Density and Diversity in African Cities*

Andre Gray†

October 31, 2025

Abstract

The impact of migration is not just a function of how many people migrate, but where they come from. Migrants carry region-specific identities, traits and skills that shape outcomes in receiving areas. In rapidly urbanizing African cities, the composition of migrants may play a negative role, as ethnic and linguistic divisions drive conflict and counteract classic agglomeration forces. This paper disentangles the effects of migrant flows and migrant composition on productivity in destinations. I build a subnational panel of internal migrant flows across Africa and develop a nonlinear shift-share instrument that identifies shocks to both levels and the birthplace composition of migrants. Using exogenous variation from climate, commodity and conflict shocks, I identify changes to the size and composition of migrants. I find that cities that receive migrants from more diverse birthplaces have lower short-run growth, but experience long-run urbanization benefits. The effects of migrant composition are heterogenous, with more diverse cities experiencing higher ethnic conflict, but also higher rates of structural transformation. The methods proposed have broad applications to identifying nonlinear effects of migration, when relative group sizes matter for outcomes.

*I would like to thank my faculty advisors Sam Bazzi and Sara Lowes for their guidance and support.
†Department of Economics, UC San Diego, email: adgray@ucsd.edu.

1 Introduction

Measuring the impact of migrant workers is usually done in levels. The size of migrant inflows, or the average level of their skill, changes labor demand and supply, prices and traffic. Past work has argued that as the flow of migrants increases, they can raise the productivity of destinations through agglommeration effects. An increase in the raw number of people in a city stimulates competition and generates new ideas. If agglomeration effects exist, then the level of migrant flows can scale-up city populations and in turn productivity growth (Glaeser and Gottlieb, 2009). But to achieve these effects, migrants have to interact with each other and with the local population. These interactions might be smooth or difficult, and the particular skills, backgrounds and social networks that migrants bring with them may be more or less complementary. The complementarity of migrant workers is a function of their composition; the relative sizes of different groups at destination. In this paper I disentangle these two effects, density and diversity, in the context of African rural-urban migration. I find that both increasing migrant labor size and diversity produce lower per-capita city growth, indicative of short-run congestion effects from ethnic conflict. However, higher migrant diversity also predicts increases to non-agricultural labor share, suggesting that migrant diversity plays an important role in structural transformation.

The urban share of population in Africa rose from 31 to 54% between 1990 and 2020 (Christiaensen et al., 2025). Across African cities about a third of the urban labor force is composed of migrants from rural parts of the country (Christiaensen et al., 2023). In developed countries, the most productive cities are both dense and cosmopolitan. New York or London host a diversity of industries, amenities and services that benefit from a wide range of workers with a healthy mix of skills. Whether developing country cities grow under the same conditions is an open question. In the context of Africa, there are reasons to be skeptical that cities benefit from either density or diversity. First, African cities suffer from high congestion costs and poor urban infrastructure. Traffic, pollution, poor housing and small industrial sectors may all prevent cities from taking advantage of increased density (Castells-Quintana, 2017). Second, African countries are saddled with ethnic and religious conflict. A literature in political economy documents correlations between measures of ethnic and linguistic diversity, low GDP growth and high incidence of violent conflict (Arbatli et al., 2020; Alesina et al., 2003; Robinson, 2020). While the mechanisms are not well understood, microeconomic evidence has shown that ethnic divisions can directly lower the productivity of firms (Hjort, 2014).

This paper offers a framework to synthesize the literatures on agglomeration and ethnic conflict. Density and diversity are jointly determined by the size of the labor force, and the mix of worker types. As migrant flows increase or decrease, they increase or decrease the level of density in a city. If migrants are all from the same origin, then they lower the relative diversity of the workforce. If migrants come from many different origins, the relative diversity of the migrant pool increases. Capturing the effects of migrants on urban Africa requires an empirical strategy that can disentangle the effects of migrant levels and composition. To identify these parameters separately, I construct a shift-share instrument that predicts linear changes in migrant flows, and nonlinear changes in migrant diversity. Then, I consider the role of pull and push shocks in driving changes in migrant levels and composition. I estimate the impact of climate, commodity prices and conflict on the size and spatial pattern of migrant flows in Africa.

The paper begins by creating a proxied origin-destination panel of African migrant flows. To overcome data gaps in the measurement of African migration, I use the universe of publically available African censuses and leverage a mix of worker birthplace and language data to identify worker origins. While past work on African migration has relied on cross-sectional data or small household panels, this origin-destination panel captures aggregate changes in population size and composition by origin between censuses.

Destinations are exposed to migrants from different birthplaces according to a pre-period settlement share of workers from a given origin o living in destination d. Each origin birthplace has a country-wide outmigration "shift", measured as the total number of people from that birthplace

observed living outside of that homeland across census periods. The combination of the aggregate shifts with the destination-specific shares is an instrument for migrant labor size, or the "levels" of migration. To capture the effects of composition, I add a second shift-share instrument. This "composition" instrument captures how changes in the distribution of the outmigration shifts predictably alter a Herfindahl Index (HHI) of migrant shares. The result is a linear and nonlinear shift-share instrument to predict changes in both migrant levels and birthplace composition, as measured by an HHI of birthplace shares. The instruments capture predicted changes in migrant levels and composition that are plausibly exogenous to contemporaneous labor demand shocks.

To address concerns about destination productivity changes being engodenous to birthplace-level outmigration shifts, I build on the shift-share strategy by replacing the outmigration shifts with predicted outmigration rates. Following work on historical immigration to the US by Boustan et al. (2010), I predict outmigration from each origin using a set of plausibly exogenous shocks related to climate, conflict and international prices. In a zero stage I show that drought reduces outmigration, while conflict and high commodity prices boost outmigration rates. Leveraging these variables to predict our shifts, I re-estimate the 2SLS model for levels and composition.

While migrant settlement shares are often treated as an endogenous component in Bartik-style migration instruments, serial correlation between pre-period settlement shares and contemporary labor demand may bias results. In a last exercise I attempt to isolate exogenous variation in shares that comes from a zero stage regression of origin-destination distances and historical destination characteristics. These "pull characteristics" predict a destination's overall attractiveness to migrants based on historical productivity shocks that are plausibly unrelated to contemporary demand. "Pull characteristics" can be thought of as historical instruments for agglomeration. Examples include distance to colonial railroads, historical mineral deposits, and portage sites.

The outcomes of interest measure different aspects of a destination's urban growth. I use different functions of changes in satellite night-light luminosity as proxies for changes in city size and GDP per capita. I also use census measured non-agricultural labor share as a proxy for urbanization, and a principle component of housing characteristics as a measure of local average household wealth. Results show that in-migration to a destination increases city light density, but not per-capita growth rates. Higher migrant levels scale up city sizes, but don't yield higher productivity growth. A higher diversity of migrants coming from more origins decreases both light density and per-capita growth. These effects are large. Relative to other developing countries, the negative lighte density growth effects of migrant labor size are 4 times larger in an African context, and the congestion effects of diversity are twice as large. These results suggest that congestion forces created by migrants may constrain the potential benefits of increasing density. However, I find heterogenous effects when examining non-light based measures of productivity. More diverse migrant pools increase the non-agricultural labor share, an indicator of urbanization and structural transformation. Migrant labor size by itself has a modest and negative effect on non-agricultural labor share. This heterogeneity points to the importance of separately identifying size and composition effects of migration.

Increasing migrant labor size and birthplace diversity reduce short-run per-capita light density growth. I consider several mechanisms to explain these results, including ethnic conflict, linguistic diversity and firm size. I find evidence consistent with the congestion effects of ethnic conflict described in the cross-sectional political economy literature. Both increasing migrant labor size and diversity predict higher rates of conflict in receiving cities. In a final section of the paper, I explore the possible long-run benefits of migrant diversity suggested by the positive impact on non-agricultural labor share. I leverage my constructed historical agglomeration pull factors in a long-run version of the empirical strategy to explore the cross-sectional relationship of diversity, population and productivity across African cities. I find that cities located in more diverse areas benefitted more from historical productivity shocks in the long-run. I conclude that migrant diversity brings short-term costs, but may play a beneficial role in long-run structural change.

This paper makes a contribution to the identification of linear and nonlinear effects of migration. The literature on developed country migration regularly quantifies the effects of migrant flows on

growth, often parsing results by skill level. Recently, researchers have begun to consider the importance of broader categories of worker type in explaining the heterogeneous effects of immigration on destinations. Alesina et al. (2016) directly studies the role of birth country diversity on destination outcomes using a gravity model to predict both the share and diversity of immigrants. They find that immigrant diversity is positively associated with measures of TFP and patent intensity. However, their estimation strategy does not consider the simulteneous role of migrant labor size, and risks conflating the effect of birthplace diversity with migrant labor size. A more recent literature has sought to microfound the role of particular ethnic or national groups in productivity growth through knowledge spillovers or home-country connections (Boberg-Fazlić and Sharp, 2024; Choi et al., 2024; Burchardi et al., 2019; Imbert et al., 2022). Another strand of work has studied the impact of ethnic enclaves and different assimilation rates on the migrant labor market (Albert et al., 2021). Assimilation rates, ethnic enclaves and knowledge-spillovers are all a function of the composition of migrants. My paper develops a framework to aggregate these ideas into a shift-share design that can disentangle the effects of pure increases in labor size from compositional changes in migrant worker types.

In developing contexts, a few papers have considered the direct effect of birthplace-specific human capital on destination productivity Bazzi et al. (2016). Most work on ethnicity and migration in development has been interested in estimating linguistic or cultural distance as a migration cost, rather than evaluating the aggregate effects on destinations (Wang, 2024). In political economy, studies of African diversity have been mostly correlational or leveraging long-run instruments (Alesina and Ferrara, 2005; Alesina et al., 2003; Arbatli et al., 2020; Ashraf and Galor, 2013). This work has generally found that more linguistically or culturally diverse countries see higher conflict, lower public goods and lower GDP growth. At subnational levels, Montalvo and Reynal-Querol (2021) find this correlation flips direction, hinting at the idea that agglomeration benefits and diversity play joint roles in city-level outcomes. This correlational work has suffered from two key empirical concerns: (1) historical measures of local diversity are correlated with local productivity factors and geography. The distribution of ethnic groups across space is related to many unobservable productivity fundamentals (Michalopoulos, 2012). (2) Diversity is intimately related to population size – historically diverse places tend to also have relatively more people. Montalvo and Reynal-Querol (2021) note that many historical trade centers, now larger cities, were founded at the intersection of ethnic boundaries. My approach addresses these concerns first by estimating a differenced equation that leverages changes in population and diversity within destinations over time. Second, I design two shift-share instruments to separately identify the effects of migrant labor size and birthplace diversity.

This paper also speaks directly to the literature on the returns to migration. While this literature is large, work on Africa in particular is relatively sparse due to data gaps. Few surveys capture internal migrant flows, or even area of previous residence. Most results in African settings have relied on household panels that track individuals over several waves, or cross-sectional data with simple markers of migrant status. Young (2013) and Gollin et al. (2021) use Demographic Health Surveys (DHS) to quantify large differences in consumption and amenities between rural and urban areas. Hamory et al. (2021) and Lagakos et al. (2020) use individual fixed effects in household panels to compare migrant returns before and after migrating from rural to urban areas, and find relatively lower estimates of urban premia. Related work looks for evidence of agglomeration forces, showing increasing returns to density across African cities (Henderson et al., 2021; Castells-Quintana, 2017).

Across these papers migrants are identified in coarse categories as coming from a “city", “town" or “rural area". The data is either not large enough or not detailed enough to consider origin-destination pairs, making it hard to estimate migration costs. Relatedly, while these papers consider heterogeneous migration costs across origins, which may include linguistic differences, there's no conception of complementarities between migrants of different origins. This paper contributes in two ways. First, I use a combination of birthplace data and ethnic linking methods to tie workers at destination to specific regional origins. This novel proxied origin-destination panel is able to answer more specific questions about the role of particular origins on migration returns. I show that my

proxied origin-destination migration panel can replicate key patterns in standard migration gravity models and estimates of migration returns. Second I bring shift-share methods to an African setting, allowing me to calculate a density premium by leveraging exogenous shifts in migration levels, rather than indvidual returns.

Finally, I contribute to a separate literature that considers the role of shocks on migration. My paper uses climate, commodity price and conflict shocks as plausibly exogenous intstruments for shifts in outmigration. Kamuikeni and Naito (2024) and Henderson et al. (2017) find drier conditions in African countries increase the rate of internal migration and urbanization respectively. Henderson et al. hypothesize that cities that specialize in non-agricultural production provide a safe-haven for rural migrants. McGuirk and Nunn (2024) links drying conditions with increased conflict in Africa between pastoralists and farmers. Few papers study the effects of price shocks on migration levels or composition in African countries (Brückner, 2012). Gollin et al. (2016) consider the role of commodity exports on urbanization, finding that higher value exports also drive urbanization at the country level. My work most relates to a separate thread in the political economy of conflict that identifies the association of granular price shocks of agriculture and minerals with conflict (Bazzi and Blattman, 2014; McGuirk and Burke, 2020). This paper is a more granular study of push shock dynamics, using origin-destination decisions and subnational exposure to shocks as a first-stage to predict migrant flows.

This paper brings a novel perspective to the impact of shocks on migration. Suppose that where migrants come from is an important variable in the production function of a destination. If both levels and composition of migrants matter, then the level and composition of shocks matters as well. An intense climate or price shock that affects a singular region of a country may drive many migrants that are relatively homogenous in terms of language or skill. A shock that is broader, affecting many locations at different intensities will create a more diverse flow of outward migrants. I show that different push-shocks contribute differently to migrant flows, affecting both levels and composition simultaneously.

A key innovation of this paper is the use of nonlinear functions of shift-share instruments to capture nonlinear effects of migration. The methods used to disentangle size and composition of migrant flows can be applied to a number of other contexts where the impact of flows have linear and nonlinear components, such as in the study of skill-complementarity, assimilation, segregation or industry concentration (Lewis, 2011). Immigration shocks do not only cause changes in population levels or average skill level. The relative sizes of immigrant populations can affect the rate at which group assimilate, the formation of enclaves, rates of crime and conflict and the organization of firms.

2 Data

An ideal dataset to study the aggregate impacts of migration would capture changes in migrant flows by destination and origin over time. Few African data sources contain granular migration flows within or across countries. To measure changes in migrant composition, I construct a proxy for origin-destination panel data by linking people at a destination to their origin birthplace or ethnic homeland. The main analysis uses the subset of available African censuses that include subnational information on household location and either birthplace, ethnicity or mother tongue. The spatial granularity is at the second administrative level, which corresponds to counties or districts. While the main tables use censuses to capture changes in flows over time, I supplement this evidence in different parts of the paper with geolocated Demographic Health Surveys (DHS) and Afrobarometer Surveys. These cross-sectional household surveys collect data on fewer households but contain more variables and years. Both DHS and Afrobarometer contain information on workers' ethnicity, mother tongue or tribe. Using the ethnic linking methods described below, I can identify the spatial origins of individuals in DHS and Afrobarometer, allowing me to construct a panel at the sample cluster to ethnic homeland pair level.

Next I discuss the inputs used in producing exogeneity in the outmigration shifts and settlement

shares. “Push shocks" are used to predict changes in outmigration rates for different origins. The variables used include drought events, agricultural and mineral price shocks, and geolocated conflict events. “Pull characteristics" are used to predict settlement shares. These are characteristics that are correlated with early settlement in a destination, but are unrelated to contemporary labor demand shocks. I use colonial rail locations, mineral deposits, and river characteristics to predict which destinations attracted migrants historically.

Next I discuss the productivity outcomes used. I use a variety of functions of night light density that been used in past work on urban productivity. I also leverage census data that captures measures of industry-specific labor shares, and individual's housing characteristics as a wealth proxy. The last part of this section explores observational returns to migration using the proxy origin-destination dataset, and finds patterns that are consistent with more standard origin-destination panels of migration.

2.1 Proxied Origin-Destination Panels

2.1.1 Census data

When available, African censuses are taken from IPUMS. In a few cases I supplement with census data from the World Bank microdata portal or government statistical websites. Cesuses are typically spaced about 10 years apart. The majority of the censuses are 10% random samples. The census years cover the period between 1970 and 2020. The main analysis will focus on the period 1990 - 2020, which is the window of time in which satellite data of nighttime light density is available.

A subset of censuses record a worker's birthplace at the second administrative level. I link each individual at a given destination and year to their birthplace, which I call their “origin". Because subnational birthplace is reported for migrants moving within country, this paper focuses on the effects of internal migrants. Aggregating the data to the origin-destination-year level, I can observe decade level shifts in the number of people from a given birthplace, at a given destination. This panel is unbalanced, as each country completed censuses in different years. For a given country, the unit of analysis is changes in migrant labor and composition across census years for a given destination-year.

I supplement the birthplace data using an ethnic linking strategy. In some censuses workers report their ethnicity or native language, but not their birthplace. Because ethnic groups were historically organized into regional territories, I can use this information to connect them to their ethnic homeland of origin. I use the Linking Ethnographic Database (LEDA) by Müller-Crepon et al. (2022) to map each reported ethnicity to an ethnic homeland region on the Murdock Map, a common ethnographic resource used in African political economy. This linking procedure allows me to match a worker to a spatial origin or “homeland". Figure 17 shows an example of the groups plotted in the Murdock map, most of which resemble counties in size. I can run the same analysis using these ethnic homelands as our “origins", while destinations are still reported at the second administrative level. See Table B1 for a list of the censuses used in the analysis, including both samples with birthplace or ethnicity-linked origin data.

2.1.2 DHS and Afrobarometer Panel data

The shift-share results are based on the constructed panel of censuses. To supplement some analyses, I use standard geolocated household surveys. The Demographic Health Surveys (DHS) and Afrobarometer are household surveys with detailed information on assets, housing quality, identity and the geolocation of households (BenYishay et al., 2017; Boyle et al., 2024). Afrobarometer includes various questions about ethnic and economic attitudes. These variables are useful in understanding city-level and individual-level outcomes over shorter time horizons. What the surveys lack is detailed information on migration behavior [1]. Since most of the surveys ask about ethnicity or home lan-

[1] A subset of DHS surveys ask individuals about their region of previous residence. Typically these regions are listed at the first administrative level such as provinces, or may be as broad as identifying North vs. North-West of the country. Ethnic homelands are a more granular and consistent spatial unit.

guage, I can use the same LEDA linking procedure as described above for censuses. Leveraging the ethnic information in DHS and Afrobarometer as an indicator of migrant origin allows me to study migration with a detailed origin subnational unit.

2.2 Push Shocks to Outmigration

In an exercise to isolate exogeneity of shifts in our shift-share design, I consider a variety of shocks that may push migrants from origins. Past empirical work on migration has used instruments for income shocks that affect an origin region's outmigration rate but are unrelated to productivity changes at destination. I explore several potential candidates for push shocks at origin based on past work. These include climate shocks, international commodity prices and conflict events. While all these subnational shocks have been used in other contexts, my paper is among the first to study these shocks in the context of an origin-destination migration panel for Africa.

2.2.1 Climate

A number of recent papers have studied climate change effects on migration in developing countries (Desmet and Rossi-Hansberg, 2024; McGuirk and Nunn, 2024; Kamuikeni and Naito, 2024). The hypothesis guiding this work is that drier conditions create negative income shocks in agricultural areas, which then affects migration decisions. For instance, a major drought in East Africa in 2011 forced a reported 920,000 people from Somalia to flee to neighboring countries, many of whom settled in cities like Nairobi. The main climatic variable of interest is a measure of drought conditions. I use data from the Standardised Precipitation-Evapotranspiration Index (SPEI), which measures drought intensity monthly by combining temperature and precipitation data (Vicente-Serrano et al., 2010). The data provides monthly estimates of drought intensity at a 0.5x0.5 degree cell resolution from 1900-2022, which I aggregate into yearly estimates. For each subnational unit, I calculate the average drought experience over time. A month-cell is under normal conditions when the SPEI index is around 0, which means there is balance between the precipitation rate and the potential evapotranspiration. I code the month as "in drought" if the index is a standard deviation lower than zero. This is the threshold for extreme dryness suggested by the index (Vicente-Serrano et al., 2010). For each region, I 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). As secondary climate variables, I also consider rainfall and temperature separately.

2.2.2 Agricultural and Mineral Commodities

Many rural areas in Africa produce commodities for export, including cash crops like maize and minerals like gold or diamonds. As international prices fluctuate, producers in different regions are differentially exposed to these potential income shocks. Migration may then respond positively or negatively to exogenous changes in international commodity prices, weighted by local commodity-specific exposure. To construct a subnational measure of price exposure for key commodities, I use data on both the share of production in commodities and international prices. I construct a yearly time series of the global market prices of key commodities spanning the years 1960-2024. I start from a price list assembled by (Bazzi and Blattman, 2014) which tracks major agricultural and mineral commodity prices from 1960-2009. I then manually extend this list using available data from the IMF Primary Commodities price system, and the World Bank Pink Sheet (Group, 2025; IMF, 2025). For mineral prices, I mainly rely on the US Geological Survey's (USGS) "Historical Statistics for Mineral and Material Commodities", which covers US prices of major minerals back to 1900 (Kelly et al., 2010).

I create a measure of subnational exposure to various agricultural commodities using FAO Global Agro-Ecological Zones (GAEZ) production maps, which estimate the cell-level average hectares dedicated to a set of major commodity crops across the continent (Berman and Couttenier, 2015). These

production areas are estimated in the year 2000, and are used as baseline exposure. Product prices are normalized to 100 in 2000 and summed with weights by the hectare area. Exposure is weighted by the share of productive hectares dedicated to that crop. For each crop product p grown in region o:

$$AgriculturalPriceExposure_{ot} = \sum_{p} Price_t * HectareShare_{op} \tag{1}$$

To account for price shocks over longer time horizons, I construct different measures of lagged price exposure. A 10 year lagged exposure is an average of price exposure between censuses:

$$PriceShock_{ot} = \frac{1}{10} \sum_{t-10}^{t} PriceExposure_{ot} \tag{2}$$

For mineral commodities, I weight exposure to given minerals using the "USGS Compilation of Geospatial Data (GIS) for the Mineral Industries and Related Infrastructure of Africa" (Kelly et al., 2010). This dataset contains geolocated mineral facilities and their estimated capacity. For each mineral, I measure the total capacity across countries. Then each mineral producing region is exposed to a given mineral price according to its relative share of total production capacity in that mineral. The total mineral price exposure for a region is then:

$$MineralPriceExposure_{ot} = \sum_{m} Price_t * RelativeCapacity_{om} \tag{3}$$

Where relative capacity in mineral m for region o is given by:

$$RelativeCapacity_{om} = \frac{Capacity_{om}}{Capacity_m} \tag{4}$$

2.2.3 Conflict

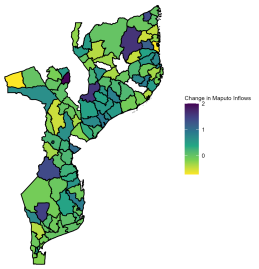
The Armed Conflict Location & Event Data (ACLED) provides geolocated conflict events across Africa[2] The ACLED data covers events between 1997 and 2025. The data is based on the manual coding of news articles, and provides basic information on each event including approximate dates, estimated death toll, and the types of actors involved. For each region, I aggregate all conflict events that appear in that region to produce an annual count of conflict events. Conflict intensity in a given year is the average number of conflict events per month that year. As a secondary outcome, I produce an index of average conflict faitalities in a given region-year.

2.3 Pull Characteristics to Predict Immigration

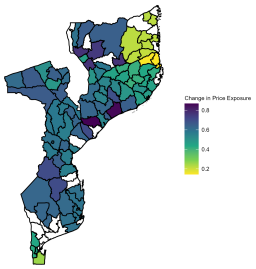
Traditional migration shift-share instruments rely on endogenous settlement shares of different immigrant groups from past years. If settlement shares are serially correlated over time, its possible that the settlement shares are endogenously related to contemporary labor demand shocks. For robustness, I'll also consider a group-specific settlement share that is instrumented by a collection of destination and origin-destination characteristics. In particular I instrument for past settlement using the interaction of the distance between origin-destination pairs and a set of "pull characteristics". These pull characteristics are designed to predict the attractiveness of destinations but are plausibly uncorrelated with current period demand shocks. I consider several historical shocks that predict urban formation, including the presence of mineral deposits, colonial railroads and portage sites. Data for mineral deposits is taken from the "USGS Compilation of Geospatial Data (GIS) for the Mineral Industries and Related Infrastructure of Africa". Data on colonial railroad projects is taken from the universe of colonial rail projects collected by Jedwab and Moradi (2016). Last, I construct an instrument for portage leveraging data on river flow from the HydroSHEDS database

[2]Different events have different degrees of confidence in the precise geolocation. I only include events where the researchers have marked having confidence in the geolocation at the second administrative level.

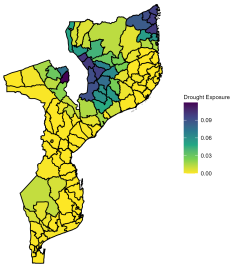
Figure 1: Data Examples from Mozambique 1997-2007



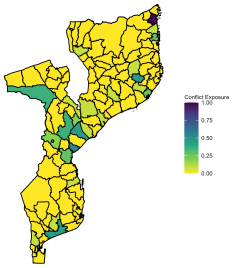
(a) Change in Migrant In-Flow to Maputo



(b) Change in Agricultural Price Exposure



(c) Drought Exposure (SPEI)



(d) Conflict Exposure

Notes: This figure shows examples of the data from one census pair, specifically Mozambique 1997 and 2007. Panel A shows the change in migrant flows to Maputo city from each birthplace between 1997 and 2007. Panel B shows the change in the weighted agricultural price exposure between 1997 and 2007. Panel B shows the average drougth intensity by region between 1997 and 2007. Panel D shows the average annual conflict events in each region over this time period. Data for the change in migrant in-flows comes from the publically avaialble African censuses on IPUMS International. Data for agricultural price exposure is taken from IMF Primary Commodities and the World Bank Pink Sheets, combined with FAO production exposure. Data for drought exposure is calculated from an SPEI index (Vicente-Serrano et al., 2010), and Conflict Exposure data is from ACLED.

(Lehner and Grill, 2013). The details of these instrument constructions are described in the empirical strategy.

Last, I leverage a set of common geographical characteristics as controls in cross-sectional regressions. These include the ruggedness index from Nunn and Puga (2012), malaria suitability from Kiszewski et al. (2004), Tse Tse suitability from Alsan (2015) and soil suitability from Ramankutty et al. (2002).

2.4 Destination Productivity Outcomes

I use a variety of subnational proxies for regional productivity. I consider two functions of changes in light density as measures of urban extent and GDP per-capita growth. I also leverage census-level measures that capture structural transformation and urbanization characteristics including non-agricultural and services labor shares. Most censuses also include data on information on housing quality, such as the material of the roof and floor of the surveyed household. I construct a principle component measure of these material characteristics as a proxy for household wealth. All outcomes are measured in terms of differences between census years.

2.4.1 Light Density

Nighttime light density is a commonly used proxy for economic development, especially at subnational levels. A variety of papers have shown correlations between measures of luminosity and other measures of wealth from either surveys of household income or administrative level human capital and non-agricultural labor share (Michalopoulos and Papaioannou, 2013; Chiovelli et al., 2023). There are three issues that must be dealt with to use luminosity as a proxy for wages or productivity. First, the light density data produced before 2013 and after 2013 are at different resolutions,

which complicates comparisons over time [3]. I use a harmonized dataset produced by Li et al. (2020) that performs an inter-calibration to combine the datasets and generates a DN-style output at 30 arc-seconds which ranges from 0 to 63.

The second issue is dealing with blooming, where high light density in a given cell may bleed into neighboring cells producing unwanted spillovers, say from a large city to a small neighboring town. The harmonized dataset partially adjusts for this, and administrative regions are large enough to somewhat mitigate this concern. Additionally, in the robustness analysis I use spatially correlated standard errors to address contamination across regions.

The third issue is how to construct an appropriate measure of economic growth from the pixel-level light data. Different papers use different constructions. Chiovelli et al. (2023) aggregate pixels by summing over their regions of interest, and then including controls in their regression for log population and region area. Montalvo and Reynal-Querol (2021) use changes in log light density divided by a gridded population estimate as a measure of local economic growth. I consider two light density constructions in differences to capture the size and growth of destinations. The first is the standardized level change in average light density in a given administrative region. This measure captures a change in light density levels, and can be thought of as a proxy for city size growth or extent. The second measure is the change in the log of light density over population. This measure captures the proportional change in light density per capita, and is used as a proxy of GDP per capita. As cities approach the maximum value of measured light density at the pixel level, there may be frontier effects. Large productive cities may show low light density growth because there is no further improvement to be made. I try to mitigate these concerns by controlling for pre-period light density in the differenced regressions. I also consider separate measures of productivity from asset data and industry shares.

$$\Delta Log(Lights/Capita)_t = Log(Lights/Capita)_t - Log(Lights/Capita)_{t-1} \tag{5}$$

2.4.2 Assets and House Quality

I follow the spirit of Young (2013) using the ownership of durables and housing conditions as proxies for household wealth. DHS surveys include a rich set of variables that describe ownership of different durable assets like a TV, bicycle, car or microwave. I use these dummy indicators for assets to construct a wealth score as a first principal component of these assets. This procedure is similar to the DHS's own measured wealth score, which leverages a first principal aggregation of asset categories. I only include the subset of assets available in all surveys, to ensure comparability across survey samples. In the census data, assets are more sparse. However, most censuses include information on the house, wall and roof materials. Leveraging this fact, I replicate the procedure for DHS estimations of wealth by taking a first principal component of a set of house quality measures for each census where it's available. Averaging these estimates across individuals in a district gives us a subnational measure of housing quality. Table 1 shows that these principal component measures are positively correlated with education and age, suggesting that the method accurately captures relative wealth.

2.4.3 Industry and Urbanization

As Gollin et al. (2016) highlights, growth in cities can take place in both tradable and non-tradable sectors. A common refrain about African cities is that they seem to grow without an expansion in the manufacturing sector; so-called "consumption cities" (Jedwab et al., 2025). I use two measures of urbanization growth as outcomes: the total non-agricultural labor share and the non-tradable services

[3] Before 2013, satellite imaging comes from the Defense Meteorological Satellite Program (DMSP)/Operational Linescan System (OLS), which provides digital number (DN) values at a resolution of 30 arc-seconds. After 2013, the Visible Infrared Imaging Radiometer Suite (VIIRS) on the Suomi National Polar-orbiting Partnership satellite provides light density data at 15 arc-seconds, and at a higher radiometric resolution. This means it can detect smaller differences in light density relative to the past technology.

labor share. Each outcome is derived from individual census data which reports general industry classifications such as agriculture, mining, wholesale trade, manufacturing, financial services, etc. These industries are aggregated into measures designed to reflect growing urbanization – as cities grow, we expect them to increase their non-agricultural share. For cities experiencing service-led growth rather than manufacturing growth, we expect "services share" to increase.

These outcomes are considered in terms of differences across census years. This is particularly important in an African data context because informal work is often underreported in census occupational data, even though it makes up a siginificant fraction of the labor. The industry measures used capture changes in the industry shares among the subset of individuals who report formal employment. The benefit of our differenced equation design is that baseline difference in formalization rates are differenced out.

2.5 Validating a Proxy Origin-Destination Panel

Using origin-destination panel data from ethnicity and birthplace differs in significant ways from typical origin-destination panels. The data does not capture flows of migrants over time, but rather the stock of individuals from particular birthplaces or homelands for a set of unbalanced years. Observing an increase in people from a given birthplace o in a destination d may reflect increased migrant flow into the destination, or decreased return migration to o. Observing an increase in people from a given ethnicity may also reflect differential changes in the population growth of that ethnicity, rather than changes in migration from an ethnic homeland.

As a first exercise I estimate observational returns to migration, and compare the results to other cross-sectional estimates as a validation exercise. I use DHS panel data, where individuals are linked to ethnic homelands. I use DHS panel for this exercise because of the detailed catalogue of individual assets, which can be used to construct a proxy for migrant wealth as an outcome. Individuals are identified as "migrants" if they are residing in an administrative region outside of their ethnic homeland. Leveraging the DHS's survey of individual assets, I construct a "durable assets" and "house quality" score. The procedure for doing this follows closely the DHS's own method for estimating wealth, transforming each asset category into a dummy variable and calculating a first principal component measure. The outcome "Durables" is the first principal component of several binary variables for different assets, including electricity, phone, car, fridge, television and bicycle. The outcome "House Quality" is the first principal component of reported wall, roof and floor material of the house. Using the constructed principal component measures of durable assets and housing quality, I compare consumption between migrants to non-migrants as:

$$\begin{aligned} Wealth_{it} = \beta_1 \text{Migrant}_i + \beta_2 \text{Migrant}_i \cdot \text{Distance}_{od} \\ + \beta_3 \text{Migrant}_i \cdot \text{CoethnicShare}_{od} + Z_i + X_{od} + W_d + \upsilon_{st} + \gamma_o + \epsilon_{it} \end{aligned} \tag{6}$$

Where $Wealth_{it}$ is the wealth score for individual i in survey year t. $Dist_{od}$ is the log distance between the ethnic homeland and the destination, $Migrant_i$ is a dummy for migrant status, and $CoethShare_{od}$ reflects the fraction of individuals in d that are from ethnic homeland o. I include fixed effects for country-year υ_{st}, and γ_o to isolate variation within an ethnic group. The controls for the individual Z_i include age and schooling, while X_{od} includes the level of o-d distance and coethnic share, and W_d includes destination log population. Wealth is a measure of either durable assets or housing quality. Subsequently, I compare outcomes of migrants to natives within a destination by replacing the ethnicity fixed effect with a destination fixed effect. Finally, I isolate the sample to only migrants, and compare the outcomes of migrants within a destination from different homelands.

Table 1 shows the results for different variations of equation 6 with different fixed effects. While the average consumption benefit to migrants is weakly positive in terms of assets, columns 1-3 suggest that migrants that travel further have higher returns. This is consistent with a model of

Table 1: Observational Returns by Migrant Status and Distance

	(1)	(2)	(3)	(4)	(5)	(6)
	Within Ethnicity		**Within Destination**		**Migrants Only**	
	Durables	House Quality	Durables	House Quality	Durables	House Quality
Migrant==1	0.152	-0.121	0.059	0.125		
	[0.079]*	[0.083]	[0.080]	[0.082]		
Migrant*Population	-0.050	-0.032	-0.015	0.000		
	[0.008]***	[0.008]***	[0.008]*	[0.008]		
Migrant*Distance	0.064	0.098	0.024	-0.009		
	[0.011]***	[0.012]***	[0.012]**	[0.013]		
Migrant*CoethnicShare	-0.251	-0.208	-0.221	-0.270		
	[0.033]***	[0.035]***	[0.038]***	[0.042]***		
ln(O-D Distance km)	0.024	-0.026	0.030	0.027	0.064	0.028
	[0.008]***	[0.008]***	[0.010]***	[0.011]**	[0.008]***	[0.009]***
ln(Population)	0.227	0.203				
	[0.008]***	[0.008]***				
Age	0.007	0.005	0.006	0.004	0.006	0.004
	[0.000]***	[0.000]***	[0.000]***	[0.000]***	[0.000]***	[0.000]***
School Years	0.101	0.075	0.081	0.058	0.082	0.057
	[0.001]***	[0.001]***	[0.001]***	[0.001]***	[0.001]***	[0.001]***
Coethnic Share	-0.163	-0.115	-0.018	0.041	-0.263	-0.254
	[0.029]***	[0.031]***	[0.033]	[0.035]	[0.022]***	[0.026]***
Mean Dep. Var	-0.017	-0.017	-0.017	-0.017	0.021	0.007
Observations	359,411	395,543	359,401	395,529	183,244	199,415
Destination FE	N	N	Y	Y	Y	Y
Ethnicity FE	Y	Y	N	N	N	N
Migrant Only	N	N	N	N	Y	Y

Notes: This table estimates observational returns to migrations from different model specifications. The outcome "Durables" is the first principal component of several binary variables for different assets, including electricity, phone, car, fridge, television and bicycle. The outcome "House Quality" is the first principal component of reported wall, roof and floor material of the house. Origins are ethnic homelands, and destinations are administrative units at the second level. The first two columns include ethnic group fixed effects to isolate variation within ethnic group across destination and migrant status. Columns 3-6 include destination fixed effects to isolate variation within destination across migrant status. All regressions include country-year fixed effects. Standard errors are clustered at the DHS sampling cluster level. * $p<0.01$, ** $p<0.05$, *** $p<0.01$.

migration with heterogeneous costs in which the migrants that choose to pay high migration costs are wealthier ex-ante, or have a higher productivity draw for a particular location (Lagakos et al., 2020). Coethnic share is consistently negative across specifications, suggesting that migrants gain better returns in more diverse destinations. This is consistent with the findings of Wang (2024), which finds that migrants that move to more culturally distant destinations have higher returns in an Indonesian sample. As large productive cities are more diverse, this finding is consistent with the idea that migrants move to productive, cosmopolitan destinations. I also find that returns to migration are decreasing in destination population size, conditional on travel distance and coethnic share. Henderson et al. (2021) find similar negative effects of population scale on household incomes in their cross-sectional African data, consistent with a story about the negative effects of sprawl and slum formation on migrant returns. The modest to negative average effects of migrant status on wealth offers a preview of the paper's baseline results, where I find no per-capita growth premium from increased migrant labor size.

As an additional check, I also run a standard gravity model relating the share of migrants from origin o in d using a Poisson pseudo-likelihood (PPML) model. Table B2 shows the relation of the share of migrants arriving in d to their travel distance and coethnic share. Consistent with past work on migration, I find that migration between origin-destination pairs in my panel is decreasing in distance and coethnic share, as individuals sort towards areas that are close by and majority coethnic. This finding is consistent with standard models of migration with linear migration costs in distance.

3 Empirical Strategy

The goal of the paper is to estimate the effects of changes in migration levels and composition on destination outcomes. Both the size of the migrant labor force and the diversity of migrants are likely endogenously determined by a destination's productivity. I use a shift-share instrumental variable strategy to overcome this identification challenge. The IV strategy adds a nonlinear component to a standard migration shift-share to simultaneously instrument for migrant size and composition. In the main analysis, the *shifts* are birthplace level outmigration rates, while the *shares* are origin-specific settlement shares at destination. Changes in migrant size and diversity are predicted as linear and nonlinear combinations of these shifts and shares. In a secondary analysis, I then isolate exogenous variation in the outmigration shifts using birthplace-level push shocks that move migrants in or out of their birthplace over time. Third I use destination-specific pull characteristics to isolate exogenous variation in settlement shares.

The strategy starts from an equation of interest that relates a destination's productivity to both migrant labor size and diversity. For a given destination d in year t, the equation of interest is:

$$y_{d,t} = \beta_1 l_{d,t} + \beta_2 div_{d,t} + \upsilon_t + \gamma_d + \epsilon_{d,t} \tag{7}$$

where $y_{d,t}$ is a proxy for log wages, or another city-level outcome. The logged migrant labor force size is l_{dt}, and the diversity of the migrant labor force is represented by div_{dt}. The diversity of the labor force is a measure of the distribution of groups within the labor force. I use a Herfindahl measure of the concentration of groups. For exposition, I specifically use the negative of a Herfindahl index as the measure of "migrant diversity", which is the inverse of the concentration of birthplaces. For workers coming from a set of birthplaces $o \in O$ and arriving in destination d, I calculate div_{dt} as:

$$-div_{dt} = \sum_o \pi_{odt}^2 \tag{8}$$

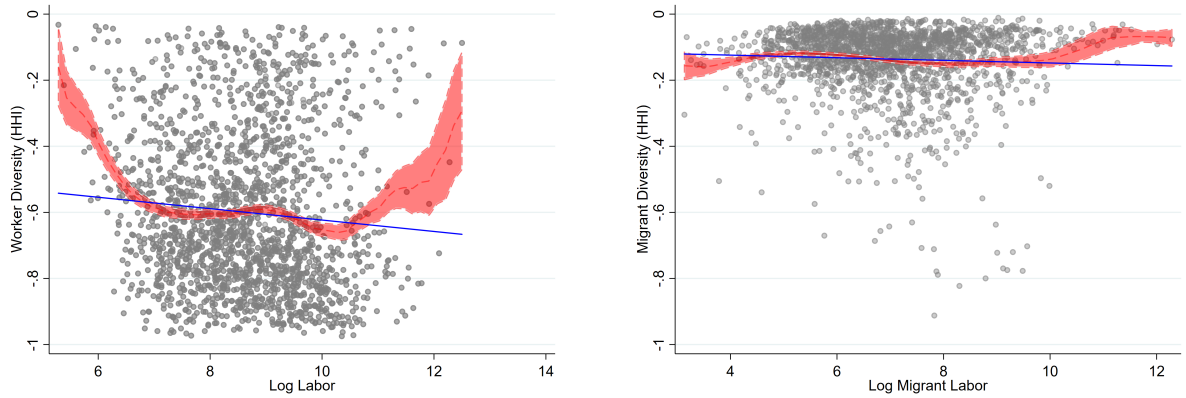
where π_{odt} is share of people in d at t from o $\frac{count_{od}}{count_d}$. As this number moves from 0 to 1, the composition of migrants moves from totally homogenous to perfectly heterogenous in terms of birthplaces. Scholars have used a variety of other indices for diversity, including fractionalization, which is the inverse of a Herfindahl Index, and polarization, which is more sensitive to large, equally sized groups. I choose to use the HHI because it is the simplest functional form of diversity to integrate into a shift-share design, and is highly correlated with other popular measures.

Theoretically we expect workers to sort into cities that offer higher wages (Combes et al., 2010). Therefore productive cities should attract a higher number of workers from a greater set of origins, creating a correlation between higher y_{dt}, higher l_{dt} and higher div_{dt}. I address the endogeneity of migrant flows and destination outcomes in the sections that follow. Section 3.1 describes a shift-share instrument that can incorporate nonlinear functions of migration like the diversity HHI. Then, I describe the exclusion restriction of the instrument in section 3.2. Sections 3.3 and 3.4 discuss extensions using pull and push shocks to predict outmigration shifts and pre-period settlement shares.

3.1 A Shift-Share Instrument for Migrant Birthplace Diversity

The goal is to develop a shift-share instrument for both the level of migration and its diversity, as measured by a birthplace HHI. I take inspiration from Schubert et al. (2024), which studies the effect of firm concentration on wages using a firm-level Bartik shock as an instrument for the change in employer labor market concentration. Similar to my setting, Schubert and co-authors instrument for an HHI variable as a function of Bartik style shifts and shares. I adapt this approach to the construction of two simultaneous instruments in a migration setting. In the case of migrant labor, the instrument is the aggregation of birthplace outmigration shifts and destination settlement shares. For example a given destination like Kigali Rwanda has a certain fraction of all migrants

Figure 3: Relationship between Diversity (HHI) and Labor Variables



(a) Worker Diversity vs. Log Total Labor

(b) Migrant Diversity vs. Log Migrant Labor

Notes: This figure shows scatterplots relating diversity and labor at the destination-year level. The left panel plots total worker diversity against total log labor size, while the right panel looks at migrant diversity against migrant log labor size. The blue line is a line of best fit across all points, while the red line is a local polynomial fit with confidence bands.

from birthplace Kirehe, in the East. The change in migrant labor from Kirehe is a combination of pre-period settlement shares of Kirehe migrants in Kigali, combined with shifts in the aggregate migration rate from Kirehe to all destinations. The change in the total number of people from Kirehe living outside Kirehe across census years constitutes a shift. To instrument for diversity, the destination settlement shares are weighted by the squared relative growth rates of each birthplace's migrant flow. The key is that this aggregation is a nonlinear combination of the individual shift-shares. The nonlinearity of the function mapping the shift-shares to aggregate diversity allows for the same set of shift-shares to predict both endogenous variables.

I begin by differencing equation 7 between censuses, which are typically 10 years apart in our data:

$$\Delta y_{d,t-(t-10)} = \beta_1 \Delta \ell_{d,t-(t-10)} + \beta_2 \Delta div_{d,t-(t-10)} + \epsilon_{d,t} \tag{9}$$

I instrument for $\Delta \ell_{dt}$ and Δdiv_{dt} using decade level shifts from different ethnic homelands and baseline ethnic shares.

Each origin o has a number of out-of-homeland migrants $count_{o,t}$, which I define as the aggregate number of people from o observed outside of that homeland in a given census year. Our shifts are the growth rates in the number of people from o observed outside of their origin, which we define as $g_{o,t} = \frac{count_{o,t} - count_{o,t-10}}{count_{o,t-10}}$.

Each destination d is exposed to these shifts weighted by their baseline settlement share of people from that origin in the previous census period. For a given year these settlement shares are the fraction of people from o in d relative to the full migrant population from o: $Shares_{odt} = \frac{count_{odt}}{\sum_d count_{odt}}$

We can predict the change in labor demand in the region d as the sum of the shifts and shares:

$$\Delta \widehat{L}_{dt} = \sum_o \frac{count_{od,t-10}}{\sum_d count_{od,t-10}} * g_{o,t} * count_{o,t-10} \tag{10}$$

In the analysis, we take a log of this predicted value as our measure $\Delta \widehat{\ell}$.

To predict a change in diversity, I start from a definition of the change in birthplace concentration. Following the HHI formulation above, change in the relative concentration of migrants from different o living in destination d is:

$$-\Delta div_{d,t} = \sum_o \pi^2_{odt} - \sum_o \pi^2_{od,t-10} \tag{11}$$

This equation can be written as a function of the initial concentration in the base period, and

the respective growth rates of migrants from o to d, and the total change in the migrant labor force at d. The equation becomes:

$$-\Delta div_{d,t} = \sum_{o} \pi^2_{od,t-10} \left(\frac{(1+g_{odt})^2}{(1+g_{dt})^2} - 1 \right) \tag{12}$$

Where mechanically the change is a function of the growth in the number of people o in d, g_{odt} relative to the total growth of population in d, called g_{dt}. The negative term at the front of the equation reverses the order, such that increasing levels represent lower Herfindahl concentration and higher birthplace diversity.

Both g_{odt} and g_{dt} are likely correlated with contemporary productivity in the destination. An instrument replaces the endogenous current period div_{dt} by substituting aggregate shifts for the two growth rates in equation 12. In particular, our instrument is:

$$-\Delta \widehat{div_{dt}} = \sum_{o} \pi^2_{od,t-10} * \left(\frac{(1+g_{ot})^2}{(1+\widetilde{g_{odt}})^2} - 1 \right) \tag{13}$$

Where g_{ot} is again the aggregate growth of migrants out of an origin and $\widetilde{g_{odt}}$ is a predicted growth of migrant labor in destination d defined as $\sum_o \pi_{od,t-10} * g_{ot}$.

3.1.1 Functional Relationship of Composition, Population and Productivity

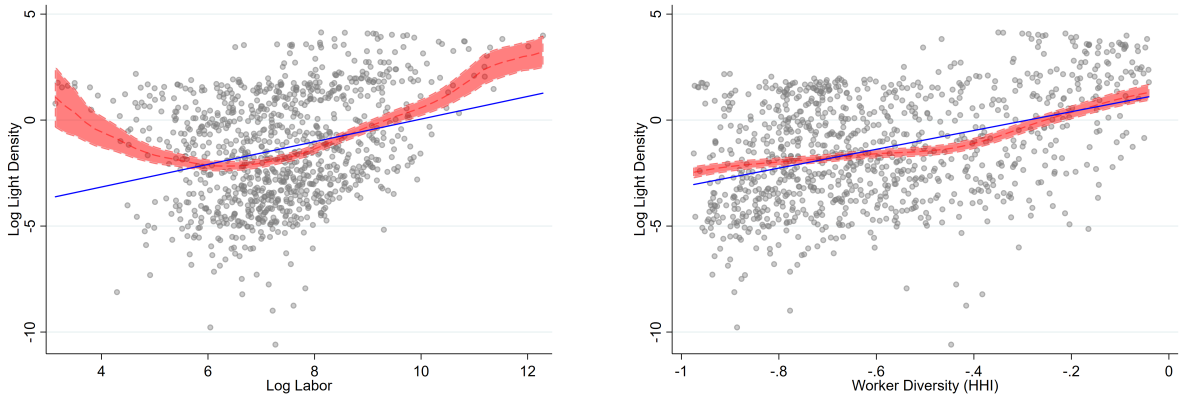
The estimation strategy leverages two functional forms for an instrument that aggregates shifts g_{ot} and shares π_{odt} across origins. This is possible because of the nonlinear relationship between the migration levels and their composition. Two potential issues emerge from this. First, the linear correlation between migrant levels and composition reduces the power of the estimates. Intuitively, the more related migrant levels and composition are, the less added variation is introduced by the second instrument. I show this in a simulation exercise in Appendix A, in which I calculate a distribution of beta coefficients as the linear correlation between migrant labor size and migrant diversity varies. In Figure 3 I plot the cross-sectional relationship between migrant diversity and migrant labor size. In general while the variables are related, the relationship is not strongly linear. This fact allows us to be more confident that migrant level changes do not linearly predict composition changes, and that both instruments provide identifying variation to the empirical estimates.

Second, we don't know ex-ante the functional relationship between city diversity and productivity. Panel B of Figure 5 presents the cross-sectional relationship between city diversity and light density in destinations. In the cross-section, I find a positive and linear relationship between worker diversity and light density. Cities that are bigger and less homogenous are higher income. This cross-sectional result is consistent with the subnational associations studied by Montalvo and Reynal-Querol (2021). In their paper, the authors argue that, while country-level regressions show a negative relationship between higher diversity and light density, at the subnational level this relationship is reversed. Their work leverages historical diversity measures from anthropological maps, rather than census-reported ethnic mix. However, the mechanism that drives their correlation is presumably the same – certain places that are high productivity attract more people, from more diverse sources. The goal of this paper is to move beyond these cross-sectional associations, and study a causally motivated estimate in first differences.

3.2 Exclusion Restriction

The exclusion restriction requires that the drivers of aggregate shifts out of a birthplace are not correlated with productivity shifts at particular destinations. For example, this would be violated if productivity gains in certain major cities like Kigali drive aggregate migration trends out of certain birthplaces. Borusyak et al. (2022) describes the exclusion restriction assumption in shift-share designs in which exogeneity comes from the individual shocks. In our setting the shocks are the estimated g_{ot} outmigration rates from origins o in period t. Borusyak and co-authors show that the

Figure 5: Migrant Size and Composition on Light Density



(a) Migrant Labor Size and Light Density

(b) Migrant Diversity and Light Density

Notes: This figure shows scatterplots relating migrant diversity, migrant labor size and light density at the destination-year level. The left panel plots migrant labor size against contemporary logged light density, while the right panel looks at migrant diversity against contemporaneous log light density. The blue line is a line of best fit across all points, while the red line is a local polynomial fit with confidence bands.

exclusion restriction in this case is equivalent to an orthogonality condition at the birthplace level. In particular, given regional exposure weights s_o and a distribution of unobservables ϕ_o:

$$\mathbb{E}[g_{ot}|\phi_o] = \mu \tag{14}$$

A second requirement is that the shocks are independent and exposure is dispersed across birthplaces. In particular:

$$\mathbb{E}[(g_{1t} - \mu)(g_{2t} - \mu)|\phi_1, \phi_2] = 0 \tag{15}$$

The dispersion condition can be defined as a Herfindahl index that goes to zero as the number of birthplaces increases: $\sum_o s_o^2 \to 0$.

While I can't directly test these conditions, I perform several tests at the birthplace level. First, I derive shift-share standard errors that address the possible covariance between g_o and ϕ_o following Borusyak et al. (2022). These estimates are presented in Figure B5. Second, I check pre-trends by running birthplace-level regressions of the exposure-weighted residuals of destination characteristics on our growth rate shocks. This analysis is produced in Figure B7. The destination characteristics I include are geographical characteristics such as disease ecology and soil suitability, as well as past-period light density. I show that my birthplace shocks are uncorrelated with pre-period destination characteristics. Another way to address endogeneity concerns is to directly leverage plausibly exogenous variation in shifts and shares. I describe two potential strategies in the following sections that utilize migration push shocks and pull characteristics.

3.3 Predicting Outmigration Shifts with Push Shocks

In the standard migration shift-share, the shift terms g_o are the real change in the number of people leaving origin o across years. The argument for exogeneity comes from an assumption that past shares are unrelated to current period changes in outcomes, except through the labor and composition channels. Since our specification is in differences, this is comparable to a parallel trends assumption in a difference-in-differences design, but with many individual treatment exposures to the aggreage shift treatment (Borusyak et al., 2025; Goldsmith-Pinkham et al., 2020). This may be violated in cases where changes in particularly large labor markets drive g_{ot} or when country-wide shocks affect both the total outmigration rate and particular destinations (Jaeger et al., 2018). Many migration paper have considered instrumenting for shifts in the outmigration rates from origins using push shocks (Mullins and Bharadwaj, 2021; Boustan et al., 2010; Bazzi et al., 2023; Kamuikeni and Naito,

2024). If the push shocks are plausibly uncorrelated with destination outcomes, then a zero-stage regression of outmigration on push shocks can isolate exogenous variation in migration shifts. I consider the total outmigration from a given origin o and year t as a function of a set of past shocks $shock_{o,t-y}$, where the shocks are measured at some lag y from the current period:

$$g_{ot} = \omega_1 log(Shock)_{o,t-y} + \upsilon_o + \delta_s t + \epsilon_{ot} \tag{16}$$

I include a linear time trend $\delta_s t$ and origin fixed effects υ_o[4]. I substitute these predicted outmigration rates $\widehat{g_{ot}}$ for the raw shift g_{ot} in equations 10 and 13.

Past work on the historic US has leveraged a variety of local economic conditions and weather variables (Boustan et al., 2010; Bazzi et al., 2023). In developing countries, estimates of migration elasticities to income or other shocks vary by context. For example, income shocks have been shown to move migration both positively and negatively depending on the country and time period (Marchiori et al., 2012; Bazzi, 2017; Shrestha, 2017). My strategy is to consider a set of plausible local shocks that may move outmigration. In a zero stage, I consider each shock and its lags separately, and then perform a joint estimation of the relevant shocks to predict the outmigration rate. I consider climate variables, conflict and international commodity price changes, weighted by local exposure.

3.4 Predicting Settlement Shares with Pull Characteristics

Even with a predicted flow estimated for g_{ot}, another concern may be that previous period settlement patterns are correlated with current period labor demand shocks in destinations. This may be the case if a given settlement of migrants was driven by a previous labor demand shock, and demand shocks are serially correlated in destination (Jaeger et al., 2018). Instead of using past settlement patterns as exposure shares, I consider replacing the shares with a gravity-model prediction of which destinations migrants are likely to choose. In particular I construct instruments for settlement shares of origins in destinations as a function of the o-d distance between regions and the interaction of distance with a set of "pull characteristics".

I estimate migrant shares of o in d as

$$\sigma_{odt} = \omega_1 Pull_d * log(Dist)_{od} + \omega_2 log(Dist)_{od} + \omega_3 log(Dist)_{od} * Z_d + \mu_d + \upsilon_o + \gamma_t + \epsilon_{odt} \tag{17}$$

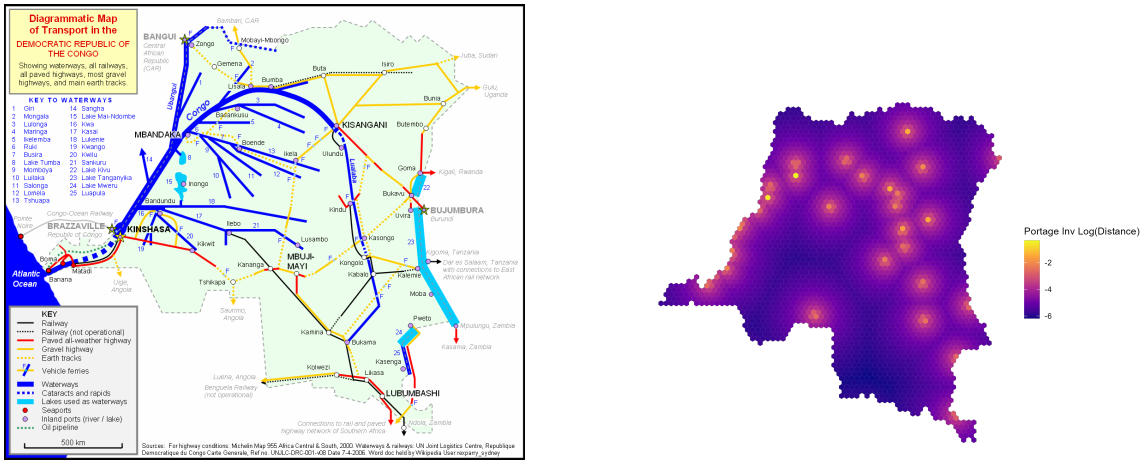
Where σ_{odt} represents the migrant shares $\sigma_{odt} = \frac{count_{odt}}{\sum_d count_{odt}}$. The vector Z_d represents a set of historical destination characteristics that predict the attractiveness of destinations. We can think of these as historical shocks that predict which locations are more predisposed to urban growth. The key is that these pull characteristics Z are not correlated with contemporary shocks to demand. In particular, we use distance to colonial rail lines, mineral deposits and portage sites as predictors of destination attractiveness. We run the same regression to separately estimate π_{odt} for our prediction of diversity change.

3.4.1 Historical Pull Characteristics as Predictors of Agglomeration

My pull characteristics draw on historical sources of agglomeration unrelated to contemporary productivity. 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. 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). In Jedwab and Moradi (2016) the authors show

[4] Another option would be to use year fixed effects in this regression. As noted by Mullins and Bharadwaj (2021), this isolates variation in the shock between counties in the same year. But this specification considers only short term adjustments to weather. Because my migration adjustments are across decades, I substitute the year fixed effects for state specific linear time trends. This allows for detrended variation within origins over time to predict the migration rate.

Figure 7: Portage Sites, DRC Example



(a) River Network Map (b) Portage Propensity

Notes: This figure shows an example of the portage score estimation for the portage pull characteristic. The left panel shows a map of the Congo river network, with locations of cities and rapids. On the right I plot my estimated portage points, which are located in bright yellow. All colors reflect nearest distance to these estimated portage sites, which is our portage score.

that colonial rails continue to predict urban agglomeration, long after the rail lines fell into disuse. I calculate a distance to the nearest colonial rail line as a predictor of a destination's attractiveness, or "agglomeration potential".

I next consider a destination's propensity to be a portage site (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 create a prediction of rapids and waterfall locations along African rivers. Non-navigable river segments are predicted using HydroSHEDS data on river discharge. In particular I measure the change in river discharge (cubic meters per second) along each river network. I define possible portage sites as places where the size class of a river, which is a logarithmic function of discharge, changes. Each potential portage site is then saved as a point in space, and joined to the administrative census regions. Figure 7 shows an example of predicted portage sites along the Congo River, along with a map of the river network and associated cities. We see for example that the mouth of the river in the West shows a high portage propensity near where the Congo River has large rapids that precede the city of Kinshasa.

Last I consider historical mineral deposits as a predictor of past urban growth. This intuition follows Combes et al. (2010) in using geological factors as an instrument for contemporary agglomeration. While deposits should be related to historical roads and infrastructure, they should not be related to contemporary demand shocks. Using interactions of the pull characteristics and origin-destination distance, I instrument for settlement shares for each origin-birthplace pair in a baseline period. I then merge these predicted shares with the outmigration shifts from the baseline analysis.

In an additional section, I explore a second use for these historical pull characteristics in estimating long-run dynamics. The main results of the paper consider contemporary changes in diversity and migrant labor across 10-year census periods. To understand the longer term relationship between migrant diversity and urban growth, I need a strategy that can be estimated in a cross-section, and doesn't rely on data that captures differences over census years. The ideal experiment would place urban centers at random locations across Africa, and expose them to different levels of migrant worker diversity. Then I would compare the growth prospects of cities located in more or less dense

areas, with more or less worker ethnic diversity. To proxy for this idealized experiment, I run a cross-sectional regression that interacts my historical instruments for agglomeration with historical measures of regional diversity. I use my agglomeration insturments as predictors of city location, and evaluate outcomes in cells that were historically likely to agglomerate. The interaction between the agglomeration instrument and regional diversity measures allowes me to compare two cells that were both predicted to agglomerate, but were exposed to different ex-ante levels of worker diversity. The empirical strategy is discussed in greater detail below.

4 Results

In this section I will first discuss the results of the nonlinear migration shift-share instrument using standard outmigration shifts and pre-period settlement shares. I will then discuss results from different predictions for shifts and shares using pull and push shocks. I show zero-stage regressions of push shocks predicting outmigration rates at the birthplace level, and then show the results for the first and second-stage SSIV.

In Section 4.1, Table 2 and Table 3 show the first and second stage results of a regression of outcome in destination d and time t, instrumenting for $\Delta\ell$ and Δdiv:

$$\Delta y_{d,t} = \alpha\Delta\widehat{l_{d,t}} + \gamma\Delta\widehat{div_{d,t}} + \epsilon_{d,t} \tag{18}$$

I find that level increases in migrant labor size predict increases in city size, as measured by average light density change, but do not correpsond to changes in per-capita light density growth. Increased migrant labor causes cities to grow, but does not map to clear effects on either productivity growth or urbanization. As the migrant labor pool becomes more diverse, cities grow slower and have lower productivity in terms of light density changes. However, increasing diversity of migrants predicts higher rates of urbanization, as captured by an increasing non-agricultural labor share. This suggests that migrant composition may have a positive benefit on industry mix, and yield long-term gains even if there are short-run costs. This argument is substantiated in the long-run estimates of diversity and city growth produced in Table 12. In the long-run, I find evidence of positive interactions between diverse migrant labor and historical productivity shocks, suggesting that benefits of migrant composition manifest over longer periods.

In section 4.2 I first consider the effect of push shocks as an instrument for the measured birthplace-level shifts g_{ot}. I find that drought predicts lower outmigration, while conflict predicts higher outmigration from birthplace. I find heterogeneous effects of prices across commodity type and time horizon. Leveraging a combination of push shocks to predict g_{ot}, I re-run our 2SLS regression and find that the first stage succeeds in predicting changes in labor size and composition. I find that the effects of migrant labor size are consistent, and that migrant diversity continues to predict higher urbanization rates.

In section 4.3 I instrument for settlement shares using the interaction of distance with the measured destination-level pull characteristics. I show consistent results of migrant labor size causing lower productivity growth in terms of light density per capita. However, the joint F-statistics in this final exercise are too weak to draw strong conclusions.

In section 4.4 I consider heterogeneity in the returns of migrant diversity by evaluating a natural experiment. I leverage the fact that South Africa's apartheid regime generated high migration barriers, which were suddenly dropped in the 1990s. The end of migration restrictions for black workers provides a plausibly exogenous shift in the outmigration rates (g_{ot}) from South Africa's native homelands. Following the baseline shift-share model, I instrument for changes in migrant labor size and diversity at destination. In this exercise I isolate variation in the composition of the black migrant composition, rather than the relationship between the black and white population. While the estimates are weaker due to low sample sizes, I find that the impact of migrant diversity in this context is inverted. Black migrant diversity in South Africa contributes to higher city size

Table 2: Census Birthplace Shift-Share First Stage

	(1)	(2)
	Migrant Flows	
	$\Delta\ \ell$	Δ div
Predicted $\Delta\ \ell$	0.452	-0.004
	[0.027]***	[0.001]***
Predicted Δ div	1.847	0.558
	[0.447]***	[0.110]***
Mean Dep.	7.58	-0.01
Observations	829	829

Note: This table estimates the first-stage regression of predicted on real changes in migrant labor and composition $real_{d,t} = \alpha\widehat{\Delta l_{d,t}} + \gamma\widehat{\Delta div_{d,t}} + \upsilon_t + \mu_c$. Δl is the logged difference in total migrant labor in destination d between census years. The predicted value is the shift-share instrumented change in migrant labor. Δdiv is the difference in the negative HHI between census years for all non-native residents in location d. The predicted value is the estimated change in HHI based on the relative aggregate shifts and shares across origins. Regressions include fixed effects for country and year. Data for this table comes from an origin-destination panel of workers in African IPUMS Census samples. * p<0.01, ** p<0.05, *** p<0.01.

growth and city productivity. I interpret this as evidence that some African states are able to overcome linguistic and ethnic differences and benefit from diverse birthplace composition.

4.1 Baseline Results: Standard Shift-Share with Nonlinearity

Table 2 presents the results for the first stage of the aggregate shift-share, as described in equation 9. The predicted variables are the aggregations of baseline shares in the previous census years, and aggregate shifts across census years. The linear and nonlinear shift-share instruments are strong predictors of the real differences in migrant labor and migrant diversity. Surprisingly, predicted l also predicts lower diversity (higher HHI), which means larger total migration shifts are typically also more homogeneous. We'd normally expect larger shifts in migration to require a broader composition, as migrants come from further and further locations to meet a destination's labor demand. The inverted correlations between diversity and migrant labor provide further evidence that the relationship between the two instruments is nonlinear. The jointly calculated F-statistics and Sanderson-Windmeijer values are presented in Table 3.

Panel A in Table 3 gives the OLS estimates of a regression of migrant labor and diversity on light outcomes. Column 1 measures the standardized change in average light density, while column 2 measured the change in the log of per-capita light density. These measures capture different aspects of light density change at destination. The first is related to city size and extent, as expanding cities increase average light density. The second has historically been compared to GDP/capita growth. As other measures of urban productivity, I include changes in the non-agricultural labor share and changes to the housing quality index, both measured from census data. I find that a 1% increase in migrant labor size corresponds to a 0.19 SD increase in average light density, and a .76% decrease in light density per capita growth. The diversit term is measured as changes in a 0 to 1 measure of diversity. Moving from a completely homogeneous to a perfectly heterogenous migrant labor pool is associated with a 1.3 standard deviation drop in average light density and 3.4% reduction in light density per capita growth. I find that an increase in migrant diversity is also related to a substantial increase in non-agricultural labor share. Going from 0 to 1 on a diversity measure is associated with 0.86 percent point increase in the non-agricultural labor share.

Table 3: Census Birthplace Shift-Share Second Stage, Migrant Flows

Panel A: OLS Results

	(1) Δ Lights	(2) Δ Log(Lights/Capita)	(3) Δ Non-Agriculture Share	(4) Δ Housing Quality
Δ ℓ	0.189	-0.763	-0.030	0.039
	[0.043]***	[0.079]***	[0.016]*	[0.041]
Δ div	-1.375	-3.397	0.860	-0.453
	[0.360]***	[1.080]***	[0.278]***	[0.512]
Mean Dep.	-0.00	3.31	0.09	0.01
Observations	829	829	634	671

Panel B: Shift-Share IV

	Δ Lights	Δ Log(Lights/Capita)	Δ Non-Agriculture Share	Δ Housing Quality
Δ ℓ	0.296	-2.081	-0.060	0.126
	[0.100]***	[0.173]***	[0.036]	[0.092]
Δ div	-2.091	-10.304	1.310	-0.135
	[0.703]***	[2.757]***	[0.586]**	[0.695]
Mean Dep.	-0.00	3.31	0.09	0.01
Observations	829	829	634	671
Kleibergen-Paap Fstat	16.462	16.462	13.716	26.970
Sanderson-Windmeijer ℓ	245.593	245.593	141.084	106.168
Sanderson-Windmeijer div	27.686	27.686	23.386	19.570

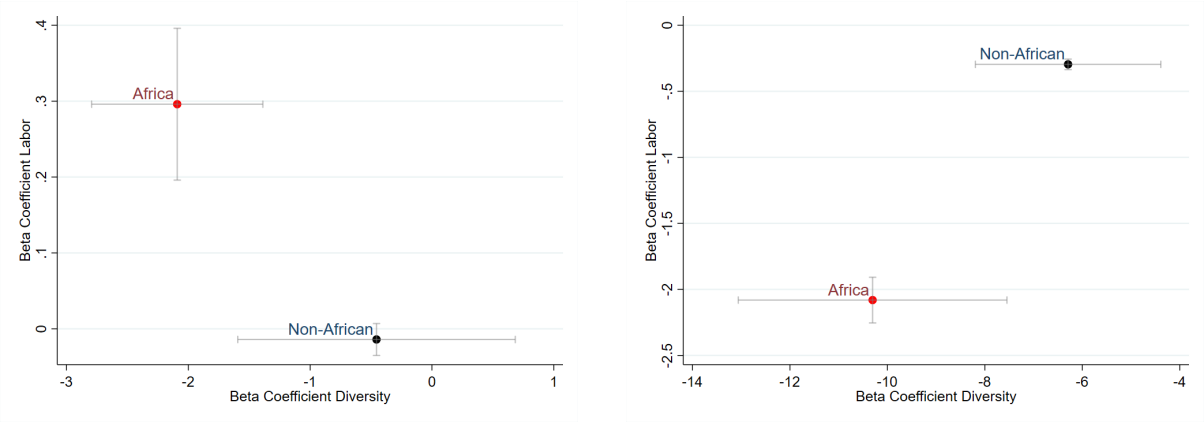
Note: This table presents the results of a second-stage regressions of instrumented changes in migrant labor size and composition on destination productivity outcomes. Panel A presents the OLS estimation of changes in migrant labor and diversity on light density outcomes. $y_{d,t} = \alpha\Delta l_{d,t} + \gamma\Delta div_{d,t} + v_{tc}$. Panel B presents the IV 2SLS estimation of changes in migrant labor and diversity on light density outcomes. The original harmonized range of light density is from 0 to 63. The first column outcome is the standardized change in average light density, while the second column is the log change in light density per capita. The third column outcome is the change in labor share in non-agricultural industries, and the fourth column measures changes in housing quality, measured by a principle component of housing characteristics. All regressions include country-year fixed effects. Data for this table comes from an origin-destination panel of workers in African IPUMS Census samples. * p<0.01, ** p<0.05, *** p<0.01.

Panel B in Table 3 shows the second stage regression using the aggregate shift-share instruments for migrant labor and diversity. The OLS and IV results are broadly consistent. Total migrant labor size is positively related to city size growth, but negatively related to per-capita productivity growth. The size of the coefficients are larger. A 1% increase in migrant labor size yields a .3 SD increase in average light density, and a -2% drop in per-capita light growth. A 0 to 1 increase in migrant diversity reduces average light density by -2 SD, and per-capita growth by -10%. The change in the non-agricultural share is more than a 1 for 1 increase.

To better interpret these measures, I take a set of censuses from non-African developing countries and replicate the shift-share exercise using available birthplace data from these countries. I then measure light density changes across census years for these countries. The strength of the linear and nonlinear shift-share instrument varies across these countries. I take the estimated coefficient on changes in migrant labor and migrant diversity, and plot them in two dimensions to compare with the African sample. I find that my esitmated impact of migrant labor size corresponds to higher city size growth and lower productivity growth relative to other developing countries. This suggest that level changes in migrant labor increase African city sprawl and extent without translating to faster per-capita growth. Across both light density measures, the African estimated impacts of migrant diversity are much lower than the non-African sample. The impact of migrant diversity on lower city night light density and per-capita growth is nearly double the estimates of non-African developing countries.

Table B3 shows the same regressions but for alternative destination outcomes. I include the levels of log light density, the change in services labor share, and the levels of housing wealth. share of the native workforce dedicated to services. I find that higher migrant diversity predicts higher services share, lower average light density in levels, and higher average levels of housing wealth. These level relationships between migrant diversity and light and housing levels suggest that migrant diversity is associated with destinations that may be smaller but wealthier on average.

Figure 9: Comparing Migration Estimates to Other Samples



(a) Δ Lights (b) Δ Log(Lights/capita)

Notes: This figure compares the shift-share migration estimates for $\Delta\ell$ and Δdiv to a sample of other countries available from IPUMS. The coefficients of the SSIV are plotted in 2-d space. The same shift-share estimate represented in Table 3 is estimated at the second administrative level for Cambodia, Chile, Indonesia, Mexico, Peru and Thailand. For each country, the available birthplace data is used to calculate migrant labor size and diversity. Birthplace is at the state level for Mexico, province for Peru, Thailand and Indonesia, commune for Chile, municipality for Colombia, and district for Cambodia. IV regressions include country-year fixed effects. Gray bars plot the standard errors from the SSIV. Horizontal bars represent the SEs on the diversity coefficient, while the vertical bars represent the SEs on the labor coefficient. Coefficients for labor and diversity in Table 3 are presented in red. The original harmonized range of light density is from 0 to 63. The first panel outcome is the standardized change in average light density, while the second panel is the log change in light density per capita. Data for this table comes from an origin-destination panel of workers in African IPUMS Census samples.

4.1.1 Robustness Checks of Shift-Share

I perform a variety of robustness checks for our results. To address concerns about the correlation of outcomes like light density across administrative regions, Figure B9 and Figure B10 show robustness to Conley spatial standard errors across a variety of distance bandwidths. Next I test the sensitivity of our results to changes in our sample. Figure B11 and B12 show how the estimates vary when dropping individual countries from the sample. Last I check the results when replacing the real estimated birthplace shocks with randomly distributed placebo shocks. Figures B13 and B14 show how the IV results appear in response to placebo shocks.

In these robustness results, I find some sensitivity to country drops, due to the short-panel structure of our data. Results are robust to spatially correlated errors across bandwidths. The placebo shocks help us understand how our shift-share results leverage variation from the shifts and shares portions of our instrument construction. While placebo shocks produce zero-results for our diversity estimator, the logged light density per capita outcome in the labor shift-share estimate is biased upward under placebo shocks. This suggests that some identifying variation is coming from the settlement shares. The prediction of settlement shares leveraging distance and pull characteristics will allow us to isolate exogenous variation in settlement shares.

4.2 Instrumenting Outmigration Shifts with Push-Shocks

Next I look at the estimation results when instrumenting for aggregate shifts using push-shocks. As described in the empirical strategy, I predict the aggregate shifts in migrant labor using shocks, or the changes between origin-destination pairs. We present the results of this analysis in three parts. First, we estimate zero-stage relationships between the raw number of people leaving birthplaces and different potential migration shocks. We find that climate and price variables do predict population movements at different lags from the contemporary period. Leveraging this result, we then estimate a first-stage regression that leverages predicted g_{ot} variables in our shift-share design, producing push-shock motivated instruments for migrant labor size and migrant composition. We find that the push shock version of our first stage is still able to predict changes in both migrant labor and migrant composition. Next, we leverage this first stage to look at our destination productivity outcomes.

While the results for migrant labor effects remain consistent, our second stage for diversity is too weak to consistently estimate an effect. We conclude that the benefits of migrant labor density outweigh potential costs of migrant diversity, at least among migrants motivated by birthplace shocks.

4.2.1 Zero-Stage Regressions of Migrant Flows on Shocks

Table 4 shows the results of regressions of the log population from o in destination d on shocks at different lags from the current period. The first column includes country fixed effects, the second column includes birthplace fixed effects, and the third column includes 10-year lags of the given variables. The outcome in columsn 1-3 is the number of people living outside of the given birthplace in a given year. I find that conflict consistently predicts greater outmigration. Contemporary drought conditions prevent migration, which is consistent with a model where negative wealth shocks prevent households from migrating (Bazzi, 2017). High agricultural prices in the contemporary period also encourage migration, which is also consistent with a model where positive wealth shocks increase migration. Different shocks to birthplaces may not only drive different levels of outmigration, but also may change the skill-level composition of the marginal migrant. If the type of workers who leave in response to high commodity prices are different from workers that leave in response to conflict, then changes in migrant outflows also reflect average skill level changes. Columns 4-6 show the results when the population counts are weighted by human capital, defined as schooling years for workers 18 and above. These labor efficiency units for a given origin-destination pair are an aggregation of these weighted population counts, defined as $le_{od} = \sum_i^N h_i$, where h_i are schooling years for an individual i from o living in d. The results are fairly consistent when using labor efficiency unites rather than log counts, although the standard errors are larger as the weighting introduces more dramatic variation. In column 6 there is evidence that lagged conflict has a negative impact on labor efficiency units, which is a flip in the sign relative to column 3. While past period conflict events move more individuals out of origins, the marginal migrant is has lower schooling years. As expected, different types of shocks have different implications for the human capital of the marginal migrant. In the section on mechanisms, I directly consider the impact of outmigration shifts on changes to the average skill level at destination.

To better understand the dynamic impact of different shocks on outmigration, Figure 11 plots the beta coefficients of individual regressions of migration counts against particular shocks at different lagged periods. Each dot represents a particular shock variable at a particular time horizon relative to the migration count year. We find that, when estimated separately, the effects of drought and conflict are consistent across time horizons, but that commodity prices show heterogeneity. Recent price spikes increase migration, while historical price spikes decrease migration. This suggests that long-run trends in hig prices for key commodities convince more people to stay in their birthplace.

4.2.2 First and Second Stage Estimates using Predicted Shifts

Panel A in Table 5 shows a first stage regression of equation 9 where the g_{ot} growth rates are instrumented using the predicted outmigration shifts. The prediction model used leverages contemporary, 1 and 10-year lags of the selected push shock variables drought, conflict and commodity prices. I also include a linear time trend in our prediction model. I find that the shift-share with predicted shifts successfully predicts changes in migrant labor size and composition. This result suggests that push-shocks have an effect on both migrant flows and the mix of migrants that arrive. Panel A of Table 6 gives the second stage regression of equation 9 using the predicted migration shifts. I find that migrant labor size is positively related to growth in city size, but negatively related to per-capita growth in light density. I also find that increases to migrant labor size predict decreasing levels of housing quality at destination. This is consistent with a story of increasing migration driving higher housing prices and slum formation at destination. The estimates for diversity impacts on light density are too noisy to make an inference. However, I find consistent evidence that incrasing migrant diversity predicts higher non-agricultural labor share. This lends further evidence that migrant diversity has

Table 4: Predicting Outmigration with Contemporary and Lagged Shocks

	(1)	(2)	(3)	(4)	(5)	(6)
	Migrant Flow			*Labor Efficiency Units*		
	log(N)	log(N)	log(N)	log(le)	log(le)	log(le)
Drought	-1.308	-0.288	0.002	-14.324	-23.666	-23.926
	[0.393]***	[0.263]	[0.259]	[0.984]***	[1.393]***	[1.423]***
Agricultural Price	0.168	0.287	0.676	-0.010	-0.043	1.002
	[0.067]**	[0.064]***	[0.077]***	[0.186]	[0.288]	[0.398]**
Mineral Price	-0.010	-0.048	-0.131	0.012	0.040	-0.229
	[0.062]	[0.068]	[0.067]*	[0.144]	[0.283]	[0.282]
Conflict	0.018	0.016	0.002	0.000	-0.016	0.019
	[0.004]***	[0.004]***	[0.004]	[0.010]	[0.015]	[0.018]
Year	7.068	6.815	5.503	4.483	7.173	2.533
	[0.944]***	[0.578]***	[0.604]***	[2.458]*	[2.916]**	[2.958]
$Year^2$	-0.002	-0.002	-0.001	-0.001	-0.002	-0.001
	[0.000]***	[0.000]***	[0.000]***	[0.001]*	[0.001]**	[0.001]
Lagged Drought			0.472			13.785
			[0.694]			[3.625]***
Lagged Agricultural Price			-0.652			-1.840
			[0.087]***			[0.438]***
Lagged Conflict			0.029			-0.070
			[0.006]***			[0.024]***
Mean Dep. Var	7.264	7.447	7.447	8.534	8.492	8.492
Observations	2,252	1,944	1,944	1,473	1,423	1,423
Country FE	Y	Y	Y	Y	Y	Y
Origin FE	N	Y	Y	N	Y	Y

Notes: Each column is a joint regression of migrant flows in terms of population or labor efficiency units on a set of contemporary and lagged shocks. The drought index is the SPEI indicator for drought intensity at an annual level. The agricultural prices are a weighted regional price exposure aggregated from individual crop shares and international prices normalized to 2000. Mineral prices are weighted by mineral and mine capacity in the origin location, and conflict represents the number of ACLED events. Lagged shocks represent average changes in the variable in the last 10 years. All regressions include country fixed effects. Year is included as a linear and squared value.

consistent impacts on urbanization and structural transformation in destinations.

4.3 Instrumenting Settlement Shares with Pull Characteristics

The placebo shocks exercise from the headline shift-share estimation showed that some variation in the instrument for labor size changes may be coming from endogenous shares. In this section I leverage pull characteristics interacted with origin-destination distances to isolate an exogenous component of settlement shares. The implicit assumption is that the historical pull characteristics, including distance to colonial rail, mineral deposits and portage sites are unrelated to contemporary labor demand shocks. In a zero stage regression, I predict the following settlement share variables:

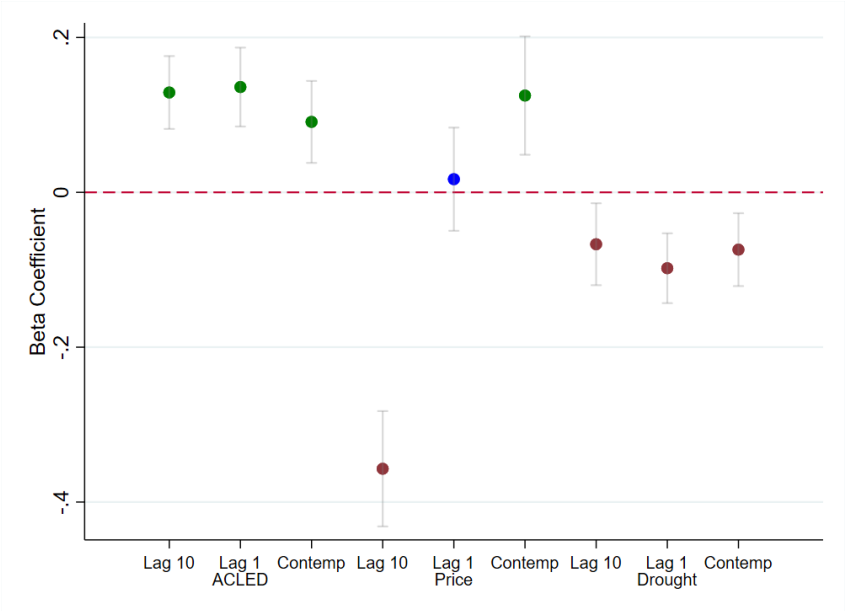
$$\sigma_{odt} = \omega_1 Pull_d * log(Dist)_{od} + \omega_2 log(Dist)_{od} + \omega_3 log(Dist)_{od} * Z_d + \mu_d + \upsilon_o + \gamma_t + \epsilon_{odt} \quad (19)$$

$$\pi_{odt} = \omega_1 Pull_d * log(Dist)_{od} + \omega_2 log(Dist)_{od} + \omega_3 log(Dist)_{od} * Z_d + \mu_d + \upsilon_o + \gamma_t + \epsilon_{odt} \quad (20)$$

Where σ_{odt} represents the migrant shares $\sigma_{odt} = \frac{count_{odt}}{\sum_d count_{odt}}$. π_{odt} is the baseline Herfindahl estimate, which behaves as the settlement share for the diversity instrument. I use the same growth rates g_{ot} as in the baseline shift-share strategy above. Panel B of Table 5 shows the results of the first-stage, where the predicted shifts are used to produce the instrument aggregates. I find that the diversity predictor is weaker than in the main shift-share, but still statistically significant at the 5% level.

Panel B of Table 6 gives the second stage regression of equation 9 using the predicted shifts. I find that the labor size estimate remains remains negative and significant for per-capita light density

Figure 11: Predicting Outmigration with Individual Shocks Across Horizons



Notes: This column shows beta coefficients from individual regressions of birthplace-level outmigration rates on individual shocks at different horizons. The lags are at the 1 and 10-year horizon.

growth. I interpret this as strong evidence that level changes in migrant labor size slow productivity growth, even as they increase city size and extent. The estimates of diversity are noisy. This is partially a result of a relatively weak instrument. I also consider this suggestive evidence that the effects of migrant labor size outweigh the effects of diversity.

Across estimation strategies, I find evidence that increasing migrant labor size causes cities to grow in light density, but lowers per-capita GDP growth. Increasing migrant diversity decreases city growth, both in level and per-capita terms, but increases the non-agricultural labor share. Migrant labor size does not have consistent effects on industry mix, reflecting the importance of identifying both level and composition parameters to understand the aggregate impacts of migration. My results are broadly consistent with cross-sectional work that finds limited evidence of agglomeration returns in African cities, and negative impacts of ethnic diversity. My findings on the effects of migrant diversity on labor allocation suggest long-run urbanization benefits of migrant diversity, which I explore further in my long-run estimation strategy below.

Using a battery of shocks to birthplaces and historical characteristics of destinations, I predict both shifts and shares. I find that migration shocks do successfully predict both the size and composition of migrant labor. While the effects of labor size remain consistent, the effects of migrant diversity on light density outcomes are not statistically significant. I find consistent evidence that migrant diversity increases non-agricultural labor share. The exercise predicting settlement shares yields a weak first stage. However, I do find that the negative effects of migrant labor size on per-capita growth rates remains consistent and statistically significant.

4.4 Evidence from a Natural Experiment: Ending Apartheid

South Africa represents a particular case of internal sorting. The end of Apartheid brought a sudden lifting of migration restrictions that had forced the black population to remain in "ethnic homelands". These homelands, established by the Native Land Act in 1913, were loosely tied to distinct historical tribes. Transkei, for example, was a homeland attached to the Xhosa people. While KwaZulu, around Durban, was the homeland for the Zulu people. The Pass Laws required black South Africans to carry internal passports to regulate their movement outside of native lands (Amodio and Chiovelli, 2018). While not all black South Africans lived in these ethnic homelands, millions are believed to

Table 5: First Stage Prediction of Migrant Size and Composition with Predicted Shifts and Shares

Panel A: Predicted Shifts

	(1)	(2)
	Migrant Flows	
	$\Delta\ \ell$	Δ div
Predicted $\Delta\ \ell$	0.463	-0.006
	[0.034]***	[0.003]**
Predicted Δ div	2.213	1.607
	[2.629]	[0.345]***
Mean Dep.	7.88	-0.01
Observations	661	661
Fstat	93.081	12.106

Panel B: Predicted Shares

	$\Delta\ \ell$	Δ div
Predicted $\Delta\ \ell$	0.214	0.005
	[0.042]***	[0.002]*
Predicted Δ div	-22.876	3.721
	[12.033]*	[0.901]***
Mean Dep.	7.77	-0.01
Observations	710	710
Fstat	28.672	9.009

Note: This table estimates the first-stage regression of predicted on real changes in migrant labor and composition $real_{d,t} = \alpha\widehat{\Delta l_{d,t}} + \gamma\widehat{\Delta div_{d,t}} + \upsilon_t + \mu_c$. Δl is the logged difference in total migrant labor in destination d between census years. In Panel A, shifts are taken from predicted changes in outmigration across birthplaces, while shares are pre-period settlement shares from each birthplace. Shifts are instrumented by drought events, conflict events and commodity price shocks. In Panel B shifts are taken from the baseline changes in outmigration across birthplaces, while shares are predicted settlement shares from each birthplace, leveraging historical productivity shocks and origin-destination distance. Historical productivity shocks include distance to colonial rail, distance to mineral deposits, and distance to portage sites. Δdiv is the difference in HHI between census years for all non-native residents in location d. Regressions include fixed effects for country and year. * p<0.01, ** p<0.05, *** p<0.01.

have been forcibly resettled in the homelands between 1960 and 1991 (Lochmann et al., 2023). The Pass Laws were repealed in 1986, followed by the Native Land Act in 1991. These events amount to a sudden reduction in migration costs for black South Africans from different ethnic origins, who could now move freely to new destinations. Leveraging municipality level data from censuses between 1991-2022, we can study the long-term labor market effects of this migration shock.

Figure 12 shows the distribution of African ethnic homelands during the Apartheid era in South Africa. The panel on the right shows the distribution of language families, as measured in the Ethnologue dataset of languages. These figures give a sense of the correspondence between the homelands and particular black ethnic groups or linguistic families. The regions outlined in black are local municipalities as of 2022. These regions will serve as our units of analysis when studying regional outcomes after Apartheid.

Between 1986 and 1991, many of the restrictions on internal migration were lifted for black South Africans. Using census data from 1991 - 2022, we can trace the trajectory of diversity and population size over this period. While the South African censuses don't contain consistent information on subnational birthplace, they do identify native language or mother tongue of respondents. Using

Table 6: Second-Stage Effects of Migration on Productivity with Predicted Shifts and Shares

Panel A: Predicted Shifts

	(1) Δ Lights	(2) Δ Lights/Capita	(3) Δ Non-Agriculture Share	(4) Δ Housing Quality
Δ l	0.256	-1.387	0.012	-0.299
	[0.076]***	[0.157]***	[0.070]	[0.130]**
Δ div	-1.849	0.406	2.104	-2.012
	[1.521]	[4.300]	[1.265]*	[2.034]
Mean Dep.	-0.18	2.43	0.35	0.03
Observations	661	661	465	472
Kleibergen-Paap Fstat	12.788	12.788	13.289	11.121
Sanderson-Windmeijer ℓ	71.251	71.251	43.754	43.862
Sanderson-Windmeijer div	12.810	12.810	13.497	11.174

Panel B: Predicted Shares

	Δ Lights	Δ Lights/Capita	Δ Non-Agriculture Share	Δ Housing Quality
Δ l	-0.283	-3.150	0.612	0.223
	[0.264]	[0.717]***	[0.185]***	[0.347]
Δ div	4.546	6.553	-1.331	0.068
	[11.405]	[8.927]	[2.099]	[2.731]
Mean Dep.	-0.13	2.70	0.23	0.01
Observations	710	710	517	553
Kleibergen-Paap Fstat	9.051	9.051	10.804	4.734
Sanderson-Windmeijer ℓ	11.744	11.744	7.938	8.339
Sanderson-Windmeijer div	9.329	9.329	10.576	5.742

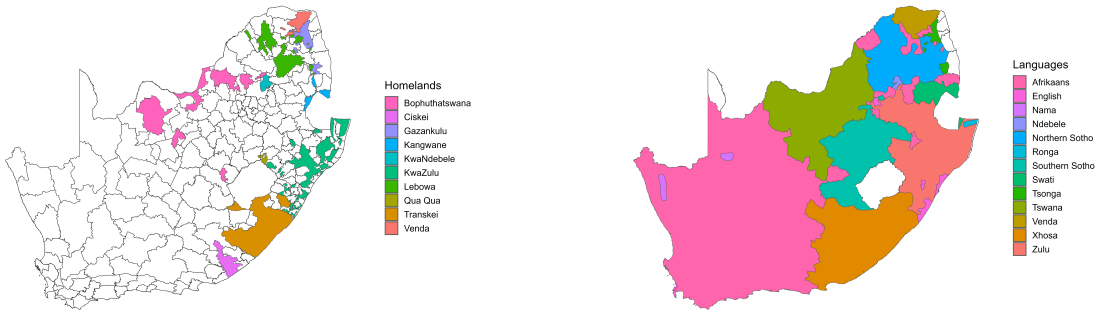
Note: This table presents the 2SLS estimation of changes in migrant labor and diversity on light density outcomes. $y_{d,t} = \alpha\Delta l_{d,t} + \gamma\Delta div_{d,t} + \upsilon_{tc}$. In Panel A, shifts are taken from predicted changes in outmigration across birthplaces, while shares are pre-period settlement shares from each birthplace. In Panel B shifts are taken from the baseline changes in outmigration across birthplaces, while shares are predicted settlement shares from each birthplace, leveraging historical productivity shocks and origin-destination distance. The original harmonized range of light density is from 0 to 63. The first column outcome is the standardized change in average light density, while the second column is the log change in light density per capita. The third column outcome is the change in labor share in non-agricultural industries, and the fourth column measures changes in housing quality, measured by a principle component of housing characteristics. All regressions include country-year fixed effects. Data for this table comes from an origin-destination panel of workers in African IPUMS Census samples. All regressions include country-year fixed effects. * p<0.01, ** p<0.05, *** p<0.01.

this language information, we can link black respondents to their ethnic homeland using the same language-ethnic group linking procedure leveraged in the main analysis. For each destination, we can calculate an ethnic HHI index as the relative diversity of the black population. We don't include English or Afrikaans speakers in this index.

Figure 14 shows the change in diversity over time in destinations. The left figure splits destinations in terms of average distance to ethnic homelands, with the furthest quantile representing destinations on the west coast of the country. The y-axes capture the negative Herfindahl index of black population diversity by native homeland, with higher values representing increased diversity. The right figure considers major urban destinations, and plots the change in the HHI index, rather than the levels. I find heterogeneous fluctuations in black ethnic diversity across destinations. In levels, the closest quantile of municipalities to ethnic homelands is the most diverse in terms of the black population's ethnic mix. After the abolishment of pass laws, this group experiences a decrease in diversity, as South Africans depart for further off destinations. The furthest quantile of municipalities experienced a long increase in diversity from 2001 to 2011, nearing the levels of the native homelands. Cape Town, one of the major towns in the furtherst quantile group, saw a strong increase in diversity from 1991-1996. This is shown in Panel B, which captures changes to the negative HHI over time. Other major cities saw clear increases in diversity, represented as points above 0 in Panel B. Though the rates are less dramatic in terms of proportional changes.

I replicate the baseline shift-share strategy from Table 3, applied to the South Africa case. The shifts g_{ot} are the post-Apartheid growth in migrants from origin o, as defined by their native language, living outside of o. The settlement shares are the Apartheid shares of each migrant group living in a particular municipality. One might instead think to use "distance to homeland" as an instrument

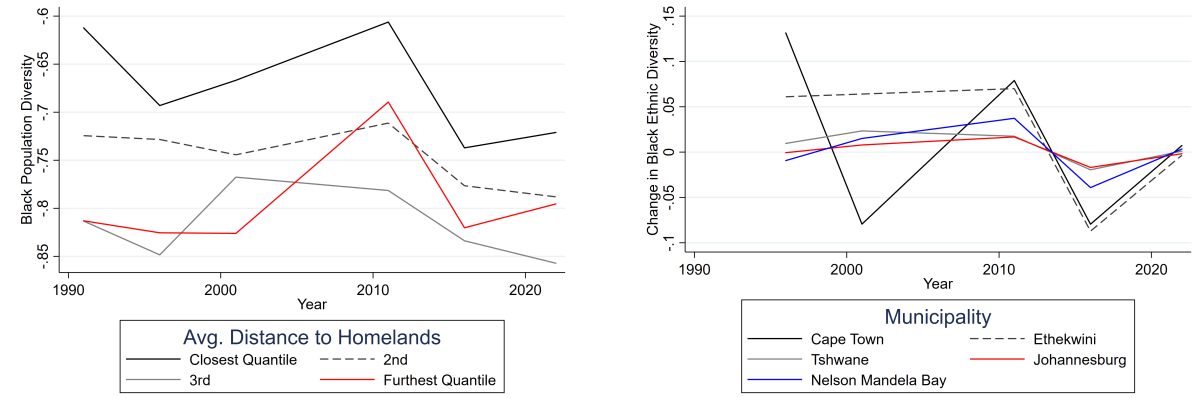
Figure 12: Ethnic Homelands under Apartheid, South Africa



(a) Ethnic Homelands (b) Language Families (Ethnologue)

Notes: These figures show the location of South Africa's ethnic homelands, where the majority of the black population were forced to settle. The figure on the right shows the distribution of major language families, according to the Ethnologue dataset of languages.

Figure 14: Black Diversity in South Africa After Apartheid



(a) Ethnic -HHI by Homeland Distance (b) Δ Ethnic -HHI in Major Cities

Notes: These figures show change in the homeland composition of the black population in destinations over time. The left figure shows the HHI score for destinations at different average distances from the native lands in the eastern part of the country. The right figure shows the change in this HHI measure for major cities, from the end of Apartheid onward.

for the settlement shares. However, as seen in the trend figures, migrants did not necessarily travel to cities close to homelands. Many major cities like Cape Town, far from the native homelands, experienced sharp shifts in the migrant population.

Table 7 shows the results of a second-stage shift-share. The linear effects of migrant labor flows are consistent with the headline results. Migrants increase city sizes, but migrant flows do not produce per-capita light density growth. However, the compositional effects are inverted. Increasing black migrant diversity causes larger cities and higher per-capita growth in light density. These results show that the effects of migrant diversity are heterogenous across locations. South Africa's particular history of apartheid may have played a role in mitigating harmful interactions across black ethnic groups.

5 Mechanisms

In this section I consider a few possible mechanisms for the headline result of the paper. Migrant flows increase city sizes but slow productivity growth. Migrant birthplace diversity causes lower city sizes and lower productivity growth, but increases urbanization. The negative impact of diversity is indicative of a nonlinear congestion force that dampens agglomeration benefits in the short-run, but may yield positive long-run benefits in terms of structural transformation. The "birthplace

Table 7: South Africa Shift-Share Second Stage, Migrant Flows

	Panel A: OLS Results	
	(1)	(2)
	Δ Lights	Δ Lights/Capita
Δ l	0.047	-0.283
	[0.016]***	[0.033]***
Δ div	1.646	1.582
	[0.301]***	[0.662]**
Mean Dep.	0.07	-0.15
Observations	798	798
	Panel B: Shift-Share IV	
	Δ Lights	Δ Lights/Capita
Δ l	0.083	-0.215
	[0.026]***	[0.046]***
Δ div	4.524	10.525
	[1.640]***	[3.161]***
Mean Dep.	0.07	-0.15
Observations	798	798
Kleibergen-Paap Fstat	9.037	9.037
Sanderson-Windmeijer L	374.608	374.608
Sanderson-Windmeijer Div	22.360	22.360

Note: Panel A presents the OLS estimation of changes in migrant labor and diversity on light density outcomes. $y_{d,t} = \alpha\Delta l_{d,t} + \gamma\Delta div_{d,t} + \upsilon_{tc}$. Panel B presents the IV 2SLS estimation of changes in migrant labor and diversity on light density outcomes. Light density is included with various specifications. The original harmonized range is from 0 to 63. The first column gives the log level in the contemporary period, the second is the logged difference across census years, the third is the non-logged distance and the fourth divides the light density value by the population of the administrative region in that period. All regressions include country-year fixed effects. * p<0.01, ** p<0.05, *** p<0.01.

composition" measure, or Δdiv, is a broad index of diversity. Because it is derived from workers' differences in birthplaces, there could be many types of diversity captured by this measure. Diversity in birthplace may be correlated with linguistic diversity, ethnic or religious diversity. To the extent that different origins may have different skills or experiences, it also encompasses diversity of skill. I consider a few examples of how birthplace diversity may create short-run negative externalities through ethnic conflict, difficulties in cooperation related to linguistic distance, or differences in skill complementarity.

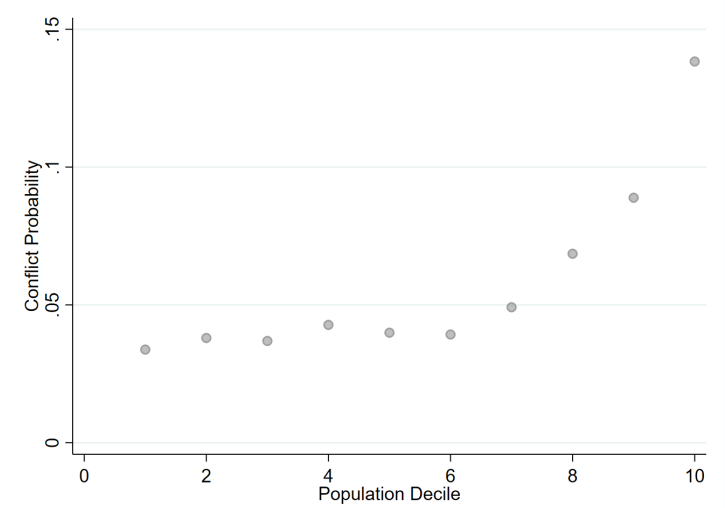
5.1 Ethnic Conflict

The short-run costs of migrant diversity may be a result of urban conflict generated between disparate groups. A standard result in political economy links ethnic diversity in Africa to ethnic conflict (Arbatli et al., 2020) [5]. Most of the evidence of this relationship comes from cross-sectional associations of diversity measures on geolocated battle events from ACLED or UCDP.

Beyond the normal causality concerns, another challenge for this conflict literature is understanding the role of population size on conflict. Indeed, in the cross-section population size is strongly positively correlated with conflict. To the extent that diversity and population move together, it's difficult to disentangle the role that diversity plays independent of population size. If larger population centers are naturally more diverse, increased conflict or crime could be due to either force. The construction of my empirical strategy is able to disentangle these effects.

[5] There are debates about the particular functional form this relationship takes, with scholars arguing for different ways to index diversity and weight relative group sizes as a metric for polarization, fractionalization, etc. The relationship between relative group size and conflict is not obvious. Many ethnic conflicts, like the Rwandan genocide, seem to emerge without the need for many groups, or equally sized groups. I continue to rely on changes in a negative Herfindahl index, for consistency with the baseline specification

Figure 16: Conflict and Population Size in the Cross-Section



Notes: This figure shows the cross-sectional relationship between the number of ACLED battle-events in a region and population size, ordered into deciles. The regions defined in this analysis are grid-cells of the African continent of approximately $1000km^2$. Population is calculated from the Worldpop estimates for 2013. The y-axis measures the average probability of observing a conflict event in that grid-cell in a given year between 1997 and 2025.

Following the Bartik style shift-share strategy elaborated in the empirical methods section, I study the effect of changes in migrant labor size and migrant diversity on conflict outcomes. Table 8 shows the results of this analysis. Conflict is measured as the number of battle events that occur in a given destination-year, according to the ACLED dataset. I also weight the conflict events by estimated average fatalities, which is represented as "deaths" in the table columns.

Table 8: Shift-Share Second Stage, Conflict Outcomes

Panel A: OLS Results

	(1) Conflict	(2) Deaths	(3) Δ Conflict	(4) Δ Deaths
Δ ℓ	0.410	-0.241	0.347	-0.023
	[0.268]	[0.296]	[0.140]**	[0.445]
Δ div	2.573	-3.249	7.122	1.159
	[1.081]**	[4.983]	[2.074]***	[3.681]
Mean Dep.	1.19	1.26	0.63	0.14
Observations	829	829	581	581

Panel B: Shift-Share IV

	Conflict	Deaths	Δ Conflict	Δ Deaths
Δ ℓ	0.719	-1.214	0.738	0.543
	[0.588]	[0.747]	[0.257]***	[0.882]
Δ div	4.041	-7.763	10.991	-4.733
	[2.022]**	[6.253]	[3.349]***	[8.157]
Mean Dep.	1.19	1.26	0.63	0.14
Observations	829	829	581	581
Kleibergen-Paap Fstat	16.462	16.462	12.030	12.030
Sanderson-Windmeijer ℓ	245.593	245.593	156.499	156.499
Sanderson-Windmeijer div	27.686	27.686	22.170	22.170

Note: Panel A presents the OLS estimation of changes in migrant labor and diversity on conflict at destination. $y_{d,t} = \alpha\Delta l_{d,t} + \gamma\Delta div_{d,t} + \upsilon_{tc}$. Panel B presents the IV 2SLS estimation of changes in migrant labor and diversity on conflict at destination. Outcomes include the number of ACLED battle events, which we call "conflict", as well as the estimated number of conflict deaths in a destination across years. We include these outcome both as levels and in differences. The differenced terms are in raw changes in the number of conflict events or deaths over time. All regressions include country-year fixed effects. * p<0.01, ** p<0.05, *** p<0.01.

Column 3 presents results for the differenced change in conflict at destination, in response to changes in migrant labor size and diversity. A 1% increase in migrant labor size adds 0.347 average conflict events per year. A .1 point increase in migrant diversity increases the number of conflict events by 0.7 per year. This is strong evidence of a congestion force that affects destinations as larger and more heterogenous migrants arrive in destinations. Ethnic conflict may be one mechanism that reduces light density growth in increasingly diverse destinations.

5.2 Linguistic Distance

A natural problem for migrants moving to new destinations may be learning a new language. If migrants arriving in a destination speak many different languages, these linguistic differences could pose challenges for coordination in firms, households, or neighborhoods. A mostly US literature on assimilation of migrants suggests that closing cultural gaps is an important part of unlocking returns to migration (Abramitzky et al., 2020). So far none our measures of diversity have tried to weight differences across birthplaces in terms of cultural or linguistic distance. Two points for why its not ex-ante obvious that cultural distance is the important mechanism driving our results: (1) This paper looks at internal migration. While many states have an abundance of local indigenous languages, its also true that most states have a condensed set of national languages. Take Mozambique for example. While the country hosts a wide variety of local languages and dialects, most people in urban centers speak Portuguese fluently. (2) There are many examples of conflict or animosity between groups that are on paper culturally similar (Posner, 2004). Differences in groups over space are often the result of historical, geographical and political factors that may arise even when groups speak the same language.

To consider the role of linguistic distance, I match each administrative region to its majority spoken language, as according to the Ethnologue, a large geographically coded dataset of world languages. A useful feature of the Ethnologue is that each language is coded within a nested tree of language families and sub-families. Oko, for example, is a Nigerian language spoken in Edo state and part of the Niger-Congo family. Within the Niger-Congo linguistic family, its part of the Volta-Niger linguistic subgroup. Between any two languages, we can then create a simple measure of linguistic distance based on how many linguistic families and sub-families they have in common. Oko is related to Ibo, another Volta-Niger language, but distant from the northern Nigerian Hausa, an Afro-Asiatic family language.

For each migrant from an origin o in destination d, I calculate the linguistic distance between the majority language in that given origin and destination pair. Weighting by the total migrants from each origin into a destination d, I calculate an average linguistic distance index that measures the average cultural distance between migrants and destination natives. Table 9 presents results from an OLS regression in first differences, relating destination productivity to changes in migrant labor size and changes in the average linguistic distance of migrants. As above, these differences are calculated across census periods. I find that increasing linguistic distance is only weakly related to changes in light density, but strongly related to an increasing share of non-agricultural labor.

Is increasing linguistic distance captured up by the diversity shift-share instrument used in the analysis? In Table 10, I regress the measure of average linguistic distance against my first-stage predictors from Table 2. In levels, higher ethnic concentration is in fact associated with a lower level of linguistic distance. However, in differences I find the opposite of the expected relationship. Increasing diversity of birthplace composition predicts decreases in average linguistic distance. That is, our predictor of changes in birthplace diversity does not simultenously predict changes in linguistic distance. Therefore, while we do observe an association between changes in linguistic distance and higher agricultural labor share, my shift-share strategy does not directly map to a levels change in average linguistic distance.

Table 9: Change in Linguistic Distance and Productivity

OLS Results

	(1) Δ Lights	(2) Δ Log(Lights/Capita)	(3) Δ Non-Agriculture Share	(4) Δ Housing Quality
Δ ℓ	0.217	-0.779	-0.027	0.027
	[0.049]***	[0.077]***	[0.016]*	[0.036]
Δ Linguistic Distance	-2.812	3.582	4.600	1.701
	[1.442]*	[5.019]	[1.505]***	[2.344]
Mean Dep.	0.01	3.32	0.13	0.03
Observations	881	881	685	722

Note: This table presents the OLS estimation of changes in migrant labor and average linguistic distance on productivity at destination. Average linguistic distance is calculated as the inverse of the similarity between a migrant's language and the majority language at destination, weighted by flow size across all migrant groups. Similarity is defined according to Ethnologue linguistic families. This measure is then differenced across census periods. All regressions include country-year fixed effects. Clustering is at the second administrative level. * p<0.01, ** p<0.05, *** p<0.01.

Table 10: First Stage Relationship of *div* and Linguistic Distance

First Stage Correlation

	Linguistic Distance	Δ Linguistic Distance
Predicted Δ l	0.015	-0.002
	[0.001]***	[0.000]***
Predicted Δ div	0.094	-0.027
	[0.022]***	[0.010]***
Mean Dep.	-0.00	-0.00
Observations	944	944

Note: This table presents a first-stage regression relating the predicted migrant labor and predicted diversity change estimates from our shift-share design to average changes in linguistic distance at the destination. Average linguistic distance is calculated as the inverse of the similarity between a migrant's language and the majority language at destination, weighted by flow size across all migrant groups. Similarity is defined according to Ethnologue linguistic families. The first column outcome is the average linguistic distance of migrants to destination natives, weighted by the group size of migrants. The second column outcome is this measure differenced across census periods. All regressions include country-year fixed effects. Clustering is at the second administrative level. * p<0.01, ** p<0.05, *** p<0.01.

5.3 Skill-Level, Segregation and Industry Concentration

The changes in migration discussed in this paper are in terms of the number of people over 18 that have settled in a destination. While the empirical model is in differences, its possible that the effects are being partially driven by differential trends in the human capital of the migrants arriving in destination. Changes in migrant diversity may be related to changes in the average skill level of the workforce. Another possible channel is through other features of the labor market. If migrants arriving in destinations bring many diverse skills, they may reduce the industrial concentration of cities and in-turn diversify production at destination. Lower industry concentration may be a channel through which migrant diversity creates long-run structural transformation.

To study these mechanisms, I consider the impact of instrumented migrant levels and composition on changes to average worker skill level, industry concentration and industry segregation. Industry concentration is measured as the HHI of industry labor-share across industries within a region. Industry segregation is a measure of the extent to which migrants from different origins sort into particular industries. The measure captures the extent to which industry labor shares for a given migrant birthplace deviate the population average in that region (**?**). For a given migrant origin group $o \in O$ living in destination d and working in industry $i \in I$, segregation is defined as:

$$Segregation_d = \frac{1}{O-1} \sum_{o=1}^{O} \sum_{i=1}^{I} \frac{N_o}{N_d} \frac{(\pi_{io} - \pi_{od})^2}{\pi_{od}} \tag{21}$$

Where π_{od} is the fraction of group o in destination d, and π_{io} is the fraction of group o in industry i of destination d. N_o is the total population of group o in destination d and N_d is the total population in destination d. Higher values of this segregation index correspond to greater segregation of ethnic groups across industries within a destination.

Table B4 presents the OLS and SSIV results for changes in average worker skill level, industry concentration, and industry segregation by birthplace. The model is the same baseline shift-share examined in Table 3. I find that both increasing migrant labor size and diversity lowers industry concentration, as measured by an HHI of labor-shares. This suggests that migrant flows and composition introduce increase diversity to the industrial mix of a destination. I also find evidence that increasing migrant labor size reduces average worker skill level. This is a possible mechanism by which increasing migrant labor size does not translate to higher per-capita productivity growth. Low skilled migrants bring down the average skill level of workers, compete for low-skilled wages, but may not geneate the ideas, businesses and networks necessary to create agglomeration benefits. I find no evidence that increasing migrant diversity affects the average skill level of workers or the segregation of groups across industries. While increasing migrant diversity does introduce diversity to the industry mix, we don't see evidence that particular migrants specialize within certain industries. The industry categories are broad, therefore its possible that given more granular occupational categories different specialization patterns would emerge.

5.4 Productivity Evidence from Firms

Past work has suggested that ethnic divisions within firms can lower productivity in team-based production (Hjort, 2014). This suggests a mechanism via poor coordination or discrimination within firms, rather than outright conflict. In this section, I consider evidence on changes in diversity within firms, and the subsequent impact on productivity. An ideal microeconomic dataset to study the role of migrants in productivity would include firm-level information on input-output, as well as detailed characteristics of the firm's labor force. While some enterprise level surveys exist for African countries, few provide information on workers, including ethnic identity or birthplace. Two exceptions come from the "Regional Programme on Enterprise Development (RPED)", led by the World Bank in collaboration with the Centre for the Study of African Economies (CSAE). In the 90s and early 2000s a panel of manufacturing firms was collected for Ghana and Tanzania which, for some waves, include data on the ethnic composition from a sample of workers.

These panels were collected between 1992 and 2003, recording basic characteristics of the firm such as wages and labor size. In addition, a worker supplement is collected for each firm in which up to 10 workers are interviewed and asked about their experience and background. I leverage this data to examine associations between firm-level productivity, number of workers and worker ethnic concentration. Table 11 shows the results of firm-level regressions of productivity measures on firm labor size and worker ethnic concentration. Outcomes include log wages per worker, and log manufacturing output value per worker. Columns 1 and 3 include fixed effects for country and year, while columns 2 and 4 look at within-firm variation over time using firm fixed effects. I find some evidence that in the cross-section, firm productivity is increasing in labor size and decreasing in ethnic concentration. This is the opposite association of what we'd expect to see if ethnic diversity hinders firm productivity. Looking within firms over time, we see evidence that firms are becoming less productive as they grow in size, with no evidence of an effect by changing ethnic HHI.

These samples are small, and the analysis is not causal. However we don't see evidence of an ethnic diversity penalty within or across firms that has been posited by microeconomic papers.

Table 11: Ethnicity and Firm Productivity in Ghana, Tanzania

	Log(Wages/Worker)	Log(Wages/Worker)	Log(Output/Worker)	Log(Output/Worker)
Log(labor)	0.299	-0.903	0.361	-1.326
	[0.053]***	[0.136]***	[0.104]***	[0.511]**
Ethnic HHI	-1.067	0.126	-0.457	0.868
	[0.269]***	[0.224]	[0.530]	[0.843]
Mean Dep.	15.19	15.78	12.82	13.44
Observations	511	332	496	330
Country FE	Y	N	Y	N
Wave FE	Y	N	Y	N
Firm FE	N	Y	N	Y

Notes: This table shows the result of a regression of log wages or output per worker on firm size and worker ethnic concentration $log(\frac{y}{\ell})_{ist} = \beta_1 \ell + \beta_2 HHI + \epsilon_{ist}$. Where i is firm, s country, and t is survey wave. The first column in each pair includes country and survey wave fixed effects, while the second includes firm fixed effects. Output represents the total monetary value of manufacturing output, while wages reflect the total wage bill. Each regression also controls for the number of workers that data was collected on for the HHI variable, up to 10 workers. * p<0.01, ** p<0.05, *** p<0.01.

5.5 Migration and Ethnic Attitudes

Are migrants more or less prone to ethnic conflict relative to natives or non-migrants? It may be that the act of migrating itself is associated with increasing or decreasing tribalism among ethnic groups. For example, a migrant from a minority group may choose to move to a city, and begin to experience economic disenfranchisement or discrimination. In turn, the migrant becomes more allied with their ethnic group, and deepens their coethnic preference. The inverse of this process is also possible. Work with panel data in Kenya has shown reduced tribalism and increasing national identity in individuals after they migrate to cities (Kramon et al., 2022). This suggests that urban centers play a diffusing role in ethnic tensions. Using Afrobarometer data, I explore observational differences between migrants and non-migrants in terms of their reported national identity and future economic expectations. I use the same model of observational returns leveraged in Table 1. I replace our outcome of interest with a measure of ethnic allegiance or economic disenfranchisement:

$$\begin{aligned} Attitudes_{it} = \beta_1 \text{Migrant}_i + \beta_2 \text{Migrant}_i \cdot \text{Dist}_{od} \\ + \beta_3 \text{Migrant}_i \cdot \text{CoethShare}_{od} + Z_i + X_{od} + W_d + \upsilon_{st} + \gamma_o + \epsilon_{it} \end{aligned} \tag{22}$$

Where $Dist_{od}$ is the log distance between the ethnic homeland and the destination, $Migrant_i$ is a dummy for migrant status, and $CoethShare_{od}$ reflects the fraction of individuals in d that are from ethnic homeland o. We include fixed effects for country-year υ_{st}, and γ_o to isolate variation within an ethnic group. The controls for the individual Z_i include age and schooling, while X_{od} includes

the level of o-d distance and coethnic share, and W_d includes destination log population. $Attitudes_{it}$ is a measure of ethnic or economic attitudes, "Nationalism" or "Expectations". National identity, or "nationalism" is an individual's response to a question about how much the individual identifies with the nation relative to their ethnic group. We use it here as a measure of ethnic allegiance. Future expectations is a question styled after the Michigan Consumer Confidence Survey, which asks individuals if they expect economic conditions to improve in the next year. We use this question as a measure of economic disenchantment or distress. Table B5 shows these results, with no significant differences between migrants and natives in terms of national identity or economic expectations.

6 Urban Growth and Diversity in the Long-Run

The analysis in this paper has focused on decade level changes in migrant labor size, composition and productivity. The positive effects of migrant diversity on non-agricultural labor share suggests that migrant diversity may affect fundamental characteristics of the economy at destination. It may be that the benefits of such effects manifest over longer time horizons. The problem with studying longer-run effects of labor size and diversity on growth is a lack of data. Most early African censuses, where they exist, don't include ethnic or birthplace identifiers. In addition, few measures of subnational productivity exist in earlier periods.

In this section I leverage the historical pull shocks constructed for the shift-share instrument to estimate a long-run effect of labor size and diversity on urban development. The historical productivity shocks, including colonial rail lines, portage sites and mineral deposits serve as agglomeration instruments, which can be used to predict contemporary city locations. 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.

I start from the following equation. For each region i, the long-term relationship between diversity, population and productivity is described as:

$$y_i = \beta_0 + \beta_1 \ell_i + \beta_2 div_i + X_i + \epsilon_i \tag{23}$$

where ℓ_i and div_i capture a destination's size and diversity in the long-run, and X_i is a vector of geographic controls. This cross-sectional regression is similar to the associations estimated by Montalvo and Reynal-Querol (2021). What I add to this framework is a causal inference strategy. Since there is no long-run panel data on diversity div_i over time, I use a historic fractionalization index that captures each region's exposure to historical ethnic groups defined by anthropological maps. I use the Murdock Map, and define each region's historical diversity exposure as the distribution of land occupied by different ethnic homelands. A region's contemporary diversity is proxied by the interaction of this historical exposure to different groups, and historical labor demand. Intuitively, areas become more diverse as labor demand shocks compel them to draw in labor from surrounding areas. The more inherently diverse this potential pool of labor (due to the historic spatial distribution of groups), the more diverse the city will be. Substituting historic diversity $HistDiv_i$ for div_i, I estimate:

$$y_i = \beta_0 + \beta_1 \ell_i + \beta_3 HistDiv_i + \beta_4 HistDiv_i * \ell_i + X_i + \epsilon_i \tag{24}$$

where $HistDiv_i$ captures a region's potential exposure to diverse migrants based on the Murdock Map, and ℓ_i captures a measure of employment density. The interaction of fractionalization and labor is the variable of interest. The interaction captures how historic fractionalization affects the labor demand elasticity, and in turn output and productivity. $HistDiv$ represents a fixed regional quality, and β_4 represents an elasticity of urban growth with respect to this quality.

Population density is an endogenous variable. In addition, $HistDiv_i$ by itself will be related to a

variety of geographic fundamentals that governed the distribution of groups over space (Michalopoulos, 2012). 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. I use these historic productivity shocks as agglomeration instruments, and predict both labor size $\hat{\ell}$ and the interaction of labor size and diversity $\ell * \widehat{HistDiv}$:

$$\hat{\ell}_i = \alpha + \beta_1 \Delta A_i + \beta_2 \Delta A_i * HistDiv_i + \omega_s + \epsilon_i \tag{25}$$

$$\ell_i * \widehat{HistDiv}_i = \alpha + \beta_1 \Delta A_i + \beta_2 \Delta A_i * HistDiv_i + \omega_s + \epsilon_i \tag{26}$$

ΔA_i represents one of the three historical pull shocks, either distance to colonial rail, distance to portage site, or distance to a mineral deposit. ω_s are state fixed effects. Using these regressions as two first-stages, I will then estimate the second stage effect on contemporary development outcomes for each region in the cross-section. The exclusion restriction requires that a historical productivity shock ΔA_i is uncorrelated with unobserved other factors in ϵ_i that drive outcomes of light density, wealth, lights/capita,and conflict in the contemporary period. Note we do not need that $HistDiv_i$ is uncorrelated with ΔA_i in this estimation.

I split the African continent into equally sized hexagonal grids of approximately $1000km^2$, 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. Figure 17 shows an example of how the Murdock Map is used to calculate a historical diversity index for each region. This is a fractionalization index of the relative share of each grid taken up by a particular Murdock group. Higher values imply more diversity.

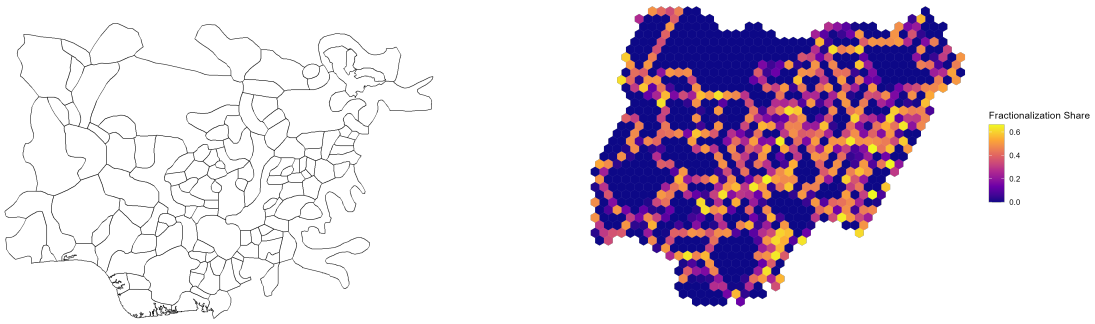
I use a dummy for labor size ℓ that marks regions as cities (or high density) if they overlap with a city of over 20k inhabitants in 2015, as defined by the urban database Africapolis. Table B6 shows the results of OLS regressions for different development outcomes y following equation 24. Panel A shows results using a city dummy for ℓ, while panel B uses an interpolated population measure at the grid level, taken from the Worldpop dataset. Urban agglomerations have higher light density, and higher conflict. Due to the high relative population size, per capita measures in these regressions are negative relative to less populated regions. In the cross-section, I find that cities located in more diverse areas have a higher lights/capita value, but also higher incidence of conflict as measured by a summation of ACLED conflict events.

Table 12 shows results for our three instrumental variable strategies following the historical pull characteristics described in the Empirical Strategy section. I find that across strategies, high diversity places that receive a productivity shock have higher values of our preferred development measure, lights/capita. The baseline negative coefficient on diversity is in line with the correlations studied in past political economy work on the relationship of diversity and development. While baseline diversity has a negative coefficient, the elasticity of urban growth with respect to diversity is positive, suggesting that in the long-run diversity of ethnic groups can be a boon for urban growth.

7 Conclusion

This paper disentangles the effects of migration levels and composition on urban African productivity. I build a panel that proxies for origin-destination migration flows using census data on reported birthplace and ethnicity. Then I implement a shift-share instrument that simultaneously identifies the linear change in migration labor size and nonlinear birthplace composition. I find that destinations that receive more migration labor grow in light density, but are not more productive in per-capita terms. This is consistent with a story of growing African cities with limited productivity returns or urban structural transformation (Jedwab et al., 2025). Migrants that are more diverse cause lower growth in levels and per-capita terms. However, increasing migrant diversity also yields higher non-agricultural labor force shares in destinations. Migrant diversity creates short-run costs, but

Figure 17: Data Examples from Nigeria



(a) Murdock Map (b) Grid Fractionalization

Notes: This figure shows examples of the grid-level data constructed from a historical map of diversity. Fractionalization is calculated as the dispersion in the share of each grid taken up by different ethnic homelands.

seems to generate long-run benefits to destinations in terms of urbanization. The results are broadly consist when I instrument for shifts using plausibly exogenous push shocks to outmigration, including international commodity prices, conflict events and drought conditions. In zero stage regressions, I find evidence that these channels move outmigration from origins, and have predictable effects on migrant labor size and composition. This fact has implications for how we study the impact of migration shocks, as the spatial dispersion of shocks changes both the size and composition of migrants.

I explore several mechanisms to better understand the effect of migrant labor size and diversity on destinations. I find evidence that increased migrant diversity and migrant labor size increase urban conflict. I don't find evidence that diversity hurts productivity at the firm level in a panel of firms. The mixed results for diversity across different outcomes lead me to pursue two further exercises. In a study of the Apartheid period of South Africa, I leverage the repeal of the Pass Laws as a migration shock event that generates high outmigration shifts from native homelands. Measuring the size and composition of the black migrant labor force and estimating the same baseline shift-share yields an inverted effect of diversity. Productivity of South African municipalities is increasing in birthplace diversity, suggesting that South African cities benefit from the diversity of the black population. The particular history of South Africa's conflict between the black and white populations may have diffused tensions across different black African ethnic groups.

Lastly, I consider the long-run effects of diversity. I leverage my pull characteristics as "agglomeration predictors", which forecast optimal locations for long-run urban growth. By interacting these predictors with historical measures of diversity from anthropological maps, I estimate an elasticity of regional diversity with respect to these historical productivity shocks. I find that in a long-run cross-section, diverse areas that experienced a historical productivity shock fared better than less diverse areas who experienced comparable shocks. This pattern is consistent across different historical pull characteristics. I conclude that migrant labor diversity is a long-term boon for African cities, and and may outweigh congestion costs over time.

Table 12: 2SLS of Historical Diversity and Productivity Instruments

Panel A: Colonial Rail

	(1) Light Density	(2) Lights/Capita	(3) Growth Lights/Capita	(4) Conflict
City	4.537	-2.447	0.294	34.455
	[0.209]***	[0.254]***	[0.257]	[10.677]***
Diversity	0.154	-0.713	0.116	21.367
	[0.127]	[0.155]***	[0.157]	[6.509]***
City*Diversity	-1.265	3.300	0.265	-136.214
	[0.910]	[1.106]***	[1.121]	[46.520]***
Mean Dep.	-1.92	-10.90	-1.71	8.50
Observations	16,322	16,321	16,321	16,322
Cragg-Donald F-Stat	179.838	179.830	179.830	179.838
Sanderson-Windmeijer City	201.228	201.276	201.276	201.228
Sanderson-Windmeijer City*Div	186.097	186.088	186.088	186.097

Panel B: Mineral Deposit

	Light Density	Lights/Capita	Growth Lights/Capita	Conflict
City	7.526	-1.325	0.538	105.514
	[0.271]***	[0.271]***	[0.340]	[13.407]***
Diversity	0.212	-1.058	-0.002	16.038
	[0.148]	[0.148]***	[0.186]	[7.319]**
City*Diversity	-2.012	3.390	0.359	-116.340
	[1.008]**	[1.009]***	[1.268]	[49.963]**
Mean Dep.	-1.78	-9.88	-1.80	8.07
Observations	22,682	22,680	22,680	22,682
Cragg-Donald F-Stat	306.111	306.087	306.087	306.111
Sanderson-Windmeijer City	318.667	318.634	318.634	318.667
Sanderson-Windmeijer City*Div	421.934	421.894	421.894	421.934

Panel C: Portage Propensity

	Light Density	Lights/Capita	Growth Lights/Capita	Conflict
City	-1.424	-25.776	-24.468	-11.948
	[1.350]	[7.386]***	[7.549]***	[90.547]
Diversity	0.539	-2.138	-0.102	1.241
	[0.154]***	[0.845]**	[0.863]	[10.345]
City*Diversity	-5.528	15.076	1.820	24.295
	[1.478]***	[8.097]*	[8.276]	[99.161]
Mean Dep.	-1.97	-9.34	-1.76	6.89
Observations	32,933	32,931	32,931	32,933
Cragg-Donald F-Stat	5.441	5.451	5.451	5.441
Sanderson-Windmeijer City	5.897	5.908	5.908	5.897
Sanderson-Windmeijer City*Div	91.601	91.596	91.596	91.601

Notes: This table presents cross-sectional regressions of instrumented urban growth and the interaction of instrumented urban growth and historical diversity on contemporary productivity outcomes. All regressions include state fixed effects. Light density outcomes are calculated in 2013, for comparison to Montalvo and Reynal-Querol (2021). Column 1 measures log light density, column 2 is a log measure of light density over a Worldpop estimated population for the grid in 2010, column 3 measures the change in log lights/capita from 1992 to 2013. Column 4 measures the number of conflict events in the grid since 1997, measured in ACLED battle events. "City" is an indicator marked as 1 if an Africapolis city is located within the grid and the population is above 20,000. "Diversity" is a historical measure of diversity calculated as the fractionalization of land share of different Murdock ethnic groups in the grid cell. All regressions control for distance to coast, malaria and TseTse suitability, ruggedness, distance to a major river, agricultural land productivity and a historical estimate of population size in 1800. * p<0.01, ** p<0.05, *** p<0.01.

References

Abramitzky, R., Boustan, L., and Eriksson, K. (2020). Do immigrants assimilate more slowly today than in the past? *American Economic Review: Insights*, 2(1):125–141.

Albert, C., Glitz, A., and Llull, J. (2021). Labor market competition and the assimilation of immigrants.

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

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

Alesina, A., Harnoss, J., and Rapoport, H. (2016). Birthplace diversity and economic prosperity. *Journal of Economic Growth*, 21:101–138.

Alsan, M. (2015). The effect of the tsetse fly on african development. *American Economic Review*, 105(1):382–410.

Amodio, F. and Chiovelli, G. (2018). Ethnicity and violence during democratic transitions: evidence from south africa. *Journal of the European Economic Association*, 16(4):1234–1280.

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.

Bazzi, S. (2017). Wealth heterogeneity and the income elasticity of migration. *American Economic Journal: Applied Economics*, 9(2):219–255.

Bazzi, S. and Blattman, C. (2014). Economic shocks and conflict: Evidence from commodity prices. *American Economic Journal: Macroeconomics*, 6(4):1–38.

Bazzi, S., Ferrara, A., Fiszbein, M., Pearson, T., and Testa, P. A. (2023). The other great migration: Southern whites and the new right. *The Quarterly Journal of Economics*, 138(3):1577–1647.

Bazzi, S., Gaduh, A., Rothenberg, A. D., and Wong, M. (2016). Skill transferability, migration, and development: Evidence from population resettlement in indonesia. *American Economic Review*, 106(9):2658–2698.

BenYishay, A., Rotberg, R., Wells, J., Lv, Z., Goodman, S., Kovacevic, L., and Runfola, D. (2017). Geocoding afrobarometer rounds 1-6: Methodology & data quality. *AidData. Available online at http://geo. aiddata. org*.

Berman, N. and Couttenier, M. (2015). External shocks, internal shots: the geography of civil conflicts. *Review of Economics and Statistics*, 97(4):758–776.

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

Boberg-Fazlić, N. and Sharp, P. (2024). Immigrant communities and knowledge spillovers: Danish americans and the development of the dairy industry in the united states. *American Economic Journal: Macroeconomics*, 16(1):102–146.

Borusyak, K., Hull, P., and Jaravel, X. (2022). Quasi-experimental shift-share research designs. *The Review of economic studies*, 89(1):181–213.

Borusyak, K., Hull, P., and Jaravel, X. (2025). A practical guide to shift-share instruments. *Journal of Economic Perspectives*, 39(1):181–204.

Boustan, L. P., Fishback, P. V., and Kantor, S. (2010). The effect of internal migration on local labor markets: American cities during the great depression. *Journal of Labor Economics*, 28(4):719–746.

Boyle, E. H., King, M., and Sobek, M. (2024). Ipums demographic and health surveys: version 11. *(No Title)*.

Brückner, M. (2012). Economic growth, size of the agricultural sector, and urbanization in africa. *Journal of Urban Economics*, 71(1):26–36.

Burchardi, K. B., Chaney, T., and Hassan, T. A. (2019). Migrants, ancestors, and foreign investments. *The Review of Economic Studies*, 86(4):1448–1486.

Castells-Quintana, D. (2017). Malthus living in a slum: Urban concentration, infrastructure and economic growth. *Journal of Urban Economics*, 98:158–173.

Chiovelli, G., Michalopoulos, S., Papaioannou, E., and Regan, T. (2023). Illuminating africa? *Institute for International Economic Policy Working Paper Series, IIEP-WP-2023-11, George Washington University*, 3(5).

Choi, J., Hyun, J., and Park, Z. (2024). Bound by ancestors: Immigration, credit frictions, and global supply chain formation. *Journal of International Economics*, 147:103855.

Christiaensen, L., Lozano-Gracia, N., et al. (2023). Migrants, markets, and mayors [migrants, marchés et maires]. *World Bank Publications-Books*.

Christiaensen, L., Lozano-Gracia, N., et al. (2025). Africa's urbanisation dynamics 2025: Planning for urban expansion. *OECD*.

Combes, P.-P., Duranton, G., Gobillon, L., and Roux, S. (2010). Estimating agglomeration economies with history, geology, and worker effects. In *Agglomeration economics*, pages 15–66. University of Chicago Press.

Desmet, K. and Rossi-Hansberg, E. (2024). Climate change economics over time and space. *Annual Review of Economics*, 16.

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.

Glaeser, E. L. and Gottlieb, J. D. (2009). The Wealth of Cities: Agglomeration Economies and Spatial Equilibrium in the United States. *Journal of Economic Literature*, 47(4):983–1028.

Goldsmith-Pinkham, P., Sorkin, I., and Swift, H. (2020). Bartik instruments: What, when, why, and how. *American Economic Review*, 110(8):2586–2624.

Gollin, D., Jedwab, R., and Vollrath, D. (2016). Urbanization with and without industrialization. *Journal of Economic Growth*, 21:35–70.

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

Group, W. B. (2025). World bank pink sheet data. *World Bank Group*.

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

Hamory, J., Kleemans, M., Li, N. Y., and Miguel, E. (2021). Reevaluating agricultural productivity gaps with longitudinal microdata. *Journal of the European Economic Association*, 19(3):1522–1555.

Henderson, J. V., Nigmatulina, D., and Kriticos, S. (2021). Measuring urban economic density. *Journal of Urban Economics*, 125:103188.

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.

Imbert, C., Seror, M., Zhang, Y., and Zylberberg, Y. (2022). Migrants and firms: Evidence from china. *American Economic Review*, 112(6):1885–1914.

IMF (2025). Imf primary commodities database. *IMF*.

Jaeger, D. A., Ruist, J., and Stuhler, J. (2018). Shift-share instruments and dynamic adjustments: The case of immigration. *NBER Working Paper*, 24285.

Jedwab, R., Ianchovichina, E., and Haslop, F. (2025). The employment profile of cities around the world: Consumption vs. production cities and economic development. *World Development*, 188:106883.

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.

Kamuikeni, C. and Naito, H. (2024). The effect of climate change on internal migration: Evidence from micro census data of 16 sub-saharan african countries.

Kelly, T. D., Matos, G. R., Buckingham, D. A., DiFrancesco, C. A., Porter, K. E., Berry, C., Crane, M., Goonan, T., and Sznopek, J. (2010). Historical statistics for mineral and material commodities in the united states. *US Geological Survey data series*, 140:01–006.

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.

Lagakos, D., Marshall, S., Mobarak, A. M., Vernot, C., and Waugh, M. E. (2020). Migration costs and observational returns to migration in the developing world. *Journal of Monetary Economics*, 113:138–154.

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.

Lewis, E. (2011). Immigration, skill mix, and capital skill complementarity. *The Quarterly Journal of Economics*, 126(2):1029–1069.

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

Lochmann, A., Rao, N., and Rossi, M. A. (2023). The long-run effects of south africa's forced resettlements on employment outcomes. Technical report, Harvard's Growth Lab.

Marchiori, L., Maystadt, J.-F., and Schumacher, I. (2012). The impact of weather anomalies on migration in sub-saharan africa. *Journal of Environmental Economics and Management*, 63(3):355–374.

McGuirk, E. and Burke, M. (2020). The economic origins of conflict in africa. *Journal of Political Economy*, 128(10):3940–3997.

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

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

Michalopoulos, S. and Papaioannou, E. (2013). Pre-colonial ethnic institutions and contemporary african development. *Econometrica*, 81(1):113–152.

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.

Müller-Crepon, C., Pengl, Y., and Bormann, N.-C. (2022). Linking ethnic data from africa (leda). *Journal of Peace Research*, 59(3):425–435.

Mullins, J. T. and Bharadwaj, P. (2021). Weather, climate, and migration in the united states. Technical report, National Bureau of Economic Research.

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.

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.

Posner, D. N. (2004). The political salience of cultural difference: Why chewas and tumbukas are allies in zambia and adversaries in malawi. *American political science review*, 98(4):529–545.

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.

Robinson, A. L. (2020). Ethnic Diversity, Segregation and Ethnocentric Trust in Africa. *British Journal of Political Science*, 50(1):217–239. Publisher: Cambridge University Press.

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

Schubert, G., Stansbury, A., and Taska, B. (2024). Employer concentration and outside options. *Available at SSRN 3599454*.

Shrestha, M. (2017). Push and pull: A study of international migration from nepal. *World Bank Policy Research Working Paper*, (7965).

Vicente-Serrano, S. M., Beguería, S., and López-Moreno, J. I. (2010). A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of climate*, 23(7):1696–1718.

Wang, Y. (2024). Linguistic distance, internal migration and welfare: Evidence from indonesia. *Internal Migration and Welfare: Evidence from Indonesia*(February 01, 2024)*.

Young, A. (2013). Inequality, the urban-rural gap, and migration. *The Quarterly Journal of Economics*, 128(4):1727–1785.

A Simulations of Shift-Share Instruments with Nonlinearity

In this section, I run a simulation of the instrumental variable strategy to observe how the double instrumenting procedure performs when the true coefficients for labor size and diversity are known. The set-up of the simulation is as follows. Suppose there are 500 regions, sourcing population from 50 homelands or origins. Each destination d receives a flow of migrants from homeland o, according to:

$$X_{od} = \alpha_1 Z_{od} + \alpha_2 U_d + \theta V_d + \psi W * V_d + \epsilon_{od} \tag{27}$$

Where Z_{od} is an exogenous component of origin flows that is normally distributed, but origins vary in size and volatility $Z_{od} \sim N(10 * o, 3 * o)$. As o is assigned numerical values 1 to 50, higher values have higher means and standard deviations. U_d is an unobserved omitted variable that is specific to a given region, $U_d \sim N(10, 5)$, and epsilon is $\epsilon_{od} \sim N(0, 1)$. We can think of U_d as a bias that captures the universal "attractiveness" of particular regions for people migrating from any homeland. In addition, I include two parameters that govern the relationship between total labor and total diversity in region d. V_d is distributed normally $V_d \sim N(0, 1)$, and W is a weighting vector that assigns 1 to the first homeland, and 0 to the rest, creating a skew in flows towards homeland 1. The strenght of these forces are governed by θ and ψ parameters. The relative strength of Z_{od} and the unobserved variable U_d are governed by α_1 and α_2.

Given these numbers, the real labor supply and diversity for a destination city d can be calculated as:

$$L_d = \sum_{o}^{O} X_{od} \tag{28}$$

As before, diversity is calculated as the herfindahl index (HHI), which is a nonlinear function of X_{od}.

$$Div_d = \sum_{o}^{O} \left(\frac{X_{od}}{L_d} \right)^2 \tag{29}$$

These aggregate components map into a city-level outcome Y_d following:

$$Y_d = \beta_1 log(L_d) + \beta_2 log(Div_d) + \beta_3 log(U_d) + \epsilon_{od} \tag{30}$$

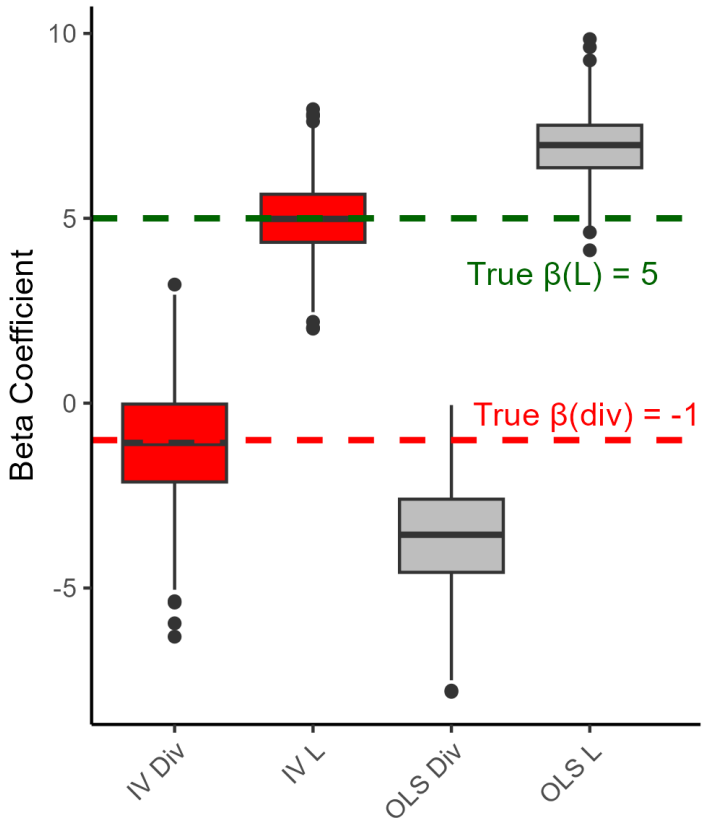
Each of these aggregates will introduce bias into the second stage equation via the unobserved feature or "attractiveness" of destinations d captured by U_d and V_d. Using 2SLS I predict L_d and Div_d using just the Z_{od} components, which are used for instruments in a zero stage as follows:

$$\widehat{X_{od}} = \alpha_3 Z_{od} + \epsilon_{od} \tag{31}$$

Using the predicted $\widehat{X_{od}}$ from this equation, I then instrument for the aggregates L_d and Div_d by replacing the real X_od flows with the predicted values. Figure A1 shows the results comparing OLS and SSIV beta coefficients in a model with 500 regions, 50 homelands, and a true β_1 of 5, and a true β_2 of -1. I also set α_1 and α_2 to 0.5, which sets the strength of the instrument Z_d relative to bias U_d. The correlation between L_d and Div_d is contorlled by parameters $\theta = 1, \psi = 2$.

Using two instruments requires that the instruments should not be linearly correlated with each other. In practice, L_d and Div_d may be correlated, as more attractive destinations also draw in a wider array of migrant groups. To explore how the beta estimator changes as the relationship between the instruments becomes stronger, I alter the model to create variation in the amount of correlation between L_d and Div_d. I draw θ and ψ from a gamma distribution such that $\theta \sim$ Gamma$(60, 10)$ and $\psi \sim$ Gamma$(200, 50)$. I also alter the weighting vector such that W is $[1, 2, 5, 0, 0, 0...]$. Figure A2 shows the distribution of beta coefficients for the SSIV model of the diversity variable. The boxplots

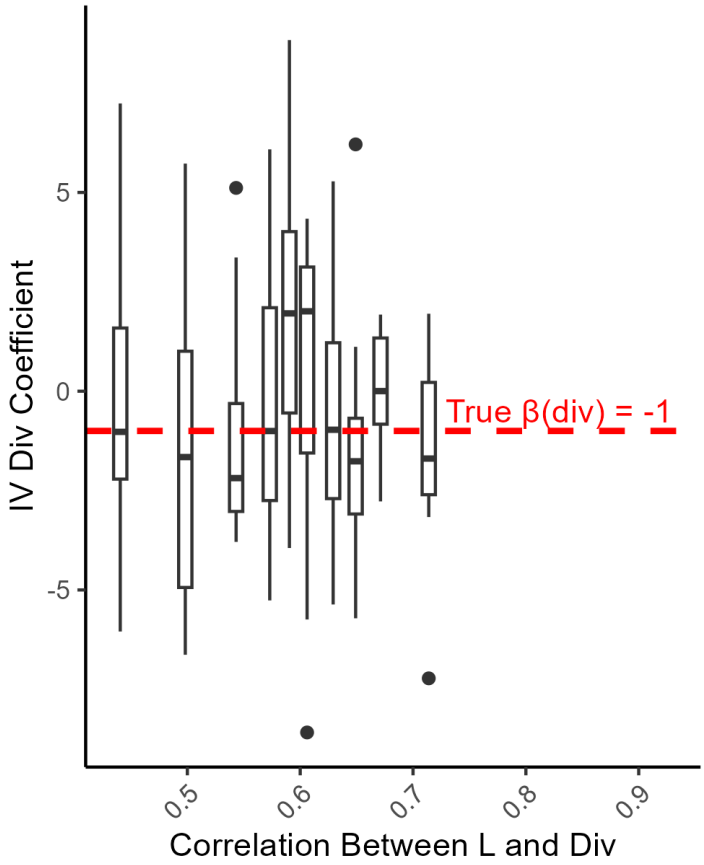
Figure A1: Simulation of OLS and SSIV with Ommitted Variable Bias



Notes: This figure shows the distribution of beta coefficients for migrant labor size L_d and migrant diversity Div_d, using both SSIV and OLS. The simulation is run 500 times, and the dotted lines show the real values for the beta coefficients, which are 5 for labor and -1 for diversity. The OLS model regresses $Y_d = \beta_1 log(L_d) + \beta_2 log(Div_d)$, while the SSIV uses aggregates calculated from predicted X_{od} flows. 500 destinations and 50 origins are included. The parameters used are $\alpha_1 = \alpha_2 = 0.5$, $\theta = 1$, $\psi = 2$. The weighting vector W is $[1, 0, 0...]$.

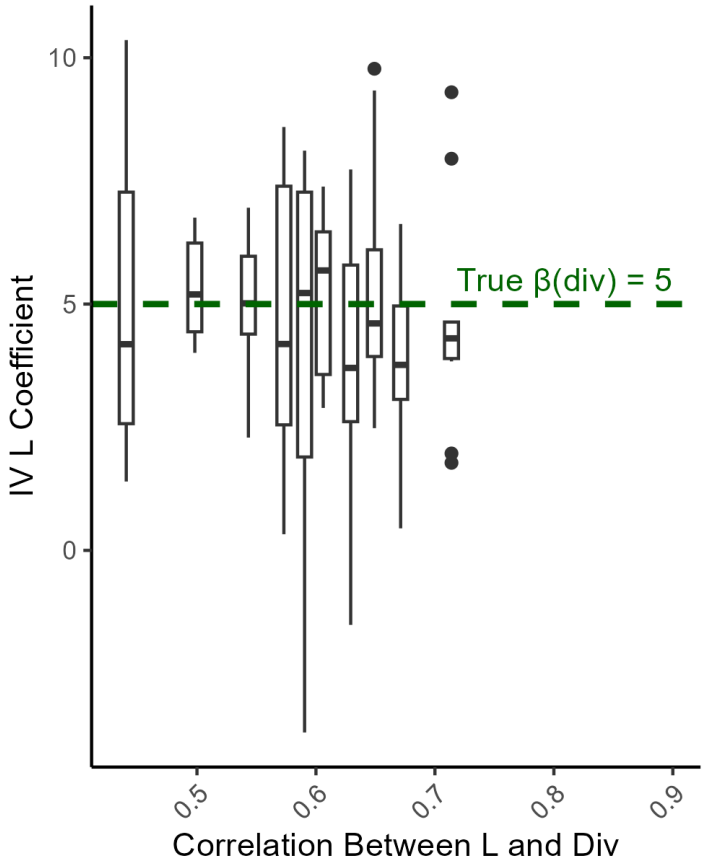
are separated into quantiles of the absolute value of the correlation between L_d and Div_d. As the correlation between the two aggregate variables moves towards 1 the accuracy of the SSIV estimator is reduced. Figure A3 shows the same result for the labor beta coefficient. Figre A4 plots how the F-statistic for the diversity instrument declines as the correlation between L_d and Div_d increases. Moving the linear correlation between the instruments from 0.5 to 0.7 reduces the F-statistic by nearly half.

Figure A2: SSIV Diversity Beta Coefficients with Varying Instrument Correlation



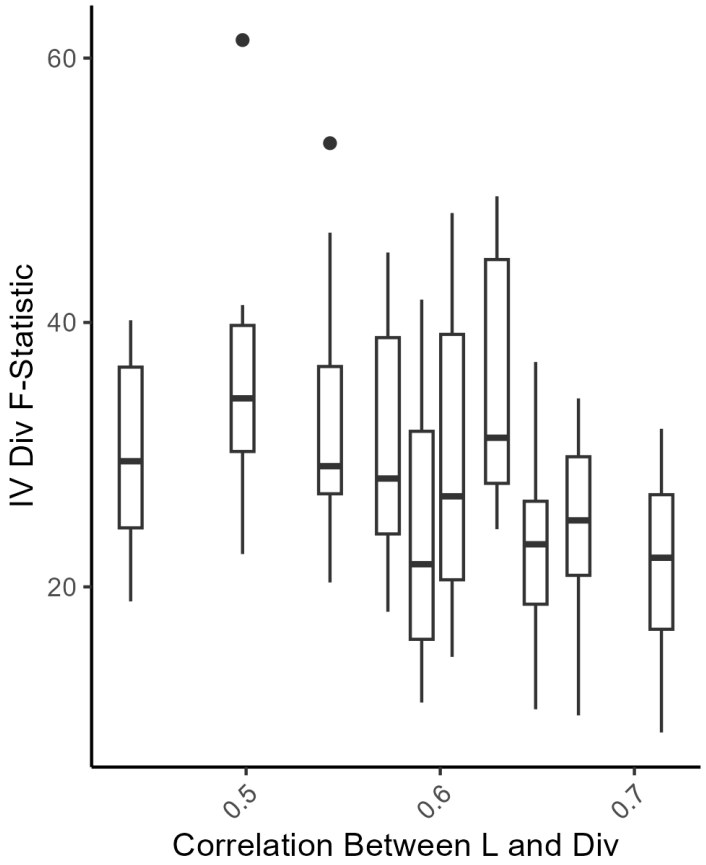
Notes: This figure shows the distribution of beta coefficients for migrant diversity Div_d, using SSIV as the correlation between the linear and nonlinear instrument varies. Box plots are grouped by quantiles of the absolute correlation coefficient between L_d and Div_d. The average of the absolute value of the correlation coefficient is shown on the x-axis. The simulation is run 500 times, and the dotted lines show the real values for the beta coefficient. The SSIV model regresses $Y_d = \beta_1 log(L_d) + \beta_2 log(Div_d)$, where the aggregates calculated from predicted X_{od} flows. 500 destinations and 50 origins are included. The parameters used are $\alpha_1 = \alpha_2 = 0.5$. The correlation between L_d and Div_d is governed by $\theta \sim$ Gamma$(60, 10)$ and $\psi \sim$ Gamma$(200, 50)$. The weighting vector W is $[1, 2, 5, 0, 0, 0...]$.

Figure A3: SSIV Labor Beta Coefficients with Varying Instrument Correlation



Notes: This figure shows the distribution of beta coefficients for migrant labor size L_d, using SSIV as the correlation between the linear and nonlinear instrument varies. Box plots are grouped by quantiles of the absolute correlation coefficient between L_d and Div_d. The average of the absolute value of the correlation coefficient is shown on the x-axis. The simulation is run 500 times, and the dotted lines show the real values for the beta coefficient. The SSIV model regresses $Y_d = \beta_1 log(L_d) + \beta_2 log(Div_d)$, where the aggregates calculated from predicted X_{od} flows. 500 destinations and 50 origins are included. The parameters used are $\alpha_1 = \alpha_2 = 0.5$. The correlation between L_d and Div_d is governed by $\theta \sim$ Gamma$(60, 10)$ and $\psi \sim$ Gamma$(200, 50)$. The weighting vector W is $[1, 2, 5, 0, 0, 0...]$.

Figure A4: F-Statistic by Varying Instrument Correlation



Notes: This figure shows the distribution of F-statistics for the instrument of migrant diversity Div_d, using SSIV as the correlation between the linear and nonlinear instrument varies. Box plots are grouped by quantiles of the absolute correlation coefficient between L_d and Div_d. The average of the absolute value of the correlation coefficient is shown on the x-axis. The simulation is run 500 times, and the dotted lines show the real values for the beta coefficient. The SSIV model regresses $Y_d = \beta_1 log(L_d) + \beta_2 log(Div_d)$, where the aggregates calculated from predicted X_{od} flows. 500 destinations and 50 origins are included. The parameters used are $\alpha_1 = \alpha_2 = 0.5$. The correlation between L_d and Div_d is governed by $\theta \sim \text{Gamma}(60, 10)$ and $\psi \sim \text{Gamma}(200, 50)$. The weighting vector W is $[1, 2, 5, 0, 0, 0...]$.

B Additional Tables & Figures

B.1 Supplementary Analysis

The following section presents additional tables and figures that are referenced in the text.

B.2 Birthplace-Level Tests Following Borusyak et al. (2022)

In this section, I conduct the robustness checks of the shift-share model suggested in Borusyak et al. (2022). The authors recommend estimating an IV coefficient from a birthplace-level regression, relating exposure-weighted outcome residuals on exposure-weighted treatment residuals. The procedure accounts for a single instrument model, so I proceed by estimating a labor-only version of the model. The purpose of this exercise is to analyze the identifying shock-level variation, so while the coefficients do not match our headline estimates due to the exclusion of the diversity parameter, they are informative about our shocks g_{ot}.

When generating the birthplace level data, I consider the standardized change in migrant labor size as the treatment, while the shocks are the growth rates at the birthplace level g_{ot}, and the instrument Z is the predicted labor change according to the shift-share. I first plot our residualized average treatment residual across destinations against the birthplace growth shocks in a binned scatterplot. Figure B1 shows a strong first stage relationship between the growth shocks and the average treatment residual. Next I plot the average outcome residual against the same binned birthplace shocks. These figures, by outcome, are shown in Figure B3. Again, I see a clear positive relationship, as higher level shocks relate to higher residualized outcomes. The benefit of these scatterplots is that we can visualize the distribution of the treatment effect across the distribution of birthplace growth shocks. We see evidence of nonlinearity, with paritcularly high growth shocks driving a disproportional amount of the variation in our treatment.

Next we calculate adjusted standard errors, as recommended in Borusyak et al. (2022), which are standard errors in a birth-place IV regression of outcome on treatment, weighted by exposure shares. Figure B5 plots the normal clustered robust standard errors from the IV regression, along with the adjusted standard errors. The gray bars represent confidence intervals. As expected, the standard errors are higher in the adjusted version, but the coefficients remain significant.

The last use of the birthplace-level dataset is to test correlations between the shocks and baseline characteristics of destinations, weighted by exposure. Our exclusion restriction requires that there is no significant correlated between pre-period characteristics and the birthplace shocks. To test this, we regress a set of baseline characteristics, averaged across destinations and weighted by exposure shares,

Table B1: African Censuses with Geolocation, Birthplace or Ethnicity

Country	Available Census (ADM2)	Birthplace	Ethnicity	Mother Tongue
Benin	1979 1992 2002 2013	1979 1992 2002 2013	1979 1992 2002 2013	2013
Botswana	1991 2001 2011			
Burkina Faso	1985 1996 2006	1985 1996 2006		2006
Cameroon	1976 1987 2005	1976 1987 2005		
Côte d'Ivoire	1988 1998	1988 1998	1988 1998	
Ethiopia	1984 1994 2007			1994 2007
Guinea	1983 1996 2014	1983 1996 2014		
Ghana	1984 2000 2010		2000 2010	
Kenya	1969 1979 1989 1999 2009 2019	1969 1979 1989 1999 2009 2019		
Malawi	1987 1998 2008 2018		2008 2018	
Mali	1987 1998 2009	1987 1998 2009		1987 1998 2009
Mozambique	1997 2007 2017	1997 2007 2017		2007 2017
Rwanda	1991 2002 2012	2002 2012		
Senegal	1988 2002 2013		1988 2002 2013	1988 2002 2013
Sierra Leone	2004 2015	2004 2015	2004 2015	2004 2015
South Africa	1996 2001 2011 2016		1996 2001 2011 2016	
Tanzania	1988 2002 2012	1988 2002 2012		
Togo	1960 1970 2010		1960 1970 2010	
Uganda	1991 2002 2014	1991 2002 2014	1991 2002 2014	
Zambia	1990 2000 2010	1990 2000 2010	1990 2000 2010	

Table B2: Gravity Regressions of Migrant Flow using Proxy O-D Panels

	N	N	N
$ln(distance)$	-1.359	0.163	0.312
	[0.017]***	[0.093]*	[0.070]***
$ln(distance^2)$		-0.155	-0.144
		[0.009]***	[0.007]***
Coethnic Share			8.634
			[0.511]***
Mean Dep. Var	65.753	65.753	65.753
Observations	83,661	83,661	83,661
Destination FE	Y	Y	Y
Origin FE	Y	Y	Y
Year FE	Y	Y	Y

Notes: This table shows results from a gravity regression $\pi_{odt} = exp(\mu_{dt} + \gamma_{ot} + \beta_1 CoethShare_{odt} + \beta_2 Dist_{od} + \beta_3 Home_{od}) + \epsilon_{odt}$. Where π_{odt} is an estimated probability defined as the fraction of individuals from o that appear in destination d at time t. That is, $\pi_{odt} = \frac{M_{odt}}{L_{ot}}$ where $_{odt}$ is the number of people from o in d. The regression includes destination-year and origin-year fixed effects, μdt and γ_{ot} respectively. Regressions are estimated using Poisson pseudo-likelihood (PPML). Distance is calculated as the log kilometer distance between the centroids of origin and destination administrative regions.

against the birthplace-level shocks. The characteristics include pre-period light density measured in 1993, as well as geographic characteristics including soil suitability, TseTse fly suitability, and malaria suitability. The data sources for these variables are enumerated in the data section of the paper. Figure B7 plots the beta coefficients from this regression, run separately for each outcome of interest related to light density.

B.3 Robustness Checks

In this section we present a series of figures that show robustness for our main shift-share specification. We perform a variety of robustness checks for our results. Figure B9 and Figure B10 show robustness to Conley spatial standard errors across a variety of distance bandwidths. Figure B11 and B12 show how our estimates vary when dropping individual countries from the sample. Last we simulate a series of random outmigration shocks and apply them in our shift-share design. Figures B13 and B14 show how the IV results appear in response to placebo shocks.

B.4 Pass-through of International Commodity Prices

Our push shock instrument leverages origin-specific shocks to predict outmigration rates. These predicted outmigration shifts are then used as an instrument to estimate the impact of migration on destination outcomes. We might be concerned that these push shocks are not only correlated across space, but also affect productivity in destinations through other channels. For instance, if a price shock hits an origin area and affects the trade of crops to a port destination, this may be realized in light or wealth growth measures, unrelated to the price shock effects on migration. International price shocks may also drive up the cost of food in destination areas, changing wealth and labor supply in destination unrelated to the migration channel.

In this section, we study the impact of plausibly exogenous push shocks on local price behavior in order to test how shocks may create differences in the relative attractiveness of urban and rural locations. In particular we estiamte impulse response functions from local projections that show the impact of international price changes on local urban/rural price dispersion. If international price shocks create wedges in urban/rural prices for affected commodities, this may be evidence of direct effects on destinations and a violation of the exclusion restriction.

Our data on local prices is gathered at the crop-month level from the Famine Early Warning system (FEWS NET) and World Food Program (VAM) (Porteous, 2019). These datasets provide crop-month price observations at a set of geolocated markets acrosss African countries. We code

Table B3: Census Birthplace Shift-Share Second Stage, Migrant Flows

Panel A: OLS Results

	Log(Lights)	Δ Services Share	Housing Quality
Δ ℓ	1.036	1.021	0.049
	[0.113]***	[1.218]	[0.039]
Δ div	-9.999	28.617	0.553
	[1.632]***	[17.441]	[0.360]
Mean Dep.	-2.49	5.01	-0.17
Observations	829	595	822

Panel B: Shift-Share IV

	Log(Lights)	Δ Services Share	Housing Quality
Δ ℓ	2.529	1.305	0.377
	[0.228]***	[1.892]	[0.096]***
Δ div	-10.425	10.640	2.612
	[2.979]***	[4.442]**	[0.850]***
Mean Dep.	-2.49	5.01	-0.17
Observations	829	595	822
Kleibergen-Paap Fstat	16.462	12.581	53.053
Sanderson-Windmeijer ℓ	245.593	161.907	183.079
Sanderson-Windmeijer div	27.686	23.125	27.438

Note: Panel A presents the OLS estimation of changes in migrant labor and diversity on wealth and urbanization outcomes. $y_{d,t} = \alpha\Delta l_{d,t} + \gamma\Delta div_{d,t} + \upsilon_{tc}$. Panel B presents the IV 2SLS estimation of changes in migrant labor and diversity on wealth and urbanization outcomes. Outcomes include the levels of logged light density, change in the labor share in services, and the level of housing quality. All regressions include country-year fixed effects. * p<0.01, ** p<0.05, *** p<0.01.

markets as urban or rural based on the local population density measured by Worldpop. We then create a urban-rural price gap as the difference in prices for a crop-month across urban and rural markets in the same country. We then estimate the following local projection. For crop c in state s at month t we estimate:

$$Urban - Rural_{cs,t+h} = \omega International_{ct} + \mu_s + \gamma_c + \epsilon_{cs,t+h} \tag{32}$$

Where $Urban - Rural_{cs,t+h}$ is the MoM change in the urban-rural price difference in crop c at horizon h months from t. The variable $International_{ct}$ captures the MoM change in the international price of crop c at month t, and crop and state fixed effects are included. Figure B15 shows the result of this projection over a 12 month horizon. We don't see evidence of a systematic effect of shocks to MoM international price changes on crop specific urban-rural price differences.

Table B4: Census Birthplace Shift-Share Second Stage, Migrant Flows

	Shift-Share IV		
	Δ Industry HHI	Δ Avg. Human Capital	Δ Industry Segregation
Δ ℓ	-0.038	-1.297	-1.919
	[0.016]**	[0.290]***	[1.682]
Δ div	-0.524	1.817	-1.329
	[0.175]***	[3.346]	[6.515]
Mean Dep.	-0.06	1.59	1.36
Observations	506	529	506
Kleibergen-Paap Fstat	12.380	4.334	12.380
Sanderson-Windmeijer ℓ	122.771	94.947	122.771
Sanderson-Windmeijer div	22.630	4.739	22.630

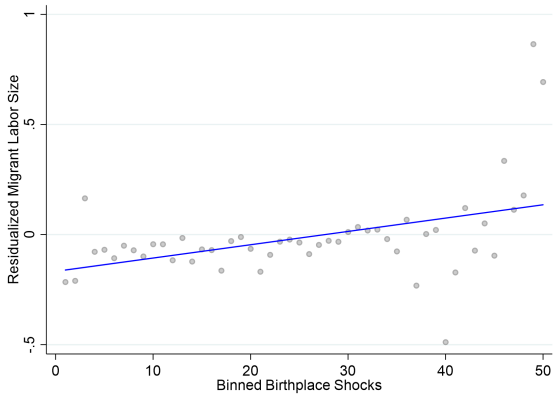
Note: This table presents the SSIV estimation of changes in migrant labor and diversity on skill and industry concentration outcomes.. $y_{d,t} = \alpha\Delta l_{d,t} + \gamma\Delta div_{d,t} + \upsilon_{tc}$. Industry concentration is measured as the change in the labor share HHI across general industry categories such as construction, mineral production, agriculture, manufacturing and retail trade. Segregation is a measure of the deviation from random of the distribution of migrants from particular origins to particular industries. Human capital changes are measured as the change in the average skill level of workers in a given destination over time, measured in terms of schooling years. All regressions include country-year fixed effects. * p<0.01, ** p<0.05, *** p<0.01.

Table B5: Self-Reported Identity by Migrant Status and Distance

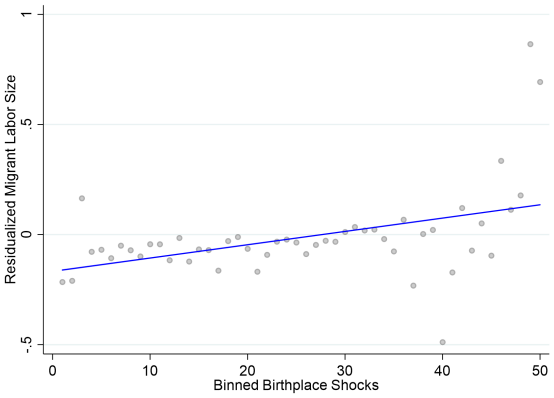
	Within Ethnicity		**Within Destination**		**Migrants Only**	
	Nationalism	Expectations	Nationalism	Expectations	Nationalism	Expectations
Migrant==1	0.166	0.188	0.201	0.089		
	[0.238]	[0.131]	[0.237]	[0.096]		
Migrant*Population	-0.038	-0.016	-0.034	-0.018		
	[0.034]	[0.016]	[0.036]	[0.016]		
Migrant*Distance	0.011	-0.028	0.011	-0.006		
	[0.033]	[0.020]	[0.027]	[0.020]		
Migrant*CoethnicShare	-0.054	0.051	-0.131	0.066		
	[0.086]	[0.070]	[0.087]	[0.061]		
ln(O-D Distance km)	0.038	0.014	0.026	0.002	0.046	-0.012
	[0.026]	[0.016]	[0.021]	[0.019]	[0.028]	[0.018]
ln(Population)	0.040	-0.008				
	[0.026]	[0.025]				
Coethnic Share	0.048	-0.074	0.104	-0.079	0.067	-0.014
	[0.079]	[0.089]	[0.078]	[0.053]	[0.066]	[0.052]
Mean Dep. Var	3.539	2.809	3.539	2.809	3.625	2.827
Observations	41,033	41,022	41,030	41,021	18,210	18,241
Destination FE	N	N	Y	Y	Y	Y
Ethnicity FE	Y	Y	N	N	N	N
Migrant Only	N	N	N	N	Y	Y

Notes: The data comes from Afrobarometer surveys, linking individuals to their origin based on reported ethnicity. The outcomes of interest are national identity and economic expectations. National identiy, or "nationalism" is an individual's response to a question about how much the individual identifies with the nation relative to their ethnic group. Future expectations is a question styled after the Michigan Consumer Confidence Survey, which asks individuals if they expect economic conditions to improve in the next year. All regressions include country-year fixed effects. Standard errors are clustered at the Afrobarometer sampling cluster level. * p<0.01, ** p<0.05, *** p<0.01.

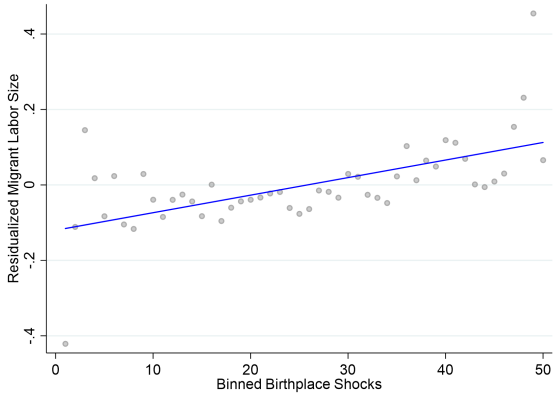
Figure B1: First Stage Scatterplots in a Birthplace-Level Regression



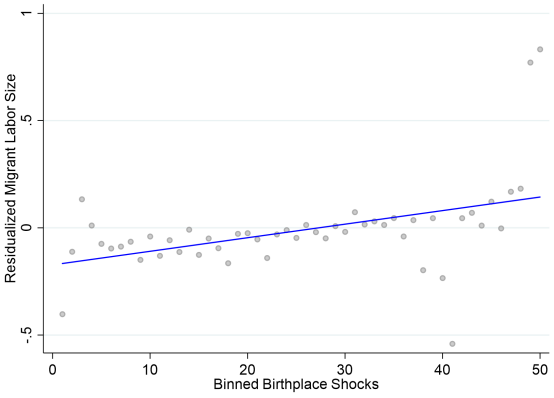
(a) Δ Lights



(b) Δ Log(Lights/Capita)



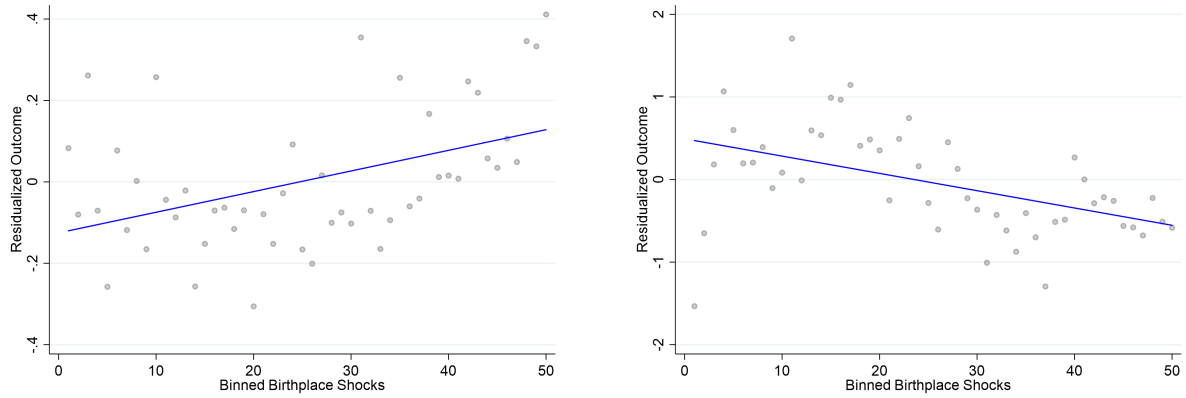
(c) Δ Non-Agricultural Share



(d) Δ House Quality

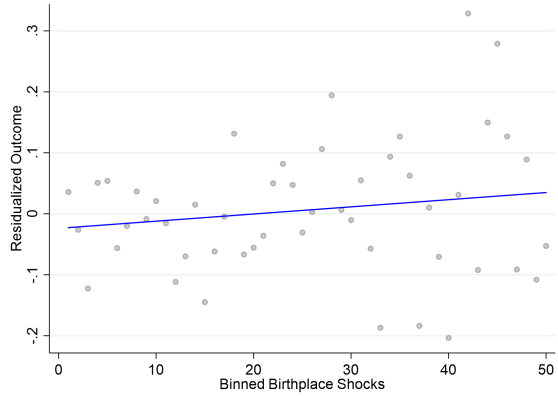
Notes: This figure shows binned scatterplots of birthplace-level treatment residuals against the birthplace-level shocks, organized in 50 bins. The OLS lines of best fit is shown in red. The residualized procedure is accomplished using the method explained in (Borusyak et al., 2022).

Figure B3: Reduced Form Scatterplots in a Birthplace-Level Regression

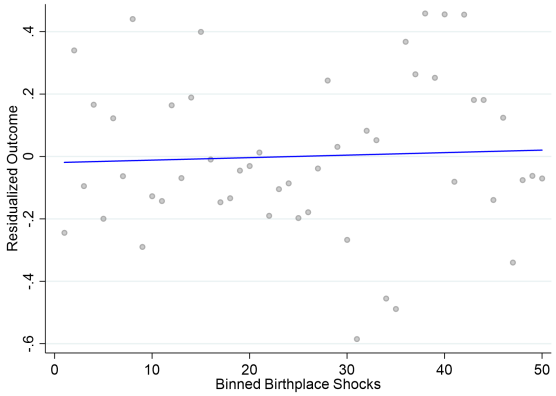


(a) Δ Lights

(b) Δ Log(Lights/Capita)



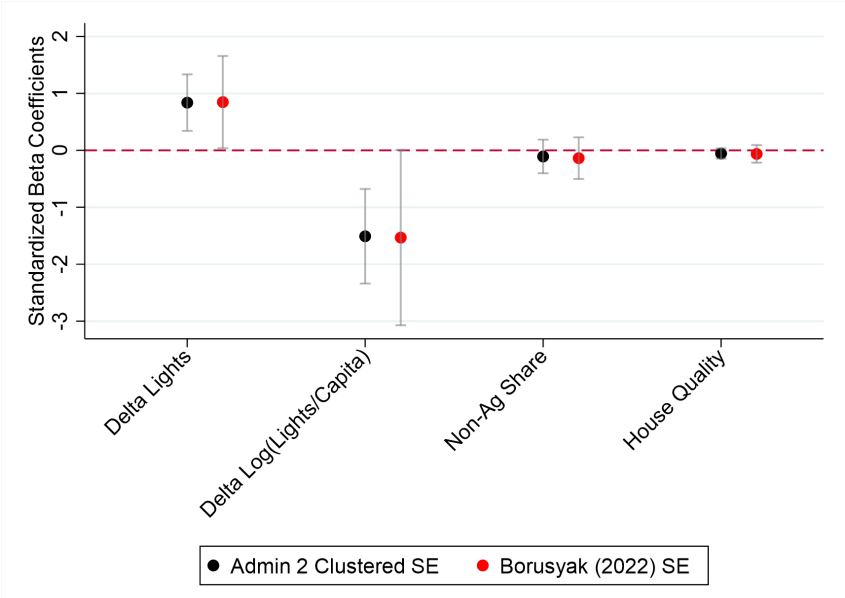
(c) Δ Non-Agricultural Share



(d) Δ House Quality

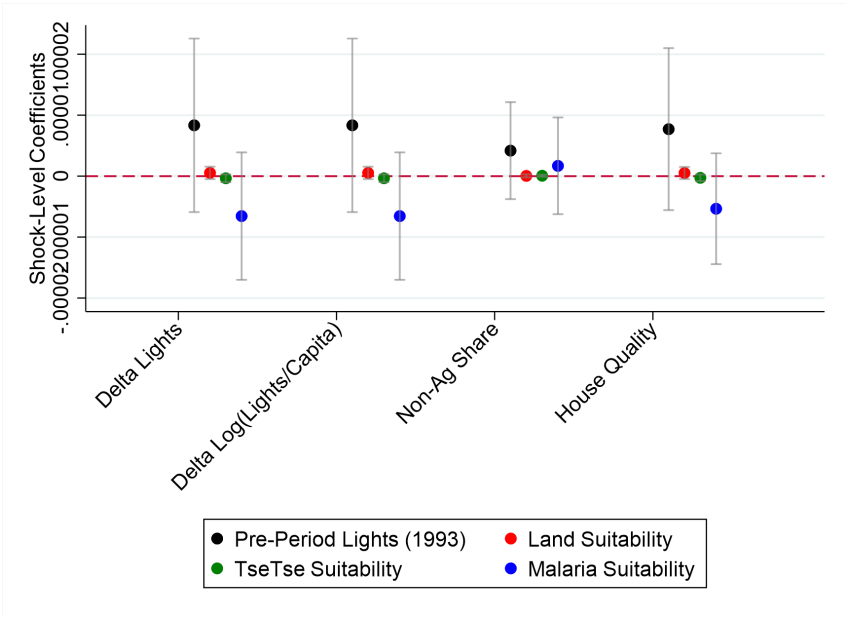
Notes: This figure shows binned scatterplots of birthplace-level treatment residuals against the birthplace-level shocks, organized in 50 bins. The OLS lines of best fit is shown in red. The residualized procedure is accomplished using the method explained in (Borusyak et al., 2022).

Figure B5: Standard Error Adjustments Following (Borusyak et al., 2022)



Notes: This figure shows beta coefficients and confidence intervals for both the normal IV regression used in the main tables, and the adjusted standard errors recommended by Borusyak and co-authors. The procedure is done separately for each light outcome of interest. The treatment are in standardized units of the raw changes in migrant labor size.

Figure B7: Balance at Baseline in Birthplace-Level Regressions



Notes: This figure shows beta coefficients from regressions of average residualized destination characteristics on birthplace shocks. The baseline destination characteristics used are pre-period light density (measured in 1993, prior to treatment for all groups), soild suitability, TseTse fly suitability, malaria suitability. Sources for these variables can be found in the data section on "Pull Characteristics". The shocks are enumerated in raw population counts.

Table B6: OLS Cross-Section of Historical Diversity and Urban Density

Panel A: Urban Dummy

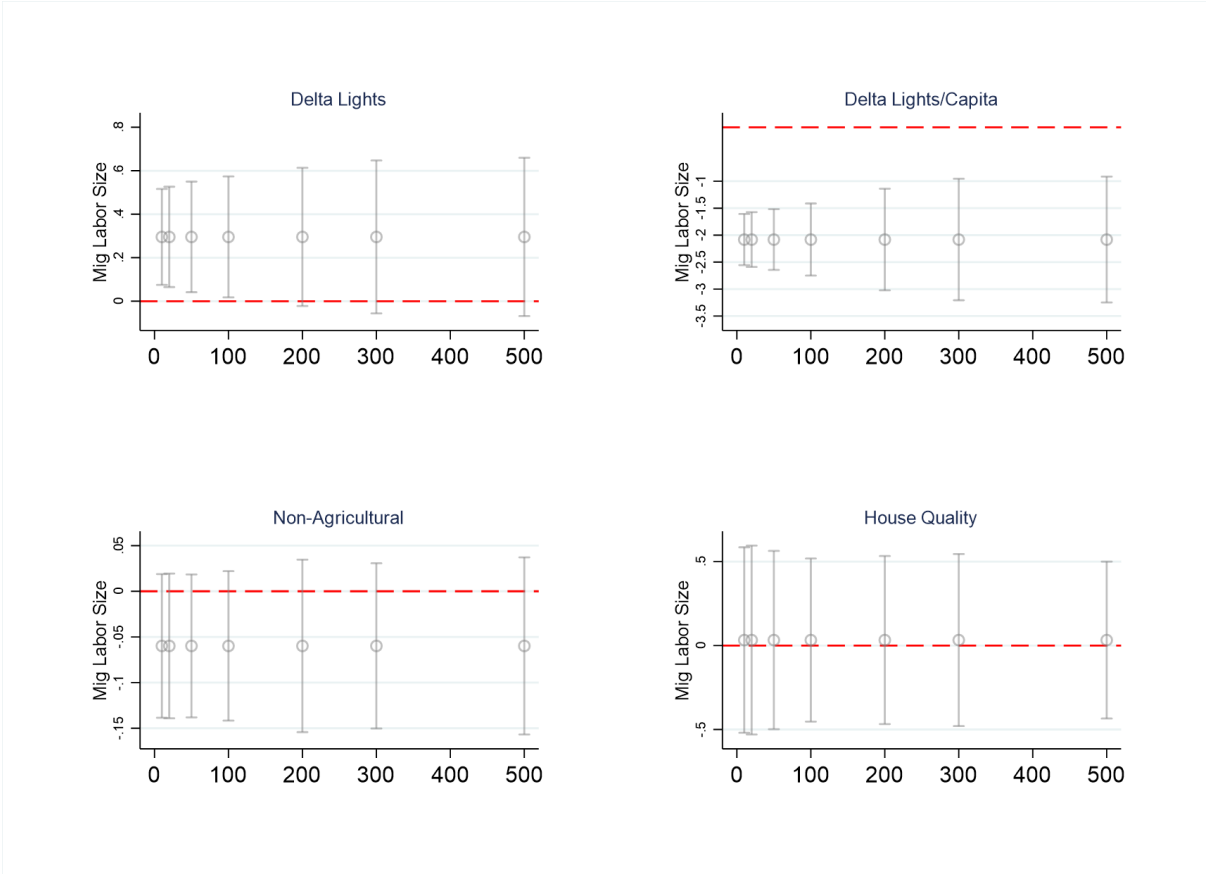
	(1) Light Density	(2) Lights/Capita	(3) Growth Lights/Capita	(4) Conflict
City	1.548	-0.465	-0.383	45.330
	[0.017]***	[0.047]***	[0.056]***	[2.153]***
Diversity	0.043	-0.542	0.200	3.718
	[0.023]*	[0.065]***	[0.079]**	[3.006]
City*Diversity	-0.720	0.460	0.030	3.560
	[0.071]***	[0.198]**	[0.240]	[9.159]
Mean Dep.	-2.00	-9.96	-1.89	7.81
Observations	28,656	28,654	28,654	28,656

Panel B: Urban Population

	Light Density	Lights/Capita	Growth Lights/Capita	Conflict
2010 Population	0.136	-0.864	-0.528	3.071
	[0.002]***	[0.002]***	[0.006]***	[0.265]***
Diversity	-0.478	-0.478	0.119	-33.902
	[0.092]***	[0.092]***	[0.250]	[10.852]***
Population*Diversity	0.043	0.043	0.037	4.164
	[0.010]***	[0.010]***	[0.028]	[1.196]***
Mean Dep.	-2.00	-9.96	-1.89	7.81
Observations	28,654	28,654	28,654	28,654

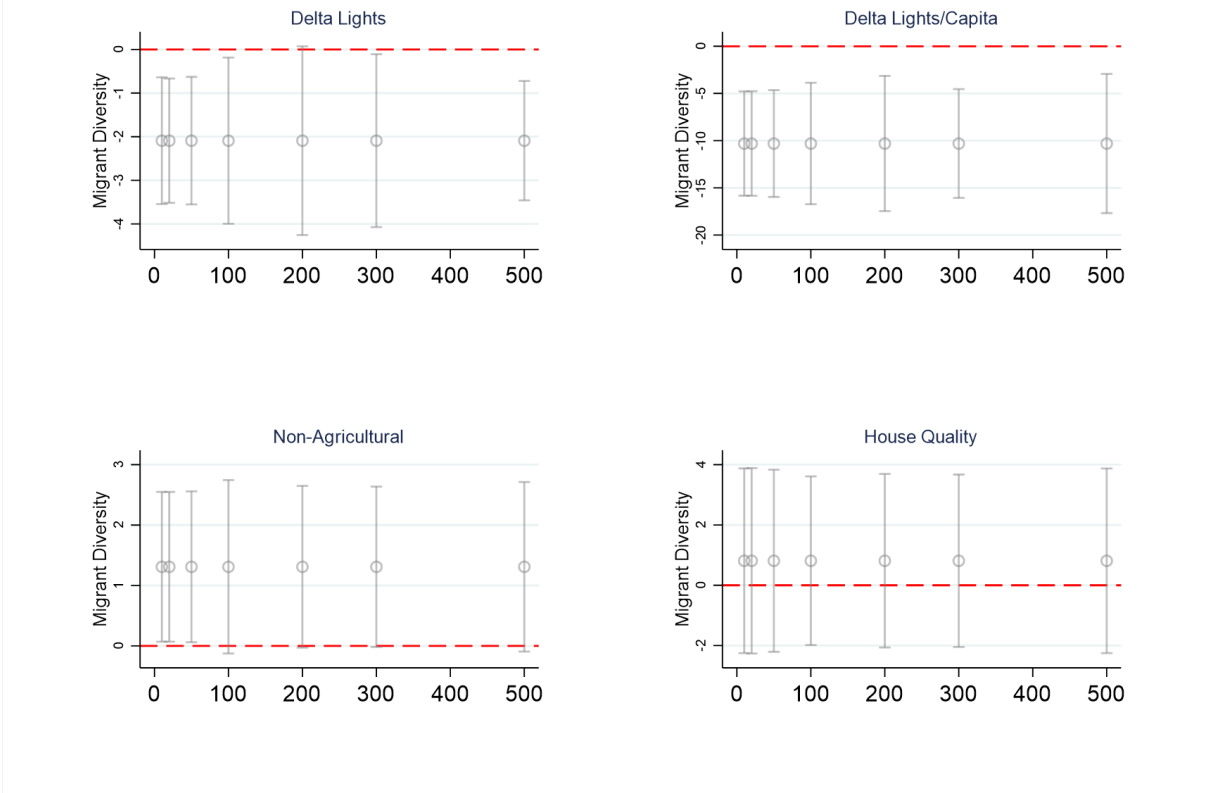
Notes: This table presents cross-sectional regressions of urban growth and the interaction of urban growth and historical diversity on contemporary productivity outcomes. All regressions include state fixed effects. Light density outcomes are calculated in 2013, for comparison to Montalvo and Reynal-Querol (2021). Column 1 measures log light density, column 2 is a log measure of light density over a Worldpop estimated population for the grid in 2010, column 3 measures the change in log lights/capita from 1992 to 2013. Column 4 measures the number of conflict events in the grid since 1997, measured in ACLED battle events. "City" in Panel A is an indicator marked as 1 if an Africapolis city is located within the grid and the population is above 20,000. In Panel B, this term is replaced by a log population estimate for 2010 from Worldpop. "Diversity" is a historical measure of diversity calculated as the fractionalization of land share of different Murdock ethnic groups in the grid cell. All regressions control for distance to coast, malaria and TseTse suitability, ruggedness, distance to a major river, agricultural land productivity and a historical estimate of population size in 1800. * p<0.01, ** p<0.05, *** p<0.01.

Figure B9: Conley Spatial Standard Errors for Labor Size Coefficient $\Delta\ell$



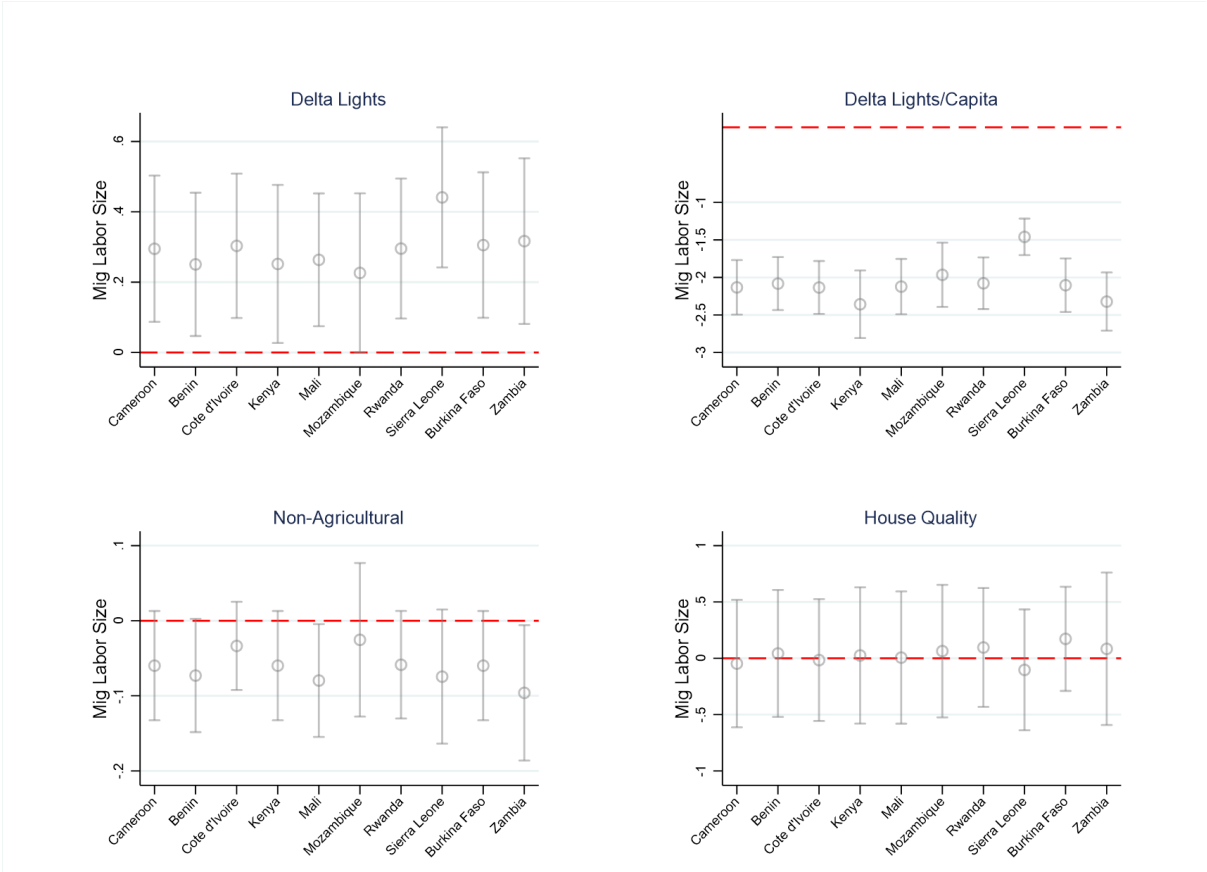
Notes: This figure plots the results of IV regressions of the main estimating equation where standard errors account for spatial autocorrelation. In particular the figure plots conley standard errors for the migrant labor size coefficient with different bounds on the decay of spatial autocorrelation.

Figure B10: Conley Spatial Standard Errors for Migrant Diversity Coefficient Δdiv



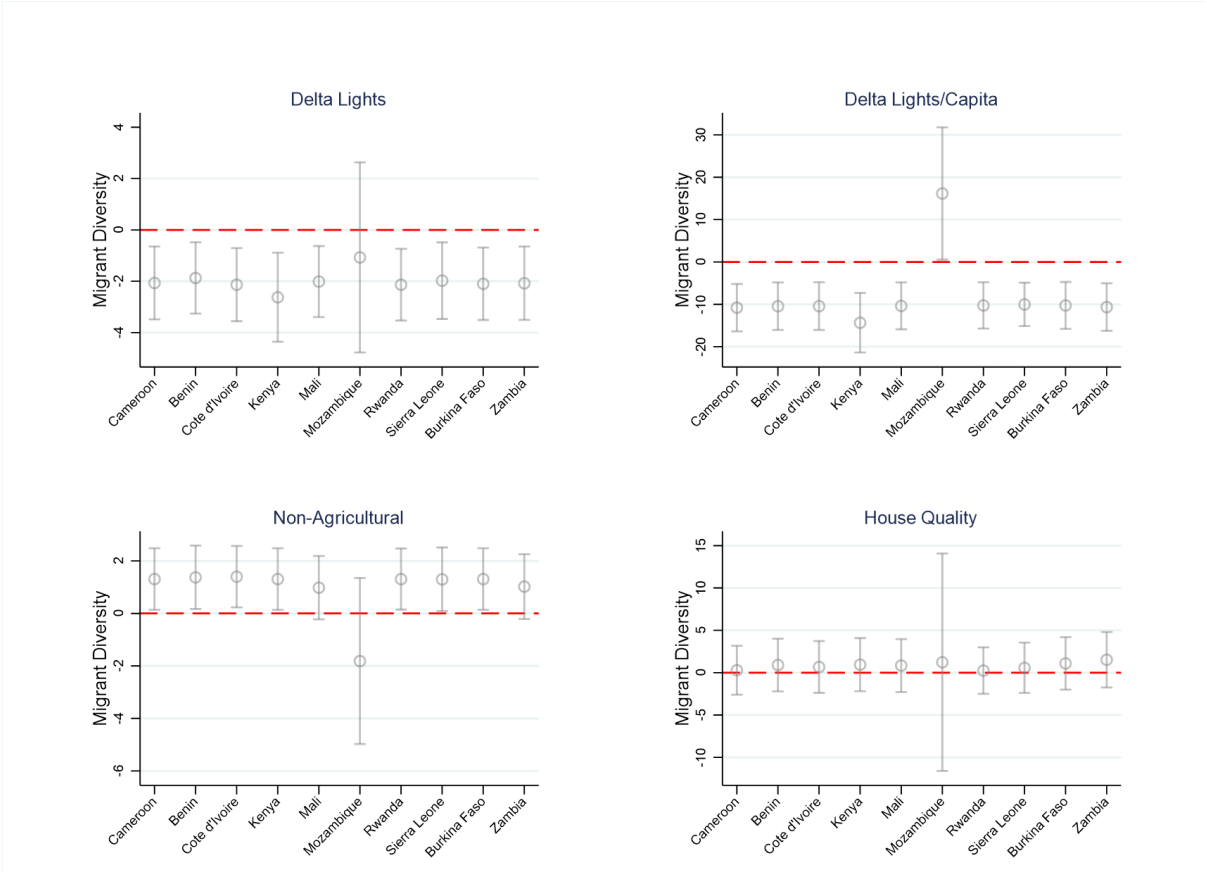
Notes: This figure plots the results of IV regressions of the main estimating equation where standard errors account for spatial autocorrelation. In particular the figure plots conley standard errors for the diversity coefficient with different bounds on the decay of spatial autocorrelation.

Figure B11: Estimates with Individual Country Drops for Labor Size Coefficient $\Delta\ell$



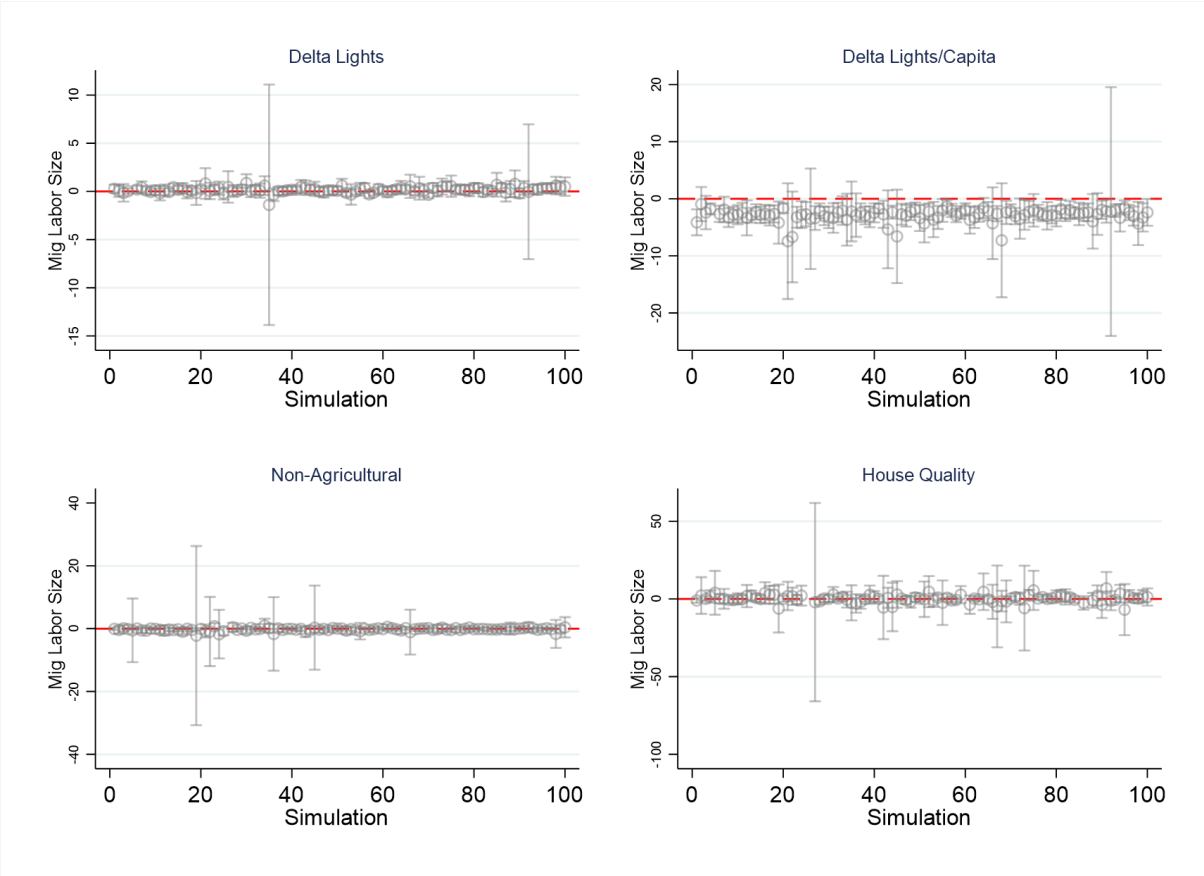
Notes: This figure plots the results of IV regressions of the main estimating equation where individual countries are dropped from the sample. This figure shows results for the migrant labor size coefficient.

Figure B12: Estimates with Individual Country Drops for Migrant Diversity Coefficient Δdiv



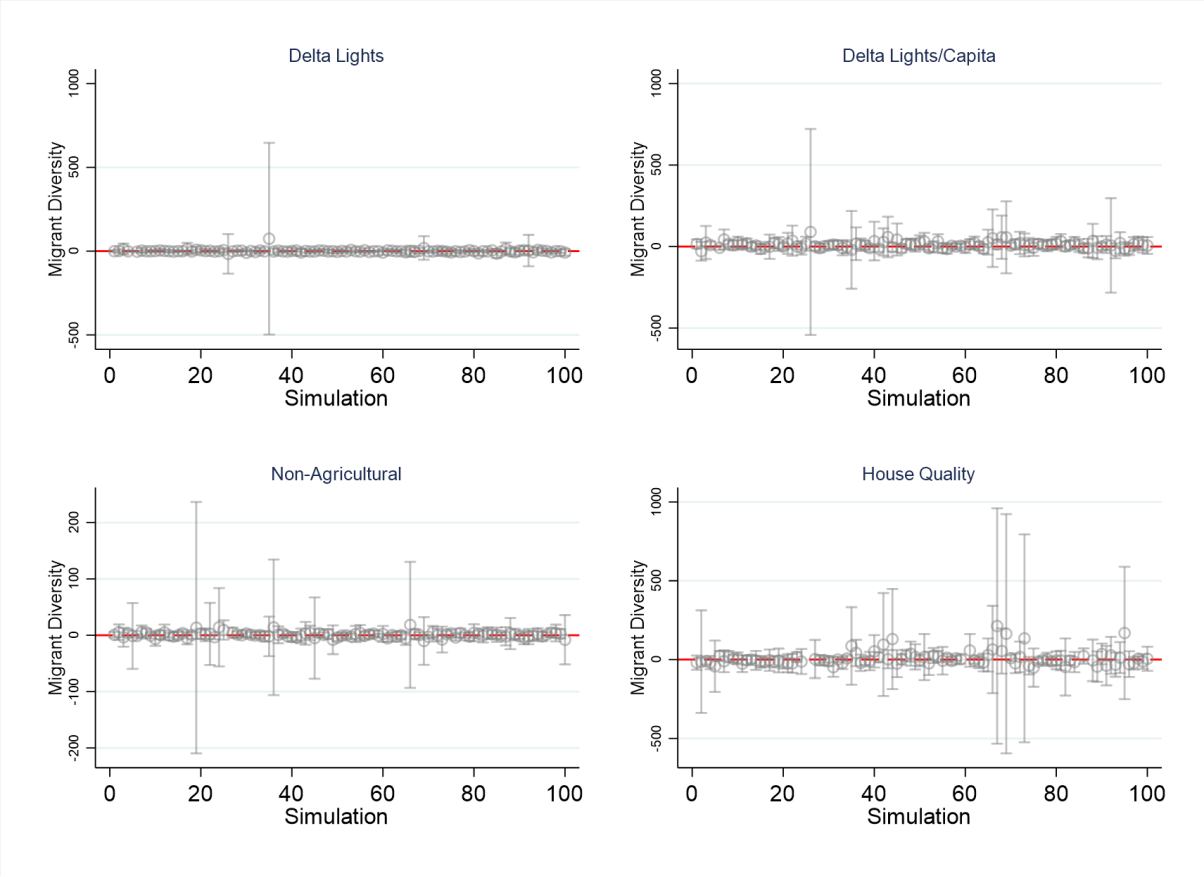
Notes: This figure plots the results of IV regressions of the main estimating equation where individual countries are dropped from the sample. This figure shows results for the migrant diversity coefficient.

Figure B13: Placebo Shocks for Labor Size Coefficient $\Delta\ell$



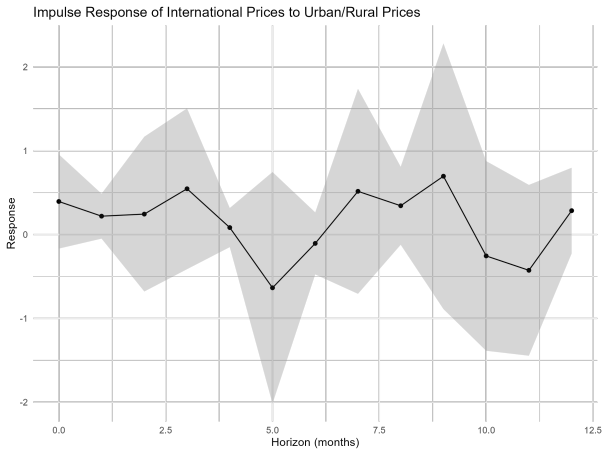
Notes: This figure plots the results of IV regressions of the main estimating equation where g_{ot} is replaced with randomly drawn shocks. This figure shows results for the migrant labor size coefficient. The x-axis tracks individual draws of the placebo shocks.

Figure B14: Placebo Shocks for Migrant Diversity Coefficient Δdiv



Notes: This figure plots the results of IV regressions of the main estimating equation where g_{ot} is replaced with randomly drawn shocks. This figure shows results for the diversity coefficient. The x-axis tracks individual draws of the placebo shocks.

Figure B15: IRF of International Prices on Local Urban/Rural Gap



Notes: This figure shows the results of a local project estimating the effect of an international price shock on the urban-rural price gap for the same commodity across African markets.