The Persistent Legacy: Colonial Railroad and Agricultural Productivity in Africa

Andu Berha

February 14, 2025

Abstract

This paper examines the long-term impact of colonial-era railroad infrastructure on agricultural productivity across 24 sub-Saharan African countries. Leveraging multiple identification strategies, including comparisons with placebo lines, spatial first differences, and spatial discontinuity designs, we find robust evidence that railroad localities exhibit significantly higher crop yields in the present day. Our findings reveal that grid cells within 20 km of historical rail lines experience an agricultural productivity premium of up to 8% relative to unconnected cells. We argue the persistence arises from path dependence: early rail access enabled regions to shift into high-value commercial crops, fostering long-term specialization and agglomeration economies. Our results remain robust under a variety of specifications, including controls for geographic fundamentals, pre-colonial ethnic homeland characteristics, and post-colonial institutional differences. These findings shed light on how initial "big push" investments can lock regions into distinct spatial equilibrium path that persist for century. In turn, policymakers may either reinforce these legacy corridors or target under-invested areas to correct historical spatial inequality.

Key Words: Railroad, Productivity, Colonialism, Rural development, Africa

1 Introduction

In the early twentieth century, European colonial powers built railroads across Africa, traversing vast, sparsely populated regions to facilitate resource extraction and military fortification. While these infrastructures were built with immediate strategic and economic interests in mind, they reshaped local economies well beyond the colonial period (Jedwab et al., 2017). In line with this view, recent literature highlights the lasting impact of colonial roads and railways on present-day economic outcomes, including urban growth (Bertazzini, 2022; Jedwab et al., 2017; Jedwab and Moradi, 2016; Okoye et al., 2019), human capital development (Huillery, 2009; Okoye et al., 2019), and modern infrastructure expansion (Huillery, 2009).

Nevertheless, the vast majority of these studies focus on urban outcomes, sidestepping the question of what this historical footprint means for contemporary rural development, particularly agriculture, the primary livelihood for two-thirds of Africa's post-colonial population (Block, 2014). Jedwab and Moradi (2016) find that colonial railroad construction in Ghana led to the specialization of nearby regions in cocoa production, transitioning these areas from subsistence to market-oriented farming in early 20th century. However, with the collapse of railroads in the 1980s and 1990s, railroad regions lost their competitive advantage. Thus, a key unresolved question is whether previously connected areas remain specialized in high-value crops due to entrenched economic geographies, or whether they revert to a pre-colonial equilibrium, such as subsistence agriculture.

To address this gap, in this study we examine the persistent impact of colonial investment in railroads on current agricultural productivity in 24 sub-Saharan African countries. Specifically, we ask: do railroads built more than a century ago continue to influence agricultural performance, even though many lines have long ceased to function? If so, through what channels?

However, to rigorously answer these questions, our empirical approach must address a major identification challenge. More specifically, empirically isolating the causal effect of colonial period infrastructure from geographic, societal, and political factors is challenging because these factors not only interact in complex ways but are also difficult to observe and quantify. For instance, colonial investment decisions were often influenced by the perceived economic potential of certain regions, such as geographic fundamentals and the presence of extractive resources, making it hard to isolate the independent effect of rail development. To ensure we are capturing a causal relationship, we must account for any unobserved factors correlated with both railroad placement and subsequent outcomes. In pursuit of a credible causal estimate, our approach combines several identification strategies, including: (a) rich controls and fixed effects; (b) placebo line as good counterfactual; (c) spatial first difference (SFD) design; and (d) spatial discontinuity design.

Drawing on these methods, we find compelling and consistent evidence that colonial-era railroad infrastructure exerts a persistent positive influence on agricultural productivity in Africa. Using both traditional agricultural survey data and satellite-derived yield measures, we find that areas closer to historical railroads exhibit higher levels of agricultural productivity today. The productivity premium for cells within 0–20 km of actual railroads relative to those near placebo lines is estimated about 8%. In alternative specification that exploits tighter local variation in treatment does, we find agricultural productivity is approximately 3% higher in areas within 0–20 km of historical rail lines compared to cells within 20–40 km of the same lines. The positive and statistically significant railroad effect remains robust whether we pursue spatial first difference, contiguous district pair regressions, alternative productivity measure, or even when restricting the sample to railroads originally built for non-agricultural purposes (e.g., mineral extraction). The findings underscore the enduring economic legacy of colonial infrastructure, highlighting how initial "big push" in transportation networks can shape development trajectories over centuries.

The enduring question is why century-old infrastructure, much of which no longer functions, still matters for contemporary agricultural performance. A growing body of work suggests that historical events shape present-day outcomes through path dependence rather than direct legacy effects (Bleakley and Lin, 2012; Jedwab et al., 2017; Jedwab and Moradi, 2016). Consistent with this view, our results indicate that the introduction of railroads effectively locked connected regions into a distinct spatial equilibrium path, characterized by specialization in high-valued crops, that persists today and influences the productivity. Intriguingly, once we explicitly control for the extent of specialization in these high-value crops, the railroad effect itself diminishes by roughly 50%, underscoring how colonial-era infrastructure triggered a lasting structural transformation that continues to drive productivity outcomes to this day.

Agriculture remains a primary source of livelihood for millions across the continent, yet progress in productivity has often been agonizingly slow. The productivity exhibits a significant heterogeneity across region. Explanations range from natural resource variability (Benin, 2016), limited technology adoption (Block, 2014), and weak institutions (Fulginiti et al., 2004; Wuepper et al., 2023) to insufficient modern infrastructure (Benin, 2016). The post-independence period in many parts of Africa has often been characterized by stagnant economic growth, weak institutions, and instability, particularly in many former colonies of West and Central Africa (Marein, 2022). Poor infrastructure remains a challenge in numerous former colonies, due to inadequate public investment, corruption, and instability (Jedwab and Moradi, 2013). Our study offers insight into another crucial driver of agricultural performance divergence, one rooted in history yet still influencing the outcomes we observe today.

Our study contributes to a broader literature that examines the lasting implications of colonialism for contemporary economic geography. One prominent strand of this literature focuses on contemporary economic and institutional differences induced by colonizer identities and their institutional arrangements such as land tenure systems and the dichotomy of direct versus indirect rule (e.g. Acemoglu et al., 2001; Austin, 2010; Banerjee and Iyer, 2005; Heldring and Robinson, 2012; Letsa and Wilfahrt, 2020; Lowes and Montero, 2016; Michalopoulos and Papaioannou, 2016). The second strand of the literature, closest work to ours, explores how colonial-era physical investments continued to shape the growth and distribution of modern economic activities (Bertazzini, 2022; Huillery, 2009; Jedwab et al., 2017; Jedwab and Moradi, 2016). Our work extends these domains by documenting how the spatial disparities in contemporary agricultural productivity can be traced back to colonial period physical infrastructure placement, a linkage often overlooked. In so doing, this paper offers some of the first quantitative evidence on the persistent effects of these legacy assets on rural development outcomes, with a particular emphasis on agriculture in post-colonial Africa.

Our study also contributes to the extensive literature on the role of market access in rural development (Adamopoulos, 2011; Aggarwal, 2018; Asher and Novosad, 2020; Donaldson, 2018; Donaldson and Hornbeck, 2016; Sotelo, 2020; Stifel and Minten, 2008). This literature consistently document that better access to modern transportation infrastructure improves agricultural performance due to improved market access (Donaldson and Hornbeck, 2016; Stifel and Minten, 2008). Our study, however, departs from these works in its focus on very longer-term question: why century ago infrastructure might continue to matter today? This is by no means obvious question, as most of the original infrastructure has become obsolete a long time ago. We argue that these historical investments created distinct spatial equilibria that persists today and shape current production decision and thus agricultural performance. In this instance, our paper relates to recent work on *path dependency*, exploring why historical events persist and influence economic development long after the event has disappeared (Bleakley and Lin, 2012; Buggle and Nafziger, 2021; Jedwab et al., 2017; Jedwab and Moradi, 2016). However, none of these studies touch agriculture or rural development in post-colonial Africa.

The remainder of this paper is organized as follows. In Section 2, we present a brief histrionical background. Section 3 provides a comprehensive overview of our data sources and a brief description of the variables of interest. Section 4 presents the empirical strategy. Section 5 presents the estimation results and discusses their implications, while Section 6 concludes the paper.

2 Historical Background

Colonialism in Africa was marked by a complex interaction of economic exploitation, political dominance, cultural assimilation, and social upheaval (Young and Brown, 1995). European powers, driven by a desire for resources, markets, and geopolitical influence, divided the African continent during the late 19th and early 20th centuries through a series of conferences and treaties. This resulted in the imposition of colonial rule on diverse societies with rich cultural heritages and long-standing political systems (Rodney, 2018). Colonial administrations often implemented indirect rule, appointing local chiefs as intermediaries while retaining ultimate control (Cooper, 1997). Socially, colonial policies frequently reinforced racial hierarchies, with Europeans occupying privileged positions of power and Africans relegated to subordinate roles (Wall, 2015). The legacy of colonialism continues to shape the political, economic, and social landscape of Africa, presenting challenges such as political instability and struggles for cultural identity (Nunn, 2008; Settles, 1996).

European colonial powers made significant investments in the expansion of the transportation infrastructure in Africa between the 1890s and 1960s, primarily to extract resources, secure military dominance, and suppress local resistance. These efforts are often regarded as the first significant development push for modern transportation infrastructure on the continent, given the scarcity of such systems during the pre-colonial era (Bertazzini, 2022). As a result, colonial railroad location had a "first-mover" competitive advantage in terms of market access compared to non-railroad regions.

Although all colonial powers sought to invest in transportation infrastructure, they pursued somewhat different approaches tailored to their specific colonial objectives, institutional arrangements, and resource endowment. For instance, french colonial administrations typically relied on more centralized approaches and direct investments by the state. The French were inclined to channel greater resources into building connectivity between key administrative hubs (the chef-lieux) and ports. By contrast, infrastructure development under British rule prioritized trade and commerce. The British invested in the development of transportation networks, including roads, railways, and river transport, to facilitate trade with neighboring colonies and overseas markets. Other colonial regimes, such as the Dutch and Belgians, focused disproportionately on regions with plantation economies or locations critical for rubber, cocoa, and coffee exports, often under highly coercive labor practices.

3 Data

For our empirical analysis, we partition the African continent into 5 arc-minute ($\approx$ 10 km by 10 km) grid cells. We then assemble a panel data comprising information on colonial period railroad and road infrastructure endowment, current agricultural production, pre-colonial economic activities, historical weather pattern, urban development, population, and other geographic and economic variables at cell level. This dataset combines information sourced from satellite imagery, historical archives, government publications, published studies, and large-scale national and regional surveys.

Agricultural production data. We collect agricultural data from three independent sources. First, we use gridded agricultural production data from the Food and Agricultural Organization's (FAO) Global Agro-Ecological Zones (GAEZ) Project (Fischer et al., 2021). The GAEZ database offers raster data for actual crop yields, production, and harvested areas for 26 major crops worldwide at a 5 arc-minute ($\approx$ 10 km by 10 km) resolution for the entire globe for the years 2000 and 2010. Specifically, we extract the total value of crop produce and the total harvested area directly using the grid extent provided by the FAO data, without creating any new grid cells.[1] We then define our key outcome variable, real yield, as total value of output per hectare for each grid cell.

The GAEZ actual yield data is derived through a down-scaling approach, where aggregate national and sub-national production statistics are disaggregated to individual grid cells (see Appendix). The down-scaling approach can introduce measurement error due to interpolation techniques, which may imperfectly capture true yield. However, under the classical measurement error framework, whereby any error in the dependent variable is assumed to be random and absorbed into the error term, this imprecision does not bias the our estimated coefficients, although it may inflate standard errors. The GAEZ database has been used in several notable recent studies in economics (e.g., Adamopoulos and Restuccia, 2022; Costinot and Donaldson, 2016; Costinot et al., 2016; Nunn and Qian, 2011; Sotelo, 2020).

Second, as an alternative measure of productivity, we use Normalized Difference Vegetation Index (NDVI) from MODIS Terra Vegetation Indices 16-Day Global dataset.[2]

[1] Using the original grid extent offers several advantages. Most notably, it eliminates the need for resampling, which can introduce interpolation errors and distort the original spatial patterns inherent in the data. By avoiding resampling, we preserve the integrity and resolution of the FAO data, ensuring that our analysis is based on the data as originally collected and processed.

[2] The use of satellite technology to measure agricultural yields hinges on a fundamental biological

We specifically extract the annual maximum NDVI for the years 2000 and 2010 at the resolution of with a pixel size of 250m by 250m (Asher and Novosad, 2020; Labus et al., 2002). Using on-board and supplementary tools, we effectively eliminate flawed pixels, cloud cover, and canopy cover, ensuring that the indexes accurately reflect crop productivity. Moreover, we overlay the FAO GAEZ Project land use raster layer onto a MODIS image and remove pixels representing non-agricultural areas such as desert, urban area, peri-urban, semi-dense urban, and dense urban centers. We finally averaged 250m by 250m pixels into a 5 arc-minute ($\approx$ 10 km by 10 km) grid cell, matching our GAEZ data spatial unit.

Finally, we assemble agricultural survey data obtained from large-scale national surveys. The primary source of agricultural output data is the World Bank's Living Standard Measurement Survey (LSMS). The LSMS is a flagship household survey program established in the 1980s aimed at improving the quality of household data collected by national statistical offices in client countries (World Bank, 2024).[3] The LSMS is a nationally representative panel survey, provides detailed information on households' demographic and socioeconomic outcomes, including agriculture, education, labor allocation, and health. The LSMS covers eight countries in Sub-Saharan Africa: Burkina Faso, Ethiopia, Malawi, Mali, Niger, Nigeria, Tanzania, and Uganda. The LSMS also provides information on geo-localized references of the Enumeration Area (the most dis-aggregated geographic unit), which enables the merging of household and community-level data with other geo-referenced data in our disposal such as infrastructure.

Railroad data. We use the railroad data from the year 1890 to 1960 provided by Jedwab and Moradi, 2016 and Nunn and Wantchekon, 2011. As reported in Table 2 and

principle: plants utilize light within the visible portion of the electromagnetic spectrum for photosynthesis, while they reflect higher-frequency light (Taiz et al., 2022). A thriving plant will thus reflect a large amount of near-infrared (NIR) light compared to red or green light, in contrast to a distressed plant. This principle forms the basis of NDVI, a widely used metric in environmental science for assessing agricultural productivity. The NDVI is calculated using the formula,

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

In recent times NDVI has become one of the central measures of agricultural outcomes in a vast literature (Asher and Novosad, 2020; Wuepper et al., 2023).

[3] The LSMS program assists countries in conducting multi-topic household surveys, which provide crucial data for measuring poverty, understanding living conditions, and informing policy decisions.

Figure 1, approximately 38,535 kilometers of railroads were built between 1890 and 1960. Nearly 68% (26,416 kilometers) of these railways were specifically developed for military or mining operations. Jedwab and Moradi (2016) further supplement this dataset with additional variables, including placebo rail lines and roads classified into distinct class for the period of 1970s to the 2010s.

Pre-colonial ethnic homeland data. Data on the pre-colonial period often comes from *Ethnographic Atlas* (EA) assembled by Murdock (1967). This dataset provides detailed settlement patterns for 1,265 societies worldwide in the late nineteenth century. Within Africa, the EA offers valuable insights into socio-economic conditions, settlement patterns, and family structures prior to European colonization. In our empirical analysis, we use 427 ethnic homeland fixed effect, thereby purging all pre-colonial social, political, and economic heterogeneity. The ethnic homeland territories are mapped by Murdock et al. (1959).

Other data. We also use several additional information, including geographic and climate conditions such as precipitation, elevation, urban centers, distance from river, distance from port, soil quality, and luminosity. Precipitation data comes from Moderate Resolution Imaging Spectroradiometer (MODIS) (NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), 2018). Elevation data is sourced from NASA's Shuttle Radar Topography Mission (SRTM) Digital Elevation dataset (Jarvis et al., 2008), while we obtain soil quality data from FAO Harmonized World Soil Database (HWSD) (Fischer et al., 2008). Additionally, we extract night light intensity from NOAA National Geophysical Data Center (NOAA, 2013).

Summary statistics. Table 2 reports the sample size and descriptive statistics. After dropping non-agricultural regions, our analysis includes approximately 47,827 gride cells for which we have complete information on key variables in year 2000 and 2010. These observations come from 24 sub-Saharan countries, 261 regions (first-level administrative divisions), and 1,1000 districts (second-level administrative division). South Africa, Lesotho, Swaziland, and Madagascar are excluded from the analysis because they had different colonial histories such as wave of independency. we also drop Mauritania, Niger, Djibouti, Burundi, Somalia, Gambia, Eritrea, Togo, Liberia because of lack of sufficient observations within predetermined distance from railroad.[4]

4 Empirical Model

Perhaps the main empirical challenge to establish a causal effect of railroad connectivity is *selection*. In particular, colonial railroad placement can correlate with geographic fundamentals, particularly the potential for agricultural production, leading to upward bias. This assertion echoes the conventional wisdom that colonial powers prioritized mineral-rich regions and fertile arable land (Roessler et al., 2022). Our data also shows that although majority of colonial railroad constructions were driven by military objectives and mining interests, about 20% of those infrastructures were explicitly designed to facilitate agricultural trade. Consequently, to establish a causal effect, it is essential to rule out spatial heterogeneity in geographic fundamentals, particularly the correlates of agriculture potential such as soil quality, climate conditions, and terrain topography. To address the concern of possible selection bias, we employ five distinct identification strategies. In what follow, we outline three of these strategy and reaming are discussed in robustness section.

Place fixed effects. The primary outcome measure in this study is agricultural productivity measured in real yield within each grid cell. In the baseline specification, we restrict our sample to grid cell within 40 km from the colonial railroad. To define the treatment indicator, we create a 20-km buffer around each railroad to serve as the treatment boundary and designate a cells as treated if centroid of the grid cell falls within 20 km buffer zone. We then estimate

$$\log(y_{it}) = \alpha + \beta Rail_i + f(lat, long) + \gamma x_i + \eta_d + \lambda_t + \epsilon_{it} \tag{1}$$

where y_{it} is real yield in cell i at year $t \in [2000, 2010]$. The term $Rail_i$ denotes the

[4]The final list of included countries is: Angola, Benin, Botswana, Burkina-Faso, Cameroon, Congo, Ethiopia, Ghana, Guinea, Ivory Coast, Kenya, Malawi, Mali, Mozambique, Namibia, Nigeria, Senegal, Sierra Leone, Sudan, Tanzania, Uganda, Zaire, Zambia, and Zimbabwe. Ethiopia, despite its unique colonial history, is included in the analysis. For our baseline models, we consider all counties with grid cells located within the predetermined distance from colonial railroads, regardless of whether railroads directly cross the counties.

treatment indicator equal to 1 if grid cell i is within 20 km from railroad. The coefficient of interest, β can be interpreted as the average productivity effect of proximity to the colonial railroad.

Our measure of railroad connectivity does not vary over time. Therefore, threats to identification can come only from spatially varying factors. As part fo effort to capture as many of them as possible, we first control for geographical variables (soil quality, precipitation, altitude, distance from cost, and distance from navigable river), and economic and demographic variables (distance from major colonial city and population density in 1900). We represent these observables by vector x. Second, we control for smooth function of latitude and longitude denoted by $f(lat, long)$, which further neutralizes unobserved spatial heterogeneity (Dell, 2010; Guarnieri and Rainer, 2021). Finally, we restrict the identifying variation to the post-colonial second-level administrative division level (commonly known as district) by controlling for district-fixed effect (η_d), a strategy we refer to as the "Place fixed effects" approach. The district-fixed effect soaks up variations across districts, thereby β is identified from the only variation observed within district. Finally, we control for pre-colonial ethnic-homeland-fixed effect, which gives us the most saturated version of specification (1). The time-fixed effect, λ_t captures time-varying factors that influence productivity continent-wide. Lastly, ϵ_{it} is an idiosyncratic error term. In our baseline specification, we cluster standard errors at the district level. We also use Conley's (1999) method with distance cutoff 50 and 100 km to account for possible spatial correlation across grid cells. We provide all estimates using Conley standard errors in the Online Appendix (Table 14).

The identifying assumption underlies the specification (1) is that that all relevant unobserved factors are common to grid cells located within 40 km of a rail line and within the same district, thus, any observed difference in outcome can be causally attributed to the treatment, colonial railroad.[5] To provide suggestive evidence in favor of our identifying assumption, we test whether the 0-20 km grid cells differ from the 20-40 km cells in terms of relevant geographic feature, particularly agricultural potential or its correlates. Specifically, we use potential yield data from Global Agro-Ecological Zones (GAEZ) project (Fischer et al., 2021) and test whether the railroad cells are sys-

[5]Districts in our sample have an average area of 4,644 square kilometers.

tematically different. The GAEZ data provide potential yields for all major crops, including those not grown in a particular cell. These potential yields combine cell-level land quality attributes with crop-specific agronomic models, for a given cultivation inputs and level of water supply.[6] Thus, potential yields measure detailed local geographical attributes in terms of total factor productivity. As a result, spatial differences in potential yields across cells reflects a measure of differences in geography total factor productivity driven by difference in soil quality, climate conditions, and terrain topography (Adamopoulos and Restuccia, 2022). We regress the potential yields on *Rail*, district fixed effects, and crop fixed effects. As shown in Table 12, we find no significant variation, providing strong support for our identifying assumption. We also compare other relevant covariates such soil quality, precipitation, and elevation.

Placebo lines as good counterfactual: During the colonial era, not all railroad lines proposed by the colonial administrators were ultimately constructed. Tightening budgets, shifting strategic interests, Great Depression, World War (I and II) and administrative challenges frequently resulted in partially planned routes that never materialized (Bertazzini, 2022; Jaekel, 1997; Jedwab et al., 2017; Okoye and Pongou, 2017). We exploit these unbuilt or *placebo* lines as a comparison group for those segments that were actually completed.

The rationale behind this approach is that the selection criteria for railroad placement, such as the perceived economic value of agricultural or mineral resources, applied to both built and unbuilt lines. Hence, locations near these planned-but-unrealized lines likely share key geographic and economic fundamentals with areas ultimately connected by the colonial railroad. This parallel makes placebo lines a credible counterfactual for evaluating the effects of rail infrastructure.

To establish practically comparable control and treatment groups, we restrict the sample to grid cells within 0–20 km of either actual or placebo railroad lines. Then a dummy variable ($Treat_i$) is defined, where $Treat_i = 1$ for cells near actual railroads and $Treat_i = 0$ for cells near placebo lines. Then we estimate canonical Regression Discontinuous (RD) type specification using distance from the lines (actual or placebo) serves

[6] GAEZ potential yields are calculated for both individual historical years (between 1901 and 2009) and average historical climate condition. We use potential yield based on the average historical climate condition.

as the running variable. Specifically, we estimate,

$$\log(y_{it}) = \alpha + \beta Treat_i + \tau(Treat_i \times Distance_i) + \psi Distance_i + \gamma x_i + \eta_{e\times c} + \lambda_t + \epsilon_{it} \quad (2)$$

Here, β captures the treatment effect at the cutoff (i.e., immediate proximity to the railroad). The running variable $Distance_i$ accounts for baseline spatial trends, while x_i represents geographic and economic controls.[7] We normalize the running variable such that distance ranges from 0 to 20 km for treated cells (actual railroads), while distance ranges from 20 to 0 km for control cells (placebo lines). To ensure that any comparison between treated and control groups occurs within shared geographic and historical contexts, we control for ethnic-homeland-country fixed effects, $\eta_{(e\times c)}$. District fixed effects are omitted due to limited number of districts with both actual and placebo lines; ethnic homelands, being broader geographic units, ensure sufficient variations. This approach enables a direct comparison of agricultural productivity between areas near actual and placebo railroads within specific geographic area, holding Euclidean distance constant and accounting for unobserved heterogeneity across pre-colonial ethnic boundaries. Finally, the time-fixed effect, λ_t captures time-varying factors that influence productivity continent-wide.

The key identifying sumption underlies specification (2) is that initial selection criteria for building a railroad, particularly agricultural potential, between places near the actual colonial railroads and the placebo segments are comparable. We test the empirical plausibility of this assumption using potential yield data from Global Agro-Ecological Zones (GAEZ) project, as we did for specification (1). Specifically, we estimate specification (2) using the potential yield as an outcome variable. Additionally, we control for crop fixed effects. As shown in Table 13, we find no significant difference in potential yield between railroad cell and placebo cells. We also find no significant discontinuous change in precipitation and measures of soil quality.

Spatial First Difference (SFD) design: This approach is particularly effective in isolating unobserved spatial heterogeneity in cross-sectional data (Druckenmiller and Hsiang,

[7] x also includes $Distance_i^2$ and $(Treat_i \times Distance_i)^2$.

2018). It exploits differences between geographically adjacent units (e.g., grid cells) to identify the treatment effect, assuming that unobserved spatial confounders are common within these pairs of adjacent units. In other words, any relevant factors besides treatment are assumed to be spatially invariant at the local level, often referred to as the "local conditional independence assumption.[8] In our setting, this assumption is likely to hold because the grid cells are fine-grained and constructed without regard to geographic attributes such as soil quality, climate conditions, or terrain topography. Formally, the Spatial First Difference (SFD) specification is given by

$$\Delta \log(y_{it}) = \beta \Delta EqDistance_i + \rho \Delta x_i + \Delta \epsilon_{it} \tag{3}$$

where Δ denotes the spatial first difference operator. The term $Distance_i$ is a continuous variable measuring euclidean distance between grid cell i and nearest railroad line. All other terms are as defined in specification (1). The coefficient of interest β can be identified from variation in key variables across two distinct dimensions: the West-East (WE) or the North-South (NS). While East-west pairs exploits variation within East-West neighboring cells located on the same latitude or distance from equator, the North-South dimension use variation within spatially adjacent grid cells along the same longitude. The appealing feature of this approach is that North-South dimensions can serve as a useful robustness check for East-West vise versa.

5 Results

In this section, we present the main results, series of robustness checks, mechanism, and some heterogeneity exercises.

[8] The Local Conditional Independence Assumption requires the following condition to hold:

$$\mathbb{E}[Y_i \mid (D_{i-1}, X_{i-1})] = \mathbb{E}[Y_{i-1} \mid (D_{i-1}, X_{i-1})] \quad \forall \{i, i-1\}$$

This condition implies that, conditional on all observable covariates, spatially adjacent units of observation with the same treatment should have the same expected outcome. The Local Conditional Independence Assumption is weaker than the standard Conditional Independence Assumption used in typical cross-sectional research designs, as it only needs to hold for spatially adjacent units. The SFD identifying assumptions can be interpreted as a generalization of the assumptions in the spatial regression-discontinuity (RD) design (Druckenmiller and Hsiang, 2018). By contrast, the SFD design does not require discontinuous change in treatment status between two adjacent units, rather the marginal effect can be identified from all variation in the outcome and treatment variables.

5.1 Main results

We begin by estimating our baseline specification, which includes district fixed effects but excludes observable covariates and ethnic homeland fixed effects. The results reported in Table 3. The agricultural productivity difference between railroad and no-railroad cells turn out to statistically and economically meaningful. Th estimate in columns (1) suggest that the railroad-connected cells (i.e. cells within 20 km from railroad) have roughly 3% higher agricultural productivity, measured by real yield. Column (2) introduces observable geographic and socioeconomic covariates to the baseline model, while Column (3) augments this by incorporating smooth function of latitude and longitude alongside these controls. Finally, Column (4) incorporates ethnic homeland fixed effects. Crucially, the estimated relationship between railroad proximity and productivity remains robust across all specifications, with coefficients retaining comparable magnitude and significance. The stability of these estimates suggests that unmeasured geographic fundamentals or historical contexts does not confound the relationship between railroad access and productivity.

Table 4 reports the result from specification (2), where we use placebo railroad cells as counterfactual. Column (1) shows a 8% productivity premium for cells within 0–20 km of actual railroads relative to those near placebo lines within the same distance. This larger effect size compared to the baseline model (3%) aligns with expectations: while estimate in specification (1) measures productivity difference between cells with different treatment does (i.e. relative proximity to the railroad), the estimate in specification (2) measures the the productivity difference between connected and unconnected cells. Column (2) compares productivity between cells within 0–40 km of actual railroads and those near placebo lines within the same distance. The estimated productivity difference remains robust, but a negligible decline in magnitude, indicating spatial decay in railroad effects.

Table 5 reports estimates drown from specification (3), spatial first difference (SFD) model. Column (1) reports marginal effect of railroad connectivity identified from variation within East-West neighboring cells, holding North-South spatial variation constant. The estimate suggest that colonial period railroad connectivity has positive and statistically significant effect on present day agricultural productivity. Results are remain robust if we identify the coefficient from alternative dimension (Column (2)). The consistency of these estimates from different dimensions bolsters our causal claim that the relationship is not driven by some unobserved feature peculiar to specific regions.

The overall findings suggest that early access to modern transportation infrastructure contributed to the better rural development to date. This finding may come as surprise even though the existing development economics literature widely document that better access to modern transportation infrastructure improves agricultural performance due to improved market access (Donaldson and Hornbeck, 2016; Stifel and Minten, 2008). This is because those century old infrastructure endowments were long gone, and thus their influence on current economic outcome is not clear. Thus, an arguably more interesting question is why century-old infrastructure influences current agricultural performance. We therefore need to trace the persistence of railroad's effect. But, before jumping to exploring potential mechanism, we challenge the robustness of our findings across various considerations in what follow.

5.2 Robustness

To further assess the robustness of our findings, we conduct a series of robustness checks, including the use of alternative identification strategies, alternative treatment definition, different sample, and alternative productivity measure. In these robustness exercises, we primarily focus on the direction of the estimates, as different strategies necessitate varying treatments of key variables, making direct comparison of coefficients less straightforward. Overall, the robustness tests provide consistent evidence that colonial railroads caused connected cells to be more productive.

Contiguous District Pair Regression: Here we pursue "natural experiment" approach that compares districts that share common border. Specifically, strategy that we pursue exploits the variation in railroad connectivity (density) that occurs between a pair of spatially matched districts to establish causal effect. As shown in Figure 2 for Nigerian case, each railroad district (in jet color) is paired with a neighboring (control) district (highlighted in battleship gray). Then we run

$$\log(y_{idpt}) = \alpha + \beta Density_d + \rho x_{idp} + \eta_{(p\times t)} + \epsilon_{idpt} \tag{4}$$

where y_{idpt} represents the outcome for cell i in district d, border pair p, at time t. The treatment variable, Density_d, measures railroad density (km of track per square km) in district d. This specification provides an additional robustness check because the treatment variable is defined in a manner that is less prone to measurement errors than in our baseline specification. The model includes time-varying district pair fixed effects ($\eta_{(p\times t)}$) to absorb all shared temporal shocks within each pair, ensuring identification comes solely from within-pair differences in railroad connectivity. Additional geographic or socioeconomic controls (x_{idp}) may be included. By design, this specification isolates the effect of railroad density using only localized variation between contiguous district pairs, mitigating confounding from broader spatial or temporal trends.

The key identifying assumption for specification (4) is that the differences in railroad connectivity between spatially paired districts are not correlated with the differences in residuals within each district. This assumption is plausible for at least following reason. First, unlike higher level of administrative divisions such as state or country, the majority of district boundaries are arbitrary (Huillery, 2009), leading to accidental variations between neighbor districts. Second, districts in our sample have an average area of 4,644 square kilometers. Agricultural potential, determined by factors such as soil quality, agro-ecological conditions, and climate amenities, is, therefore, unlikely to vary discontinuously between adjacent districts.

An individual district may share its borders with several other entities along its border segment and thus will be in the sample as many times as it can be paired with a neighbor unit across the border segment (see Figure 2). The presence of a single entity in multiple pairs may introduce mechanical correlation across pairs in border segments. To account for this source of correlation in the residual, the standard errors are clustered at the border segment level.

Table 6 shows geographical discontinuities estimates of the impact of colonial railroad connectivity on current agricultural productivity. Although the coefficients are not directly comparable, the direction of the effects are consistent with the baseline estimates, suggesting that our findings are robust to alternative measures of treatment and identification approach.

Military and Mining Rails. We further challenge the robustness of our baseline result by considering only railroads that are built for the non-agricultural purpose. Specifically, we consider railways constructed for the military and or mining purpose. In doing so, we aims at to address the concern of *selection bias* associated with the unobserved confounder that may influence crop potentials and thus shaped railroad placement decision. We strongly believe that the placement of railroads specifically targeting military outposts or mining regions are not influenced by the agricultural potential of the connected regions. Results reported in 7 show a positive and statistically significant railroad effect.

Alternative productivity measure. Considering potential measurement error, particularly in the down-scaling process for FAO's GAEZ agricultural yield data, we use the Normalized Difference Vegetation Index (NDVI) as an alternative measure of agricultural productivity. We extract annual maximum NDVI values for the years 2000 and 2010 from the MODIS Terra Vegetation Indices 16-Day Global dataset. The data is aggregated to a 5 arc-minute ($\approx$ 10 km by 10 km) grid cell resolution, matching the spatial resolution of our other datasets.

Table 8 presents the results. Column (1) reports the baseline specification without covariates, while Column (2) includes geographic and socioeconomic controls. Column (3) further incorporates smooth function of latitude and longitude, while Column (4) includes ethnic homeland fixed effects. The results indicate that proximity to colonial railroads is associated with higher NDVI values, suggesting that areas closer to historical railroads exhibit greater vegetation health and, by extension, higher agricultural productivity.

5.3 Tracing the Persistence of Colonial Railroad's Effects

An enduring puzzle in economic history is why certain historical shocks continue to influence economic outcomes long after the original conditions that triggered them have vanished. In the context of colonial infrastructure, one might expect that once the rail lines ceased to function or were substantially downgraded, their economic relevance would dissipate. However, a wealth of empirical literature indicates otherwise: key historical shocks can shape the spatial distribution of economic activity that persist for a long time, even if the immediate advantage of the shock has become obsolete (Bleakley and Lin, 2012; Jedwab et al., 2017).

Central to this discussion is the concept of *path dependence*, which suggests that once a region gains an early advantage, be it in terms of population, infrastructure, or a strategic location, it can remain "locked in" to a particular development path even if the initial reasons for that advantage no longer apply (Arthur, 1994). In many economic models that feature increasing returns to scale and agglomeration externalities (Fujita et al., 2001; Krugman, 1991), such an advantage can push an economy into a higher-level equilibrium, whereas locations that missed out on the early push remain stuck in a "bad" equilibrium. The underlying logic is that economic agents prefer to locate factors in locations where other economic activities are already placed, thereby reinforcing the initial advantage. These self-reinforcing mechanisms help explain why a historical event like colonial railroad placement can have lasting effects on regional productivity.

A key empirical example in the literature is provided by Bleakley and Lin (2012), who show that towns built around portage sites in the United States continue to flourish despite the portage advantage becoming technologically obsolete in the age of railroads and highways. They interpret the finding as evidence of path dependence. Similarly, Jedwab et al. (2017) show that colonial cities in Kenya have continued to flourish even after the factors that initially led to their establishment, such as colonial railroads, disappeared. These findings resonate with Redding et al. (2011) and Buggle and Nafziger (2021), who underscore the role of historical shocks in shaping spatial equilibria.

Building on this literature, our study asks: which specific channels connect historical railroads to modern-day agricultural outcomes? There are at least three candidate mechanisms through which historical railroad infrastructure translate into higher agriculture productivity. These are coordination problem, specialization and spatial spillover.

Coordination Problem. From the perspective of new economic geography, spatial coordination failures can arise when economic agents must choose where to locate capital among multiple potential sites (Krugman, 1991). Because establishing large-scale initial infrastructure (e.g., roads, bridges, dams) involves high sunk costs, no single private agent wants to be the "first mover" and risk underutilized capacity (Murphy et al., 1989; Rosenstein-Rodan, 1943), leading to suboptimal spatial distributions of economic activity. In such scenarios, initial sunk investments, even relatively small ones, can serve as a coordination device to overcome the equilibrium selection problem and thus shaping subsequent spatial investment decisions (Bleakley and Lin, 2012). This implies that regions with comparable resource endowments may follow divergent development trajectories if one region successfully attracts an initial "big push" in infrastructure.

To put this into perspective, a modest colonial rail spur, for example, might lead to the concentration of economic activities in the railroad regions, such as the emergence of a small railroad city, market, and thus higher population density along the route.[9] Over time, this established economic geography tends to draw further investments (Huillery, 2009), largely due to localized agglomeration economies and/or coordination failure (i.e., creating new economic clusters is costly).[10] Therefore, our premise is that regions along colonial railroads have continued to receive relatively better physical capital (e.g., feeder roads) in the post-colonial period and thus remain relatively better integrated into markets and exhibit higher agricultural productivity.

We perform two a two-stage strategy to evaluate whether the railroad's legacy persists primarily by having propelled a region onto a "good" equilibrium characterized by better modern infrastructure and market integration. First, we investigate whether proximity to a colonial railroads indeed associated with better physical infrastructure today, controlling for initial conditions. This addresses whether early rail connectivity systematically shaped the distribution of subsequent government spending on alternative infrastructure. Following Bleakley and Lin (2012); Jedwab et al. (2017), we focus

[9] The practical example of colonial cities are Accra in Ghana and Lusaka in Zambia.

[10] In contrast, a region without such infrastructure struggles to compete, even if its geographical conditions are similar, because it fails to achieve the critical "big push" necessary to transition into a new ("good") equilibrium path.

on a key outcome that proxy the spatial distribution of physical capital, paved road endowment in 2010. Using this variable, we run

$$Road_{i,2010} = \rho + \gamma Rail_i + f(lat, long) + \gamma x_i + \eta_d + \lambda_t + e_i \tag{5}$$

where $Road_{i,2010}$ is a dummy variable equal to if grid cell i has paved road in 2010, and zero otherwise. The term x represents the control variables including colonial period road connectivity and population in 1960, allowing us to control for initial differences. We control for distinct fixed effects and ethnic homeland effects.

Second, we add $Road_{i,2010}$ into our main specification to assess its role as a mediator of long-run productivity effect of colonial railroads. Formally,

$$\log y_{it} = \alpha + \beta Rail_i + \psi Road_{i,2010} + f(lat, long) + \gamma x_i + \eta_d + \lambda_t + \epsilon_{cit}, \tag{6}$$

The logic of our empirical exercise here is that if *Road* mediates the observed effect to any extent, the coefficient on the colonial railroad, β should diminish upon inclusion of *Road*. The notion is that the direct effect of the railroad alone might be small (especially given that many lines are defunct): the railroad's legacy persists primarily by having propelled a region onto a "good" equilibrium characterized by higher alternative infrastructure endowment and market integration.

The results in Colum (1) of Table 9 corroborate our hypothesis that localities historically close to colonial rail corridors have relatively higher levels of paved road development today. This finding implies that, for post-colonial elites and policymakers, reactivating or reinforcing historical corridors might be a cost-effective way to consolidate existing spatial patterns, whereas attempting to create entirely new economic hubs in historically disconnected areas requires overcoming large coordination hurdles.

The results in Column (2) and (3) of Table 9 report estimated railroad effect without and with paved road as an additional control, respectively. The rail road effect does not change up on inclusion of paved road, suggesting limited role of post-colonial period investment in alternative transportation newton in reinforcing the first-mover competitive advantage.

Specialization. A second mechanism through which colonial-era investment on modern infrastructure can continue to influence agricultural outcomes is through enduring patterns of specialization on high-value crops. In classical trade theory (e.g., (Ricardo, 1821)), a region's production decisions hinge on comparative advantage and market access. Historically, the introduction of railroads, connecting the hinterland areas with domestic and global market hubs, prompted the regions near the railroads to specialize in cash crop like cocoa, cotton, or groundnuts production, enabling these region to shift from subsistence (pre-colonial equilibrium) to market-oriented production (Jedwab and Moradi, 2016; Okoye et al., 2019). However, with the collapse of railroads in the 1980s and 1990s, cash crop regions lost their competitive advantage. The key question is whether the railroad economy continues to specialize in high-value crops due to the established economic geography, despite the disappearance of railroads, or whether it transitions to a pre-colonial type equilibrium, such as reverting to subsistence farming. Therefore, it is natural to ask: do regions near former railroads exhibit higher levels of specialization compared to their neighbors today?

Our hypothesis is that, after the collapse of the railroads, the railroad economy continued to specialize in its high-value crops because switching to an alternative equilibrium is costly or has not occurred, at least in the short term. Thus, once we control for the degree of specialization difference, we should expect the observed railroad effect to diminish. To assess whether inclusion of specialization measure alter our baseline estimate, we estimate,

$$\log y_{it} = \alpha + \beta Rail_i + \psi Specilization_{it} + f(lat, long) + \rho x_i + \epsilon_{cit}, \tag{7}$$

where $Specialization_{it}$ measures the share of land allocated to a crops identified as high-value or cash crops, including cacao, oil palm, cotton, tobacco, sugar, *inter alia*. All other terms are defined as previous specifications.

Table 10 presents the results. Notably, we find that the railroad effect diminishes substantially once the specialization variable is included in the model, as shown in column (2). Specifically, the marginal railroad effect is reduced by approximately 50 percent relative to the baseline level. This evidence provides strong support for our hypothesis that century-old infrastructure investments have continued to influence the trajectory of rural development over time. Recall that we find no meaningful difference in cash crop potential between railroad cells and their counterpart in our covariate balance cheek exercise (see Table 12). Constituently, any difference in specialization in high-value crops can largely be attributed to railroad rather than difference in agronomic conditions. Thus, we argue that by setting the railroad economy into a distinct spatial equilibrium path, characterized by cash crop production, these historical investments have established patterns of agricultural performance that persist to this day. In essence, colonial railroads induced a path-dependent shift in production structures. In theoretical terms, this aligns with Ricardian trade models augmented by sunk costs and frictions, as well as Lewis-type dual economy frameworks where market-ordained sector (e.g. cash-crop agriculture) can coexist alongside with a traditional, subsistence-based sector.

There are several frictions that potentially prevent the region from switching back from commercial monoculture to purely subsistence-based activities once they lost their competitive advantage. First, the local farmers and traders over time might invested in specialized knowledge, equipment, or facilities, for example, cocoa fermentation sheds, cotton gins. It is expensive to switch back to a new crop portfolio. Second, historical specialization often fosters the development of extension services, research stations, or private input suppliers with a specific crop focus. These networks may then reinforce continued crop specialization even when original transport advantages recede. Finally, regions might develop powerful producer cooperatives or marketing boards that reinforce crop choices. For example, in West Africa, cocoa producer unions developed into large local institutions. The costs of transition associated with dismantling these organizations would be very high. Finally, a thriving city near historical rail lines may offer localized economies in processing, marketing, and distribution to nearby agricultural regions. In areas with higher population densities, local traders can reduce per-unit transport costs, and farmers may have greater bargaining power or more stable demand for agricultural goods. We examine the notion of spatial slipover (Marshall, 1920) arises from urban-rural linkage in following sub-section.

Spatial spillover: The third potential mechanism is urban-rural linkage. Colonial railroads often facilitated the emergence of trading posts and intermediate cities (Jedwab et al., 2017). Larger urban agglomerations coupled with dense population, in turn, create thicker markets for inputs and outputs, potentially reinforcing agricultural productivity in nearby rural locations(Benziger, 1996; Christiaensen and Todo, 2013; Katzman, 1974; Oueslati et al., 2019). Urbanization can also impacts through improved capital markets (Katzman, 1974) and information and knowledge spillovers (Duranton and Puga, 2004; Glaeser, 2010).

To assess whether the spacial spillover arising from urban-rural linkage partly explain the observed railroad effect, we augment our baseline specification with proxy of concentration of modern economic activities: distance to a major city or night light intensity in 2010 (Bertazzini, 2022). If population density or urban spillovers drive the observed persistence, the railroad coefficient should diminish up on controlling for proximity to urban center or luminosity.

Column (2) in Table 11 reports the railroad effect while controlling for proximity to the nearest major urban center. Including the location of an urban center does not significantly attenuate the baseline railroad effect. In Column (3), we instead control for luminosity, which reduces the baseline railroad effect by roughly 16%. This outcome suggests that urban spillovers partially mediate the long-term influence of railroads. Nevertheless, the remaining persistence could stem from other channels, such as path-dependent in human capital, technology diffusion, or institutional legacies, among others. While these factors fall beyond the focus of this study, they present promising avenues for future research.

5.4 Heterogeneity

This section explores how the impact of colonial-era railroads on present-day agricultural productivity varies across different economic, institutional, and geographic contexts. The persistent effects of historical infrastructure should not be expected to manifest uniformly across the sub-Saharan African countries. Indeed, a host of factors, ranging from colonial administrative structures, pre-colonial political economy, post-independence institutions, and agro-ecological conditions could shape the magnitude and direction of the impact. Below, we examine two salient dimensions that potentially introduce systematic heterogeneity in the colonial railroad legacy.

Colonizer Identity and Colonial Objectives. A first dimension of heterogeneity may emerge from differences in the motives and administrative styles of various European colonial powers. Extractive colonialism in Africa inherently exhibited substantial heterogeneity in terms of governance, the strength and form of property-rights enforcement, and strategies for mobilizing local resources (Acemoglu et al., 2001; Austin, 2010; Banerjee and Iyer, 2005). With respect to investment in transpiration infrastructure, Europeans also pursued somewhat different approaches tailored to their specific colonial objectives, economic geography, institutional arrangements, and resource endowment (See section (2) for more detail).

To explore potential colonial institution-driven heterogeneity, here we augment our baseline model with the interaction of railroad and colonizers identity. Specifically, we interact the railroad variable with four major colonies dummy, including France, Belgium, Portugal and Britain. Table **??** reports the results, where the Britain dummy is omitted category. It is evident that the persistent effect of railroad shows non-negligible heterogeneity: the effect turns out to be less pronounced in Belgium colonies compared to British. In contrast, the observed legacy effect appears to be relatively stronger in the French colonies.

Post-colonial Institutions. Post-colonial institutional arrangement disparities can moderate how effectively regions leverage any competitive advantage conferred by historical railroads. Even if certain areas inherited superior transportation infrastructure, instability, poor governance or corruption in post-colonial period could attenuate or nullify these advantages (Fulginiti et al., 2004; Wuepper et al., 2023). Moreover, findings from our mechanism exercise suggest that the historical railroad infrastructure translate into higher agriculture productivity partly through post-colonial investment in alternative infrastructures such as feeder roads along defunct rail lines footprints. Therefore, the capacity of post-colonial institution to mobilize the investments in railroad regions could mediate the persistent effect of legacy assets.

Empirically, we show the heterogeneity in persistent railroad effect across post-colonial national institution. Specifically, we estimate the baseline model for each countries and plot the distribution of estimates.[11] Figure 3 depicts the distribution, indicating the majority of estimate fall right side of the zero. The overall, the finding suggests that although the magnitude of localized effect of railroad may be heterogeneous, the historical railroad infrastructure indeed translate into higher agriculture productivity in Africa. In other words, while there is post-colonial institution-driven heterogeneity in railroad effect, the average overall effect is not driven by a few countries with better institution such Nigeria and Kenya.

6 Conclusion

Colonial investments in railroads have left a persistent and uneven imprint on agricultural performance across post-colonial Africa. This paper has shown that, despite these legacy infrastructures being obsolete or minimally operational in most places, their initial placement and short-run economic roles continue to influence the spatial distribution of productive factors. By leveraging multiple empirical approaches, we find robust evidence that areas with closer proximity to colonial period rail lines enjoy substantially higher crop yields. These results hold even after accounting for a wide array of economic, geo-climatic, and institutional factors. In general, our findings suggest that although many colonial railroads were initially designed for resource extraction or militarily dominance, the residual connectivity benefits have extended to agricultural sector. This phenomenon is especially salient in contexts where subsequent post-independence institution have been strong.

We posit that these persistent effects arise through path-dependent processes. Early colonial investments in rail connectivity integrated regions to local and international markets, incentivization specialization in regions comparative advantage crops that persisted today. Moreover, the legacy assets coordinated spatial investment decisions in ancillary sectors (e.g., market centers, cities, and supply chains). Once these localized economies coalesced around former rail hubs, they continued to accumulate capital and human skills, further reinforcing the initial spatial equilibrium.

[11] To avoid noisy estimates, we restrict our sample to countries with sufficient number of observation.

Our results imply that policymakers and development practitioners should consider the historical distribution of infrastructure when planning modern rural development strategies, particularly transport projects. While reinforcing or rejuvenating older transport corridors might be a cost-effective way to exploit benefits of long-standing market linkages, supply networks, and human capital clusters, extending modern infrastructure to historically unconnected regions can potentially offer significant returns on investment (at least in the long run) because those infrastructure investments are sunk, and have potential to transform the overall economic geography of these regions. However, attempting to create new economic hubs in historically disconnected areas requires overcoming large coordination hurdles. All in all, in areas bypassed by the initial "big push" in infrastructure, there may be a need for targeted modern infrastructure extensions or complementary interventions to "catch up" to more favored regions, thereby reducing spatial inequality.

References

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). American Economic Association The Colonial Origins of Comparative Development: An Empirical Investigation. Issue: 5 Pages: 1369-1401 Publication Title: Source: The American Economic Review Volume: 91.

Adamopoulos, T. (2011). Transportation costs, agricultural productivity, and cross-country income differences. *International Economic Review*, 52(2):489–521.

Adamopoulos, T. and Restuccia, D. (2022). Geography and agricultural productivity: Cross-country evidence from micro plot-level data. *The Review of Economic Studies*, 89(4):1629–1653.

Aggarwal, S. (2018). Do rural roads create pathways out of poverty? evidence from india. *Journal of Development Economics*, 133:375–395.

Arthur, W. B. (1994). *Increasing Returns and Path Dependence in the Economy*. University of Michigan Press, Ann Arbor.

Asher, S. and Novosad, P. (2020). Rural roads and local economic development. *American economic review*, 110(3):797–823.

Austin, G. (2010). *African economic development and colonial legacies*. Number 1. Institut de hautes études internationales et du développement.

Banerjee, A. and Iyer, L. (2005). American Economic Association History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India. Issue: 4 Pages: 1190-1213 Publication Title: Source: The American Economic Review Volume: 95.

Benin, S. (2016). *Agricultural productivity in Africa: Trends, patterns, and determinants*. Intl Food Policy Res Inst.

Benziger, V. (1996). Urban access and rural productivity growth in post-mao china. *Economic Development and Cultural Change*, 44(3):539–570.

Bertazzini, M. C. (2022). The long-term impact of italian colonial roads in the horn of africa, 1935–2015. *Journal of Economic Geography*, 22(1):181–214.

Bleakley, H. and Lin, J. (2012). Portage and path dependence. *The quarterly journal of economics*, 127(2):587–644.

Block, S. (2014). The decline and rise of agricultural productivity in sub-saharan africa since 1961. In *African Successes, Volume IV: Sustainable Growth*, pages 13–67. University of Chicago Press.

Buggle, J. C. and Nafziger, S. (2021). The slow road from serfdom: labor coercion and long-run development in the former russian empire. *Review of Economics and Statistics*, 103(1):1–17.

Christiaensen, L. and Todo, Y. (2013). Urbanization and poverty reduction: the role of rural diversification and secondary towns. *Agricultural Economics*, 44(4-5):447–459.

Conley, T. G. (1999). Gmm estimation with cross section dependence. *Journal of Econometrics*, 91(2):103–138.

Cooper, F. (1997). Citizen and subject: Contemporary africa and the legacy of late colonialism.

Costinot, A. and Donaldson, D. (2016). How large are the gains from economic integration? theory and evidence from us agriculture, 1880-1997. Technical report, National Bureau of Economic Research.

Costinot, A., Donaldson, D., and Smith, C. (2016). Evolving comparative advantage and the impact of climate change in agricultural markets: Evidence from 1.7 million fields around the world. *Journal of Political Economy*, 124(1):205–248.

Dell, M. (2010). The persistent effects of peru's mining mita. *Econometrica*, 78(6):1863–1903.

Donaldson, D. (2018). Railroads of the raj: Estimating the impact of transportation infrastructure. *American Economic Review*, 108(4-5):899–934.

Donaldson, D. and Hornbeck, R. (2016). Railroads and american economic growth: A "market access" approach. *The Quarterly Journal of Economics*, 131(2):799–858.

Druckenmiller, H. and Hsiang, S. (2018). Accounting for unobservable heterogeneity in cross section using spatial first differences. Technical report, National Bureau of Economic Research.

Duranton, G. and Puga, D. (2004). Micro-foundations of urban agglomeration economies. *Handbook of regional and urban economics*, 4:2063–2117.

Fischer, G., Nachtergaele, F., Prieler, S., van Velthuizen, H., Verelst, L., and Wiberg, D. (2008). *Global Agro-ecological Zones Assessment for Agriculture (GAEZ 2008)*. IIASA, Laxenburg, Austria and FAO, Rome, Italy.

Fischer, G., Nachtergaele, F. O., Van Velthuizen, H., Chiozza, F., Francheschini, G., Henry, M., Muchoney, D., and Tramberend, S. (2021). Global agro-ecological zones (gaez v4)-model documentation.

Fujita, M., Krugman, P. R., and Venables, A. (2001). *The spatial economy: Cities, regions, and international trade*. MIT press.

Fulginiti, L. E., Perrin, R. K., and Yu, B. (2004). Institutions and agricultural productivity in sub-saharan africa. *Agricultural Economics*, 31(2-3):169–180.

Glaeser, E. L. (2010). *Introduction to "Agglomeration Economics"*. University of Chicago Press.

Guarnieri, E. and Rainer, H. (2021). Colonialism and female empowerment: A two-sided legacy. *Journal of Development Economics*, 151:102666.

Heldring, L. and Robinson, J. A. (2012). Colonialism and economic development in africa. Technical report, National Bureau of Economic Research.

Huillery, E. (2009). History matters: The long-term impact of colonial public investments in french west africa. *American economic journal: applied economics*, 1(2):176–215.

Jaekel, F. (1997). *The History of the Nigerian Railway: Network and Infrastructures*, volume 2. Spectrum books.

Jarvis, A., Reuter, H., Nelson, A., and Guevara, E. (2008). Hole-filled srtm for the globe, version 4. Available from the CGIAR-CSI SRTM 90m Database. Accessed: 2025-02-01.

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. (2013). Transportation technology and economic change: The impact of colonial railroads on city growth in africa.

Jedwab, R. and Moradi, A. (2016). The permanent effects of transportation revolutions in poor countries: Evidence from Africa. *Review of Economics and Statistics*, 98(2):268–284. Publisher: MIT Press Journals.

Katzman, M. T. (1974). The von thuenen paradigm, the industrial-urban hypothesis, and the spatial structure of agriculture. *American Journal of Agricultural Economics*, 56(4):683–696.

Krugman, P. (1991). Increasing returns and economic geography. *Journal of political* *economy*, 99(3):483–499.

Labus, M., Nielsen, G., Lawrence, R., Engel, R., and Long, D. (2002). Wheat yield estimates using multi-temporal ndvi satellite imagery. *International Journal of Remote Sensing*, 23(20):4169–4180.

Letsa, N. W. and Wilfahrt, M. (2020). The mechanisms of direct and indirect rule: Colonialism and economic development in africa. *Quarterly Journal of Political Science*, 15(4):539–577.

Lowes, S. and Montero, E. (2016). Blood rubber: The effects of labor coercion on institutions and culture in the drc.

Marein, B. (2022). Colonial roads and regional inequality. *Journal of Urban Economics*, 131:103492.

Marshall, A. (1920). *Principles of Economics*. Macmillan, London.

Michalopoulos, S. and Papaioannou, E. (2016). The long-run effects of the scramble for africa. *American Economic Review*, 106(7):1802–1848.

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

Murdock, G. P. et al. (1959). *Africa: Its peoples and their culture history*, volume 6. McGraw-Hill New York.

Murphy, K. M., Shleifer, A., and Vishny, R. W. (1989). Industrialization and the big push. *Journal of political economy*, 97(5):1003–1026.

NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC) (2018). Aster mount gariwang image from 2018. https://lpdaac.usgs.gov/resources/data-action/aster-ultimate-2018-winter-olympics-observer/. Retrieved on YYYY_MM_DD, 2024 from https://lpdaac.usgs.gov. Maintained by the NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC) at the USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota.

NOAA (2013). Dmsp ols: Nighttime lights time series version 4, defense meteorological program operational linescan system. https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html. Accessed: 2025-02-01.

Nunn, N. (2008). The long-term effects of africa's slave trades. *The Quarterly Journal of Economics*, 123(1):139–176.

Nunn, N. and Qian, N. (2011). The potato's contribution to population and urbanization: evidence from a historical experiment. *The quarterly journal of economics*, 126(2):593–650.

Nunn, N. and Wantchekon, L. (2011). The slave trade and the origins of mistrust in africa. *American economic review*, 101(7):3221–3252.

Okoye, D. and Pongou, R. (2017). The missionaries: God against the natives. *American Economic Journal: Applied Economics*, 9(4):217–244.

Okoye, D., Pongou, R., and Yokossi, T. (2019). New technology, better economy? the heterogeneous impact of colonial railroads in nigeria. *Journal of Development Economics*, 140:320–354.

Oueslati, W., Salanié, J., and Wu, J. (2019). Urbanization and agricultural productivity: some lessons from european cities. *Journal of Economic Geography*, 19(1):225–249.

Redding, S. J., Sturm, D. M., and Wolf, N. (2011). History and industry location: evidence from german airports. *Review of Economics and Statistics*, 93(3):814–831.

Ricardo, D. (1821). *On the principles of political economy*. J. Murray London.

Rodney, W. (2018). *How europe underdeveloped africa*. Verso Books.

Roessler, P., Pengl, Y. I., Marty, R., Titlow, K. S., and Walle, N. v. d. (2022). The cash crop revolution, colonialism and economic reorganization in africa. *World Development*, 158. Publisher: Elsevier Ltd.

Rosenstein-Rodan, P. N. (1943). Problems of industrialisation of eastern and south-eastern europe. *The Economic Journal*, 53(210/211):202–211.

Settles, J. D. (1996). The impact of colonialism on african economic development.

Sotelo, S. (2020). Domestic trade frictions and agriculture. *Journal of Political Economy*, 128(7):2690–2738.

Stifel, D. and Minten, B. (2008). Isolation and agricultural productivity. *Agricultural Economics*, 39(1):1–15.

Taiz, L. et al. (2022). *Plant Physiology and Development*. Sinauer Associates : Oxford University Press, New York, NY, seventh edition.

Wall, I. (2015). Imperial identities: Stereotyping, prejudice and race in colonial algeria.

World Bank (2024). Living standards measurement survey (lsms). https://www.worldbank.org/en/programs/lsms. Accessed: 2024-07-07.

Wuepper, D., Wang, H., Schlenker, W., Jain, M., and Finger, R. (2023). Institutions and global crop yields. Technical report, National Bureau of Economic Research.

Young, C. and Brown, W. H. (1995). The african colonial state in comparative perspective. *History: Reviews of New Books*, 24(1):39–40.

Table and Figure

Table 1: Colonial railroad lines lengths in Kilometers

Country	Length (km)
Angola	2892.294
Benin	550.6883
Botswana	713.2622
Burkina Faso	517.6434
Cameroon	450.4017
Congo	475.0158
Cote d'Ivoire (Ivory Coast)	633.9022
Djibouti	91.07671
Equatorial Guinea	15.57202
Eritrea	299.6477
Ethiopia	653.2147
Ghana	943.9069
Guinea	709.7694
Kenya	1865.433
Liberia	132.3724
Malawi	480.0863
Mali	621.9655
Mozambique	2627.715
Namibia	2423.369
Nigeria	2721.513
Sao Tome and Principe	20.50156
Senegal	1013.932
Sierra Leone	528.55
Somalia	13.97499
South Africa	31.2416
Sudan	4430.325
Tanzania	2259.563
Togo	417.8464
Uganda	855.1809
Zaire	5090.777
Zambia	1748.756
Zimbabwe	2410.526

Notes: The railroads were constructed between the 1890 and 1960. Data for Algeria, Egypt, Morocco, and Tunisia is currently being processed. South Africa, Lesotho, Swaziland, and Madagascar are excluded from the analysis. Ethiopia and Liberia are included in the analysis, despite their different colonial histories..

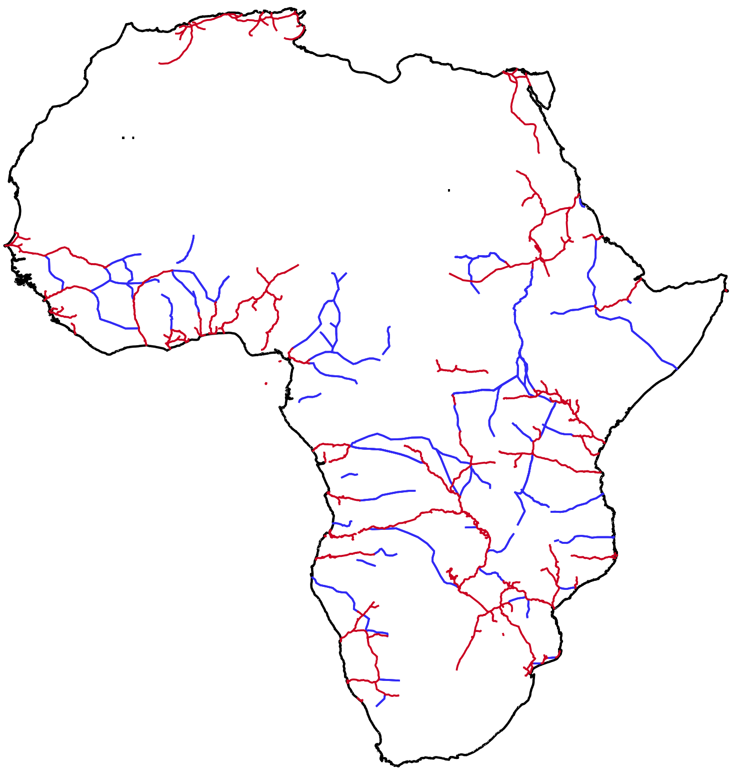

Figure 1: Map of colonial railway lines (red) and planned but never actually built (blue). Data from (Jedwab and Moradi, 2016) and (Nunn and Wantchekon, 2011).

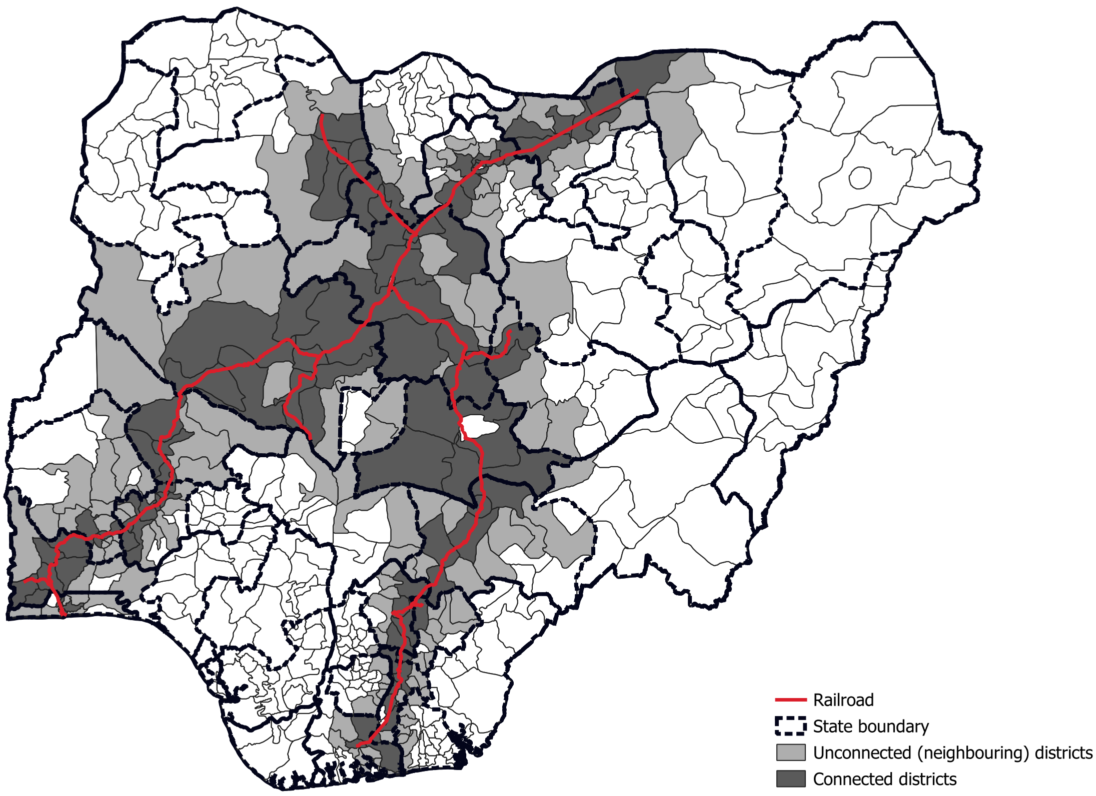

Figure 2: Contiguous District Pair

Table 2: Summary Statistics

(1)	Mean	Min	P10	P50	P90	Max	Std. Dev.
Real Yield (GK$/ha)	377.36	9.24	133.75	362.09	612.64	3011.27	206.59
Rail within 20 km (0/1)	0.54	0.00	0.00	1.00	1.00	1.00	0.50
Railroad per sqkm	0.02	0.00	0.00	0.01	0.05	0.75	0.04
Distance from paved road (km) 1960	40.28	0.06	4.02	22.80	98.71	464.73	53.15
Distance from paved road (km) 2010	29.43	0.09	4.34	22.36	59.66	178.13	27.42
Cash crop share (%)	24.91	0.00	6.03	20.44	52.67	100.00	17.54
Luminosity in 2010	1.06	0.00	0.00	0.00	1.81	63.00	4.85
Share of class 1 soils	0.06	0.00	0.00	0.00	0.18	1.00	0.20
Share of class 2 soils	0.12	0.00	0.00	0.00	0.52	1.00	0.26
Mean altitude (m)	712.98	1.32	75.59	560.84	1401.12	3663.12	538.82
Avg. annual precipitation (mm) 1960	1099.40	26.97	515.28	1048.49	1643.91	3357.44	521.12
Euclidean distance (km) to nearest river	180.08	0.03	19.65	140.68	408.28	715.33	153.66
Euclidean distance (km) to nearest city 2010	96.07	0.45	22.90	67.23	239.89	390.30	83.64
Euclidean distance (km) to nearest city 1900	355.85	0.22	53.34	269.56	777.57	1120.96	279.46
Euclidean distance (km) to nearest coast	514.81	0.26	62.46	433.16	1102.69	1747.31	400.54
Observations				47,827			

Table 3: District FE Regression

	Ln(Real Yield (GK$/ha))			
	(1)	(2)	(3)	(4)
Rail within 20 km	0.0329***	0.0322***	0.0316***	0.0301***
	(0.0095)	(0.0095)	(0.0095)	(0.0089)
Controls	No	Yes	Yes	Yes
Lat & Long	No	No	Yes	Yes
District fixed effects	Yes	Yes	Yes	Yes
Tribe fixed effects	No	No	No	Yes
Year fixed effects	Yes	Yes	Yes	Yes
N	47827	47827	47827	47613
R-squared	0.6994	0.7000	0.7002	0.7151

Notes: The dependent variable is the natural log of value of production expressed in international price weights, calculated by the FAO in terms of year 2000 US dollars per hectare (GK$/ha). Standard errors (in parentheses) are clustered at the district level. ***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Table 4: Placebo Lines

	Ln(Real Yield (GK$/ha))	
	(1)	(2)
Rail (1= actual)	0.0802**	0.0754**
	(0.0403)	(0.0372)
Controls	Yes	Yes
Tribe x Year fixed effects	Yes	Yes
Year fixed effects	No	No
N	45716	84072
R-squared	0.6159	0.6266

Notes: The dependent variable is the natural log of value of production expressed in international price weights, calculated by the FAO in terms of year 2000 US dollars per hectare (GK$/ha). Standard errors (in parentheses) are clustered at the district level. ***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Table 5: Spatial First Differences

	ΔLn(Real Yield (GK$/ha))	
	(1)	(2)
ΔDistance to railroad (km)	-0.0016**	-0.0018***
	(0.0008)	(0.0004)
District fixed effects	No	No
Tribe fixed effects	No	No
N	44260	43517
R-squared	0.0007	0.0014

Notes: The dependent variable is the spatial first difference of the natural log of value of production expressed in international price weights, calculated by the FAO in terms of year 2000 US dollars per hectare (GK$/ha). Column (1) reports the marginal effect of railroad connectivity identified from variation within East-West neighboring cells, while the estimate in Column (2) is identified from the North-South dimension. Standard errors (in parentheses) are clustered at the district level. ***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Table 6: Contiguous District Pair Regression Results

	Ln(Real Yield (GK$/ha))			
	(1)	(2)	(3)	(4)
Railroad per km^2	3.0524***	2.9819***		
	(0.6077)	(0.5989)		
Rail (0/1)			0.0354**	0.0364***
			(0.0146)	(0.0141)
Controls	No	Yes	No	Yes
Pair x Year effects	Yes	Yes	Yes	Yes
N	718491	718491	718491	718491
R-squared	0.7099	0.7139	0.7095	0.7136

Notes: The dependent variable is the natural log of value of production expressed in international price weights, calculated by the FAO in terms of year 2000 US dollars per hectare (GK$/ha). The treatment variables are measured in railroad density (Railroad per km^2) and a connectivity dummy (Rail (0/1)). Standard errors (in parentheses) are clustered at the district border segment level. ***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Table 7: Mining and military

	Ln(Real Yield (GK$/ha))			
	(1)	(2)	(3)	(4)
Rail within 20 km	0.0179**	0.0180**	0.0178**	0.0180**
	(0.0086)	(0.0087)	(0.0087)	(0.0089)
Controls	No	Yes	Yes	Yes
Lat & Long	No	No	Yes	Yes
District fixed effects	Yes	Yes	Yes	Yes
Tribe fixed effects	No	No	No	Yes
Year fixed effects	Yes	Yes	Yes	Yes
N	32576	32576	32576	32458
R-squared	0.6434	0.6438	0.6441	0.6588

Notes: The dependent variable is the natural log of value of production expressed in international price weights, calculated by the FAO in terms of year 2000 US dollars per hectare (GK$/ha). The sample is restricted to railroads constructed for mining and military purposes. Standard errors (in parentheses) are clustered at the district level. ***Significant at the 1 percent level. **Significant at the 5 percent level.

Table 8: Normalized vegetation Index (NDVI)

	Ln(NDVI)			
	(1)	(2)	(3)	(4)
Rail within 20 km	0.0167***	0.0181***	0.0175***	0.0176***
	(0.0060)	(0.0059)	(0.0056)	(0.0054)
Controls	No	Yes	Yes	Yes
Lat & Long	No	No	Yes	Yes
District fixed effects	Yes	Yes	Yes	Yes
Tribe fixed effects	No	No	No	Yes
Year fixed effects	Yes	Yes	Yes	Yes
N	43484	43484	43484	43484
R-squared	0.8640	0.8751	0.8773	0.8887

Notes: The dependent variable is the natural log of the Normalized Vegetation Index (NDVI). Standard errors (in parentheses) are clustered at the district level. ***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Table 9: Post-colonial Infrastructure

	Paved road dummy	Ln(Real Yield (GK$/ha))	
	(1)	(2)	(3)
Rail within 20 km	0.0160***	0.0316***	0.0312***
	(0.0048)	(0.0095)	(0.0096)
Paved road dummy in 1960 (0/1)	0.2814***		
	(0.0224)		
Paved road dummy in 2010 (0/1)			0.0116
			(0.0107)
Controls	Yes	Yes	Yes
Lat & Long	Yes	Yes	Yes
District fixed effects	Yes	Yes	Yes
Tribe fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
N	47827	47827	47827
R-squared	0.3884	0.7002	0.7002

Notes: The dependent variable is paved road dummy in 2010 in Column (1) and the natural log of value of production expressed in international price weights (calculated by the FAO in year 2000 US dollars per hectare, GK$/ha) in Columns (2)-(3). Standard errors (in parentheses) are clustered at the district level. *** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level.

Table 10: Cash Crop Share

	Ln(Real Yield (GK$/ha))	
	(1)	(2)
Rail within 20 km	0.0316***	0.0146**
	(0.0095)	(0.0071)
Cash crop share (%)		0.0149***
		(0.0016)
Controls	Yes	Yes
Lat & Long	Yes	Yes
District fixed effects	Yes	Yes
Tribe fixed effects	Yes	Yes
Year fixed effects	Yes	Yes
N	47827	47025
R-squared	0.7002	0.7386

Notes: The dependent variable is the natural log of value of production expressed in international price weights, calculated by the FAO in terms of year 2000 US dollars per hectare (GK$/ha). Standard errors (in parentheses) are clustered at the district level. ***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Table 11: Spatial Spillover

	Ln(Real Yield (GK$/ha))		
	(1)	(2)	(3)
Rail within 20 km	0.0316***	0.0310***	0.0265***
	(0.0095)	(0.0094)	(0.0092)
Euclidean distance (km) to nearest city in 2010		-0.0002	
		(0.0003)	
Luminosity (stable light)			0.0072***
			(0.0014)
Controls	Yes	Yes	Yes
Lat & Long	Yes	Yes	Yes
District fixed effects	Yes	Yes	Yes
Tribe fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
N	47827	47827	47827
R-squared	0.7002	0.7002	0.7014

Notes: The dependent variable is the natural log of value of production expressed in international price weights, calculated by the FAO in terms of year 2000 US dollars per hectare (GK$/ha). Standard errors (in parentheses) are clustered at the district level. ***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Table 12: Covariate Balance Check for Model (1)

	(1)	(2)	(3)	(4)	(5)
	Potential yield (kg/ha)	Share of class 1 soils in the cell	Share of class 2 soils in the cell	Mean elevation (m)	Average annual precipitations (mm) in 1900-1960
Rail within 20 km	-0.5601	0.0018	-0.0036	-5.5541	1.1734
	(5.5945)	(0.0035)	(0.0050)	(4.2255)	(1.5261)
District fixed effects	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes
Crop fixed effect	Yes	No	No	No	No
N	658786	24042	24042	24042	24042
R-squared	0.8450	0.5589	0.4010	0.9432	0.9901

Notes: Standard errors (in parentheses) are clustered at the district level. ***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Table 13: Covariate Balance Check for Model (2)

	(1)	(2)	(3)	(4)	(5)
	Potential yield (kg/ha)	Share of class 1 soils in the cell	Share of class 2 soils in the cell	Mean elevation (m)	Average annual precipitations (mm) in 1900-1960
Rail (1= actual)	-24.9448	0.0001	-0.0014	-46.8259**	-3.1530
	(19.6074)	(0.0072)	(0.0125)	(18.9854)	(13.4417)
Tribe x Year fixed effects	Yes	Yes	Yes	Yes	Yes
Crop fixed effect	Yes	No	No	No	No
N	622055	23010	23010	23010	23010
R-squared	0.8344	0.3933	0.2965	0.8777	0.9486

Notes: Standard errors (in parentheses) are clustered at the district level. ***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Table 14: Conley's standard errors

	Ln(Real Yield (GK$/ha))			Ln(Real Yield (GK$/ha))		
	(1)	(2)	(3)	(4)	(5)	(6)
Rail within 20 km	0.0329***	0.0322***	0.0316***	0.0329***	0.0322***	0.0316***
	(0.0078)	(0.0078)	(0.0019)	(0.0083)	(0.0082)	(0.0019)
Controls	No	Yes	Yes	No	Yes	Yes
Lat & Long	No	No	Yes	No	No	Yes
District fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Tribe fixed effects	No	No	No	No	No	No
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	47827	47827	47827	47827	47827	47827
R-squared	0.0017	0.0036	0.0043	0.0017	0.0036	0.0043

Notes: The dependent variable is the natural log of value of production expressed in international price weights, calculated by the FAO in terms of year 2000 US dollars per hectare (GK$/ha). Standard errors (in parentheses) are calculated using Conley's (1999) method with distance cutoff 50 in Columns (1–3) and 100 km in Columns (4–6). ***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

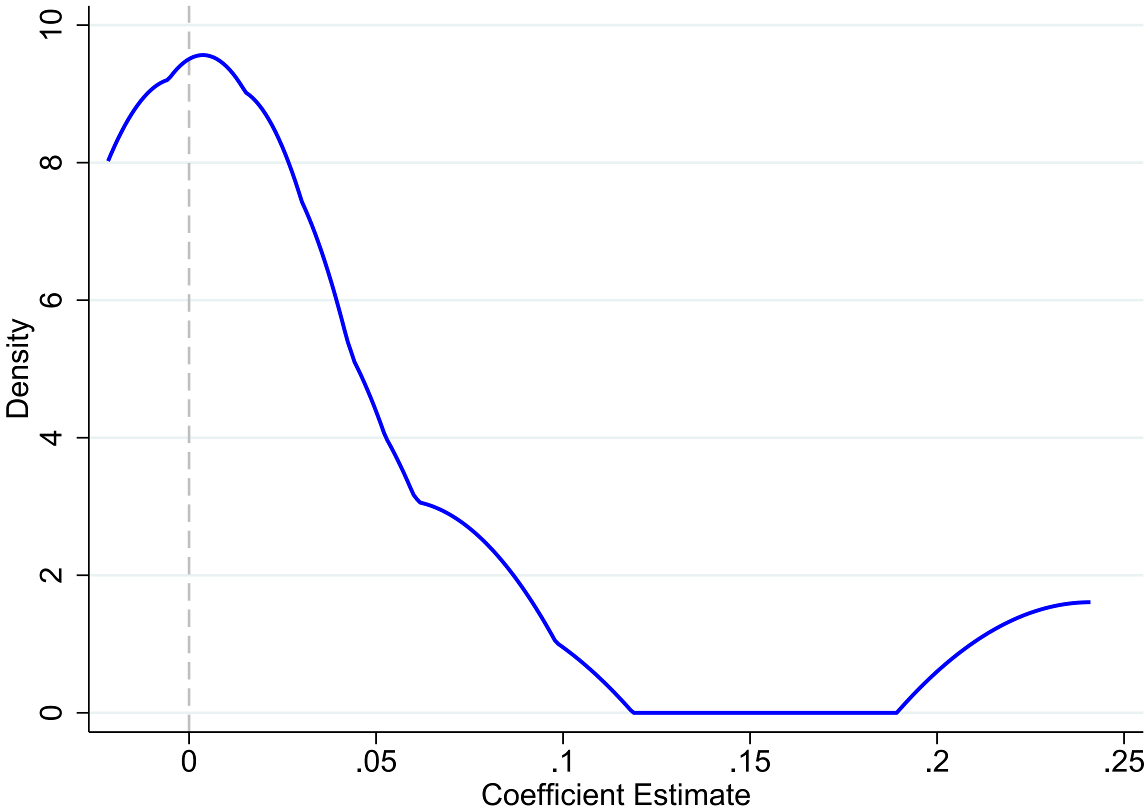

Figure 3: Density Plot of Country-level Coefficients

Appendix A

Global Agro-Ecological Zones (GAEZ) Database

The Global Agro-Ecological Zones (GAEZ) project (Fischer et al., 2021), developed by the FAO and IIASA, provides comprehensive data for assessing agricultural potential by standardizing climate, soil, and terrain characteristics that are essential for crop production. GAEZ estimates *potential yields* at a 5 arc-minute resolution by combining detailed micro-geographic information, such as soil properties, climate conditions, and topography, with crop-specific agronomic models, thereby capturing the maximum possible output per hectare under defined water supply and cultivation assumptions.

In contrast, the *actual yield* data for the years 2000 and 2010 is derived through a down-scaling approach, where aggregate national and sub-national production statistics are disaggregated to individual grid cells. This down-scaling process involves two main steps: first, calibrated cropland shares are compiled at a 30 arc-second resolution using the GLC-Share global land cover database and then aggregated to 5 arc-minute grid cells; second, crop-specific harvested areas, yields, and production figures are allocated to these grid cells based on geospatial data on land cover, soil, climate, vegetation distribution, and population density. The resulting product is a spatially detailed representation of actual yields and production for 26 major commodities, ensuring consistency with national and sub-national statistics while leveraging advanced down-scaling techniques implemented by the GAEZ team.