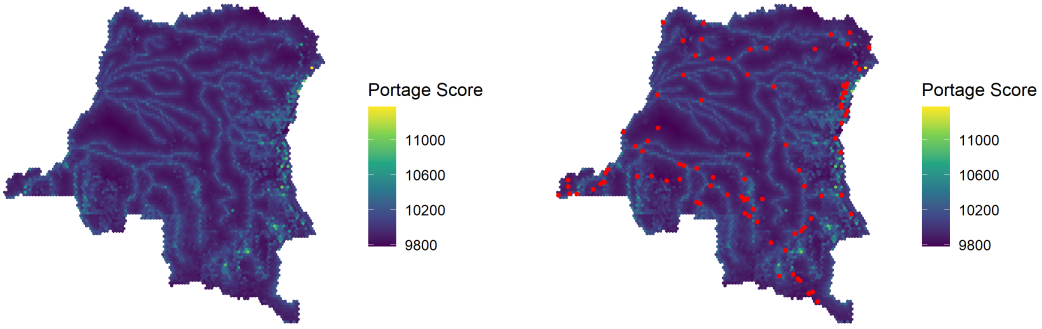
I use the portage score as an instrument to predict historical city location. Because these sites only became relevant during the colonial period (due to increased coast-to-inland transport), these sites are likely orthogonal to the interaction between city locations and the distribution of historical ethnic group homelands. Before this time, African river trade was often conducted by canoe, which were better able to navigate rapids and other elevation changes (Smith, 1970). Using the interaction of ruggedness and the river network helps to move away from the association between ruggedness and ethnic boundary lines discussed by Michalopoulos (2012). Particular features of ruggedness such as mountain ranges are likely to predict where tribes or groups begin and end, in turn affecting the underlying diversity of rugged regions. By isolating our analysis to river-adjacent ruggedness, we hopefully isolate elevation changes that predict waterfalls and rapids, which are less likely to have determined ancient migratory patterns and the initial development of ethnolinguistic diversity.

Figure 11 shows how portage score positively predicts city location and population. In the Appendix I test my portage score instrument by comparing it to hydrological data on river flows, rapids and a georeferenced sample of known portage towns in Africa.

Using the portage score P_i for a given grid i, I can instrument for city location and the interaction of city location and ethnic diversity in a similar fashion as in the rail design:

$$C_i = \alpha + \beta_1 P_i + \omega_s + \epsilon_i \tag{6}$$

Figure 10: Grid-Level Portage Score for DRC



(a) DRC Portage Score

(b) DRC City Locations

$$C_i * D_i = \alpha + \beta_1 P_i + \beta_2 P_i * D_i + \omega_s + \epsilon_i \tag{7}$$

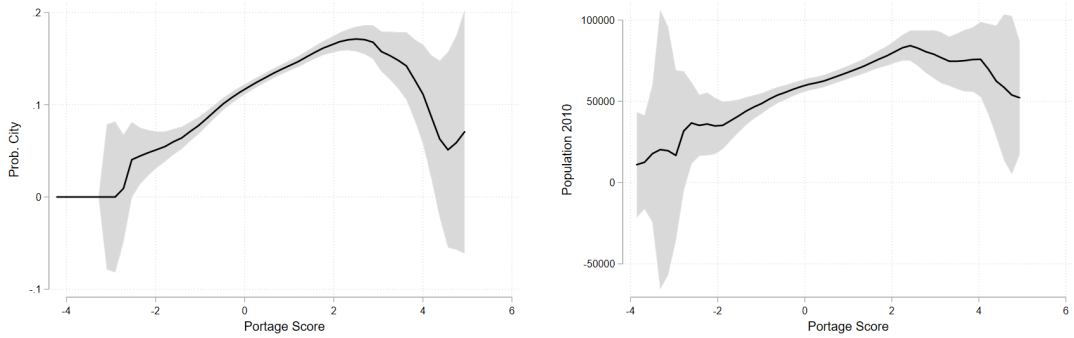
Here C_i represents a dummy for whether a grid is a city, D_i represents the grid's ethnic diversity, P_i is the grid's portage site score, and ω_s controls for state fixed effects.

I can then estimate a second stage equation to examine the effect of a portage city located in a more or less diverse area on an outcome Y_i. In the second stage we are interested in the coefficient β_3 from the equation:

$$Y_i = \alpha + \beta_1 Z_{1i} + \beta_2 D_i + \beta_3 Z_{2i} + X_i + \epsilon_i \tag{8}$$

Where Z_{1i} represents the instrumented value from equation 6 and Z_{2i} represents the instrumented value from equation 7. X_i includes the grid level controls malaria suitability and soil suitability.

Figure 11: Local Polynomial of City Formation by Portage Score



(a) Prob. of City Location

(b) City Pop. 2010

Again we might be concerned about the association between the portage locations and the historic distribution of ethnic groups. Table 3 shows regressions of the portage score measure on the fractionalization indexes. We see that higher portage scores are associated with higher diversity grids, which is going in the opposite direction as the rail city predictor (where further distance from the rail was associated with lower city probability and higher diversity). A 1-unit increase in the portage score associated with a .03 - .09 standard deviation increase in fractionalization.

Table 3: Fractionalization and Portage Probability

	Murd	Murd	Murd	Murd	PC	PC	PC	PC
Portage Score	0.003	0.004	-0.105	-0.103	0.107	0.095	0.014	0.008
	[0.011]	[0.011]	[0.014]***	[0.014]***	[0.010]***	[0.010]***	[0.013]	[0.013]
Dist to River	<100km	<100km	<50km	<50km	<100km	<100km	<50km	<50km
River FE	N	Y	N	Y	N	Y	N	Y
Mean Dep.	0.306	0.306	0.386	0.386	0.397	0.397	0.497	0.497
Observations	37,304	37,304	23,000	23,000	36,741	36,741	22,855	22,855

Notes: The fractionalization measures are standardized. The regressions include malaria suitability, land suitability, historic population, ruggedess and river distance as controls, as well as country fixed effects.

6 Results

Using both instruments, I first show the predictive power of the instrument on city location (whether a grid is a city), and then I show the second stage results from equation 5 and equation 8 for rail and portage score respectively.

6.1 Rail IV Results

6.1.1 First Stage

Table 4 shows the results of the first-stage regression of the city dummy on distance to a colonial rail line in equation 3. This suggests that within a given bandwidth around colonial rail lines, city formation is highly predicted by distance to the rail. This pattern is similar to what is shown in Figure 7, which shows a spike in the probability of city formation near colonial railways. For the rest of this analysis, we restrict the sample to locations within 300km and 100km of a colonial rail line. Table 5 shows the results of the second first stag regression on the interaction of our city dummy with ethnic fractionalization.

6.1.2 Second Stage

Table 6 shows the effect of an instrumented city-ethnic diversity interaction on light density for the years 2010-2013. Table B4 shows the same regression for light density aggregated in the years 2000-2009. In columns 1 and 2 of Table 6 we see that having a city placed in a grid with higher Murdock ethnic diversity show lower nighttime light density. The relationship is also seen in the PC measure of diversity for the lights outcome variable in years 2010-2013. Table 7 shows the effect on conflict outcomes. Conditional on having a city located in a grid, a higher level of ethnic diversity is associated with higher conflict probability between the years 1975 and 2021. This relationship is significant for the principle component measure of diversity. Table B5 shows the effect on number of conflict deaths, conditional on experiencing conflict during this period. The coefficients are positive, but not significant for the city*diversity interactions. Using the distance to colonial rail instrument, cities located in more diverse regions have lower light density and higher conflict incidence.

Table 8 shows the results for durables consumption. The observations now are at the DHS household level, rather than the grid-level. Cities placed in more fractionalized locations have consistently lower levels of durables consumption.

Table 4: Rail IV - Predict City

	City	City	City	City
Rail Distance (km)	-0.005	-0.005	-0.004	-0.004
	[0.000]***	[0.000]***	[0.001]***	[0.001]***
Rail*Murd Fract	0.001	0.001	0.001	0.001
	[0.001]*	[0.001]*	[0.001]*	[0.001]*
Coast Distance	-0.042	-0.019	-0.042	-0.019
	[0.009]***	[0.011]*	[0.009]***	[0.011]*
River Distance	-0.010	-0.014	-0.010	-0.015
	[0.005]*	[0.006]**	[0.005]*	[0.006]**
Elevation	0.000	0.000	0.000	0.000
	[0.000]***	[0.000]	[0.000]***	[0.000]
Ruggedness	0.008	0.018	0.018	0.030
	[0.008]	[0.008]**	[0.015]	[0.014]**
Land Suit	0.086	0.035	0.104	0.041
	[0.031]***	[0.036]	[0.057]*	[0.055]
Malaria Suit	0.003	-0.001	0.003	-0.001
	[0.001]***	[0.001]	[0.001]***	[0.001]
Tsetse Suit	-0.089	-0.182	-0.080	-0.166
	[0.056]	[0.062]***	[0.086]	[0.086]*
Agricultural Suit	0.008	0.008	0.008	0.008
	[0.003]**	[0.004]**	[0.003]**	[0.004]**
Animal Suit	0.000	-0.005	0.000	-0.005
	[0.002]	[0.003]**	[0.002]	[0.003]**
Pastoral Suit	-0.008	-0.008	-0.008	-0.008
	[0.003]***	[0.003]**	[0.003]***	[0.003]**
HunterGatherer Suit	0.011	0.009	0.011	0.009
	[0.003]***	[0.004]**	[0.003]***	[0.004]**
Rail*LandSuit			-0.001	-0.000
			[0.002]	[0.001]
Rail*Elevation			0.000	-0.000
			[0.000]	[0.000]
Rail*Ruggedness			-0.000	-0.000
			[0.000]	[0.000]
Rail*TsetseSuit			-0.000	-0.001
			[0.002]	[0.002]
Constant	0.000	0.000	0.000	0.000
	[0.005]	[0.004]	[0.005]	[0.004]
Mean Dep. Var	0.170	0.170	0.170	0.170
Rail FE	N	Y	N	Y
Dist to Rail km	<50km	<50km	<50km	<50km
Observations	9,753	9,753	9,753	9,753

Notes: All regressions include country fixed effects, columsn 2 and 4 include rail fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid. Conley standard errors with spatial correlation at 20km.

Table 5: Rail IV - Predict City*Diversity

	City_Frac	City_Frac	City_Frac	City_Frac
Rail Distance (km)	-0.002	-0.002	-0.002	-0.002
	[0.000]***	[0.000]***	[0.001]***	[0.001]***
Rail*Murd Fract	0.004	0.003	0.004	0.003
	[0.000]***	[0.000]***	[0.000]***	[0.000]***
Coast Distance	-0.012	-0.001	-0.012	-0.001
	[0.004]***	[0.005]	[0.004]***	[0.005]
River Distance	-0.005	-0.004	-0.005	-0.004
	[0.002]**	[0.003]	[0.002]**	[0.003]
Elevation	0.000	-0.000	0.000	0.000
	[0.000]***	[0.000]	[0.000]*	[0.000]
Ruggedness	0.008	0.011	0.015	0.019
	[0.004]**	[0.004]***	[0.008]*	[0.007]***
Land Suit	0.014	-0.007	0.038	0.011
	[0.015]	[0.017]	[0.027]	[0.025]
Malaria Suit	0.001	-0.001	0.001	-0.001
	[0.000]*	[0.001]	[0.000]*	[0.001]
Tsetse Suit	-0.065	-0.084	-0.056	-0.071
	[0.026]**	[0.030]***	[0.040]	[0.039]*
Agricultural Suit	0.002	0.001	0.002	0.001
	[0.002]	[0.002]	[0.002]	[0.002]
Animal Suit	0.000	-0.001	0.000	-0.001
	[0.001]	[0.001]	[0.001]	[0.001]
Pastoral Suit	-0.003	-0.005	-0.003	-0.005
	[0.001]**	[0.001]***	[0.001]**	[0.001]***
HunterGatherer Suit	0.008	0.005	0.007	0.005
	[0.001]***	[0.002]**	[0.001]***	[0.002]**
Rail*LandSuit			-0.001	-0.001
			[0.001]	[0.001]
Rail*Elevation			0.000	-0.000
			[0.000]	[0.000]
Rail*Ruggedness			-0.000	-0.000
			[0.000]	[0.000]
Rail*TsetseSuit			-0.000	-0.000
			[0.001]	[0.001]
Constant	0.000	0.000	0.000	0.000
	[0.002]	[0.002]	[0.002]	[0.002]
Mean Dep. Var	0.060	0.060	0.060	0.060
Rail FE	N	Y	N	Y
Dist to Rail km	<50km	<50km	<50km	<50km
Observations	9,753	9,753	9,753	9,753

Notes: All regressions include country fixed effects, columsn 2 and 4 include rail fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid. Conley standard errors with spatial correlation at 20km.

Table 6: 2SLS Rail IV - Light Density 2010s

	Lights	Lights	Lights	Lights	Lights	Lights	Lights	Lights
City*Murd Fract	-0.328	-0.320	-0.246	-0.241				
	[0.090]***	[0.105]***	[0.101]**	[0.106]**				
Murd Fract	0.049	0.039	0.033	0.028				
	[0.013]***	[0.015]***	[0.018]*	[0.018]				
City	0.247	-0.062	1.635	1.660	0.317	-0.018	1.656	1.675
	[0.151]	[0.203]	[0.118]***	[0.117]***	[0.137]**	[0.192]	[0.111]***	[0.109]***
City*PC Fract					-0.348	-0.391	-0.418	-0.355
					[0.084]***	[0.109]***	[0.141]***	[0.148]**
PC Fract					0.097	0.099	0.131	0.115
					[0.019]***	[0.023]***	[0.039]***	[0.037]***
Rail FE	N	Y	N	Y	N	Y	N	Y
Dist to Rail	<300km	<300km	<100km	<100km	<300km	<300km	<100km	<100km
F-stat	126	83	170	98	115	58	30	30
Mean Dep. Var	-0.039	-0.039	0.012	0.012	-0.039	-0.039	0.012	0.012
Observations	40,262	40,262	17,275	17,275	39,096	39,096	16,788	16,788

Notes: Controls include land suitability, malaria suitability, ruggedness. All regressions include country and rail fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid. Light density measures are also standardized after averaging across years 2000-2009 and 2010-2013.

Table 7: 2SLS Rail IV - Prob. Conflict

	P(conflict)	P(conflict)	P(conflict)	P(conflict)	P(conflict)	P(conflict)	P(conflict)	P(conflict)
City*Murd Fract	-0.004	0.001	0.005	0.006				
	[0.008]	[0.008]	[0.007]	[0.008]				
Murd Fract	0.003	0.000	0.000	-0.000				
	[0.001]**	[0.001]	[0.001]	[0.001]				
City	-0.063	-0.018	0.048	0.057	-0.054	-0.015	0.046	0.052
	[0.013]***	[0.013]	[0.008]***	[0.008]***	[0.012]***	[0.013]	[0.007]***	[0.007]***
City*PC Fract					-0.002	0.012	0.023	0.034
					[0.008]	[0.009]	[0.011]**	[0.012]***
PC Fract					0.006	0.000	-0.001	-0.007
					[0.002]***	[0.002]	[0.003]	[0.003]**
Rail FE	N	Y	N	Y	N	Y	N	Y
Dist to Rail	<300km	<300km	<100km	<100km	<300km	<300km	<100km	<100km
F-stat	126	83	170	98	115	58	30	30
Mean Dep. Var	0.013	0.013	0.012	0.012	0.013	0.013	0.012	0.012
Observations	40,262	40,262	17,275	17,275	39,096	39,096	16,788	16,788

Notes: Controls include land suitability, malaria suitability, ruggedness. All regressions include country and rail fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid. Prob. conflict is defined as the proportion of years in which the grid experienced a conflict across 1975-2021.

Table 8: 2SLS Rail IV - DHS Durables Consumption

	Durables	Durables	Durables	Durables	Durables	Durables	Durables	Durables
City*Murd Fract	-0.142	-0.141	-0.124	-0.152				
	[0.011]***	[0.014]***	[0.016]***	[0.019]***				
Murd Fract	0.044	0.046	0.045	0.063				
	[0.007]***	[0.008]***	[0.012]***	[0.013]***				
City	2.185	2.275	2.148	2.123	2.262	2.239	2.202	2.159
	[0.012]***	[0.020]***	[0.015]***	[0.019]***	[0.013]***	[0.019]***	[0.015]***	[0.019]***
City*PC Fract					-0.264	-0.098	-0.219	-0.204
					[0.010]***	[0.014]***	[0.016]***	[0.018]***
PC Fract					0.117	0.018	0.097	0.103
					[0.007]***	[0.009]*	[0.012]***	[0.014]***
Rail FE	N	Y	N	Y	N	Y	N	Y
Dist to Rail	<300km	<300km	<100km	<100km	<300km	<300km	<100km	<100km
F-stat	15769	6536	7955	6107	15166	5092	7874	5216
Mean Dep. Var	-0.000	-0.000	0.152	0.152	-0.001	-0.001	0.151	0.151
Observations	592,466	592,466	379,436	379,436	590,974	590,974	378,117	378,117

Notes: Controls include land suitability, malaria suitability, ruggedness. All regressions include DHS sample fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid. Durables consumption is a principle component measure that is then standardized.

6.2 Portage Score IV Results

6.2.1 First Stage

Table 9 shows the results of the first-stage regression of the city dummy on portage site score. We restrict the sample to be within 100km and 50km of a river. We also restrict the sample to include standardize portage score measures between -3 and 3 standard deviation units, to exclude outliers. A 1 standard deviation increase in the portage score increases the probability of a grid being a city by approximately 3.5 percentage points, from a baseline average of about 12 percentage points. This is a significant increase in the likelihood of city formation, and the relationship holds with and without fixed effects for the nearest river.

Table 9: Portage IV - Predict City

	P(city)	P(city)	P(city)	P(city)
Portage Score	0.035	0.041	0.035	0.043
	[0.003]***	[0.004]***	[0.005]***	[0.005]***
Dist to River	<100km	<100km	<50km	<50km
River FE	N	Y	N	Y
Mean Dep.	0.127	0.127	0.136	0.136
Observations	37,304	37,304	23,000	23,000

Notes: Controls include land suitability, malaria suitability. All regressions include country fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid. The "Dist" row describes the sample cutoff of distance to nearest river for that particular regression.

6.2.2 Second Stage

Table B6 and Table 10 show the effect of an instrumented city-ethnic diversity interaction on light density for years 2000-2009 and 2010-2013 respectively. Cities placed in grids with higher ethnic diversity, as measured by the principle component, do not show a consistent association with nighttime light density for all years. The results using just the Murdock fractionalization measure are negative, but insignificant.

Table 11 shows the effect on conflict outcomes. Conditional on having a city located in a grid, a higher level of ethnic diversity measured by the principal component variable is associated with higher conflict probability between the years 1975 and 2021. Using the portage score instrument and the principle component fractionalization measure, portage cities located in more diverse regions have lower light density and higher conflict incidence.

Table 12 shows the results for durables consumption. The observations now are at the DHS household level, rather than the grid-level. While the majority of specifications show a negative impact on durables consumption, the results are noisy, with some specifications show a positive impact on durables consumption. In a robustness analysis, I explore how these coefficients are affected by differing specifications of the controls and spatial correlation structure.

Table 10: Portage IV - Light Density 2010s

	Lights	Lights	Lights	Lights	Lights	Lights	Lights	Lights
City*Murd Fract	0.074	0.070	-0.014	-0.034				
	[0.154]	[0.141]	[0.265]	[0.225]				
Murd Fract	-0.014	-0.017	-0.015	-0.016				
	[0.020]	[0.019]	[0.037]	[0.032]				
City	0.010	-0.037	-0.664	-0.499	0.274	0.216	-0.645	-0.492
	[0.357]	[0.296]	[0.487]	[0.386]	[0.306]	[0.270]	[0.444]	[0.374]
City*PC Fract					-0.169	-0.128	0.042	0.030
					[0.174]	[0.163]	[0.321]	[0.253]
PC Fract					0.026	0.005	-0.016	-0.026
					[0.035]	[0.032]	[0.065]	[0.051]
River FE	N	Y	N	Y	N	Y	N	Y
Dist to River	<100km	<100km	<50km	<50km	<100km	<100km	<50km	<50km
F-stat	41	60	22	36	30	33	16	21
Mean Dep. Var	-0.015	-0.015	0.020	0.020	-0.015	-0.015	0.020	0.020
Observations	37,545	37,545	23,240	23,240	36,982	36,982	23,095	23,095

Notes: Controls include land suitability, malaria suitability. All regressions include country fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid. Light density measures are also standardized after averaging across years 2000-2009 and 2010-2013.

Table 11: Portage IV - Prob. Conflict

	P(conflict)	P(conflict)	P(conflict)	P(conflict)	P(conflict)	P(conflict)	P(conflict)	P(conflict)
City*Murd Fract	0.011	0.026	-0.000	-0.002				
	[0.013]	[0.012]**	[0.024]	[0.019]				
Murd Fract	-0.000	-0.002	0.002	0.003				
	[0.002]	[0.002]	[0.003]	[0.003]				
City	0.135	0.097	0.206	0.154	0.094	0.064	0.121	0.095
	[0.030]***	[0.021]***	[0.056]***	[0.031]***	[0.027]***	[0.019]***	[0.047]***	[0.027]***
City*PC Fract					0.055	0.063	0.071	0.049
					[0.011]***	[0.011]***	[0.019]***	[0.015]***
PC Fract					-0.007	-0.008	-0.010	-0.006
					[0.002]***	[0.002]***	[0.004]**	[0.003]**
River FE	N	Y	N	Y	N	Y	N	Y
Dist to River	<100km	<100km	<50km	<50km	<100km	<100km	<50km	<50km
F-stat	29	56	11	28	31	52	13	31
Mean Dep. Var	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014
Observations	37,310	37,310	23,006	23,006	36,747	36,747	22,861	22,861

Notes: Controls include land suitability, malaria suitability. All regressions include country fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid. Prob. conflict is defined as the proportion of years in which the grid experienced a conflict across 1975-2021.

Table 12: Portage IV - DHS Durables Consumption

	Durables	Durables	Durables	Durables	Durables	Durables	Durables	Durables
City*Murd Fract	-6.613	-1.258	0.777	0.982				
	[5.716]	[0.218]***	[0.103]***	[0.193]***				
Murd Fract	4.004	0.716	-0.551	-0.683				
	[3.508]	[0.133]***	[0.065]***	[0.121]***				
City	0.535	1.983	1.998	1.754	1.849	1.826	1.649	1.768
	[1.501]	[0.030]***	[0.043]***	[0.033]***	[0.036]***	[0.060]***	[0.030]***	[0.031]***
City*PC Fract					-0.695	1.527	-0.220	-0.541
					[0.080]***	[0.653]**	[0.042]***	[0.091]***
PC Fract					0.406	-1.080	0.089	0.317
					[0.054]***	[0.439]**	[0.029]***	[0.063]***
River FE	N	Y	N	Y	N	Y	N	Y
Dist to River	<100km	<100km	<50km	<50km	<100km	<100km	<50km	<50km
F-stat	1	43	252	81	114	7	581	135
Mean Dep. Var	-0.050	-0.050	-0.008	-0.008	-0.050	-0.050	-0.008	-0.008
Observations	472,968	472,968	308,488	308,488	472,821	472,821	308,488	308,488

Notes: Controls include land suitability, malaria suitability. All regressions include DHS sample fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid. The principle component asset score is standardized.

7 Climate and Diversity

Our IV strategy shows the effects of cities located in relatively more or less diversed areas on long-term development. However we may also be interested in how contemporary changes to diversity across cities affects year-to-year changes in urban growth and productivity. Again there are two empirical challenges: (1) Disentangling the role of agglomeration, or increases in density, versus changes in the composition of the workforce. (2) The distribution of ethnic groups is a function of geographic fundamentals and long-term land productivity. Our strategies so far have considered "pull shocks", where migrants are drawn into regions that expreience a temporary productivity shock due to their proximity to a railroad or portage site.

Climate shocks can be considered a "push" shock, driving agricultural workers towards other regions, and potentially towards urban centers that may offer refuge. Empirical work has shown that bad weather shocks drive urbanization in African cities that specialize in non-agricultural products (ex. higher manufacutring share) (Henderson et al., 2017). Further, empirical work has shown that the groups that droughts affect have implications for conflict, for example when pastoralists are pushed into adjacent farmland (McGuirk and Nunn, 2024; Kramon et al., 2022). For a given city, their exposure to climate-induced migration is a function of all potential origin regions, weighted by distance. The composition of migrant flows into a city will then be a function of how many individuals choose to migrate differentially from each origin.

Here we will leverage both the intensity of drought shocks, and their distribution across space to capture changes to size and composition of migration flows into cities. In particular we will instrument for city population L_i and city fractionalization C_i in a regression of city-level productivity on population and contemporary diversity (see equation 9).

$$Log(Y)_i = \beta_0 + \beta_1 L_i + \beta_2 C_i + X_i + \epsilon_i \tag{9}$$

While city-level population data is easy to come by, subnational data on contemporary ethnicity is rare in Africa. Our estimates of fractionalization will utilize aggregations of DHS data, as well as broader census estimates from administrative regions. For a given DHS sample year, we will estimate city level fractionalization using ethnic data from every DHS sample point within 20km of the city. Because DHS sampling clusters typically collect data within a neighborhood across nearby houses, we may expect any single cluster of observations to mismeasure city-level diversity if there is residential segregation across groups. The accuracy of our diversity measure then relies on having multiple sample points per city. For states that have multiple DHS rounds, we can construct within-city changes in our diversity measure across sample years.

For a broader view, we will also use census data at the second administrative level. While these regions are larger than any given city, the representativeness gives us a more accurate

picture of subnational contemporary diversity. These censuses also provide an urban dummy that will allow us to parse urban and rural areas within an administrative region. For a few countries we have multiple census rounds (ex. Benin, Ghana, Mauritius) and can create a measure of fractionalization change across a decade.

7.1 Measuring Moments of Drought Shocks

We will use data from the Standardised Precipitation-Evapotranspiration Index (SPEI), which measures drought intensity monthly by combining temperature and precipitation data. The data provides monthly estimates of drought intensity at a 0.5x0.5 degree cell resolution from 1900-2022, which we aggregate into yearly estimates. Layering our various ethnic map sources over this data, we calculate the average drought experience over time for each ethnic group separately. For each murdock group g we calculate a yearly drought index from monthly dummy variables that signal whether the month was a drought based on the SPEI (ie. the share of year in drought). We aggregate our yearly measures to decade level as the sum of drought intensity across 10 years. Higher values are more drought years, and higher proportion of year spent in drought. Using this measure, we construct the following instruments for a city's exposure to changes in population L_i and diversity C_i:

1. Drought Intensity = For a given region i, and decade t, drought exposure $\mu_i t$ is the sum of decade-level drought-intensity across all murdock groups within a 300km boundary, weighted by distance. For example, a 1970 drought intensity measure for region i is:

$$\mu_{1970i} = \sum_{g}^{G} \frac{1}{log(Dist)} Drought_{1960-1970} \tag{10}$$

   Note this measure is a function of how many nearby murdock groups there are within the boundary, so we condition on this variable when needed.

2. Drought Distribution = For a given region i and decade t, drought distribution σ_{it} is the relative distribution of drought severity across tribes:

$$\sigma_{it} = \sum_{g}^{G} (\frac{\frac{1}{log(Dist)} Drought_{gt}}{\mu_{it}})^2 \tag{11}$$

   A higher value suggests that the drought shock is more concentrated in particular groups.

7.2 Cross-Sectional Evidence with Drought Instruments

For our cross-sectional sample of contemporary fractionalization and population estimates, we use aggregated historical drought instruments. In particular we take the years 1900 to

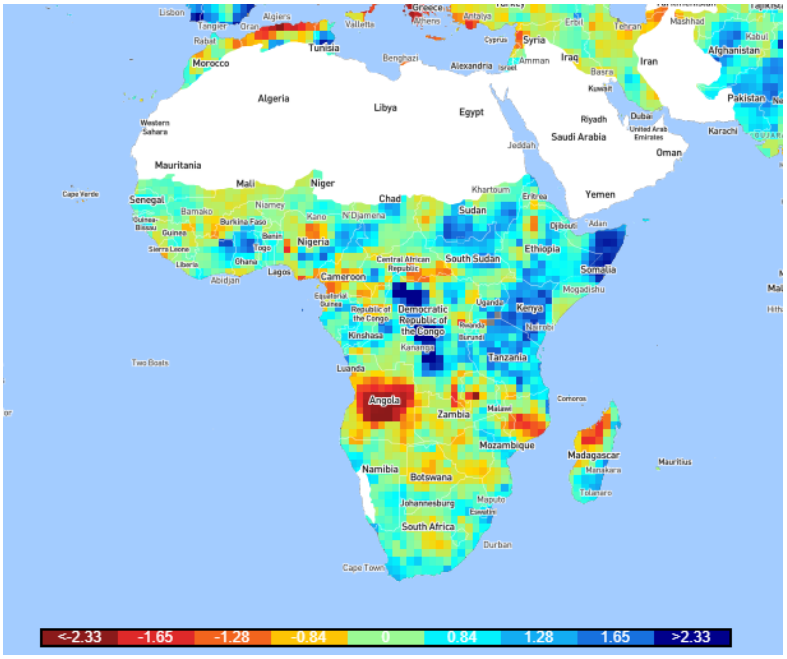


Figure 12: Drought Event Example, 1990

1950 and calculate measures of drought intensity μ_i and drought distribution σ_i. In each regression we control for historic diversity, as measured by our Murdock fractionalization index. A first stage is presented in Table 13 where we regress contemproary measures of city population and DHS fractionalization (measured between 1990 and 2015) on our drought instruments. Drought intensity in neighboring areas predicts higher population in 2000, and lower fractionalization. More concentrated drought predicts lower population.

For our first stage, we estimate the following equation:

$$Y_i = \alpha + \beta_1 \sigma_{1950i} + \beta_2 \mu_{i1950} + D_i + Pop_i 1800 + \epsilon_i \tag{12}$$

Where the outcomes Y_i are various measures of contemporary population and fractionalization, and σ_{1950i}, μ_{i1950} are drought measures aggregated across years 1900 to 1950, while D_i is historic fractionalization and Pop_{i1800} is a measure of historic population. For the census data version of the equation, we omit this last variable as we don't have a comparable variable for large administrative regions. Table 13 shows that drought intensity before 1950 predicts population growth today, as well as lower fractionalization. The coefficient on drought distribution has the correct sign – higher concentration of drought reduces fractionalization. In Table 14 has a less clear association of drought events with population for the countries where census data is available.

Table 13: Drought Shock Impact on City-Level Characteristics (Cross-Section)

	Pop 1960	Pop 2000	Pop. Growth 2000	DHS Fract
Historic Drought Distribution	-0.457	-0.538	-0.176	-0.589
	[0.141]***	[0.067]***	[0.042]***	[0.379]
Historic Drought Intensity	-0.029	0.009	0.010	-0.009
	[0.011]***	[0.004]**	[0.003]***	[0.003]***
Historic Diversity	-0.046	-0.027	-0.006	0.040
	[0.037]	[0.016]*	[0.011]	[0.010]***
Pop. in 1800	0.000	0.000	-0.000	-0.000
	[0.000]***	[0.000]***	[0.000]*	[0.000]***
Mean Dep.	10.047	10.142	0.385	0.433
Observations	858	4,191	2,728	612

Notes: Controls include population in 1800, the count of nearby murdock groups and historic diversity, measured by murdock fractionalization at a 50km boundary.

Table 14: Drought Shock Impact on Census-Level Characteristics (Cross-Section)

	log(Pop2000)	log(Pop2000)	Ethnic HHI	Ethnic HHI
Historic Drought Intensity	-2.841	-2.478	-1.676	-0.692
	[2.137]	[3.640]	[0.485]***	[0.739]
Historic Drought Distribution	0.011	0.200	-0.132	-0.112
	[0.758]	[1.160]	[0.172]	[0.235]
Historic Diversity	-0.506	-0.700	-0.217	-0.200
	[0.117]***	[0.199]***	[0.027]***	[0.040]***
Observations	2,210	897	2,210	897
Urban Only	No	Yes	No	Yes

Notes: Controls include malaria suitability, the count of nearby murdock groups and historic diversity, measured by murdock fractionalization at a 50km boundary. All regressions include country and year fixed effects. The 2nd and 4th columns isolate the sample to just areas coded as urban. Columns 1 and 3 control for urban areas as a dummy.

7.3 Panel Evidence with Drought Instruments

Next we want to leverage contemporary changes in population and fractionalization within cities, over time. We would like to estimate the equation:

$$Y_{it} = \alpha + \beta_1 \sigma_{it} + \beta_2 \mu_{it} + \gamma_t + \upsilon_i + \epsilon_{it} \tag{13}$$

Where the drought shock instruments measure decade level drought intensity and drought distribution, while our outcomes are decade level changes in fractionalization or population. We include γ_t decade fixed effects and υ_i city or region fixed effects. Figure ?? shows a scatterplot of decade-decade population changes compared to changes in fractionalization. Interstingly, we see limited correlated between our measures of population and DHS composition changes.

Table 15 shows results from a regression on changes in urban log population, measured across decades in the Africapolis dataset, and the average value of the drought shock instruments in the same decade. We find evidence that drought intensity pushes lower fractionalization in a panel setting, and more concentrated droughts lower population growth in the DHS sample. In Table 16, we find evidence that in the census regions drought intensity is associated with higher population growth.

Table 15: Drought Shock Impact on City-Level Characteristics (Panel)

	DHS Fract	DHS Fract	Pop Growth	Pop Growth
Drought Intensity	-0.010	-0.018	-0.003	-0.002
	[0.001]***	[0.003]***	[0.003]	[0.003]
Drought Distribution	-9.926	-13.352	-0.581	-0.275
	[12.219]	[12.028]	[0.135]***	[0.168]
Mean Dep.	0.475	0.475	0.406	0.406
Observations	574	574	16,033	16,033
CityFE	Yes	Yes	Yes	Yes
DecadeFE	No	Yes	No	Yes

Notes: Panel regressions include city fixed effects, while the second column adds decade fixed effects. Growth regressions include population data from 1960-2015, while the fractionalization regressions include DHS observations pooled at 1990, 2000 and 2010.

7.4 2SLS Results with Drought Instruments

The relationship of instruments to outcomes in a first stage vary across our samples. Given that our DHS sample represents more granular population and fractionalization observations at a city level, we proceed with this sample to attempt to instrument for population and fractionalization. For both the cross-section and panel datasets, we first instrument just for fractionalization in a regression on lights, controlling for population:

$$Y_{it} = \alpha + \beta_1 \hat{Fract}_{it} + \beta_2 Pop_{it} + \gamma_t + \upsilon_i + \epsilon_{it} \tag{14}$$

Table 16: Drought Shock Impact on Census-Level Characteristics (Panel)

	Ethnic HHI	Ethnic HHI	Log Pop	Log Pop
Drought Intensity	-2.248	-1.361	8.182	0.561
	[1.268]*	[1.928]	[0.582]***	[0.107]***
Drought Distribution	-0.513	0.731	11.355	0.666
	[0.412]	[2.074]	[0.189]***	[0.116]***
Mean Dep.	0.490	0.490	11.889	11.889
Observations	272	272	272	272
CityFE				
DecadeFE				

Notes: Panel regressions include city fixed effects, while the second column adds year fixed effects.

Where

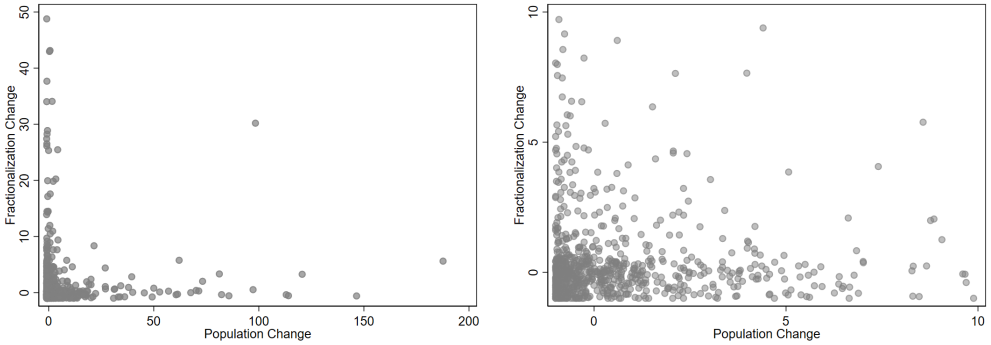
$$Fract_{it} = \alpha + \beta_1 \sigma_{it} + \beta_2 \mu_{it} + \gamma_t + \upsilon_i + \epsilon_{it} \tag{15}$$

In a second model, we instrument for both population and fractionalization. We find that fractionalziation, as predicted by decade level drought shock and distribution, predicts higher light density. That is, cities exposed to more diverse climate migrants see higher light density in that same decade.

Table 17: 2SLS Population and Diversity on Lights, DHS Data

	Cross	Cross	Panel	Panel
PopGrowth			0.002	-0.297
			[0.010]	[2.525]
Fractionalization	-0.829	-1.659	3.719	3.713
	[1.969]	[3.747]	[0.797]***	[0.841]***
Population 2000	0.758	0.595		
	[0.127]***	[0.747]		
Mean Dep.	0.091	0.091	-0.241	-0.241
Observations	513	513	817	817
F-stat	1.777	0.374	17.831	0.083
CityFE	N	N	Y	Y
Instrument Population	N	Y	N	Y

Notes: Panel regressions include city fixed effects, while the second column adds year fixed effects.



(a) Fractionalization Change by Population Change

(b) Fractionalization Change by Population Change (Zoom in)

8 City Size and Diversity

8.1 Heterogeneity by City Size

In this section I consider potential nonlinearities in the relationship between city size and regional diversity. The model of city growth posited in Appendix Section A suggests that cities grow at the cost of diversity, and that cities located in homogeneous regions are able to grow larger without absorbing these costs. For a given city, we might expect diversity costs to be a function of city size. For example, it may be the case that a large metropolis is better able to manage high levels of diversity than a smaller city. Policymakers have become increasingly interested in the role of secondary and tertiary cities in African urbanization; it's important to understand how these different city sizes are exposed to benefits and costs of migration and ethnic conflict (ADB, 2022).

Figure 14 look at the correlational relationship between diversity, conflict and growth at different population sizes. The scatterplots show the beta coefficients of a regression of a diversity measures on light density in 2000, at different quantiles of population measured by the interpolated Worldpop figures for 2000. We see that in identified cities, very large populations see a high association of diversity with growth, as measured by light density. The relationship is mixed when the analysis includes all grids, regardless of city status in the Africapolis dataset. Figure 15 shows the same regressions for the conflict probability outcome. We see that larger cities have a stronger positive association between diversity and conflict. This might be expected from the general spatial equilibrium implications of my model, where cities pay for the larger size with increased conflict.

Why might we expect larger population cities to differentially benefit from diversity? One hypothesis might be that larger cities feature higher rates of industrial or residential segregation. If different groups live and work in different areas of a city, we might expect this

to limit effects of ethnic conflict among neighbors or worker teams. Alternatively, it may be that larger cities feature different kinds of production, including more services or more complex manufactured products. These production processes may feature higher levels of labor intensity relative to more agricultural areas, forcing ethnic groups to collaborate more closely to produce output. This kind of mechanism is at work in Fiszbein et al. (2022), which shows that US regions that produce crops requiring higher labor intensity also show lower rates of individualist norms.

8.2 Occupational Segregation and City Size

Do ethnic groups in large cities work together, or in different industries? How do these patterns affect the relationship between diversity and growth? A thesis put forth in Glaeser et al. (1995) suggests that residential segregation in moderately diverse cities may reduce ethnic conflict between groups, and contribute to city growth.

To explore the dynamics of ethnic diversity and size further I use a collection of censuses available for a subset of African countries at IPUMS International. The %10 samples provide a representative measure of ethnicity, occupation and industry codes that can be aggregated to a county level (administrative level 2). For a given county I calculate statistics meant to capture the county's diversity, industrial concentration, and the segregation of ethnic groups across industries.

Given a set of industries or ethnic groups $i \in I$ in county c, industry concentration and ethnicity concentration are measured according to a standard HHI, where we sum the squared share of each group relative to the county population:

$$HHI_c = \sum_{i=1}^{I} (\frac{N_i}{N_c})^2 \tag{16}$$

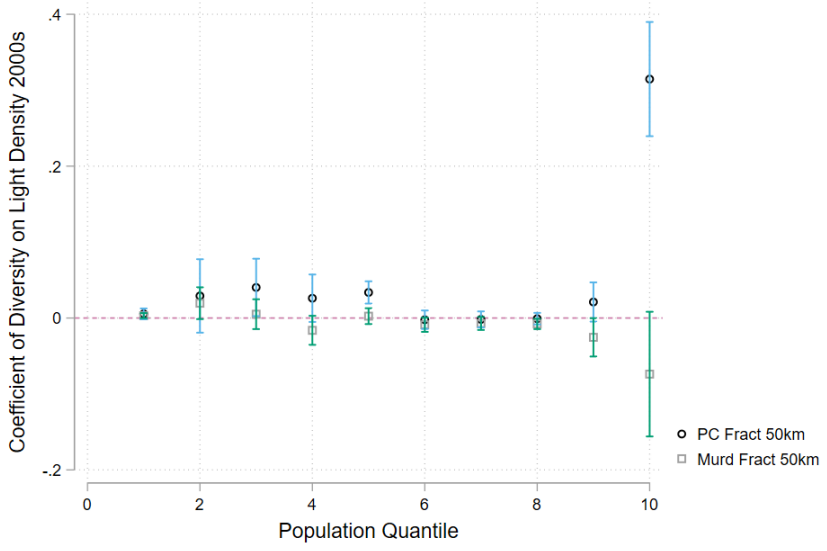
To measure the segregation of ethnic groups $j \in J$ across industries $i \in I$, I follow Alesina and Zhuravskaya (2011) and calculate industry segregation in county c of country m as:

$$Seg_c = \frac{1}{J-1} \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{N_j}{N_c} \frac{(\pi_{ij} - \pi_j)^2}{\pi_j} \tag{17}$$

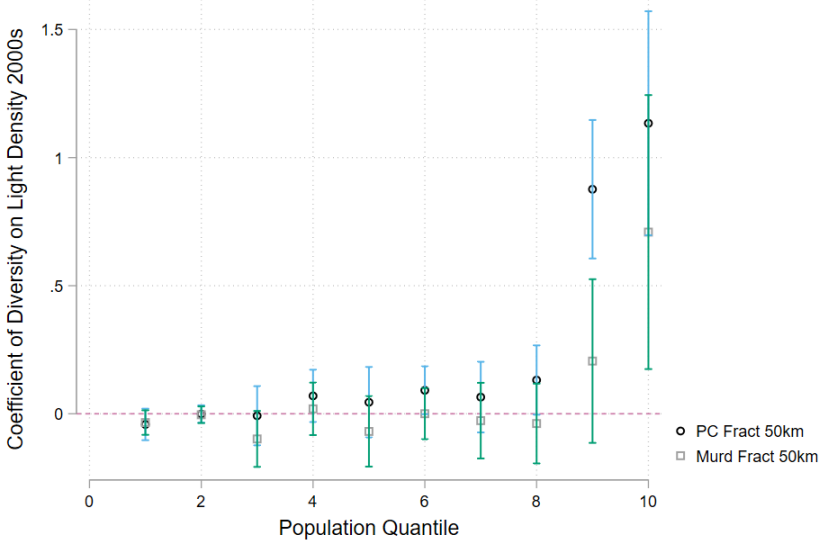
Where π_j is the fraction of group j in county c, and π_{ji} us the fraction of group j in industry i of county c. N_j is the total population of group j in county c and N_c is the total population in county c. Higher values of S correspond to greater segregation of ethnic groups across industries within a county.

In figure 16 and table 18 I regress these measures on light density. I find that high segregation is correlated with greater night lights, in particular in larger cities.

Figure 14: Diversity Association with Nighttime Lights by City Size

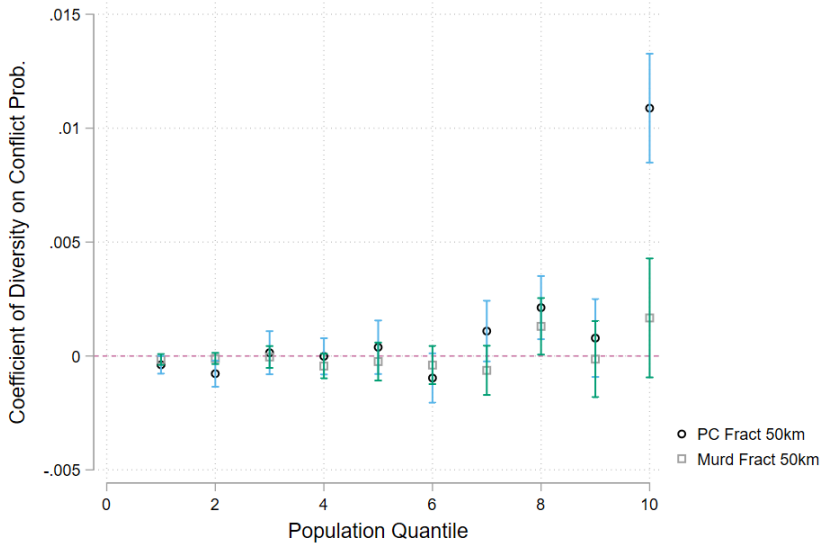


(a) All Grids

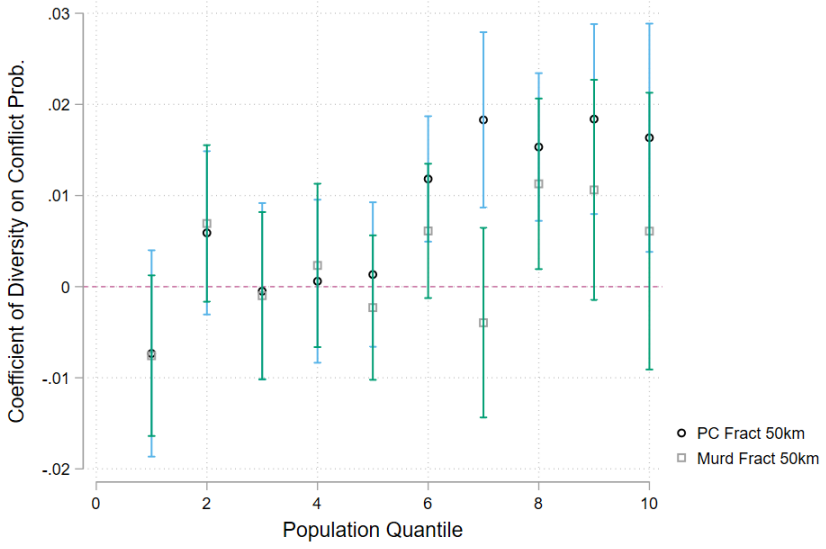


(b) Only Grids in Africapolis Cities Dataset

Figure 15: Diversity Association with Conflict Probability by City Size

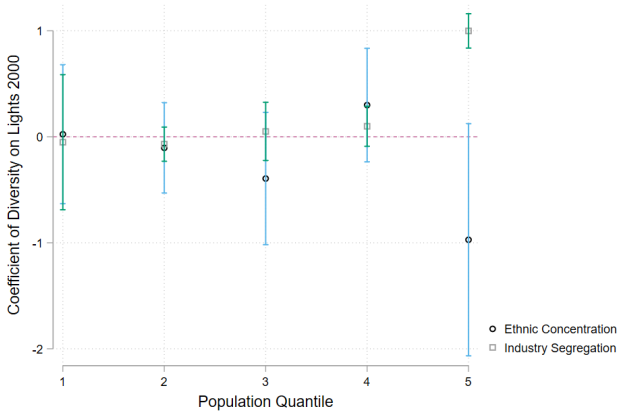


(a) All Grids



(b) Only Grids in Africapolis Cities Dataset

Figure 16: Census Diversity and Light Intensity by Population



Note: This figure shows coefficients from a regression of $Lights_c = \alpha + \beta_1 X_c + \epsilon_c$, run by population quintile. I control for census year, dummy for urban region, and fixed effects for state and country. Data is at county level (administrative level 2). The variable X is either Ethnic HHI or county level segregation respectively.

Table 18: Census Diversity and Light Intensity

	Lights 2000	Lights 2000	Lights 2000
Industry Concentration (HHI)	-2.737		
	[0.291]***		
Ethnic Concentration (HHI)		-0.368	
		[0.249]	
Industry-level Segregation			0.592
			[0.145]***
Population	0.682	0.654	0.426
	[0.180]***	[0.183]***	[0.130]***
Urban	-0.565	0.314	0.230
	[0.113]***	[0.061]***	[0.046]***
Observations	2,384	2,384	2,253

Note: This table shows coefficients from a regression of $Lights_c = \alpha + \beta_1 X_c + \epsilon_c$, controlling for current population, census year, dummy for urban region, and fixed effects for state and country. Data is at county level (administrative level 2). The variable X is the Industry HHI, Ethnic HHI and county level segregation respectively.

8.3 Residential Segregation and City Size

Segregation by residence is another way to measure the extent to which groups interact in a city. Few African countries have any data on ethnicity across neighborhoods within a city. Here we can leverage the fact that DHS sampling collects neighborhood-level data by surveying multiple adjacent households at a clustering point. Comparing different clusters within the same city consitutes a partial sampling of neighborhood-level characteristics. We can then measure residential segregation by comparing the ethnicity composition of neighborhood-level clusters to the city-level ethnic composition of all the city's sample clusters aggregated together.

Figure 17 shows an example of DHS clusters within the city of Lagos, colored by their respective fractionalization indexes. For a given sample year t, we compute residential segregation of a city i using a multi-group dissimilarity index. Given a city with neighborhoods j and ethnic groups m, this is calculated as:

$$D_{it} = \sum_{m=1}^{M} \sum_{j=1}^{J} \frac{t_j}{T} |\pi_{jm} - \pi_m| \tag{18}$$

Where π_{jm} is the fraction of group m in neighborhood j, π_m is the fraction of group m in the city, and t_j is total population in neighborhood j.

The accuracy of this exercise relies on having sufficient clusters within a city, as well as their spatial coverage of the region. To test the accuracy of this measure we can run simulations that compare this sparse sampling method to ground truth residential segregation measured using census data. I run simulations of this nature using both US and Indonesia census data, where I construct artificial DHS samples of neighborhoods that resemble our African samples. The Indonesian data is particularly helpful here because we can compare our census estimates to our simulated DHS samples, as well as to real DHS samples taken for the country.

Figure 18 shows a scatterplot of our measured residential segregation by city size. We see evidence of increased residential segregation by city size. Table 19 shows regressions of our measured residential segregation on a variety of city characteristics, including night light density, probability of conflict and historic ethnic diversity, controlling for city population. Segregation seems to strongly predict light density, even when controllingfor the city's overall ethnic fractionalization. We also find that our residential segregation measure seems unrelated to our historic measures of ethnic diversity from anthropological maps.

9 Conclusion

An open question in the urban economics of developing countries is whether increased density in poor countries has and will achieve the same agglomeration returns as in rich countries.

Figure 17: DHS Clusters within Lagos

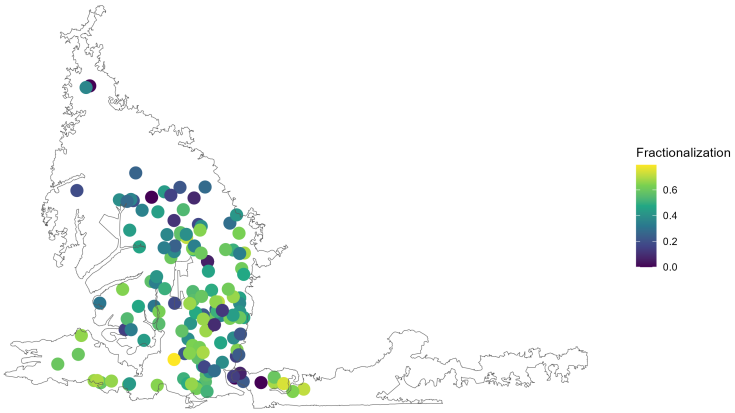


Figure 18: Residential Segregation by City Size

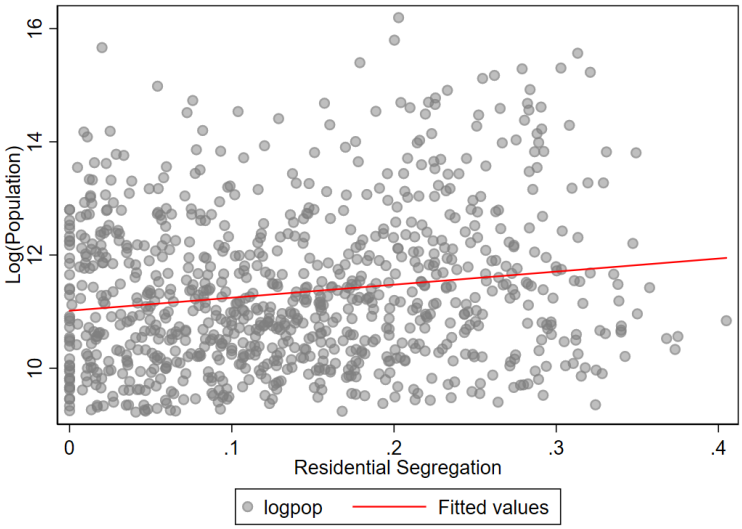


Table 19: Residential Segregation and Development

	Log(Night Lights)	Conflict Prob.	Murdock Div.
Segregation	3.152	0.287	0.001
	[0.895]***	[1.283]	[0.604]
Fractionalization	-0.307	0.467	0.867
	[0.337]	[0.483]	[0.228]***
Population	0.561	0.055	-0.146
	[0.040]***	[0.058]	[0.027]***
Mean Dep.	0.774	0.946	0.547
Observations	898	898	898

Note: All columns control for population, and contain country and year fixed effects.

This paper considers whether Africa's history of ethnic division has long term effects on the development and growth of cities. In particular, I show that cities exogenously placed in more diverse regions during the colonial period are less developed today relative to cities placed in homogeneous ethnic homelands. This paper suggests one particular mechanism, which is that diverse regions make it more difficult for cities to source labor without also generating a higher probability of conflict in the urban center. Worker ethnic diversity works as a congestion force, limiting the benefits of agglomeration in dense African cities.

I use two identification strategies to leverage exogenous variation in city location that is independent from historical regional ethnic diversity. Cities in the colonial period were more likely to emerge along colonial railways, which were often constructed according to the least-cost path between a coastal port and an inland natural resource. Cities were also likely to emerge at portage sites, where steep elevation change meets an inland river due to the need to move goods from ships to land transport. Both distance to railway and geographic propensity as a portage site predict city emergence. I use this fact to compare the development of cities placed in more or less diverse regions. Cities placed in more diverse areas, as measured by a compilation of anthropological data sources, are less developed and more prone to conflict today.

More work needs to be done to fully identify and explore the labor supply mechanism that might explain how cities are constrained by ethnic diversity. In future work, I hope to explore how city size and firm organization may be able to mitigate the congestion effects of worker diversity on city expansion.

Lastly, this work may have implications for how contemporary migrations affect the success of African cities. A next step will be to understand how contemporary "push shocks" that move migrants into cities affect fractionalization and growth. While contemporary migrants generally increase the population of urban areas, they may or may not affect fractionalization depending on the spatial nature of the shock. Modelling the political economy of ethnic conflict as a "congestion force" will help us account for these different effects.

References

ADB (2022). *Dynamics of Systems of Secondary Cities in Africa*. Cities Alliance, Brussels.

Adhvaryu, A., Fenske, J., Khanna, G., and Nyshadham, A. (2021). Resources, conflict, and economic development in Africa. *Journal of Development Economics*, 149:102598.

Aker, J. C., Klein, M. W., O'Connell, S. A., and Yang, M. (2014). Borders, ethnicity and trade. *Journal of Development Economics*, 107:1–16.

Alesina, A. and Ferrara, E. L. (2005). Ethnic Diversity and Economic Performance. *Journal of Economic Literature*, 43(3):762–800.

Alesina, A. and Zhuravskaya, E. (2011). Segregation and the quality of government in a cross section of countries. *American Economic Review*, 101(5):1872–1911.

Almagro, M. and Dominguez-Iino, T. (2022). Location sorting and endogenous amenities: Evidence from amsterdam. *Available at SSRN 4279562*.

Arbatli, C. E., Ashraf, Q. H., Galor, O., and Klemp, M. (2020). Diversity and Conflict. *Econometrica*, 88(2):727–797.

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

Bleakley, H. and Lin, J. (2012). Portage and Path Dependence *. *The Quarterly Journal of Economics*, 127(2):587–644.

Bleakley, H. and Lin, J. (2015). History and the Sizes of Cities. *American Economic Review*, 105(5):558–563.

Bryan, G. and Morten, M. (2019). The aggregate productivity effects of internal migration: Evidence from indonesia. *Journal of Political Economy*, 127(5):2229–2268.

Diamond, R. (2016). The determinants and welfare implications of us workers' diverging location choices by skill: 1980–2000. *American Economic Review*, 106(3):479–524.

Duranton, G. and Puga, D. (2020). The Economics of Urban Density.

Eaton, J. and Kortum, S. (2002). Technology, geography, and trade. *Econometrica*, 70(5):1741–1779.

Esteban, J., Mayoral, L., and Ray, D. (2012). Ethnicity and Conflict: An Empirical Study. *American Economic Review*, 102(4):1310–1342.

Fiszbein, M., Jung, Y., and Vollrath, D. (2022). Agrarian origins of individualism and collectivism. Technical report, National Bureau of Economic Research.

Fortes-Lima, C. A., Burgarella, C., Hammarén, R., Eriksson, A., Vicente, M., Jolly, C., Semo, A., Gunnink, H., Pacchiarotti, S., Mundeke, L., et al. (2024). The genetic legacy of the expansion of bantu-speaking peoples in africa. *Nature*, 625(7995):540–547.

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

Gisselquist, R. M., Leiderer, S., and Niño-Zarazúa, M. (2016). Ethnic Heterogeneity and Public Goods Provision in Zambia: Evidence of a Subnational "Diversity Dividend". *World Development*, 78:308–323.

Glaeser, E. L., Scheinkman, J., and Shleifer, A. (1995). Economic growth in a cross-section of cities. *Journal of monetary economics*, 36(1):117–143.

Gollin, D., Kirchberger, M., and Lagakos, D. (2021). Do urban wage premia reflect lower amenities? evidence from africa. *Journal of Urban Economics*, 121:103301.

Guedes, R., Iachan, F. S., and Sant'Anna, M. (2023). Housing supply in the presence of informality. *Regional Science and Urban Economics*, 99:103875.

Gören, E. (2014). How Ethnic Diversity Affects Economic Growth. *World Development*, 59:275–297.

Harari, M. (2020). Cities in bad shape: Urban geometry in india. *American Economic Review*, 110(8):2377–2421.

Heinrigs, P. (2020). Africapolis: understanding the dynamics of urbanization in africa. *Field Actions Science Reports. The journal of field actions*, (Special Issue 22):18–23.

Henderson, J. V., Storeygard, A., and Deichmann, U. (2017). Has climate change driven urbanization in africa? *Journal of development economics*, 124:60–82.

Hjort, J. (2014). Ethnic divisions and production in firms. *The Quarterly Journal of Economics*, 129(4):1899–1946.

Jedwab, R., Kerby, E., and Moradi, A. (2017). History, Path Dependence and Development: Evidence from Colonial Railways, Settlers and Cities in Kenya. *The Economic Journal*, 127(603):1467–1494.

Jedwab, R. and Moradi, A. (2016). The Permanent Effects of Transportation Revolutions in Poor Countries: Evidence from Africa. *The Review of Economics and Statistics*, 98(2):268–284.

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

Kramon, E., Hamory, J., Baird, S., and Miguel, E. (2022). Deepening or diminishing ethnic divides? the impact of urban migration in kenya. *American Journal of Political Science*, 66(2):365–384.

Lehner, B. and Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, 27(15):2171–2186.

Li, X., Zhou, Y., Zhao, M., and Zhao, X. (2020). A harmonized global nighttime light dataset 1992–2018. *Scientific data*, 7(1):168.

Lowes, S. (2017). Matrilineal kinship and spousal cooperation: Evidence from the matrilineal belt. *Unpublished manuscript. URL: https://scholar. harvard. edu/files/slowes/files/lowes_ matrilineal. pdf.*

McGuirk, E. F. and Nunn, N. (2024). Transhumant pastoralism, climate change and conflict in africa. *Review of Economic Studies*, page rdae027.

Melitz, J. and Toubal, F. (2014). Native language, spoken language, translation and trade. *Journal of International Economics*, 93(2):351–363.

Michaels, G., Rauch, F., and Redding, S. J. (2012). Urbanization and Structural Transformation *. *The Quarterly Journal of Economics*, 127(2):535–586.

Michalopoulos, S. (2012). The origins of ethnolinguistic diversity. *American Economic Review*, 102(4):1508–1539.

Montalvo, J. G. and Reynal-Querol, M. (2021). Ethnic Diversity and Growth: Revisiting the Evidence. *The Review of Economics and Statistics*, 103(3):521–532.

Monte, F., Redding, S. J., and Rossi-Hansberg, E. (2018). Commuting, migration, and local employment elasticities. *American Economic Review*, 108(12):3855–90.

Mueller, H., Rohner, D., and Schönholzer, D. (2022). Ethnic Violence Across Space. *The Economic Journal*, 132(642):709–740.

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

Notowidigdo, M. J. (2020). The incidence of local labor demand shocks. *Journal of Labor Economics*, 38(3):687–725.

Nunn, N. and Puga, D. (2012). Ruggedness: The blessing of bad geography in africa. *Review of Economics and Statistics*, 94(1):20–36.

Paolillo, J. C. and Das, A. (2006). Evaluating language statistics: The ethnologue and beyond. *Contract report for UNESCO Institute for Statistics*.

Pérez-Sindín, X. S., Chen, T.-H. K., and Prishchepov, A. V. (2021). Are night-time lights a good proxy of economic activity in rural areas in middle and low-income countries? examining the empirical evidence from colombia. *Remote Sensing Applications: Society and Environment*, 24:100647.

Porteous, O. (2019). High trade costs and their consequences: An estimated dynamic model of african agricultural storage and trade. *American Economic Journal: Applied Economics*, 11(4):327–366.

Ramankutty, N., Foley, J. A., Norman, J., and McSweeney, K. (2002). The global distribution of cultivable lands: current patterns and sensitivity to possible climate change. *Global Ecology and biogeography*, 11(5):377–392.

Saiz, A. (2010). The geographic determinants of housing supply. *The Quarterly Journal of Economics*, 125(3):1253–1296.

Schlebusch, C. M. and Jakobsson, M. (2018). Tales of human migration, admixture, and selection in africa. *Annual Review of Genomics and Human Genetics*, 19:405–428.

Semo, A., Gayà-Vidal, M., Fortes-Lima, C., Alard, B., Oliveira, S., Almeida, J., Prista, A., Damasceno, A., Fehn, A.-M., Schlebusch, C., et al. (2020). Along the indian ocean coast: genomic variation in mozambique provides new insights into the bantu expansion. *Molecular Biology and Evolution*, 37(2):406–416.

Smith, R. (1970). The canoe in west african history1. *The Journal of African History*, 11(4):515–533.

Sundberg, R. and Melander, E. (2013). Introducing the ucdp georeferenced event dataset. *Journal of Peace Research*, 50(4):523–532.

Ullman, E. L. (1970). A Theory for the Location of Cities. In *A Geography of Urban Places*. Routledge. Num Pages: 10.

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

A Model

What follows is a spatial equilibrium model with endogenous amenities that are a function of many groups.

- set S locations indexed by i
- workers come from J discrete groups, where $g_{ji} \in (g_{1i}, g_{2i}...g_{Ji})$ denotes the number of workers in group j in region i
- each group has fixed total size $\bar{L}^j$ st. $\sum_{i=1}^{S} g_{ji} = \bar{L}^j$
- workers have total mass $\bar{L} = \sum_{j=1} \bar{L}^j = \sum_{i=1}^{S} L_i$

A.1 Labor Demand

- Production is conducted by many identical firms with free entry at each location producing a homogeneous good according to $y_i = A_i L_i$
- Productivity of a firm d at location i is driven by number of workers, and can face a conflict cost C defined at the city-level as a function of group mix in location i: $A_{di} = \bar{A}_i L_{di}^{\alpha} C_i(g_{1i}, g_{2i}...g_{Ni})^{-\gamma}$
- Because conflict is defined at the city level, each individual firm takes this cost as given and simply chooses a number of L workers such that

$$W_{di} = (\alpha + 1)\bar{A}_i L_{di}^{\alpha} C_i(g_{1i}, g_{2i}...g_{Ni})^{-\gamma} \tag{19}$$

Adding up across firms in location i we have total labor demand in city i:

$$ln(W_i) = ln(\alpha + 1) + ln(\bar{A}_i) + \alpha ln(L_i) - \gamma ln(C_i(g_{1i}, g_{2i}...g_{Ni})) \tag{20}$$

Note here that labor demand is impact by the agglomeration effects of added labor L_i, as well as the relative composition of that labor in the function C_i, where $\sum_{t=1}^{N} g_{tn} = L_i$. In order to disentangle these effects, we need to add structure to the particular equation for city-wide conflict. We can add structure to the equation for conflict at location i as proportional to the fractionalization or more simply the HHI across groups:

$$C_i = \omega \sum_{x=1}^{N} (\frac{g_{xi}}{L_i})^2 \tag{21}$$

A.2 Labor Supply

- Workers from ethnic group j are born in region o and decide where to reside i, where they receive the equilibrium wage w_i.

- Workers in i also receive a location-specific amenity that is negatively related to total population $L_i^{-\beta}$, and positively related to their relative group's share in the local labor force $(\frac{g_j i}{L)i})^{\upsilon}$
- Each individual worker t also receive an idiosyncratic preference shock for each region ϕ_{it} that is distributed Frechet $F(z) = e^{-z^{-\theta}}$
- Moving from origin o to i incurs a migration cost τ_{oi}
- Given the above, worker t from group j and birthplace o receives the following total utility if they move to i:

$$U_{jiot} = \frac{w_i}{P_i} L_i^{-\beta} (\frac{g_{ji}}{L_i})^{\upsilon} \tau_{oi} \phi_{it} \tag{22}$$

  Where P_i is the local price index.

Given the extreme value distribution, we can follow Eaton and Kortum (2002) to get an expression for the migration flows between an origin o and destination i for a given group. The proportion of people from origin o and type j who choose to work in i is:

$$\pi_{io}^{j} = \frac{(\frac{w_i}{P_i} L_i^{-\beta} (\frac{g_{ji}}{L_i})^{\upsilon} \tau_{oi})^{\theta}}{\sum_i (\frac{w_i}{P_i} L_i^{-\beta} (\frac{g_{ji}}{L_i})^{\upsilon} \tau_{oi})^{\theta}} \tag{23}$$

Then we can add these shares across origins to get:

$$L_{ji} = \sum_o \pi_{io}^{j} L_{jo}^{0} \tag{24}$$

Where L^0 represents the initial distribution of workers across groups and origins.

$$L_i = \sum_j L_{ji} \tag{25}$$

Note that the total labor demand L_i for a region is a function of two particular matrices of parameters, L^0 and τ. L^0 is a $J \times S$ matrix marking the distribution of initial workers by group J and origin region S. τ is the $S \times S$ matrix of migration costs between each origin and destination. For a given region i, $L^0 \times \tau_i$ produces a $J \times 1$ vector that captures the region's exposure to each ethnic group, weighted by migration costs. This vector can be summarized as a region's potential diversity exposure, which we proxy in the empirical exercise with a fractionalization index within a given radius.

A.3 Production

- Following Bryan and Morten (2019), I assume that a representative firm maximizes the economy-wide production Y that aggregates the regional varieties y_i according to

:

$$Y = (\sum_{i=1}^{S} y_d^{\frac{\sigma-1}{\sigma}})^{\frac{\sigma}{\sigma-1}} \tag{26}$$

The individual prices for regional products p_d are pinned down by the representative firm maximizing the production of Y subject to costs $\sum_d p_d y_d$. This gives us $p_d = (\frac{Y}{y_d})^{\frac{1}{\sigma}}$. Individual workers consume the final good Y, and it enters linearly into utility. We take this price as the numeraire.

Under perfect competition, we have that firms offer a wage equal to the marginal product of labor: $w_i = A_i$.

A.4 Identifying Parameters with Productivity Shocks

The labor demand equation 20 includes C_i, an endogenous function of contemporary diversity in region i. In the empirical section, we use a proxy for this measure using the region's fixed exposure to diverse potential workers from the historic distribution of ethnicities, which we call an area's fractionalization D_i. This index can be thought of as a kind of summary statistic for a region's diversity exposure, $L^0\tau$. The interaction of the fractionalization index and labor demand L_i proxies the increase in diversity generated by increased labor demand. Therefore the adapted labor demand equation we estimate is:

$$ln(W_i) = ln(\alpha + 1) + ln(\bar{A}_i) + \alpha ln(L_i) - \gamma Frac_i * ln(L_i) + \omega Frac_i \tag{27}$$

In lieu of granular wage data, we utilize light density as a measure of output y. We can use the portage and rail instruments as productivity shocks to identify the labor demand equation 27. The endogeneity concern stems from the unobserved fundamental productivity $ln(\bar{A}_i)$. Both the rail and portage instrument can be thought of as a region-specific productivity shock that shifts the region's fundamental productivity and the associated labor demand independent of the underlying distribution of ethnic groups. This means that given a regional productivity shock Z_i, we can instrument for L_i. We can also instrument for $Fract * ln(L_i)$ using the interaction $Frac * Z_i$.

The city dummy used in the empirical exercise is effectively capturing a threshold of labor demand, and is therefore a proxy for historical labor demand L_i.

A.5 Computing Spatial Equilibrium

We have a labor market clearing condition so that total labor $\bar{L} = \sum_i L_i$ In addition, each group j has a set number of workers $\bar{L}^j$ such that $\bar{L} = \sum_j \bar{L_j}$. We reach a spatial equilibrium by first producing a guess for the J matrices Π_{io}^j, which are $S \times S$ giving the proportion of people working in i for each origin o in group j (each element in the matrix is a realized π_{io}^j).

I then proceed as follows:

- Calculate the new shares π_{io}^{j}.
- Summarize the shares across groups j to create a total origin destination matrix $\Pi_{io} = \sum_j \Pi_{io}^{j}$
- Iterate until convergence of Π_{io}

B Portage Score Validation

B.1 Calculating Portage Score

The portage score is calculated from the interaction of a grid's ruggedness and distance to the nearest river. I standardize the log of the variable ruggedness and distance to river, setting the mean to 100 to avoid negative values. The distance to the river is multiplied by -1, so that higher values correspond to more ruggedness and closer to river. These two variables are then interacted and standardized, creating score such that higher values mean a grid is more rugged and closer to a river.

B.2 Validating Portage Score with Hydrological Data

The portage score is meant to capture points at which a river network transitions from navigable water to rapids or waterfalls. The intersection of land ruggedness and the river is meant to capture these points, assuming that waterfalls and rapids are generated by rapid elevation changes along a river. This approach is very different from that of Bleakley and Lin (2012), which identifies portage sites using the interaction of the US river network and the Atlantic Seaboard Fall Line. To validate that my portage score is capturing rapids and waterfalls, I relate the score to hydrological measures taken from the HydroSHEDS database on rivers, which gives a granular record of average discharge at points along the river, as well as a rating of flow volume on a categorical scale. Using these variables, I calculate for each river segment the variation in discharge and flow by collecting all river points within a 30km radius. The flow variation and discharge variation of a river segment capture any substantial changes in river speed and volume at a given point, which may be related to the presence of rapids or other sharp changes to river navigability.

Figure B1 shows an example of the flow change measure on the river network of the Democratic Republic of Congo. The overlapping red dots mark cities identified in the Africapolis dataset. We see high values for flow variation near the mouth of the river, where Kinshasa is located after a series of rapids that limit navigability from the coast. Figure B2 shows a transport map for the DRC, combining highway and waterway info from Michelin and the UN. We can see that Kinshasa, Kiangani and Kasongo are all located before or

after significant rapids. We also see that several major cities such as Mbandaka are located at important river forks.

Table B1 shows the grid-level relationship between portage score and the flow and disharge variation of the river segment closest to the grid. The analysis is restricted to grids that lie near a river segment. We see that portage score is significantly associated with the river segment's variability in terms of flow and discharge, suggesting that ruggedness along a river network is predictive of the river's hydrological variability, which in turn is associated with rapids and waterfalls.

Figure B1: Flow Variation in DRC River Network

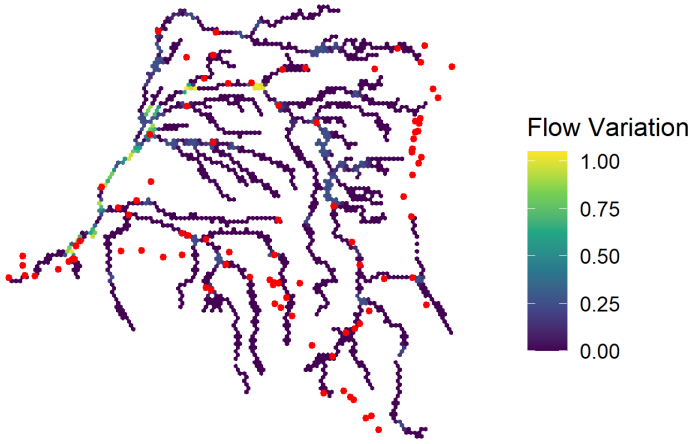


Table B1: Portage Score and Hydrological Features

	Portage Score	Portage Score	Portage Score	Portage Score
Discharge Variation	0.021		0.021	
	[0.003]***		[0.003]***	
Flow Variation		0.252		0.442
		[0.068]***		[0.064]***
Dist to River	<50km	<50km	<100km	<100km
Mean Dep. Var	1	1	0	0
Observations	23,219	23,219	37,460	37,460

add

C Additional Figures & Tables

Figure B2: DRC River Transport

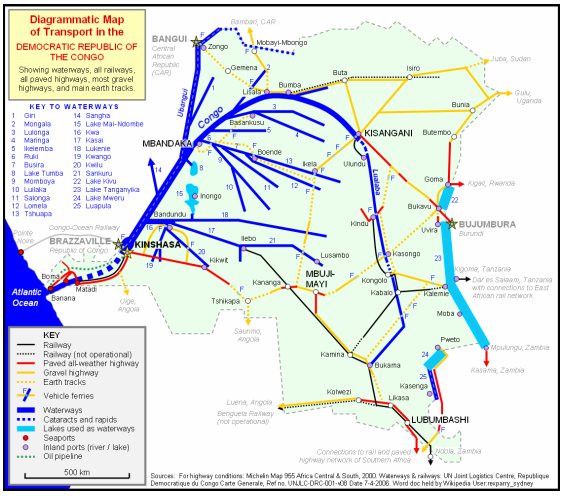


Table B2: Correlation across Measures of Ethnic Diversity

	Murd. F	GREG F	Lang. C	Murd. C	GREG C	Lang. C	PC
Murd. F	1.00						
GREG F	0.43***	1.00					
Lang. C	0.44***	0.49***	1.00				
Murd. C	0.80***	0.47***	0.50***	1.00			
GREG C	0.33***	0.59***	0.54***	0.41***	1.00		
Lang. C	0.38***	0.43***	0.69***	0.52***	0.71***	1.00	
PC	0.72***	0.73***	0.80***	0.80***	0.78***	0.81***	1.00
Observations	85511						

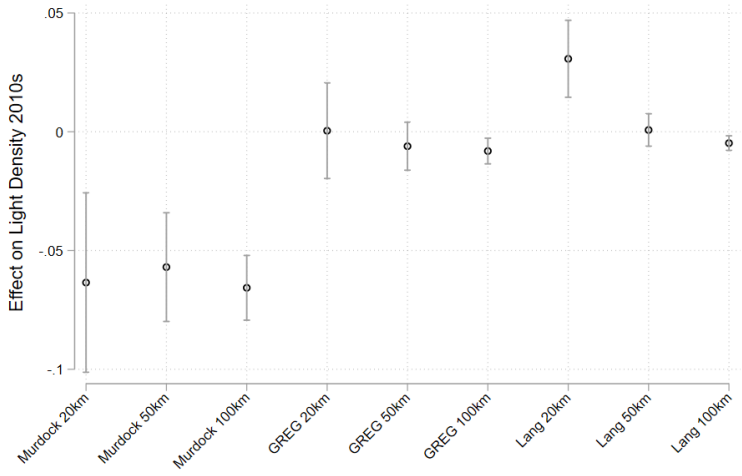
Notes: All measures are aggregated using the 50km radius around grid centroids. "F" labels denote measured fractionalization using shares of land (except the DHS measures, which uses shares of people). This is calculated as Fract $= \sum_1^m n_i(1-n_i)$ where n_i is the proportion of area covered by group m within the area of grid i's buffer. "C" denotes a count of the number of intersecting ethnic groups within the grid's buffer range. "PCA" is the principal component of the murdock, GREG and language fractionalization and count measures within the 50km boundary.

Table B3: Relationship of Diversity Measures to Census Diversity

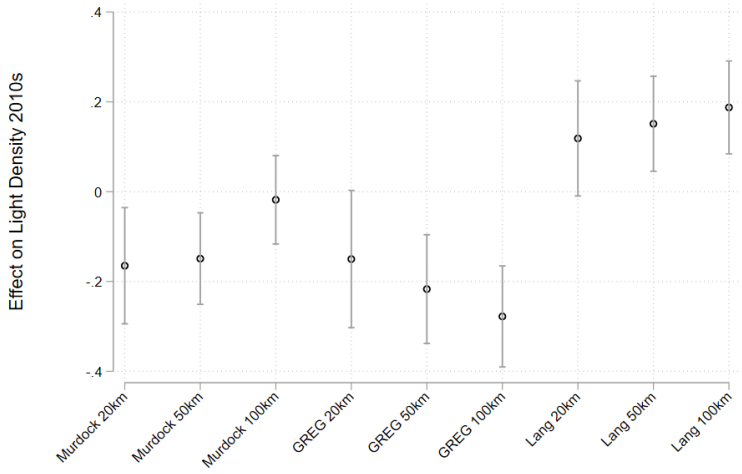
	Murd Fract	PC Fract	Murd Count	Fract Lang	Fract GREG
Census Ethnic Concentration	-0.691	-0.394	-0.491	-0.104	-0.092
	[0.080]***	[0.075]***	[0.088]***	[0.018]***	[0.016]***
Observations	2,384	2,384	2,384	2,384	2,384

Notes: The census sample includes %10 samples from Benin (1979,1992,2002,2013), Ethiopia (1994), Ghana (2000,2010), Guinea (2014), Malawi (2008), Mali (2009), Mauritius (2000,2011), Morocco (2014), Senegal (2013), Sierra Leone (2004), Togo (2010), Uganda (2002), Zambia (2000, 2010). Ethnic fractionalization for country i is calculated as $\sum_{j=1}^{J}(\frac{g_j}{N_i})^2$ where g_j is the number of people from ethnic group j, and N_i is the total sampled population of the county. Higher levels of ethnic concentration imply less diversity. Regressions include country fixed effects.

Figure B3: Effect of Diversity Across Diversity Definitions



(a) Count Measures of Diversity



(b) Fractionalization Measures of Diversity

Table B4: Rail IV - Light Density 2000s

	Lights	Lights	Lights	Lights	Lights	Lights	Lights	Lights
City*Murd Fract	-0.176	-0.187	-0.053	0.010				
	[0.082]**	[0.093]**	[0.097]	[0.103]				
Murd Fract	0.021	0.019	-0.004	-0.016				
	[0.012]*	[0.013]	[0.019]	[0.019]				
City	0.754	0.378	1.137	1.077	0.712	0.375	1.160	1.076
	[0.098]***	[0.118]***	[0.109]***	[0.112]***	[0.102]***	[0.119]***	[0.109]***	[0.110]***
City*PC Fract					0.029	-0.094	-0.114	-0.009
					[0.066]	[0.082]	[0.099]	[0.104]
PC Fract					0.015	0.032	0.064	0.019
					[0.015]	[0.017]*	[0.029]**	[0.029]
Rail FE	N	Y	N	Y	N	Y	N	Y
Dist to Rail	<300km	<300km	<100km	<100km	<300km	<300km	<100km	<100km
F-stat	353	193	312	267	358	181	222	197
Mean Dep. Var	-0.032	-0.032	-0.005	-0.005	-0.030	-0.030	-0.002	-0.002
Observations	41,436	41,436	17,763	17,763	40,257	40,257	17,268	17,268

Notes: Controls include land suitability, malaria suitability, ruggedness. All regressions include country and rail fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid. Light density measures are also standardized after averaging across years 2000-2009 and 2010-2013.

Table B5: Rail IV - Conflict Deaths

	Deaths	Deaths	Deaths	Deaths	Deaths	Deaths	Deaths	Deaths
City*Murd Fract	6.003	16.103	5.657	30.030				
	[40.768]	[43.704]	[50.469]	[51.353]				
Murd Fract	-5.341	-8.688	-5.541	-17.157				
	[16.404]	[17.511]	[24.694]	[25.755]				
City	-47.038	-59.089	-66.664	-69.749	-53.936	-61.160	-29.250	-32.314
	[41.747]	[61.540]	[45.654]	[50.135]	[40.932]	[57.106]	[50.313]	[49.577]
City*PC Fract					13.051	18.103	-74.428	-54.633
					[32.793]	[40.108]	[80.348]	[73.967]
PC Fract					-9.194	-11.403	42.699	33.182
					[15.946]	[19.388]	[46.046]	[44.846]
Rail FE	N	Y	N	Y	N	Y	N	Y
Dist to Rail	<300km	<300km	<100km	<100km	<300km	<300km	<100km	<100km
F-stat	57	35	29	29	51	25	8	9
Mean Dep. Var	31.800	31.800	24.635	24.635	31.807	31.807	24.635	24.635
Observations	4,143	4,143	1,752	1,752	4,142	4,142	1,752	1,752

Notes: Controls include land suitability, malaria suitability, ruggedness. All regressions include country and rail fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid.

Table B6: Portage IV - Light Density 2000s

	Lights	Lights	Lights	Lights	Lights	Lights	Lights	Lights
City*Murd Fract	0.178	-0.064	-0.019	-0.342				
	[0.219]	[0.185]	[0.450]	[0.310]				
Murd Fract	-0.050	-0.010	-0.065	0.001				
	[0.029]*	[0.025]	[0.064]	[0.045]				
City	-0.830	-0.125	-2.422	-0.731	-0.723	0.084	-2.302	-0.690
	[0.436]*	[0.301]	[0.866]***	[0.465]	[0.370]*	[0.257]	[0.733]***	[0.413]*
City*PC Fract					0.212	-0.239	0.124	-0.205
					[0.179]	[0.161]	[0.358]	[0.246]
PC Fract					-0.033	0.045	-0.010	0.036
					[0.037]	[0.033]	[0.072]	[0.049]
River FE	N	Y	N	Y	N	Y	N	Y
Dist to River	<100km	<100km	<50km	<50km	<100km	<100km	<50km	<50km
F-stat	40	67	17	35	42	62	20	38
Mean Dep. Var	-0.006	-0.006	0.035	0.035	-0.004	-0.004	0.037	0.037
Observations	37,310	37,310	23,006	23,006	36,747	36,747	22,861	22,861

Notes: Controls include land suitability, malaria suitability. All regressions include country fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid. Light density measures are also standardized after averaging across years 2000-2009 and 2010-2013.

Table B7: Portage IV - Conflict Deaths

	Deaths	Deaths	Deaths	Deaths	Deaths	Deaths	Deaths	Deaths
City*Murd Fract	-34.416	-33.941	-87.729	-77.365				
	[57.427]	[57.676]	[69.136]	[60.047]				
Murd Fract	12.996	12.653	33.164	30.071				
	[20.843]	[21.252]	[26.672]	[23.648]				
City	85.432	78.145	43.575	50.795	93.738	92.076	37.363	44.863
	[77.610]	[67.777]	[77.198]	[64.187]	[80.959]	[72.191]	[81.734]	[70.122]
City*PC Fract					-33.993	-37.597	-85.255	-64.001
					[48.222]	[53.074]	[79.538]	[60.215]
PC Fract					13.437	14.443	40.487	31.998
					[22.909]	[25.506]	[38.831]	[30.399]
River FE	N	Y	N	Y	N	Y	N	Y
Dist to River	<100km	<100km	<50km	<50km	<100km	<100km	<50km	<50km
F-stat	12	17	7	10	9	12	3	7
Mean Dep. Var	28.609	28.609	27.572	27.572	28.619	28.619	27.572	27.572
Observations	3,826	3,826	2,483	2,483	3,824	3,824	2,483	2,483

Notes: Controls include land suitability, malaria suitability. All regressions include country fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid.

Table B8: Fractionalization and Dist to Rail

	Murd Fract	Murd Fract	PC Fract	PC Fract
Dist to Rail	0.176	1.175	0.270	0.389
	[0.152]	[0.297]***	[0.124]**	[0.054]***
Dist to Rail	<100km	<60km	<100km	<60km
Observations	17,667	11,224	17,173	29,964

Notes: Distance to rail and the fractionalization measures are standardized. The regressions include malaria suitability, land suitability, historic population and ruggedness as controls, as well as country and rail fixed effects.

Table B9: Rail IV - Predict City

	Prob. City	Prob. City	Prob. City
Dist to Rail	0.020	-1.541	-3.602
	[0.002]***	[0.059]***	[0.129]***
Dist to Rail		<100km	<60km
Mean Dep.	0.089	0.176	0.206
Observations	90,555	17,763	11,280

Notes: Controls include land suitability, malaria suitability, ruggedness. All regressions include country and rail fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid. The "Dist" row describes the sample cutoff of distance to nearest colonial rail for that particular regression.

Table B10: Rail IV - Light Density

	Lights 2000s	Lights 2010s	Lights 2000s	Lights 2010s	m_4
City*Murd Fract	0.011	-0.027	-0.057		
	[0.093]	[0.089]	[0.096]		
City*PC Fract				-0.060	-0.180
				[0.087]	[0.090]**
City	1.232	1.267	1.487	1.263	1.532
	[0.103]***	[0.102]***	[0.107]***	[0.105]***	[0.109]***
Murd Fract	-0.012	-0.008	-0.008		
	[0.021]	[0.021]	[0.022]		
PC Fract				0.065	0.066
				[0.028]**	[0.029]**
Rail FE	Y	N	Y	Y	Y
F-stat	331	360	331	346	346
Mean Dep. Var	0.021	0.021	0.041	0.024	0.044
Observations	11,280	11,280	11,280	10,987	10,987

Notes: Controls include land suitability, malaria suitability, ruggedness. All regressions include country and rail fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid. Light density measures are also standardized after averaging across years 2000-2009 and 2010-2013. The sample is restricted to grids within 60km of a colonial rail line.

Table B11: Rail IV - Conflict

	Prob. Conflict	Avg Deaths	Prob. Conflict	Avg Deaths
City*Murd Fract	0.005	15.493		
	[0.005]	[15.225]		
City*PC Fract			0.021	0.194
			[0.005]***	[12.277]
City	0.057	11.385	0.053	18.115
	[0.006]***	[16.224]	[0.006]***	[16.228]
Murd Fract	-0.001	-4.384		
	[0.001]	[8.456]		
PC Fract			-0.004	1.181
			[0.002]**	[7.942]
Rail FE	Y	Y	Y	Y
f-stat	331	50	346	53
Mean Dep. Var	0.013	19.116	0.014	19.116
N	11,280	1,213	10,987	1,213

Notes: Controls include land suitability, malaria suitability, ruggedness. All regressions include country and rail fixed effects. Fractionalization measures are standardized, and defined using a 50km buffer from the grid centroid. Prob. conflict is defined as the proportion of years in which the grid experienced a conflict across 1975-2021. The sample is restricted to grids within 60km of a colonial rail line.