Chugging Along: Long Run Population Dynamics After Railroad Expansion in Mexico

Danielle Wilson*This work was funded by the Economic History Association

July -1, 2025

Abstract

This paper examines the long-run effects of 19th-century rail expansion in Mexico. For this analysis, a new series of digitized rail maps was created using numerous archival resources. With these maps, and harmonized locality-level population data, I assess how rail expansion altered the population distribution in Mexico and whether any resulting changes persisted into the 21st century. I apply the inconsequential units approach and use distance from a hypothetical least-cost path network to instrument rail access. I find that changes in the spatial distribution were persistent, but only for urban localities up to 1970. The lingering but concentrated effects of historical rail are also driven by lines constructed during a specific phase of expansion into Northern Mexico. This finding is notable given the destruction of lines during the Revolution and the network's increasing obsolescence and decline throughout the 20th century. The persistent influence of rail among urban localities abruptly disappears after 1970 and new infrastructure in the form of roads near rail becomes a strong determinant of population.

Keywords: Mexico, Railroads, Long-run, Population

JEL Codes: N76, N96

*Department of Economics, American University
danielle.wilson@student.american.edu
This research was supported by a Cambridge University Press Early-Stage Dissertation Grant from the Economic History Association. I am grateful for their generous support.

0.1 Introduction

Prior to the introduction of railroads, Mexico was primarily rural, and interregional travel was scarce due to precarious trails and limited waterways. After Mexico gained independence from Spain in 1821, war-related destruction and capital flight stunted economic growth for decades (Kuntz-Ficker and Tena-Junguito, 2018). Economic and social development remained slow throughout most of the 19th century (Donly, 1920; Meyer and Sherman, 1991; Sánchez, 2015), and population growth during the first half of the 19th century was less than 0.5 percent per year (McCaa, 2000). The expansion of railroads during the *Porfiriato*, the 35-year regime of Porfirio Díaz from 1876 to 1911, is credited with pulling Mexico out of its "de-industrialization phase" and connecting it to the global economy. This connectivity, in turn, helped national income levels recover to those of the colonial era (Coatsworth, 1978). However, the lasting effect of this pivotal infrastructure, which was heavily financed and controlled by foreign investors, on Mexico's population distribution and economic development remains unclear.

This study investigates whether the expansion of Mexico's rail network during the late 19th and early 20th centuries had an enduring and persistent impact on the distribution of people into the 21st century. I assess how rail access influenced population growth in the decades following expansion and over a century later, with a particular focus on urban localities and urbanization. I also investigate whether the effects were uniform or concentrated within specific regions.

To address these questions, I digitized a series of rail maps using a portfolio of cartographic records sourced from various archives. These maps are combined with locality-level population data, which were geographically harmonized from 1900 to 2020, to investigate whether localities near rail grew from their size in 1900 and how they fare a century later. Localities were harmonized or linked over time using a comprehensive catalog provided by Mexico's National Institute of Statistics and Geography (INEGI), which

documents all localities that ever existed from 1900 onward. This analysis is conducted at both the locality and grid cell levels, where the population of each grid cell is the sum of all locality populations it contains. The grid-cell analysis examines spatial patterns of population and density, while detailed locality-level data disentangle the types of localities driving spatially aggregate results.

Although transportation networks are historically associated with economic development (Redding and Turner, 2015), their effect may vary across space and depend on local geographic, cultural, and economic characteristics. Given the context of how and where Mexico's rail network expanded, it's thus far unknown whether only certain regions benefited in the long run from connectivity. Like in other Latin American and developing countries, rail expansion in Mexico was largely driven by the money, technology, and preferences of foreign investors (Wells, 1992; Kuntz-Ficker, 2000). In the case of Mexico, rail may have facilitated export specialization in raw goods and induced a "reversal of fortunes" in connected regions by leaving them reliant on foreign demand and unable to independently develop (Goldfrank, 1976; Haber, 1989; Coatsworth, 1981; Herranz-Locán, 2013). On the other hand, connectivity may have encouraged domestic demand, facilitated local linkages, and fostered industrialization and overall growth (Kuntz-Ficker, 1995).

This analysis finds changes in the spatial distribution of population across Mexico between 1900 and 1921 following rail expansion. Population density was higher near the network and primarily driven by increases in urban localities near lines that survived the Revolution, which lasted ten years and ended in 1920. Initially, the average marginal effect of rail access on density is concentrated within 20 kilometers of the nearest line. However, for urban localities, a positive association between proximity and population persists up to 60 kilometers from the network.

After a century, OLS results show a diminishing association between historical rail

access and locality population, particularly after 1970 when modern roads became more influential. However, transportation infrastructure is rarely placed randomly. Railroads were routed based on pre-existing geographic, political, or economic conditions, making access endogenous and OLS estimates potentially biased. To address this, I construct an instrument for rail access using the "inconsequential units" approach (Redding and Turner, 2015; Gonzalez-Navarro et al., 2023). Following related studies (Faber, 2014; Bogart et al., 2022), I use ArcGIS to generate a least-cost path (LCP) network that connects major historical cities and customs stations by minimizing distance and topographic slope. Distance to this LCP network serves as an instrument for rail access. To ensure exogeneity, I exclude from the estimation sample any localities that were used to construct the LCP or are located near these origin nodes. Because proximity to the LCP reflects network optimization rather than unobservable characteristics correlated with population, the instrument isolates plausibly exogenous variation. I test alternative sets of historical nodes to confirm the instrument's relevance and robustness. Using this approach, I find that rail access significantly shaped the spatial distribution of population, but only for urban localities and primarily before 1970. After 1970, distance to the modern road network becomes a stronger predictor of population, which is consistent with the growing importance of roads relative to rail during the mid-twentieth century.

Using two dozen cartographic records from various archives, I also observe the expansion of historical rail between 1874 to 1910 across ten time periods. I use this panel to test for heterogeneous effects from the timing of connectivity. My findings suggest that the persistent effects of historical rail for urban localities are largely driven by connections made during a particular phase during the *Porfirato*.

Comparable studies, including those that consider heterogeneity in transportation access in other countries, have found overall urban population dispersion and greater increases in population among cities that were initially smaller and more isolated (Jedwab

and Storeygard, 2021; Fenske et al., 2023). The Table A.1 in the Table Appendix provides an overview of recent work on population and historical infrastructure, including similar applications of the inconsequential units approach.[1] I similarly find that new and initially small urban localities in the north near rail had larger populations later on compared to initially larger, historical cities in the south.

This paper contributes to the expanding literature on the long-run effects of railroads (Jedwab et al., 2017; Berger and Enflo, 2017; Bogart et al., 2022; Banerjee et al., 2020; Jedwab and Storeygard, 2021; Fenske et al., 2023; Barsanetti, 2023b; Esteban-Oliver, 2023), particularly in Latin America (Pérez, 2018; Forero et al., 2021; Fajgelbaum and Redding, 2022; Marein, 2022; Américo, 2022). Although these studies are similar, contextual differences provide insight as to when and under what conditions history matters in the short and long-run (Lin and Rauch, 2022). This analysis adds to this literature and offers new empirical evidence that formally tests the impact of historical infrastructure on Mexico's population distribution into the 21st century.

The combination of natural endowments, initial conditions and political economy in Mexico make its setting unique and worth investigating. Due to the relatively barren terrain, the initial cost savings from rail was over six times greater than savings in the United States (Herranz-Locán, 2006), which had plentiful waterways and alternative modes of reliable transportation (Donaldson and Hornbeck, 2016). Although the geography of Mexico is similar to that of Spain, which too has few waterways and whose center is separated by extensive mountain ranges (Esteban-Oliver, 2023), it neighbored the relatively wealthy and developed United States. American investors, motivated by their own business interests and "expansion schemes" (Kuntz-Ficker, 2000), largely funded the network's development (Wasserman, 1979), were allowed to decide the timing and direction of expansion, and often extended existing American lines directly into Mexico (Kuntz-

[1]This table is an extension of the work in Table A.19 of Barsanetti (2023a).

Ficker, 2000). Consequently, rail lines, especially in the north, preceded demand, failed to prioritize existing urban centers, and were arguably ill-suited to meet domestic needs.

This work also contributes to the empirical literature on persistence and long-run economic outcomes. Moreover, it engages with the wider discussion on whether "first nature geographies" or natural characteristics (Caruana-Galizia et al., 2021), such as landscape and access to water, are deterministic, or if history can have permanent, path-dependent effects on development. The benefit of long-run analyses is that they provide insights into the processes and evolution of economies, and help explain why, for example, people are clustered where they are today or why certain regions took off while others lagged behind (Hanlon and Heblich, 2022). This work thus parallels the literature on the interaction of geography and history (Sokoloff and Engerman, 2000; Nunn and Puga, 2012; Dell, 2010; Dell, 2012).

So far, it is unclear whether history or natural endowments played a larger role in determining the current spatial distribution of people in Mexico. I thus investigate whether the dictatorship's most notable achievement, the successful financing and development of the rail system, had any persistent effect or redirected the long-run trajectory of people and cities across Mexico. In this context, persistence is the lasting influence of historical events, even after the original determinants of such persistence have diminished or no longer exist (Allen and Donaldson, 2022). For path dependence to then emerge, small initial advantages must unevenly accumulate and redirect the development of affected regions.

Infrastructure can provide relative spatial advantages that may remain influential for decades. In some cases, history has successfully altered the trajectory of long-run development through changes in trade, production, and agglomeration (Juhász, 2018; Bleakley and Lin, 2012; Jedwab et al., 2017). However, in other cases geographic advantages are deterministic, location fundamentals or "first order geographies" prevail, and the spatial

distributions of economies revert to their prior equilibrium (Davis and Weinstein, 2002; Miguel and Roland, 2011). Studies like this help empirically identify and test under what conditions history matters and when inter-regional investments can permanently redirect a country's development.

More specifically, this work contributes to the empirical literature on infrastructure investment and long-run interregional development in Mexico, and builds on the economic history literature on railroads (Coatsworth, 1981; Haber, 1989; Kuntz-Ficker, 1995). In regard to public infrastructure and growth, few studies have regionally compared the short-run effects of infrastructure on growth (i Font and Rodriguez-Oreggia, 2005; Chiquiar, 2005). These studies have found that regions with better infrastructure in the late 20th century fared better during Mexico's trade liberalization era, and that additional investments were correlated with reductions in inequality within richer, more industrialized regions.

The long-run population analysis in this paper is most closely related to that of Alix-Garcia and Sellars (**?**), who similarly construct a harmonized panel of Mexican localities to examine patterns of urbanization over the twentieth century. Other existing empirical work on long-run outcomes in Mexico has primarily focused on national trends or broad regional differences (Campos-Vazquez et al., 2017; Campos-Vazquez et al., 2014), rather than spatial heterogeneity at the subnational level. Other studies have investigated long-run economic divergence between regions in Mexico (Mendoza-Velázquez et al., 2020; Aguilar-Retureta, 2015; Serra et al., 2006), but have stopped short of identifying any potential explanatory sources or mechanisms. Some work has also explored whether history in the form of culture or institutions can account for patterns in long-run development. For example, Artega (2018) studies the long-run development of municipalities previously near pre-colonial indigenous communities, and Waldinger (2017) and Alvarez-Villa and Guardado (2020) focus on the effects of colonial Catholic missionary orders and (legal

and illegal) trading ports. I contribute to this literature by providing a long-run analysis of post-colonial development during Mexico's first industrialization era.

0.2 Historical Background

In 1837, the first concession to build a rail line connecting Mexico City to the port of Veracruz was authorized by the government of President Anastasio Bustamante (Pletcher, 1950). The line eventually spanned approximately 260 kilometers, and after numerous failed contracts and delays, was inaugurated 36 years later in 1873 by President Sebastián Lerdo de Tejada (Pletcher, 1950; Gilly, 2005). For comparison, the United States had begun railroad construction in the 1830s and between 1839 to 1859 had expanded the network by approximately 26,000 miles (Fogel, 1962).

Although the completion of the Veracruz line was deemed an engineering marvel and captured the attention of journalists and politicians, little progress was made in expanding the line or building new lines elsewhere (Matthews, 2017; Knapp, 1952). During Lerdo's presidency, there was strong criticism of foreign capital, especially from the United States, which had annexed Texas from Mexico a few decades prior as a result of the Mexican-American War. Lerdo himself was wary of American investors, favored European capital, and was against building rail lines that would directly connect Mexico to the United States (McNeely, 1964). Critics feared that American ownership, construction, and management of integral infrastructure could threaten Mexico's autonomy. The country's nationalist Congress subsequently failed to approve additional concessions to foreign investors (Matthews, 2014).

Díaz's primary objective during his first presidency from 1877 to 1880 was to change the global perception of Mexico as politically unstable. He worked to quell rebellions, eliminate unnecessary bureaucracy, and enforce taxation (Meyer and Sherman, 1991). Díaz had no uniform policy on railroads but once in office repealed a controversial Central

Rail contract, which was thought to be too generous to foreign investors (McNeely, 1964). During his first term, railroad expansion was initially decentralized. Individual states had the responsibility of negotiating concessions for lines that would later be incorporated into a national system (Garner, 2011). Under this structure, little progress was made in expanding the network and there was no national policy to supplement failed foreign contracts.

When Díaz returned to power in 1884, his primary focus was on modernizing Mexico via railroads. His new agenda heavily relied on capital from foreign backers who were now willing to invest in a more stable Mexico. Critics continued to object to foreign ownership of vital infrastructure and influence over domestic affairs. However, under increased political pressure from the United States, the Díaz government successfully rallied support for concessions by equating railroads to civilization and progress. Through extensive propaganda, the *Porfirian* regime emphasized the potential of an expansive railroad network to physically unite the country and bolster its national identity (Matthews, 2014). Regardless of how the project was funded, its success became synonymous with nationalism and economic development.

Although railroads provided the first reliable mechanism of inland travel, little evidence suggests that lines were intentionally built to support domestic growth. Regardless, the dictatorship claimed that the rail network would create an integrated national market and support domestic industrialization. In hopes of achieving these goals, politicians argued for the prioritization of east-west rail lines that would connect Atlantic and Pacific ports and potentially facilitate trade with Europe and Asia (Matthews, 2017). However, rail lines blatantly followed foreign (often American) interests, branched out from the capital to mines and ports, and often ran south to north towards the United States (Pindell and Mallis, 1997; Herranz-Locán, 2013). The network had expanded from less than 600 kilometers in 1873 to over 19,205 kilometers in 1910, yet construction of lines often pre-

ceded demand (Kuntz-Ficker, 2000) and lines were arguably ill-placed to serve domestic needs (Gilly, 2005).

By 1892 rail connections between Mexico City and key cities including eight state capitals had been established. By the turn of the century, railroads had subsequently centralized political power in Mexico City. However, José Limantour, a technocrat and Díaz's Secretary of Finance, later acknowledged that resulting rail network may have been poorly suited for fostering economic development. Limantour recognized that the federal government's concession policies were likely too reliant on foreign demand (Lewis, 2017), and as a result, the network was potentially better suited to serve investors' interests rather than domestic ones. In other words, the same market forces that facilitated rapid expansion and determined the pattern of railroads in Mexico (Coatsworth, 1981) may have failed to meet the needs of the people.[2]

By 1911, the order of the dictatorship had ended. The Díaz regime was followed by ten years of revolution with multiple counter-rebellions and split insurgent factions. Revolutionaries often used rail lines for counteroffensives either near or along the tracks (Donly, 1920; Haber et al., 2004), and rallied railroad workers to provide logistical and transportation support. The network was hit hardest during the height of the war from 1914 to 1917. From 1915 onward, railroad operations ceased except for those of a "destructive nature" (Hardy, 1934), as revolutionaries obstructed bridges, tore up tracks (Knight, 1986), and occasionally sent locomotives loaded with dynamite into enemy territory (Agredano, 2019). Without an alternative interregional transportation network, trade and the military's ability to deploy troops into revolutionary territory were largely inhibited. Industries reliant on the network, such as the textile, mining and manufactur-

[2]The Yucatán Peninsula may be an exception to this pattern. Wells (1992) argues that rail expansion in the Peninsula was distinct from that in the rest of Mexico. He notes that rail in this state was instead constructed and financed by local capitalists, who had close ties or were part of the henequen industry.

ing, were unable to continue operations without the delivery of inputs and unsuccessfully pleaded for federal help (Haber et al., 2004).

Rail companies went into debt trying to maintain freight operations and without a stable government in power, the system largely fell into in disarray. Parts of the network were restored after the Revolution and industries dependent on rail successfully achieved or outpaced prewar production (Haber et al., 2004). However, the highway system built during the late 1920s provided an economical substitute for long-distance inland transport. Although the rail industry was fully nationalized by the late 1930s, the system was unable to compete with new roads. Administration of the rail network was chaotic, control over the system passed from one organization to the next, and operations were reliant on credit. After numerous deadly accidents and significant income losses, the rail network effectively collapsed by 1940 (Michaels, 1970).

McCaa defines the period between Mexico's independence in 1821 and the start of the Revolution in 1910 as the "national" era (McCaa, 2000). During this time, population growth initially stagnated but began to accelerate shortly before the *Porfiriato*. In the decade after independence, annualized average growth rates in Southern and Central Mexico, the most populous regions in the country, were negative. These rates began to reverse and were once again positive even before Díaz's first presidency, ranging from 0.1 percent in Southern Mexico to 2.2 percent in Central Mexico. After Díaz assumed power, Mexico's population grew more rapidly, increasing by 32 percent from 11.5 million in 1890 to 15.2 million by 1910, the year the Revolution began (McCaa, 2000). However, population growth during the *Porfiriato* was far from uniform. Between 1880 and 1910, annualized average growth rates exceeded two percent in Northern Mexico. During the rail boom from 1890 to 1900 (Figure A.1 in the Figure Appendix), states within region, such as Durango, Chihuahua, and Coahuila experienced annualized growth rates of 3.4, 3.2, and 4.8 percent, respectively. This study investigates the impact of rail expansion

on the distribution of people across Mexico, and whether the network played a persistent role in driving the uneven regional growth that emerged during this period.

0.3 Data

0.3.1 Creating the Rail Network Database

The rail network series constructed for this analysis references a variety of historical and modern sources. Dozens of maps were collected from archives and online libraries dating from 1874 onward. Archival maps of Mexico were collected from *El Centro de Documentación e Investigación Ferroviarias* (CEDIF) and the National Archives Records Administraction at College Park (NARA). Additional maps were sourced from online libraries including David Rumsey's Historical Map Collection and the Historical Map and Chart Collection of the Office of Coast Survey at National Oceanic and Atmospheric Administration.

The Data Construction Appendix contains a detailed description of how each archival source was used to identify and document changes in the network over time. I was able to digitized rail maps from 1874, 1878, 1882, 1884, 1886, 1892, 1896, 1900, 1904 and 1910. Unfortunately, unlike in other studies (Bogart et al., 2022; Barsanetti, 2023b), rail stations could not be consistently identified thus access is defined using distance to the nearest rail line. Figure 0.1 illustrates the evolution of the rail network in 1874, 1884, 1892, and 1910.

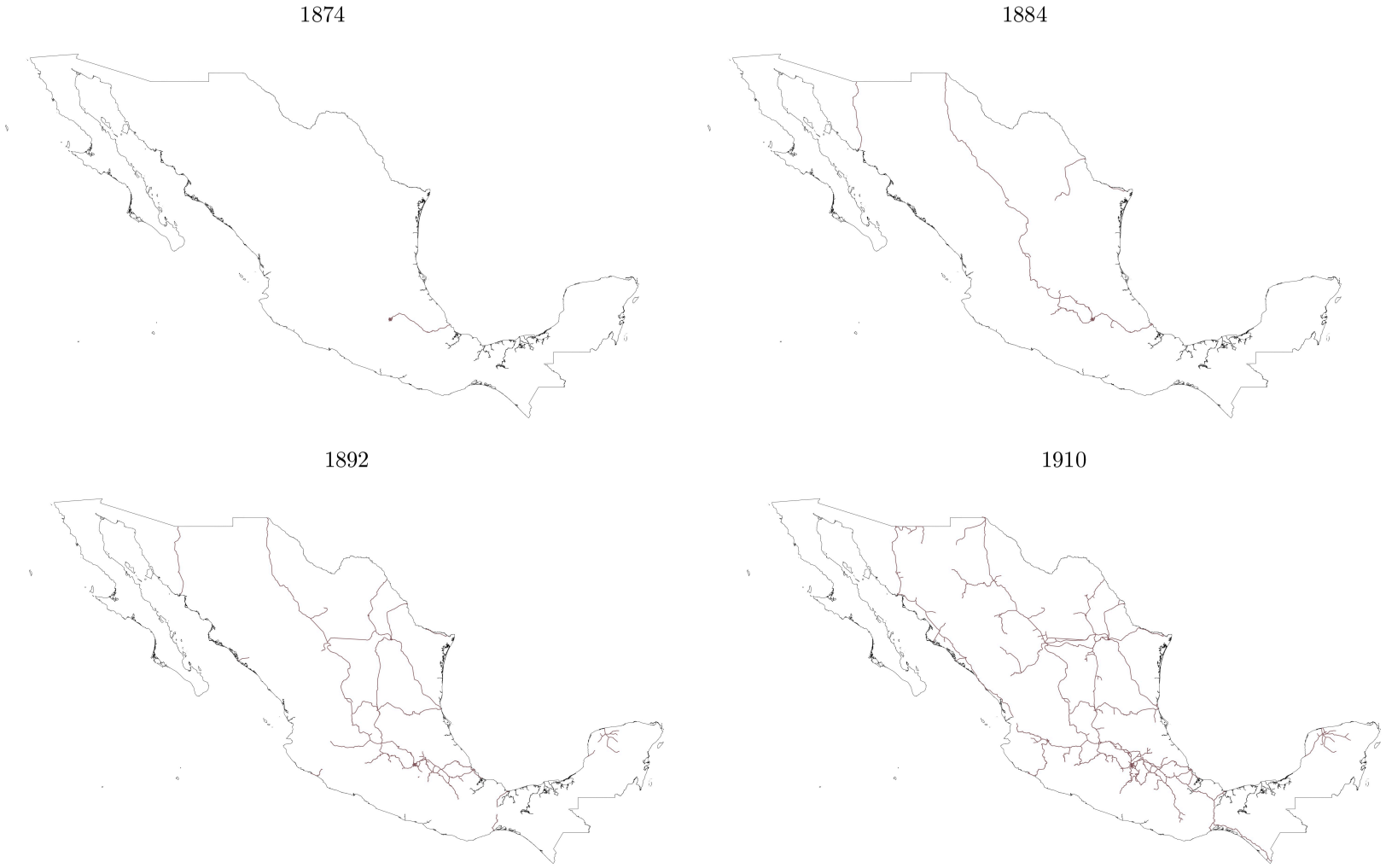
Modern, digitized maps in the form of shapefiles were used as a base for constructing the network. By using modern maps in this way, I implicitly assume that modern lines do not significantly differ from their original or prior placement as the cost of moving or changing a line connecting the same two locations would be high. Modern rail files were sourced from MEXLIST, [3] the United States Geological Survey, *Instituto Nacional*

[3]MEXLIST is a community of railroad scholars, researchers and enthusiasts who discuss and study

de Estadística y Geografía (INEGI), OpenStreetMap and Esri.

multiple facets of Mexico's railroad network. Doug Spurling and MEXLIST freely provide digital maps for users, some of which contain historical details and references.

Figure 0.1: Expansion of Rail from 1874 to 1910



For details on how this series was constructed and additional years within the rail panel, see Appendix Data Construction Appendix. Larger maps illustrating the extent of rail in all observed years from 1874 to 1910 are provided in this appendix.

0.3.2 Unit of Analysis

Due to data limitations, previous work on railroads often relies on municipal population data or use aggregated population estimates at the grid cell level. This study, however, benefits from granular and disaggregate population data at the locality level. Although locality-level data are rich, the information associated with them is spatially concentrated into a single point. In reality, population spans physical space. The space covered by localities varies in density and size, and often bleeds into or intersects neighboring localities, especially in metro areas (Blackman et al., 2024). Most locality coordinates are provided by INEGI, but some have a margin of error due to changes in locality size or political boundaries over time. Localities without INEGI-provided coordinates may have a larger margin of error, as their locations were estimated using their name and corresponding municipality and state information. These coordinates are approximated from publicly available geospatial data but have not been validated by a statistical organization.

To account for these limitations and provide comparable estimates to other studies, I aggregate locality-level data to the grid cell level and re-estimate results accordingly. Grid cells are 0.1 x 0.1 decimal degrees in size and about 17 thousand span Mexico. There are on average about 70 localities within each grid cell but as shown in Figure 0.2, the variation between grid cells is large (i.e., the standard deviation is approximately 75 localities).

To differentiate coastal and landlocked areas, grid cells are weighted by the proportion of land they contain. Fully landlocked cells are assigned a weight of one, while coastal cells receive a weight less than one, based on the proportion of land within the cell. This weighting prevents coastal cells, which are largely comprised of water, from disproportionately influencing results related to land-dependent characteristics like population, as people cannot reside or agglomerate in the ocean.[4]

[4]For context, about 5.8 percent of all grid cells contain more ocean area than land area. Baseline

There are various benefits to using grid cells. First, population is not assumed to be concentrated into a single point. Second, given that all cells are of the same size, they can also serve as a more intuitive unit of analysis when discussing density. Lastly, unlike municipalities, the boundaries of cells are not politically determined and they remain the same over time.

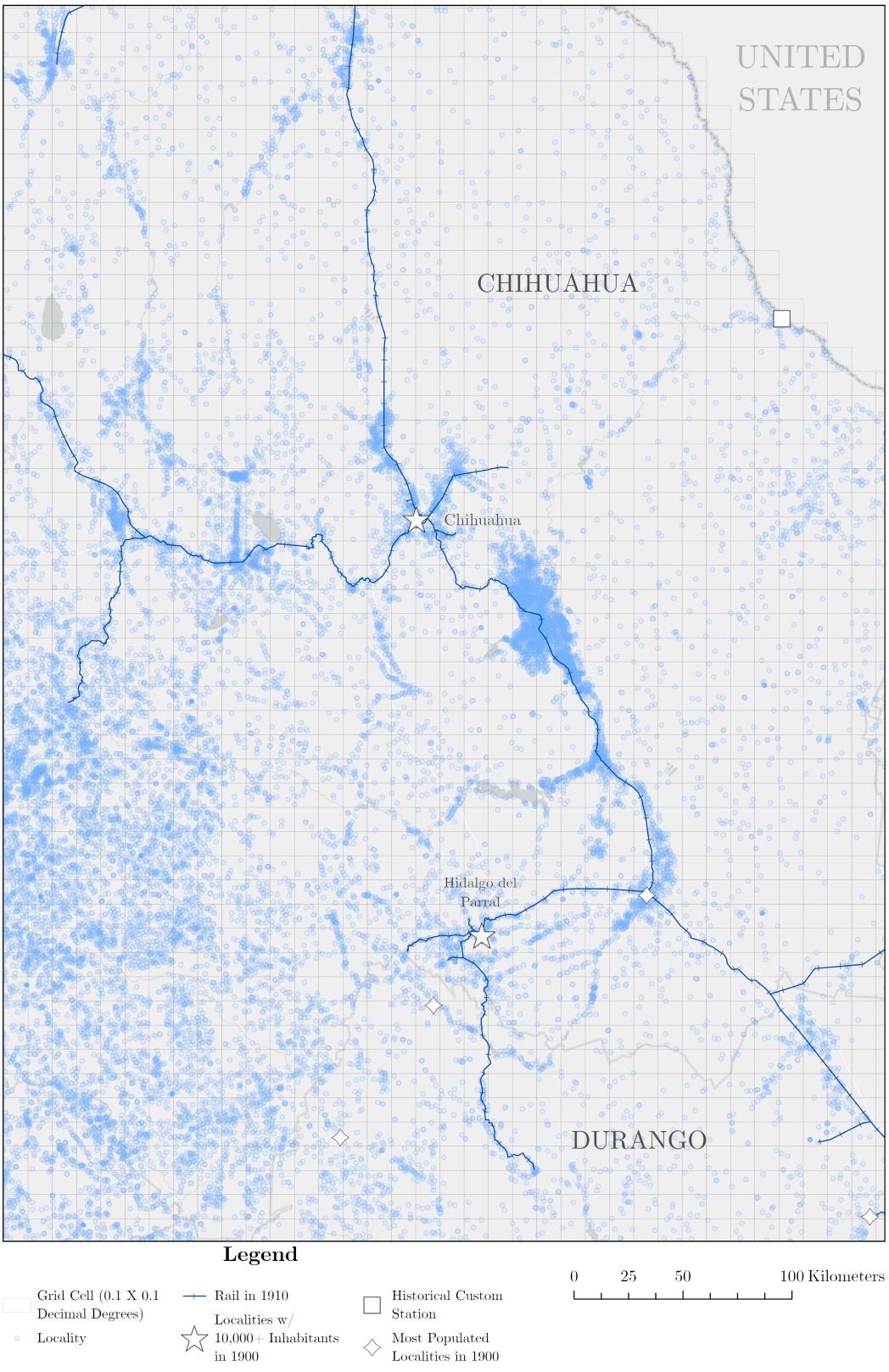
There are, however, several limitations. As with any polygon-based spatial unit, cells are defined by sharp boundaries that do not account for communities that may span multiple cells. In this analysis, a cell's population is the total number of people associated with each coordinate point within that cell. Even if a locality spans multiple cells, its entire population will only be assigned to the cell containing its associated coordinates. Results may thus be sensitive to the size of grid cells. Cells of varying sizes were tested to consider the influence of boundaries and results remained consistent.

Throughout the analysis, results at both the locality and grid cell level are provided. Results at the locality level do not replace, supersede, or contradict those at the grid cell level. The two units of analysis are meant to provide different perspectives and can be interpreted together.

---

results were not sensitive to this weighting.

Figure 0.2: Localities, Grid Cells and Historical Rail in Chihuahua, Mexico



This map is of Northern Mexico, focusing on the states of Chihuahua and Durango. Grid cells (0.1 × 0.1 degrees in size, approximately 11 × 11 at the equator) are overlaid along with localities, represented by blue circular markers. Blue lines show the historical rail network as of 1910. Diamond markers indicate the top 1% of the most populated localities, with star markers highlighting a subset with at least 10,000 inhabitants. Historical customs stations are marked with squares.

0.3.3 Population Data from 1900 to 2020

The primary outcome of interest is population. In this context, population estimates indicate where people, and inherently firms, tend to agglomerate or cluster across space. A key determinant of agglomeration is transportation costs; populations converge toward regions where these costs are relatively low (Redding and Turner, 2015; Krugman, 1991). Lower transportation costs attract firms by expanding market access and reducing input costs, thereby increasing local labor demand. In response, population or labor supply in connected areas may increase to meet the growing demand. Labor then can continue gravitating toward local economies with large manufacturing sectors (Fujita et al., 1999), sustained productivity (Rosen, 1981), or productive amenities (Roback, 1982). Thus, population, as a measure of labor supply and a key input for production, can serve as a proxy for economic activity.

Furthermore, during the 19th and much of the 20th century, Mexico can be characterized as being in the post-Malthusian regime (see Table A.3 in the Table Appendix). In this context, population density reflects economic activity, as gains in productivity are absorbed through population growth rather than increases in income per capita (Galor, 2005). Population is also often one of the few historically available estimates of economic activity. In Mexico, more direct measures like output, exports, or the number of firms are typically available only at the state level. With only 32 states, population remains the most disaggregated variable throughout the 20th century that reflects economic activity.

A population panel at the locality level was created using census data from 1900 to 2020. INEGI defines localities as areas that can be uniquely recognized by name and contain at least one dwelling (INE, 2019). Localities were historically identified as being, for example, either a congregation, mine, ranch, village, or city. Although rail expansion and the dictatorship of Porfirio Díaz began in the 19th century, few census publications exist before 1900. Among publications that do exist before 1900, population estimates

are rarely reported at the locality level and are instead inconsistently provided at more aggregate political levels. Consequently, the time span of the locality panel used for this analysis is limited to 1900 onward. It's also important to note that any analyses using data from the 1900, 1910 and 1921 Census should be interpreted with caution. The accuracy of population data before 1930 has been questioned due to the inconsistency of aggregate estimates and the unexpectedly large decrease in population from 1910 to 1921 after the Mexican Revolution (McCaa, 2003).

At the start of the 20th century, population data were collected by the *Dirección General de Estadística* (DGE), the first federal entity under the Ministry of Development responsible for collecting and publishing national statistics on topics ranging from population to agricultural and industrial production (Medeles, 2016). The DGE fostered relationships with local politicians to oversee the collection of data by field representatives. These representatives were meant to be local "well accepted" members of their community and were responsible for canvassing all homes, regardless of the respondent's literacy (García, 2022). These practices, along with the level of detail in the data, suggest that each census was designed to be exhaustive. A balanced panel can thus be created under the assumption that localities with no population data are assumed to either not yet exist or be uninhabited, and thus were assigned a population of zero.

Data from early publications were initially transcribed by state each decade and harmonized using information from Mexico's *Encyclopedia of Municipalities*. During this project, INEGI continued to publish digitized locality population data with unique identifiers, including those for the 2020 decennial. These unique numeric identifiers consist of state, municipal and locality codes, but are not constant over time. If a locality's boundary or political status changed at any point between 1900 and 2020, INEGI assigned it a new unique locality identifier. The creation and assignment of all new identifiers is cataloged in an index within the database of localities as a "movimiento" or "movement".

Movements are often documented alongside details on the date and corresponding laws that initiated changes in political boundaries or locality status.

Although population data were available for each census year, changes in locality boundaries and identifiers over time made it difficult (or impossible) to track the same localities across decades. Ignoring these changes would have incorrectly suggested locality births or deaths that were, in reality, reclassifications or boundary shifts. To address this, I used INEGI's index of "movimientos" or "movements" (i.e., locality-specific changes in political boundaries) to extract and interpret qualitative descriptions of changes, and constructed a crosswalk of identifiers spanning the sample period.

This harmonization process involved extensive parsing of unstructured text using regular expressions and lexical pattern recognition; systematic recovery of ambiguous or incomplete identifiers; and manual review of unmatched localities. The result of this process was a crosswalk linking original and new locality identifiers. When necessary, new aggregate identifiers were created to form the most stable geographic units possible, a process frequently applied to localities in the Distrito Federal, now Mexico City. The resulting crosswalk was then applied to the full population dataset to harmonize localities consistently over time. Figure A.2 in the Figure Appenidx summarizes the steps used to process and harmonize the raw data files from INEGI into a usable locality-level population panel.

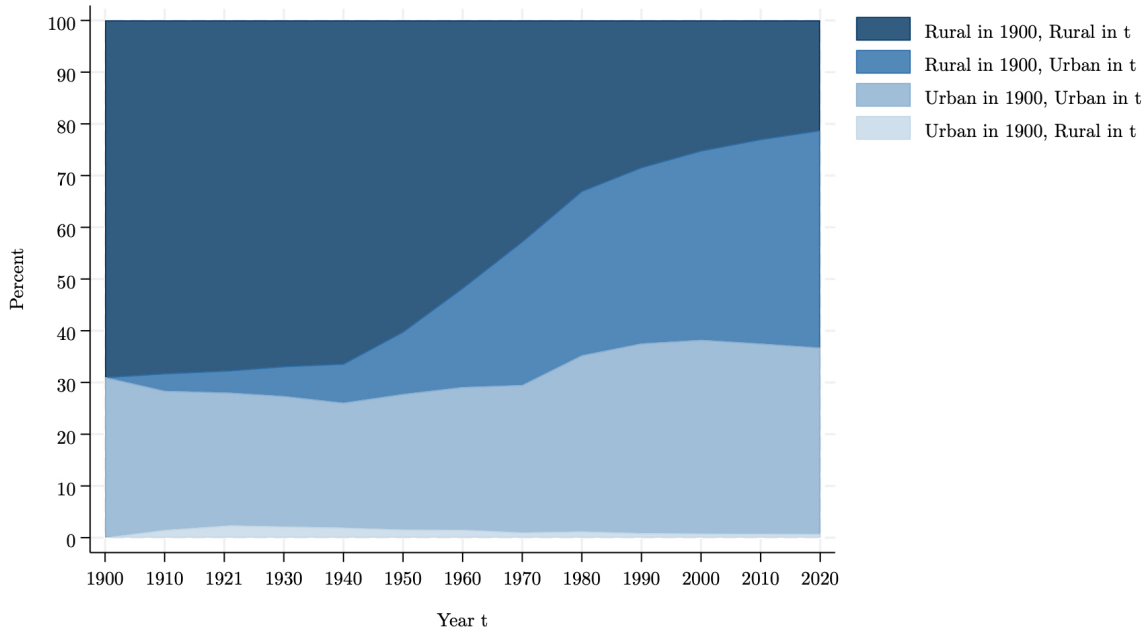
The resulting harmonized dataset includes approximately 350,000 unique localities. About 85 percent of these localities had coordinates provided by INEGI's locality database. For the remaining localities, I used GeoPy, a Python-based geocoding application, to approximate their locations based on name, municipality, district (when available), and state information.[5]

[5]Through GeoPy, I use the ERSI ArcGIS API geocoder to gather latitude and longitude coordinates.

In 1900 there were approximately 32,000 inhabited and documented localities. By the 21st century, there were almost 200,000. Figure A.3 in the Figure Appendix maps the location and population of localities in 1900 and 2020. Most of these localities were clustered around Central Mexico. However, throughout the century, the spatial distribution of localities changed, with population sizes increasingly varying across regions. In 1900, almost all localities had fewer than 10,000 inhabitants (in orange). By 2020, large urban localities emerged in northern Mexico, beyond where population and cities were historically clustered.

For this analysis, I follow INEGI's definition of an urban locality as one with at least 2,500 inhabitants (INE, 2019). Figure 0.3 shows Mexico's rural-to-urban transition from 1900 to 2020, categorizing localities into four groups: those that (1) were urban in 1900 and remained so, (2) were urban in 1900 but later declined, (3) were rural in 1900 but became urban, and (4) remained rural throughout.

Figure 0.3: Population of Urban and Rural Localities Over Time



Urban localities are those with at least 2,500 inhabitants. Only the sample of inhabited localities was used in each decade. Source estimates can be found in Table A.4 of the Table Appendix.

In 1900 and 1910, about 70 percent of the population lived in rural localities. By 1921, urban population levels remained stable at roughly 30 percent due to an increasing share of people in new urban localities, offset by a decreasing share in preexisting ones. After 1940, urban growth accelerated, led by the emergence of new urban localities. By 1950, half of the population lived in urban areas, one-third of whom had resided in rural areas in 1900. By 2020, nearly 80 percent of the population lived in urban localities, less than half of which were urban at the start of the century.

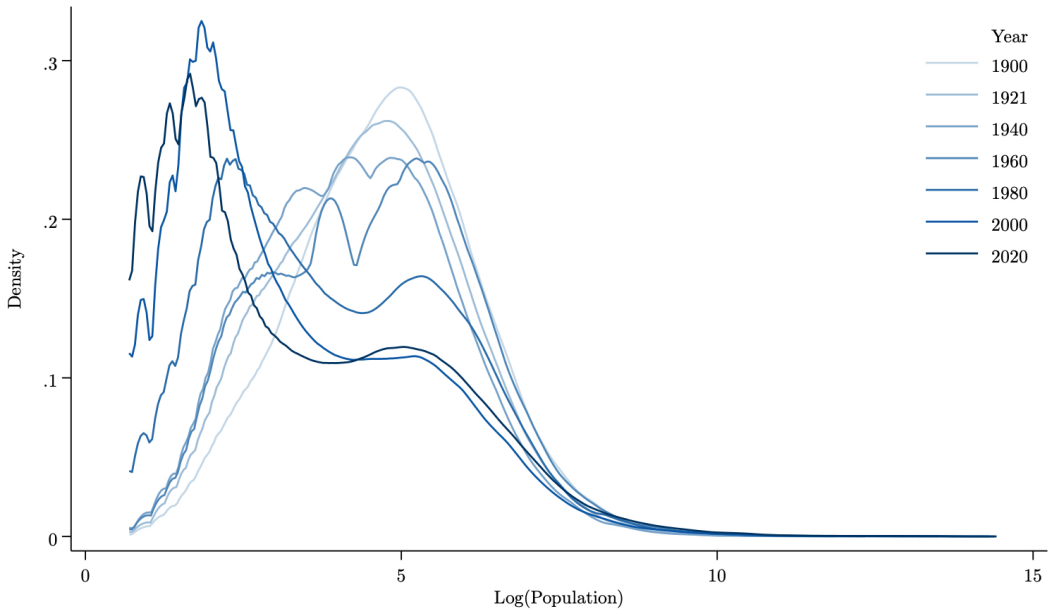
Over time, population became more unevenly distributed. From 1900 to 2020, the average locality size nearly doubled while the median fell, and the coefficient of variation rose sharply after 1940. Figure 0.4 shows that the distribution of log population widened, shifted left, and became increasingly bimodal. The left peak, representing small rural localities, grew more prominent, while the right peak, corresponding to larger urban settlements, became flatter after 1980. These changes reflect both the expansion of rural marginalization and increasing concentration among urban localities.

Figure 0.5 replicates this analysis using grid cells, which provide a second unit of analysis for measuring spatial dispersion. The bimodal pattern persists but is less pronounced. The right (urban) peak remains higher throughout, and the overall distribution flattens in 2020, suggesting increased dispersion in density.

Figures 0.4 and 0.5 present patterns of uneven growth at the locality level alongside increased dispersion in population density at the grid cell level. This combination suggests that over the 20th century, urban growth has spread beyond a few centralized cities, fueling the concentration of people around multiple urban centers. The growth and densification of historically significant primary cities, along with emerging secondary cities, likely played a role in shaping the bimodal distribution observed in the locality-level density plots.

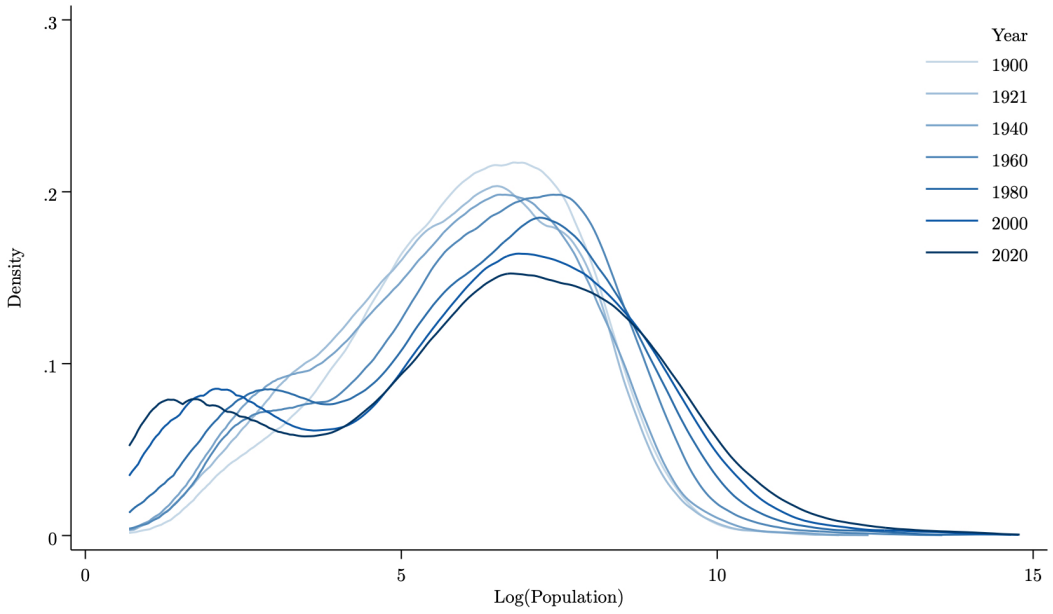
At the same time, urban agglomeration appears to be driving shifts in population toward and around these urban hubs, likely resulting in the broader dispersion patterns

Figure 0.4: Kernel Density of Log Population from 1900 to 2020 at the Locality Level



See Table A.2 in the Table Appendix for corresponding summary statistics. Densities use the Epanechnikov kernel with bandwidth selected to minimize mean integrated squared error.

Figure 0.5: Kernel Density of Log Population from 1900 to 2020 at the Grid Cell Level



Corresponding summary statistics are provided in Table A.5. Densities reflect inhabited cells and use the same kernel specification as in Figure 0.4.

observed at the grid cell level. Differences in patterns between localities and grid cells may arise because the former capture population at specific points, while the latter capture population density across space. The point-based measure of population at the locality level may not capture urban sprawl or expansion, especially around newly urbanizing areas. This is evident in the grid cell density plots, which remain centered but flatten over time, suggesting that population has dispersed more evenly across space, even as growth remains uneven at the level of named localities. This paper investigates the extent to which historical transportation infrastructure contributed to, or perhaps drove, these observed spatial patterns.

0.3.4 Historical and Geographic Controls.

To account for historical, environmental, and economic confounders, I include a set of additional controls measuring proximity to colonial trade routes and ports, mineral resource deposits, climatic conditions, Revolutionary-era insurgency, and pre-1900 indigenous settlement patterns, which I use as a proxy for initial population before the 20th century. These variables are defined in detail in Appendix A.3. Their inclusion helps ensure that estimated relationships between rail access and population are not driven by omitted historical or geographic factors.

Summary statistics associated with these variables at the locality and grid cell level are shown in Table A.6 in the Table Appendix. Continuous variables, such as distance to the nearest ocean and average caloric index for maize, generally have similar distributions across both units. However, the average number of indigenous communities is a discrete variable that is nearly five times larger at the locality level than at the grid cell level. This difference likely reflects the variation in unit size and the correlation between the pre-1900 population proxy and the locations of 20th-century localities.

0.4 Main Empirical Specification

To account for potential confounding effects, I use the following specification to test the short-run relationship between rail and population in each decade:

$$\ln(\text{Population}_{g,t} + 1) = \alpha_t + \beta_\tau Rail_g^\tau + X_g'\delta_t + \epsilon_{g,t} \quad (1)$$

where the dependent variable is the logged population[6] of geography g, either a locality or grid cell, at decade t, from 1900 to 2020. The primary variable of interest is R_g^τ, historical rail access in τ, a period of rail expansion observed between 1874 to 1910.

A geography is considered to have "rail access" if it's at most 10 km away from the rail network. In specification (1), $Rail_g^\tau$, thus captures whether geography g was within 10 km of the nearest rail line as the network in period τ. When using grid cells, distance to the nearest rail line is measured from the center of each cell. The 10 km cutoff for rail access is based on Coatsworth's (1979; 1981) cost estimates from the 19th century, which suggest that individuals traveling on foot averaged about 30 km per day. A 10 km indicator thus represents a reasonable distance that could be traveled to and from a rail line within a single day.

For robustness, rail access is defined using two alternative measure of rail access based on distance to the nearest rail line. The first measure categorizes distance into 10-km intervals.[7] While dichotomizing continuous variables generally leads to information loss (MacCallum et al., 2002), doing so may help capture the varying impact of distance

[6] As shown in Figure A.5 in the Figure Appendix, population is highly right-skewed at both the locality and grid cell levels. Applying a log transformation (after adding one to account for uninhabited areas) yields a distribution that is closer to normal and more linearly related to log distance to rail. For consistency, specification (1) is also used to estimate the relationship between rail access and locality population from 1900 to 1921.

[7] Indicators for distances up to 100 km were initially tested, but issues with multicollinearity and reduced sample size arose, particularly within the urban locality sample. The number of categories was therefore reduced to capture distances up to at least 30 km.

without assuming linearity. The second method uses log distance to the nearest rail line.[8]

The vector X_g captures the characteristics discussed in Section 0.3 and detailed in Appendix A.3. Baseline estimates include geographic characteristics such as such as latitude, average daily air temperature, average monthly precipitation, distance to the nearest river, distance to the nearest ocean, distance to the nearest mineral deposit, and average caloric suitability for various crops. Other variables accounting for locations of economic and cultural importance are also considered, such as distance to the U.S. border, to the nearest historical customs station, to the nearest point along the Camino Real, and to the nearest insurgent outbreak at the start of the Revolution in 1910. When estimating the relationship between rail and population from 1910 onward, log population in 1900 is included as a covariate to control for potential confounding effects between modern cities, early urban clusters, and rail expansion.

Table 0.1 compares average characteristics between localities with and without rail access in 1910 as defined by the 10 km indicator. Localities and grid cells with historical rail access had more indigenous communities, were closer to the Camino Real and early insurgent outbreaks, and were farther from historical customs stations. These locations also tended to be closer to various mineral resources, farther from the Pacific Ocean, and in regions with lower average air temperatures and less precipitation. These trends are generally consistent at the grid cell level, with some exceptions. For example, grid cells with rail access were situated at higher average elevations and in regions less suitable for cotton production. Meanwhile, localities with rail access were not at statistically different elevations than those without access and were in regions equally productive for growing cotton.

---

[8]Figure A.4 in the Figure Appendix shows that the distribution of distance to the nearest rail line is initially right-skewed but becomes less skewed over time. Figure A.6 confirms that the distribution of log-transformed distance to rail approximates a normal distribution, particularly in later years.

Table 0.1: Descriptive Statistics by Rail Access in 1910

	Using Localities					Using Grid Cells				
	No Rail Access in 1910		Rail Access in 1910			No Rail Access in 1910		Rail Access in 1910		
	Mean	SD	Mean	SD	Diff	Mean	SD	Mean	SD	Diff
Number of Pre Colonial Indigenous Communities	0.60	1.74	1.94	4.41	1.348**	0.13	0.59	0.51	1.51	0.380*
Elevation (km)	0.93	0.80	1.12	0.91	0.194	0.98	0.80	1.20	0.87	0.229*
Average Daily Air Temperature (in Celsius)										
1901	21.21	3.48	20.19	3.88	-1.01*	20.47	3.49	19.85	3.69	-0.62
2020	22.80	3.33	21.78	3.81	-1.02*	22.12	3.36	21.46	3.61	-0.66
Average Monthly Precipitation (in Millimeters)										
1901	968.49	573.47	790.33	485.05	-178.17***	720.79	532.90	671.99	417.33	-48.80
2020	1,000.45	554.97	835.64	464.18	-164.81***	760.38	541.84	720.32	423.92	-40.06
Average Caloric Index (in Thousands)										
Bananas w/ High Inputs	6.48	3.21	5.94	3.57	-0.540	5.81	3.88	5.63	3.93	-0.179
Chickpeas w/ High Inputs	5.86	4.75	5.92	4.90	-0.062	7.41	5.37	6.72	5.34	-0.692
Coffee w/ High Inputs	18.34	11.72	15.85	11.70	-2.493*	13.54	12.99	15.01	12.23	1.469
Cotton w/ High Inputs	4.39	2.38	4.39	2.60	-0.004	5.00	2.48	4.27	2.83	-0.727*
Maize w/ Low Inputs	8.48	3.87	7.82	3.68	-0.655	6.65	4.70	7.14	4.43	0.485
Distance to the Nearest (in Hundreds)										
Intermittent River (km)	0.52	0.53	0.46	0.48	-0.059	0.47	0.50	0.48	0.53	0.014
Perennial River (km)	0.39	0.82	0.32	0.59	-0.068	0.55	0.82	0.44	0.69	-0.107
Pacific Ocean (km)	2.68	2.33	3.17	2.18	0.481*	3.00	2.39	3.45	2.13	0.456
Gulf of Mexico (km)	4.86	4.69	4.14	3.39	-0.715	6.21	4.86	4.62	3.54	-1.597**
Coal Deposit (km)	3.96	2.55	3.22	2.03	-0.741**	3.57	2.59	3.18	2.16	-0.387
Copper Deposit (km)	1.01	1.49	0.85	1.18	-0.157	0.99	1.50	0.88	1.37	-0.105
Iron Deposit (km)	1.38	1.48	1.20	1.14	-0.183	1.37	1.48	1.23	1.30	-0.147
Zinc Deposit (km)	1.24	1.54	0.88	1.03	-0.361*	1.14	1.47	0.83	1.17	-0.307*
Historical Gold Mine, Wide Distribution (km)	3.08	2.29	2.36	2.02	-0.720***	2.91	2.15	2.40	1.95	-0.503*
Historical Gold Mine, Moderate Distribution (km)	2.14	2.33	1.61	1.77	-0.535*	2.04	2.27	1.63	1.89	-0.416
Historical Gold Mine, Limited Distribution (km)	3.34	1.98	2.95	1.67	-0.391*	2.87	2.05	2.78	1.78	-0.090
Historical Silver Mine, Wide Distribution (km)	2.35	2.11	1.66	1.64	-0.694**	2.25	2.06	1.69	1.75	-0.556**
Historical Silver Mine, Moderate Distribution (km)	2.31	2.26	1.50	1.65	-0.804***	2.08	2.20	1.45	1.75	-0.626**
Historical Silver Mine, Limited Distribution (km)	6.50	3.92	5.17	3.33	-1.335**	5.77	3.76	4.77	3.29	-0.999*
U.S. Border (km)	6.70	3.48	6.03	3.12	-0.668	5.39	3.60	5.37	3.05	-0.026
Historical Customs Station (km)	1.30	0.84	1.64	1.12	0.343**	1.50	0.89	1.78	1.10	0.272*
Camino Real (km)	4.12	3.30	2.68	2.58	-1.448***	4.11	3.23	2.70	2.55	-1.415***
Early Insurgent Outbreak (km)	1.39	1.19	1.08	1.09	-0.305*	1.38	1.20	1.08	1.13	-0.302***
N	246,848		97,621			14,309		2,451		

Rail access in 1910 is measured as an indicator and captures whether a geographic unit, g, or locality was within 10 km from the nearest line. For details on how characteristics are defined and on their data sources, see Appendix A.3. For locality statistics, the number of pre-colonial indigenous communities corresponds to those within a 10km radius of each locality's coordinates. Average precipitation, air temperature and caloric suitability for each product is the average within the same size radius. Standard errors used to test for significance differences are clustered at the state level for the locality sample. Standard errors used to test for significance differences are adjusted for spatial dependence following Conley (1999) using a distance cutoff of 300km. Grid cells are 0.1×0.1 degrees (i.e., approximately 11×11 km at the equator). Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

This analysis assumes that population responses to rail access or changes due to rail expansions are unlikely to occur instantaneously. Therefore, the relationship between rail and population is measured using a time lag on rail access. This lag also accounts for any uncertainty about the exact timing of expansions relative to census years, ensuring that rail access predates recorded population. This lag is approximately ten years but varies due to inconsistent intervals in the rail data, where time between observations ranges from two to six years.

Results from specification (1) are thus divided into two parts. The first part focuses on the relationship between historical rail access and population during the first two decades of the 20th century: the association between rail access prior to 1900 and population in 1900, rail access before 1910 and population in 1910, and rail access in 1910 and population in 1921. Focusing on this period allows me to test for short-run responses to rail access during and at the end of the dictatorship, when the network was still expanding, though at a slower pace than in the 19th century.

The second part extends to the remaining decades of the 20th century and early 21st century, studying the association between historical rail access as of 1910, at end of the *Porfiriato* and the final observation in my rail panel, and population. This long-run analysis focuses on how the historical rail network as of 1910,which largely resembles the extent of the network today, influenced spatial population patterns from 1921 onward. Using 1910 rail access as the reference point corresponds to a significant historical turning point, offers a consistent definition of treatment across all localities, and allows me to examine the persistence of rail-driven effects after the network was largely completed.

0.5 Rail Access and Population, 1900 to 1921

0.5.1 Grid Cell Level Results, 1900 to 1921

Table 0.2 presents grid cell-level estimates using ordinary least squares (OLS), with the rail access indicator as the primary explanatory variable. Conley standard errors (Conley, 1999) that account for spatial correlation within a 300 km radius are reported alongside regression estimates. These standard errors account for spatial correlation in the regression error terms across neighboring areas and broader regions. This adjustment typically leads to wider confidence intervals and reduces the risk of overstating statistical significance when nearby areas are affected by similar unobserved factors. Alternative distance cutoffs were also tested and resulting standard errors associated with rail access in 1910 are provided in Table A.11 of the Table Appendix.

Specification (1) regresses log population solely on rail access from roughly a decade prior. R-squared results show that between 1900 and 1910, rail access can account for approximately 4 to 8 percent of the variation in log population. Specification (2) adds pre-20th century population proxies as covariates, which collectively reduce the rail access coefficient and, based on R-squared values, account for a large portion of the variation in log population between 1910 and 1921. Specification (3) includes latitude to control for underlying spatial trends (Kelly, 2022), and specification (4) introduces additional geographic characteristics, such as elevation, climate (Harris et al., 2020), land quality (Galor and Özak, 2016a), and distances to rivers and oceans.

Specifications (5) through (9) sequentially test additional covariates to assess their influence and account for potential economic and cultural confounders. The rail access coefficient remains consistent across these specifications. In Specification (4), rail access in 1910 is associated with a 107 percent increase in grid cell population by 1921 ($e^{0.728} - 1 \approx 1.071$). Since grid cells are uniform in size, population within them represents density. In 1921, cells with rail access were therefore approximately twice as dense as those without.

Table 0.2: Rail Access and Log Population Between 1900 and 1921, OLS Results at the Grid Cell Level

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dependent Variable: Log Population in 1900								
Within 10km of Rail in 1892	2.341***	1.758***	1.704***	1.276***	1.153***	1.286***	1.227***	1.256***	
	(0.573)	(0.377)	(0.343)	(0.246)	(0.227)	(0.247)	(0.225)	(0.237)	
Number of Pre Colonial Indigenous Communities		1.283***	0.992***	0.836***	0.668**	0.800***	0.798***	0.820***	
		(0.295)	(0.284)	(0.305)	(0.278)	(0.302)	(0.308)	(0.306)	
R-squared	0.040	0.139	0.272	0.362	0.421	0.369	0.376	0.382	
	Dependent Variable: Log Population in 1910								
Within 10km of Rail in 1900	2.607***	0.428***	0.467***	0.499***	0.518***	0.525***	0.498***	0.501***	
	(0.449)	(0.096)	(0.097)	(0.120)	(0.118)	(0.114)	(0.121)	(0.119)	
Number of Pre Colonial Indigenous Communities		0.090**	0.072**	0.065**	0.050*	0.072**	0.062**	0.066**	
		(0.037)	(0.028)	(0.029)	(0.028)	(0.029)	(0.030)	(0.033)	
Log Population in 1900		0.856***	0.828***	0.801***	0.782***	0.799***	0.798***	0.792***	
		(0.023)	(0.026)	(0.023)	(0.025)	(0.023)	(0.022)	(0.024)	
R-squared	0.065	0.777	0.781	0.786	0.790	0.789	0.787	0.789	
	Dependent Variable: Log Population in 1921								
Within 10km of Rail in 1910	2.472***	0.661***	0.698***	0.728***	0.756***	0.738***	0.741***	0.730***	
	(0.335)	(0.137)	(0.135)	(0.141)	(0.150)	(0.143)	(0.141)	(0.139)	
Number of Pre Colonial Indigenous Communities		0.149**	0.129**	0.119**	0.089*	0.096**	0.114**	0.121**	
		(0.062)	(0.053)	(0.057)	(0.047)	(0.046)	(0.056)	(0.059)	
Log Population in 1900		0.797***	0.765***	0.744***	0.724***	0.739***	0.739***	0.738***	
		(0.028)	(0.026)	(0.022)	(0.020)	(0.020)	(0.020)	(0.022)	
R-squared	0.076	0.711	0.716	0.723	0.727	0.726	0.723	0.724	
Initial Population Proxy		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Latitude			Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographic Characteristics				Yes	Yes	Yes	Yes	Yes	Yes
Distance to Historical Mines					Yes				
Distance to Other Mineral Deposits						Yes			
Distance to U.S. Border							Yes		
Distance to the Camino Real								Yes	
Distance to Insurgent Outbreaks									Yes
N	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760

This table presents estimates of log population regressed on rail access (i.e., being within 10 km of the nearest rail) from approximately ten years prior. Initial population includes the number of indigenous communities and log population in 1900. Geographic characteristics include average monthly precipitation during each year, average air temperature during each year, elevation, distance to rivers and oceans, and land quality (Galor and Özak, 2016a). Distance to historical mines includes silver and gold mines documented in Panczer (1987) and distance to other mineral deposits are those documented by INEGI for copper, iron, zinc, and coal. A full table of results from these specifications can be found in Tables A.7-A.9 in the Table Appendix. Grid cells are 0.1×0.1 degrees (i.e., approximately 11 × 11 km at the equator) in size. Standard errors used to test for significance differences are adjusted for spatial dependence following Conley (1999) using a distance cutoff of 300km. Alternative distance cutoffs were tested and resulting standard errors associated with rail access in 1910 can be found in Table A.11 of the Table Appendix. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Rail access in 1900 is similarly associated with a 65 percent increase in density in 1910. The initial association between rail access and population in 1900 is larger than that in later decades. This stronger initial association should be interpreted with caution and may be due to omitted variable bias, which the pre-20th century population proxy does not fully capture.

Tables A.7-A.9 in the Table Appendix present all coefficient estimates associated with specifications in Table 0.2. These coefficient estimates show that in each decade, cell density was consistently associated with certain geographic and historical features, yet associations with other characteristics varied over time. Between 1900 and 1921, denser cells tended to be located at higher elevations and closer to perennial rivers, two characteristics that commonly coincide with central Mexico. In 1900, denser cells also tended to be closer to the U.S. border and the Camino Real, and had land that was relatively more productive for cultivating bananas and maize. In 1910 and 1921, population was strongly associated with density in 1900, and proximity to gold and silver mines. By 1910, a positive association between density and soil suitability for cotton also emerged, coinciding with rapid growth in the cotton textile industry during the *Porfiriato* (Razo and Haber, 1998). In other words, proximity to rail was not the only variable significantly associated with population density, yet remain significant even after conditioning for various regional characteristics.

While population is a non-negative, limited dependent variable, a linear model offers a reasonable approximation in this context. As a robustness check, I estimate specification (1) using a Tobit model to account for the censored nature of uninhabited grid cells. The average marginal effects from Tobit estimates are comparable to or slightly smaller than OLS estimates (see Appendix Table A.10), suggesting limited bias in the linear specification.

I also examine alternative measures of rail access using categorized and continuous

distance to the nearest rail line. Estimates by distance category show that only grid cells within 10 km of rail were consistently denser in 1900 and 1921; cells farther away showed weaker or no association (Table A.12 in the Table Appendix). Continuous specifications confirm a negative elasticity between distance and population density, with stronger effects in 1921 than in 1910. These results suggest that population became increasingly responsive to proximity to the rail network over time.

The association between early rail access in 1892 and population in 1900 is significant and stronger than in later decades, regardless of the measure used. However, these estimates should be interpreted with caution due to potential omitted variable bias, which is less of a concern in later decades when estimates are conditioned using both pre-colonial indigenous communities and population in 1900. Regardless, estimates, particularity those for log population in 1900, are not causal. Without additional context, OLS results cannot disentangle whether early rail expansion connected initially dense areas or if rail access encouraged clustering near the network.

0.5.2 Locality Level Level Results, 1900 to 1921

Unlike at the grid cell level, the proportion of zeros, or units with no inhabitants, in the balanced panel of localities is large. When using the entire balanced panel, the relationship between, for example, rail access and locality population in 1921, is estimated in conjunction with all future, unknown and uninhabited localities. As a result, the distribution of localities is extremely skewed with an abnormal number of zeros, particularly during the first half of the 20th century when many localities have yet to exist.

Using the full census of localities may subsequently bias the estimated effect of rail access, especially if the geographic characteristics of inhabited and uninhabited localities differ. To address this, I use the subset of all localities, inhabited or not, that existed up to each decade. For example, when regressing log population in 1921 on historical rail access, I include only localities that existed in 1900, 1910, or 1921, even if they "died out"

or were uninhabited by 1921.

To complement the grid cell analysis, I reproduce Table 0.2 using all inhabited localities from 1900 to 1921. Results are presented in Table A.13 in the Table Appendix and show that in 1910, localities within 10 km of the rail network in 1900 were about 38 percent more populated than those farther away ($e^{0.323} - 1 \approx 0.381$); and by 1921, this difference had grown to 45 percent ($e^{0.370} - 1 \approx 0.448$). To account for the non-negative nature of population, I re-estimate these models using a Tobit specification. Average marginal effects from Tobit models (Appendix Table A.16) are similar to OLS estimates, suggesting limited bias from using a linear approximation.

To assess the relative predictive strength of rail access compared to other factors, I estimate standardized OLS specifications and conduct a Shapley decomposition. In the standardized regressions (Appendix Table A.14), all continuous variables are standardized following Wooldridge (2012).[9] In both 1910 and 1921, initial population is the strongest predictor of subsequent population, while rail access is consistently the second most predictive factor. The Shapley decomposition of log population in 1921 (Table A.15) confirms these patterns: initial population accounts for 62 percent of explained variation, followed by geographic characteristics such as elevation, rivers, and coastal proximity (13 percent), climate (5 percent), and soil suitability (4.5 percent). Rail access explains about 5 percent of the overall variation and 8.5 percent among northern localities, suggesting that its role is comparable to that of geographic characteristics.

I also estimate the relationship between population and alternative measures of rail access. As done at the grid cell level, Table 0.3 presents results using categorized distance (in 10 km intervals) and log distance to the nearest rail line. In 1900, localities within 10 km of the 1892 rail network were 27 percent more populated than those farther away. Localities 10–40 km from the network were also denser by 18–20 percent. In 1910 and

[9]rail access is left unstandardized due to its binary nature.

Table 0.3: Distance to Rail and Log Population Between 1900 and 1921, OLS Results at the Locality Level

Dependent Variable:	Log Pop. in 1900	Log Pop. in 1910	Log Pop. in 1921
Rail Access in Year:	1892	1900	1910
	(1)	(2)	(3)
Measured Using Categorized Distance			
Within 10km of Rail	0.240**	0.408***	0.485***
	(0.091)	(0.080)	(0.117)
10 - 20km from Rail	0.166**	0.264***	0.270**
	(0.076)	(0.079)	(0.129)
20 - 30km from Rail	0.190**	0.145**	0.188
	(0.075)	(0.069)	(0.142)
30 - 40km from Rail	0.176*	0.062	0.127
	(0.096)	(0.080)	(0.110)
40 - 50km from Rail	0.109	0.060	0.073
	(0.070)	(0.075)	(0.101)
50 - 60km from Rail	0.042	0.020	-0.040
	(0.105)	(0.082)	(0.103)
R-squared	0.220	0.227	0.240
Measured Using Continuous Distance			
Log Distance from Rail	-0.077***	-0.130***	-0.165***
	(0.026)	(0.024)	(0.030)
R-squared	0.220	0.227	0.240
N	32,069	46,228	55,716

This table presents estimates of log population regressed on either categorized or log distance to historical rail (about) a decade prior, controlling for latitude, geographic fundamentals, average air temperature and precipitation, and initial population proxies. Standard errors clustered at the state level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

1921, positive associations extended up to 30 km from the network, though with weaker magnitudes beyond 10 km.

While grid-cell estimates suggest that the association between rail access and population is concentrated within a 10 km radius, locality-level estimates may reflect spillover effects extending to greater distances. By 1921, however, this association appears more spatially concentrated, with stronger correlations limited to localities within 20 km of the network.

To provide further insight, I re-estimate the locality models using subsamples of localities classified as never urban or ever urban between 1900 and 1921. Results are shown

in Tables A.17 and A.18, using categorized and log distance measures, respectively. In both cases, the overall association between rail access and population is primarily driven by urban localities. In 1910, the relationship is about 75 percent stronger among urban than rural localities.

I also estimate models using only localities that were urban in 1900. In 1910, rail access was not significantly associated with population within this subset, suggesting that new urban localities emerging after 1900 drove the observed association. By 1921, however, the pattern reverses: rail access is more strongly associated with population in historically urban localities. This shift appears concentrated among the 70 percent of these localities that remained urban after the Revolution, as roughly 30 percent "collapsed" following insurgent violence.

Overall, both grid cell and locality-level analyses confirm that, between 1900 and 1921, population is associated with proximity to the rail network. At the grid cell level, areas within 10 km of the network are notably denser than those farther away. At the locality level, while populations are generally more concentrated within the same 10 km range, the positive association extends up to 40 km, though the effect diminishes with distance. The positive association between rail access and population is also predominantly being driven by urban localities, although rural localities are not completely exempt from this trend.

0.6 Rail Access and Population, 1921 to 2020

This section investigates whether the initial association between historical rail access and population persisted throughout the 20th century. Between 1910 and 1921, the coefficient associated with rail access increased, suggesting that people were progressively clustering around the network and that this clustering was reshaping the broader spatial distribution. However, clustering near rail lines may reflect an endogenous relationship

between population growth and infrastructure, rather than an effect driven solely by rail access.

The prior analysis using historical rail access as of 1910, the end of the rail-building boom and Díaz's dictatorship, is extended using population data up to 2020. To address the endogeneity between population growth and infrastructure, I apply an instrumental variable approach. This method disentangles the extent to which rail access, rather than pre-existing relationships or local characteristics associated with rail, drove differences in density. Because the instrument is constructed to reflect rail access as of the end of the dictatorship in 1910, I do not apply it to the short-run analysis for 1900 and 1921, which focuses on associations during the period of rail expansion and prior to 1910. For robustness, I then conduct a placebo analysis to confirm that spurious correlations do not account for the significant association between historical rail access and population.

If the spatial distribution of population after the rail expansion era persists, then the coefficient for rail access or distance to the nearest rail line should remain similar throughout the 20th century. As defined earlier, persistence is the lasting influence of historical events even if their original determinants have diminished or no longer exist. Such persistence might weaken if the initial benefits of clustering near rail lines are quickly outweighed by the costs, which may come in the form of geographic constraints, overcrowding, or limited resources.

However, the relationship between population and rail may *slowly* decay over time if the network was initially able to attract local investment, but unable to maintain this inertia in the long run (Allen and Donaldson, 2022). Persistent clustering near historical infrastructure could therefore be driven by either (1) ongoing benefits from connectivity via rail, or (2) the slow depreciation of rail infrastructure and related sunk investments, such as factories built near the network (Bleakley and Lin, 2012). See Bertazzini (2021) for further discussion on the four possible mechanisms of persistence. This paragraph focuses

on the first two mechanisms of persistence, while the next addresses two additional self-reinforcing processes.

Strong persistence can also develop if historical rail access continues to attract economic activity, even without the ongoing use of the original infrastructure. This can occur if (3) initial clustering near rail lines encourages economic activity that further perpetuates clustering, or if (4) the network serves as a coordination mechanism that guides future investments, thus sustaining economic growth. Regions historically connected by rail may have developed enough during the *Porfiriato* to support a continued pattern of clustering near lines, even as the network gradually became obsolete.

In other words, this early momentum might have been strong enough to support self-reinforcing growth (Fujita et al., 1999). It is also possible that, due to the high sunk costs of infrastructure, new investments in the form of interstate highways either co-located with or replaced outdated rail lines. Consequently, the post-rail spatial distribution of population may have persisted as a result of these overlapping investments (Bleakley and Lin, 2012). Section 0.8 explores whether this was the case.

Given that the rail network fell into relative obsolescence after the Revolution, its continued advantages are unlikely. The slow depreciation of investments into, or associated with, the rail network is possible given the networks initial role as the sole form of reliable inter-regional transportation. However, during the rail-building boom, lines were foreign-owned, vertically integrated, and heavily reliant on foreign inputs and supplies (Coatsworth, 1979; Coatsworth, 1981; Van Hoy, 2008). As a result, weak forward linkages and limited domestic investments may have been insufficient to generate persistent, long-term effects.

There are also several reasons to expect either weakly increasing or decreasing returns to rail access, where the benefits of the network diminish over time. First, the unique geography of Mexico may limit the potential influence of history, particularly in

the context of infrastructure expansion. Mexico has few inland waterways, is mostly arid in the north, and is divided by two expansive mountain ranges. This distinct geography might limit where people can settle and make clustering in regions initially connected to rail too expensive over time. In other words, high geographic costs may hinder infrastructure's ability to permanently change the spatial distribution of economic activity, thereby limiting the role that history can play (Davis and Weinstein, 2002; Allen and Donaldson, 2022).

As previously noted, rail lines may have also supported export dependence over industrialization. Development during the 20th century has been characterized as extremely uneven with few "islands" of wealth conjoining underdeveloped and "backward" regions (Kemper and Royce, 1979). If rail lines were primarily designed to move raw materials to international trading hubs (Haber, 1989), then connected regions between intended origin and destination points may have been left undeveloped after the network's demise. Even locations intentionally connected to rail may have failed to spur sufficient forward and backward linkages outside of either the rail or primary exporting industry that rail connected.

0.6.1 Instrumenting for Historical Rail Access

Building the Instrument

Results thus far have characterized short-run population dynamics following rail expansion but have not established causality between rail access and changes in population. Infrastructure expansion is often inherently endogenous, meaning that the locations where these investments occur may be influenced by factors such as local productivity, population levels, or labor supply. As a result, the estimated relationship between rail access and population likely suffers from bias due to reverse causality, particularly in historically significant or large urban areas.

To address endogeneity, I use the inconsequential units approach to construct an

instrument for rail access (Redding and Turner, 2015; Bogart et al., 2022; Fenske et al., 2023). This method involves creating a least-cost path network that connects key locations across the country, acknowledging that culturally and economically significant locations were often prioritized during the construction of transportation networks. Distance to this least-cost path network is used to instrument rail access.

Localities or grid cells that fall along or near least-cost paths do so "inconsequentially," solely because they lie between major nodes. Endogenous factors, such as political preferences, priorities, or strategic investments, may cause infrastructure to deviate from what would otherwise be an efficient network. These deviations reflect unobserved influences on infrastructure, which the instrument accounts for by focusing on localities that were incidentally connected along the least-cost path network. To satisfy the exclusion restriction, the key localities directly involved in constructing the network are excluded from the estimation sample, ensuring that any effect captured is not driven by their prior economic or historical significance.

Conditional on available qualitative and quantitative data, researchers must make context-specific assumptions regarding which nodes to use when constructing a network. These nodes can vary and often include locations historically significant to investors, engineers, or politicians during the construction process. For example, Banerjee, Duflo, and Qian (2020) use major historical cities and four Treaty Ports as network nodes when constructing an instrument for infrastructure in China. These Treaty Ports were established by the League of Eight Nations after the First Opium War and were strategically chosen for their military advantage as well as their role as international trading hubs. Similarly, Fajgelbaum and Redding (2022) examine Argentina's integration into the global economy through rail expansion in the late 19th century. They construct an instrument using the shortest routes between each district and the nearest 16th-century city. The resulting instrument for rail access is the proportion of each geographic unit that lies along these

routes.

For this study, various least cost path networks were constructed using different sets of nodes to ensure relevance. Nodes were identified using different combinations of the following: the most populated localities in 1900, the most prolific gold and silver mines from Panczer (1987), historically important cities documented in the 1892 Census, and historically documented customs stations from the same census. These networks were constructed using either straight-line routes to minimize distance "as the crow flies" or terrain-based routes, similar in spirit to Bogart et al. (2022), which minimize changes in elevation while avoiding barriers such as lakes and oceans. Thus, both a straight-line and a terrain-based least-cost path network were created for each set of nodes. Log distance to the least-cost path network was then calculated and used as an instrument for historical rail access.

For context, Figure 0.6 shows a terrain-based least-cost path network overlaying the spatial layer used to assess the relative cost of traveling across Mexico. The figure focuses on the region spanning the states of Chihuahua, Durango, Coahuila, and Nuevo León. Lighter yellow areas indicate steeper slopes, which are more costly to traverse.[10] The resulting terrain-based least-cost network crosses the Chihuahuan Desert, about 100 miles east of the Sierra Madre Occidental mountain range (left side of Figure 0.6), and connects to various customs stations by passing through the low-lying areas of Nuevo León and Coahuila (right side of Figure 0.6). Figure A.8 in the Figure Appendix illustrates a least-cost path network over the sample region introduced in Section 3, and Figure A.9 provides an overview of the entire least-cost path network across Mexico.

The inconsequential units approach produces an instrument that satisfies the rele-

[10]The Distance Accumulation tool in ArcGIS produces an accumulative cost raster and a Euclidean back-direction raster, using an input elevation raster to calculate slope along with various barrier layers. These two rasters are then used as inputs to generate the optimal network connecting the specified localities, customs stations, and mines.

vance and exogeneity conditions. The strongest instrument for rail access is distance to the least-cost path network based on the most populated cities in 1900 and historical customs stations, consistent with nodes used in prior studies (Banerjee et al., 2020; Fajgelbaum and Redding, 2022). The relevance of this instrument is confirmed by the Kleibergen-Paap rk Wald F statistic, produced from a first-stage OLS model that regresses rail access on distance to the least-cost path network. Table A.19 in the Table Appendix compares this statistic across alternative instruments produced using different sets of nodes at the locality and grid cell level. The instrument based on the most populated cities in 1900 and historical customs stations has a sufficiently large Kleibergen-Paap rk Wald F statistic, confirming that it is not weak and is the strongest among the alternatives. Exogeneity is assumed to be satisfied because the least-cost path network is derived from an optimization procedure that minimizes construction costs, independent of local demographic or economic characteristics.

Using the Instrument

When using an indicator to measure rail access, the endogenous explanatory variable is binary. A linear first-stage model with a continuous instrument may fail to capture the non-linear relationship between the instrument and the binary treatment variable. As a result, the first stage may inaccurately estimate how the probability of rail access changes with distance to the least-cost path network. This instrument may subsequently present as weak, and lead to biased or inconsistent estimates in the second stage (Xu, 2021).

To address this potential misspecification in the first stage, the control function approach is applied. Following Heckman (1978) and Wooldridge (2010; 2015), a Probit model is used to estimate the probability of being within 10 km of the rail network based on distance to the least-cost path network and baseline geographic controls. The resulting coefficients are then used to generate predicted probabilities of rail access. These probabilities reflect the role of geography and proximity to the most efficient network. Economic

characteristics, initial population levels, and other unobserved cultural or historical factors do not influence the estimated probability in this setting.

In the second stage, the predicted probabilities from the first stage are uncorrelated with the error term, which captures the influence of unobservable predictors of population. By design, these probabilities reflect only the geographic and least-cost path determinants of rail access. As a result, the instrument isolates the exogenous variation in rail access, helping to disentangle its effect from bias introduced by endogenous confounding factors. This approach thus provides a more reliable estimate of the impact of rail access on population.

Table A.20 in the Table Appendix compares the second-stage estimates associated with rail access and log population in 1921 using either the linear 2SLS or control function approach at the grid cell level. When the functional form is misspecified in the first stage of a 2SLS model, the coefficient estimates and standard errors may be inconsistent, biased, and inflated. In this case, the estimated coefficient for rail access on population is twice as large when using 2SLS compared to the control function approach, and the estimated standard error is over five times as large.

Another approach would be to use a continuous measure of rail access, specifically log distance to the nearest rail line, in a 2SLS model. In this approach, the first stage would use distance to the least-cost path network to estimate distance to the actual network in 1910. Table A.21 in the Table Appendix presents the resulting Kleibergen-Paap rk Wald F statistics for all instruments derived from various least-cost path networks. At the locality level, distance to the least-cost path network based on the most populated cities in 1900 and historical customs stations remains the strongest instrument. However, all instruments are significantly weaker at the grid cell level, regardless of specification. Even the F statistic for the terrain-based least-cost path network falls below the recommended

threshold of 10.[11]

Given these limitations, the control function approach may offer a more reliable method for estimating the effect of rail access at both the locality and grid cell levels. Nevertheless, results using binary rail access estimated with the control function approach and log distance to the rail network estimated with 2SLS are provided for comparison throughout this section.

0.6.2 Grid Cell Level Results, 1921 to 2020

Due to the highly skewed nature of population, I continue to log-transform population. Figure A.7 in the Figure Appendix provides a series of histograms and Quantile-Quantile plots of log population from 1921 to 2020. These plots show that throughout most of the 20th century, the distribution is approximately normal. However, from 1980 onward, the distribution becomes increasingly bimodal, suggesting a potential shift in clustering, which this section explores.

Specification (1) is again used to test the long-run relationship between historical rail access and population into the 21st century. Similar to Jedwab and Moradi (2016), rail access is measured using categorized distances up to 30 km. When using categorized distance, the primary variable of interest is being within 10 km of the nearest rail line, while the other categories capture possible spillover effects in neighboring areas.[12]

Regression estimates in Table 0.4 capture the relationship between historical rail access and density from 1921 to 2020. Columns (1) and (2) use categorized distance to the rail network, focusing on the coefficient associated with having rail access or being within 10 km of a line. Column (1) reports OLS results following specification (1). Column (2)

[11] An instrument may be strong at the locality level but relatively weak at the grid cell level due to the latter being a larger, more aggregated spatial unit. This aggregation may dilute the precision of the instrument, resulting in a weaker correlation with actual rail access.

[12] Additional distance categories up to 100 km were tested, but multicollinearity and smaller urban sample sizes required reducing the number of categories.

reports IV results obtained using the control function approach, whcih follows specification (1) in the second stage. IV results confirm that the greater density observed near the network is attributable to rail access in the early decades of the 20th century. In 1921, cells with rail access were about 67 percent ($e^{0.515} - 1 \approx 0.674$) denser than those further away.

Table 0.4: Log Population and Historical Rail Access at the Grid Cell Level, 1921 to 2020

	Within 10 km of Rail in 1910		Log Distance to Rail in 1910	
Dependent Variable Log Population in:	(1) OLS	(2) IV Control Function	(3) OLS	(4) 2SLS
1921	0.810***	0.515*	-0.267***	-0.343
	(0.155)	(0.296)	(0.043)	(0.417)
1930	0.905***	0.271	-0.329***	-0.381
	(0.185)	(0.302)	(0.051)	(0.352)
1940	1.043***	-0.257	-0.401***	-0.397
	(0.198)	(0.297)	(0.055)	(0.439)
1950	1.130***	0.170	-0.435***	-0.625
	(0.208)	(0.300)	(0.061)	(0.427)
1960	1.192***	-0.122	-0.449***	-0.631
	(0.207)	(0.348)	(0.056)	(0.440)
1970	1.235***	-0.046	-0.460***	-0.657
	(0.185)	(0.333)	(0.055)	(0.442)
1980	1.284***	-0.376	-0.482***	-0.449
	(0.186)	(0.287)	(0.061)	(0.425)
1990	1.370***	0.871***	-0.490***	-0.601*
	(0.166)	(0.320)	(0.051)	(0.353)
2000	1.356***	0.258	-0.476***	-0.276
	(0.135)	(0.329)	(0.044)	(0.347)
2010	1.438***	0.467	-0.477***	-0.306
	(0.133)	(0.299)	(0.050)	(0.397)
2020	1.498***	1.016***	-0.490***	-0.409
	(0.107)	(0.328)	(0.040)	(0.396)
N	16,760	14,820	16,760	14,820

This table presents estimates of log population regressed on either categorized or log distance to historical rail in 1910, controlling for latitude, geographic fundamentals, average air temperature and precipitation, and initial population proxies. Categorized distance to rail is defined in 10 km intervals from 0 to 40 km from the nearest rail line. Instruments for columns (2) and (4) are produced using the least-cost path network of straight lines connecting the most populated localities in 1900 and historical customs stations listed in the 1892 Census. Grid cells within 30 km of these nodes, used to construct the least-cost path network, are excluded from the estimation sample. The Kleibergen-Paap rk Wald F statistics for the control function results in column (2) range from 35.9 to 39.3, while those for the 2SLS results in column (4) range from 13.7 to 15.3. Standard errors are adjusted for spatial dependence following Conley (1999) and using a 300 km cutoff. Conley standard errors in column (2) are bootstrapped with 100 iterations in both the first and second stages. Significance levels are denoted as * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$. Grid cells are 0.1×0.1 degrees (approximately 11×11 km along the equator).

OLS results show a strong and increasing association between historical rail access and population after 1930. Meanwhile, IV results indicate little to no direct effect of historical rail on density in the following decades. In other words, according to OLS results, clustering near rail and subsequent changes in the spatial distribution appear to persist, but in the decades after 1921, historical rail access cannot directly explain this trend.

IV results suggest a resurgence in the direct effect of historical rail access on density in 1990 and 2020. This resurgence may reflect the renewed relevance of certain historical corridors in industrial or trade-linked regions, potentially driven by late 20th-century trade liberalization policies or infrastructure upgrades following rail privatization. It is unlikely that these effects stem from increased usage of the original rail lines themselves. Instead, they may reflect the layered influence of historical corridors and related industrial infrastructure that has since been reactivated or repurposed. In such cases, persistence may be reinforced not by the survival of rail itself, but by the layered influence of new infrastructure built along historically advantageous routes.

At the end of the Díaz regime, Mexico's rail network spanned approximately 20,000 km (Coatsworth, 1981). Today, it spans about 23,000 km (CIA, 2024), with the system largely resembling what was built a century earlier. That the rail network still closely mirrors its original form may help explain why historically advantageous routes continue to have an influence. Infrastructure investments, particularly in rail, often involve large sunk costs, meaning they are extremely expensive to initially establish (Redding and Turner, 2015). As a result, historical investments can lay the groundwork for future, overlapping infrastructure by reducing these initial sunk costs and preserving the former's layout and influence over time.

In addition to sunk costs, limited expansion of the historical network may also reflect institutional constraints. Article 28 of the 1917 Constitution granted the state

authority over key infrastructure, including railways.[13] This provision later served as the basis for regulations requiring government approval for new lines or expansions (OECD, 2016). Mexico's constitutional framework may have thus unintentionally constrained the development of new infrastructure while preserving the structure of the original system. Consequently, the renewed effect of historical rail on population density may reflect the long-run influence of early infrastructure, reinforced by trade liberalization policies from the 1980s that promoted exports and manufacturing investment.

Rail access is alternatively measured using log distance to the nearest rail line, which provides interpretable elasticity estimates. OLS results in column (3) suggest that in the early decades of the 20th century, a ten percent increase in distance from the network was on average associated with a 2 to 4 percent decrease in density. However, 2SLS results suggest no positive direct elasticity effect between rail and density. The lack of statistical significance with a continuous measure of rail access, but an observed direct effect with a discrete measure, suggests that the influence of rail may have been heavily concentrated along or near lines. Previous analysis also support this interpretation. Before 1930, associations between population and rail access were strongest within 10 km but diminished significantly beyond 20 to 40 km.

0.6.3 Locality Level Results, 1921 to 2020

The analysis draws on locality-level data to disentangle the aggregate findings presented earlier. Long-run trends between urban and rural localities are compared and reveal patterns otherwise overlooked at the grid cell level. Urban localities are defined as those that were ever urban or had more than 2,500 inhabitants at any point between 1921 and 2020. Conversely, rural localities are defined as those that were ever inhabited but never urban during this period.

[13] Constitución Política de los Estados Unidos Mexicanos, 2024.

Regression estimates in Table 0.5 capture the relationship between historical rail access and population among all localities from 1921 to 2020. As before, columns (1) and (2) use categorized distance to the rail network, focusing on the coefficient associated with having rail access or being within 10 km of a line. Column (1) reports OLS results following specification (1) while column (2) reports IV results obtained using the control function approach.

Both OLS and IV estimates confirm that, in the decades following the *Porfiriato*, localities with rail access were more populated than those further away and without access. IV estimates imply that throughout the first half of the 20th century, localities with rail access were between 70 and 90 percent more populated than those without access. At the locality level, IV estimates are also larger in magnitude than OLS estimates. This difference is likely due to endogeneity-related bias attenuating OLS estimates, a pattern often observed in other rail studies (see Table A.1 in the Table Appendix).[14] As shown in columns (3) and (4), these results remain consistent when rail access is measured using log distance to the nearest rail line. The 2SLS estimates in column (4) suggest that between 1921 and 1960 a 10 percent increase in distance from the network led to a 2 to 3 percent decrease in locality population.[15]

Regardless of how rail access is measured, proximity to the network has no association or direct effect on locality population after 1970. These results are not observed at the grid cell level. This discrepancy may stem from the heterogeneity among, or variety in, the types of localities across Mexico. I subsequently investigate whether rural localities, which make up the majority of the sample, are biasing the potential effects of

[14]Unlike at the locality level, IV coefficient estimates for rail access are not always larger than OLS estimates at the grid cell level. This difference may be due to heterogeneous treatment effects at the locality level or spillover effects captured by the aggregation process at the grid cell level.

[15]For reference, when using 2SLS, Marein (2022) finds that the elasticity between population and distance to historical roads in Puerto Rico is about 0.12. In other words, a ten percent increase in distance from the nearest road corresponds to a 1.2 percent decrease in population.

Table 0.5: Log Population and Historical Rail Access at the Locality Level, 1921 to 2020

	Within 10 km of Rail in 1910		Log Distance to Rail in 1910	
Dependent Variable Log Population in:	(1) OLS	(2) IV Control Function	(3) OLS	(4) 2SLS
1921	0.335***	0.687***	-0.108***	-0.205***
	(0.089)	(0.251)	(0.031)	(0.073)
1930	0.257***	0.520***	-0.096***	-0.173***
	(0.086)	(0.176)	(0.029)	(0.054)
1940	0.299***	0.558***	-0.120***	-0.204***
	(0.101)	(0.162)	(0.033)	(0.053)
1950	0.303***	0.747***	-0.116***	-0.278***
	(0.100)	(0.202)	(0.034)	(0.064)
1960	0.277**	0.769***	-0.107***	-0.304***
	(0.107)	(0.229)	(0.038)	(0.079)
1970	0.183*	0.650***	-0.071*	-0.243***
	(0.107)	(0.207)	(0.038)	(0.060)
1980	-0.045	0.231	0.024	-0.100
	(0.089)	(0.189)	(0.033)	(0.068)
1990	-0.132	0.053	0.058*	-0.028
	(0.095)	(0.215)	(0.033)	(0.062)
2000	-0.171*	0.032	0.076**	-0.019
	(0.093)	(0.191)	(0.032)	(0.061)
2010	-0.120	0.215	0.063*	-0.072
	(0.095)	(0.212)	(0.033)	(0.062)
2020	-0.106	0.264	0.059*	-0.077
	(0.092)	(0.208)	(0.032)	(0.060)

This table presents estimates of log population regressed on either categorized or log distance to historical rail in 1910, controlling for latitude, geographic fundamentals, average air temperature and precipitation, and initial population proxies. Categorized distance to rail is defined in 10 km intervals from 0 to 30 km from the nearest rail line. Instruments for columns (2) and (4) are produced using the least-cost path network of straight lines connecting the most populated localities in 1900 and historical customs stations listed in the 1892 Census. Localities within 30 km of these nodes, used to construct the least-cost path network, are excluded from the estimation sample. Additional details, including the number of localities and the Kleibergen-Paap rk Wald F statistics associated IV control function estimates, can be found in Tables A.22 and A.23 of the Table Appendix. Standard errors clustered at the state level and IV control function estimates are bootstrapped using 100 iterations in both the first and second stages. Significance levels are denoted as * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

rail. The estimates in Table 0.5 are reproduced using urban and rural localities separately. Results using the binary variable for rail access are presented in Table 0.6. Accompanying results using log distance to the nearest rail line are provided in Table A.23 of the Table Appendix.

OLS estimates by urban status confirm that, between 1921 and 1970, historical rail access was positively associated with population in ever-urban localities (column 3) but was not significantly associated with population in consistently rural localities (column

Table 0.6: Log Population and Historical Rail Access by Urban and Rural Status, 1921 to 2020

	Localities Never Urban		Localities Ever Urban	
Dependent Variable Log Population in:	(1) OLS	(2) IV Control Function	(3) OLS	(4) IV Control Function
1921	0.191*	0.416*	0.287**	0.767**
	(0.102)	(0.236)	(0.114)	(0.362)
1930	0.122	0.268*	0.252***	0.568**
	(0.100)	(0.154)	(0.088)	(0.282)
1940	0.174	0.329**	0.299***	0.275
	(0.112)	(0.167)	(0.090)	(0.310)
1950	0.169	0.494***	0.306***	0.800***
	(0.109)	(0.189)	(0.101)	(0.253)
1960	0.133	0.528**	0.332***	0.478
	(0.116)	(0.240)	(0.113)	(0.317)
1970	0.042	0.440***	0.410***	0.854***
	(0.111)	(0.158)	(0.111)	(0.295)
1980	-0.127	0.071	0.077	0.225
	(0.092)	(0.225)	(0.224)	(0.464)
1990	-0.189*	-0.062	-0.233	-0.313
	(0.095)	(0.195)	(0.253)	(0.577)
2000	-0.220**	-0.065	-0.227	-0.281
	(0.093)	(0.177)	(0.271)	(0.569)
2010	-0.174*	0.121	-0.169	-0.264
	(0.093)	(0.210)	(0.294)	(0.597)
2020	-0.162*	0.173	-0.078	-0.371
	(0.090)	(0.200)	(0.277)	(0.567)

This table presents estimates of log population regressed on categorized distance to historical rail in 1910, controlling for latitude, geographic fundamentals, average air temperature and precipitation, and initial population proxies. In each year, localities that were ever urban had at least 2,500 inhabitants at one point in time before or during that year. Categorized distance to rail is defined in 10 km intervals from 0 to 30 km from the nearest rail line. Instruments for columns (2) and (4) are produced using the least-cost path network of straight lines connecting the most populated localities in 1900 and historical customs stations listed in the 1892 Census. Localities within 30 km of these nodes, used to construct the least-cost path network, are excluded from the estimation sample. Additional details, including the number of localities and the Kleibergen-Paap rk Wald F statistics associated IV control function estimates, can be found in Tables A.22 of the Table Appendix. Standard errors clustered at the state level and bootstrapped with 100 iterations in both the first and second stages. Significance levels are denoted as * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

1). Interestingly, toward the end of the 20th century, locality population is negatively associated with historical rail access among consistently rural localities.

Meanwhile, IV estimates show that the effect of historical rail access was positive in both ever-urban and consistently rural localities but over twice as large in ever-urban ones.[16] These findings hold when rail access is measured using continuous distance. See

[16]For example, in 1970, the coefficient associated with rail access was 0.440 among localities never

Table A.23 in the Table Appendix for the corresponding OLS and 2SLS results from 1921 to 2020.

Nevertheless, after 1970, historical rail access no longer had a significant effect on population in either rural or urban localities. However, as discussed in Section 0.6.2, population density remained greater near the rail line. This contrast, with insignificance at the locality level but persistence at the grid cell level, may reflect rural-to-urban migration or the diffusion of population beyond urban centers. Consequently, the average effect of rail access among localities may have diminished, while clustering at the grid cell level persisted. These contrasting results may reflect key differences in how each unit is defined, particularly since grid cells are fixed spatial areas that inherently measure density and may better capture diffusion or peripheral growth.

Beyond differences in measurement, several behavioral or structural mechanisms may help explain the diverging patterns observed between units. One possibility is rural migration: if proximity to rail initially influenced the population of rural localities, but residents later moved to other nearby communities, then population may appear to decline at the locality level even if it remains marginally more concentrated within the broader grid cell. Migrants may have relocated to other relatively rural places that were also near rail but offered complementary advantages, such as better agricultural suitability or a station positioned advantageously within the network. For example, Barasantti (2023b) finds that municipalities near rail, particularly those with terminal stations, were more populated in the long run due to their "first-mover" advantage and role as potential hubs connecting different transportation networks.

Insignificance at the locality level, but persistence at the grid cell level, could also reflect rural-to-urban migration or the diffusion of agglomeration effects. If rail access

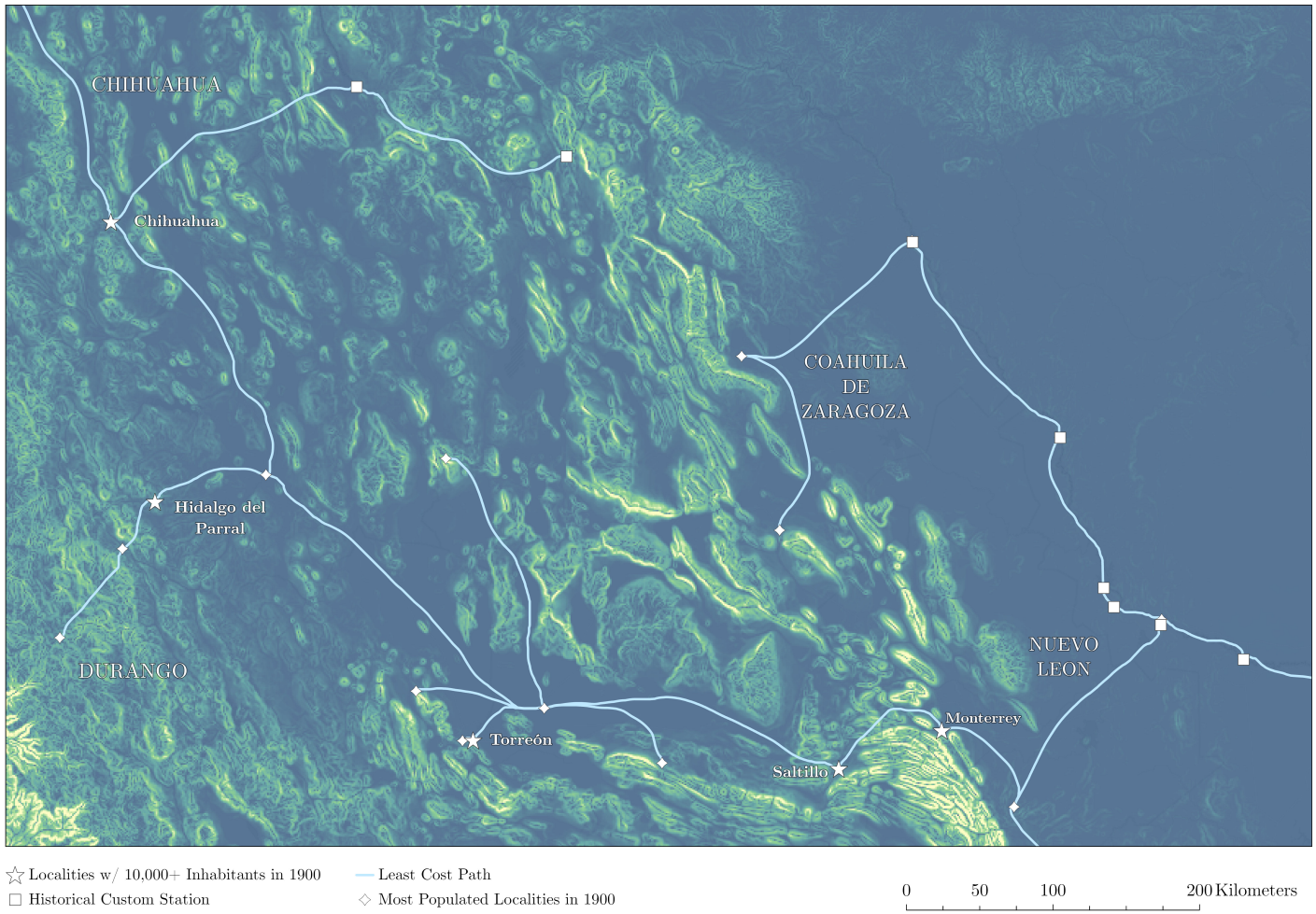
---

urban and 0.854 among localities that were ever urban. Thus the ratio of urban to rural estimates is $\frac{e^{0.854}-1}{e^{0.440}-1} \approx \frac{1.153}{0.516} \approx 2.236$.

initially spurred population growth in certain rural localities, but residents were later attracted toward nearby urban centers or regional hubs, then any long-term association between rail access and population would appear diminished at the locality level, but present at the grid cell level. Using linked 19th-century census microdata, Arteaga (2018) finds evidence of rural-to-urban migration in Argentina following rail expansion: connected areas became increasingly urbanized, and rail access encouraged children to migrate toward urban centers and away from farming, particularly in regions with low agricultural suitability. Without similar linked microdata for Mexico, I cannot confirm whether (intergenerational) migration dynamics were similarly at play.

Another possibility is that differing results between units reflect unaccounted spatial spillovers or heterogeneity, which could attenuate estimated effects at the locality level. In Section 0.7, I explore one dimension of heterogeneity, focusing on variation in the timing of rail access and phases of network expansion. The absence of locality-level significance may also result from the substitution of new infrastructure developed away from historical rail lines. In this case, population shifts between localities with new and old infrastructure could have diluted the apparent effect of historical rail, even as population density near rail remained elevated at the grid level. Without accounting for this new infrastructure, the long-run effect of rail on locality population may be obscured. In Section 0.8, I discuss the potential substitution effect of modern roads.

Figure 0.6: Cost Raster Used to Construct Least Cost Paths



This figure shows the cost raster used to construct the least-cost path network, focusing on a region in northern Mexico that includes Chihuahua, Durango, Nuevo León, and Coahuila. The least-cost path network minimizes cost while connecting historical customs stations to the top 1 percent of most populated localities in 1900.

0.6.4 Placebo Lines

As a robustness check, I conduct a placebo analysis to test whether spurious correlations are driving the relationship between historical rail access and population. This analysis confirms that differences in population are attributable to rail access and cannot be entirely explained by underlying factors that may have influenced the planning of lines. Similar to Donaldson (2018), I create digital reproductions of planned but unbuilt rail lines using cartographic records and historical references. I then replicate the primary specification by replacing measures of rail access with proximity to these placebo lines. The absence of strong and significant associations between population and placebo lines further supports the causal link between population and historical rail access.

Placebo lines are typically digitally reproduced or constructed using engineering or military surveys. In the absence of such information, I constructed placebo least-cost paths using historical sources, such as books and articles, that described proposed rail connections. Using the same least-cost path approach, I also recreated survey and reconnaissance routes produced by engineers based on information from archival maps. These projected or least-cost lines are modeled in ArcGIS, and consider elevation, slope, and natural barriers (e.g., major lakes and oceans) to minimize effective distance through terrain.

Figure A.10 in the Figure Appendix illustrates how an archival map with projected but unbuilt rail lines was digitally reconstructed to produce placebo lines. Panel A shows a section of the archival map, highlighting a planned route (marked by a red dashed line) connecting Aguascalientes and Guadalajara. Panel B displays the cost raster used to construct placebo lines.[17] Overlaid on this cost raster are solid red lines, representing the actual rail network as of 1910, and dashed red lines, representing the digitally

[17]Note that the same cost raster was used to produce the instrument for rail access, as described in Section 0.6.1.

reconstructed planned lines. These reconstructions, produced using the cost raster and least-cost path approach, incorporate locations that the planned lines were intended to pass through or connect. The resulting placebo lines in Panel B closely align with the planned routes shown in the archival map in Panel A. Additional details, including the sources used to construct various placebo lines across Mexico, are provided in the Data Construction Appendix. The results of this reconstruction, shown with political boundaries for context, are presented in Figure 0.7.

Correlations between population and distance to the nearest placebo line are presented in Table 0.7. The top panel provides estimates at the gird cell level, while the bottom panel provides estimates at the locality level. For each unit of analysis, results are shown using both categorized and continuous measures of distance to a placebo line. When using categorized distance, hypothetical rail "access" is defined as being within 10 km of the nearest placebo line, and the corresponding coefficients are provided.

At the grid cell level, being within 10 km of a placebo line was positively associated with density in 1921. However, this association was only about 20 percent as strong as that between density and actual rail access, as shown in Table 0.4.[18] After 1921, there was no significant association between density and proximity to a placebo line. Between 1921 and 1980, localities within 10 km of the nearest rail line were also no more populated than those further away. After 1980, localities near planned but unbuilt lines were *less* populated than those farther from a proposed line. This pattern holds when using the continuous measure of hypothetical "access" to a placebo line.

Replicating the primary specification with proximity to planned but unbuilt rail lines confirms that these placebos have no significant association with density or population between 1921 and 2020. The absence of a strong relationship between population and

[18]In 1921, the coefficient for population and hypothetical "access" to a placebo line was 0.193, while the coefficient for population and actual rail access was 0.810. The ratio of placebo to actual rail access is calculated as $\frac{e^{0.193}-1}{e^{0.810}-1} \approx \frac{0.213}{1.248} \approx 0.171$, which is less than 0.2.

hypothetical rail "access" suggests that previous findings on actual rail access cannot be fully attributed to preexisting conditions or spurious correlations. It's possible that proposed lines were left unbuilt for a reason, either due to infeasibility or cost. For instance, lines proposed in Baja California were abandoned due to poor geological conditions near connections like San Quintín (Heath, 2011). However, as detailed in the Data Construction Appendix, such proposals would have been unsuitable for a placebo analysis and thus were not used. Instead, reconstructed placebo lines were limited to those abandoned due to lack of funds or Revolutionary insurgency, or explicitly documented as planned in cartographic records.

0.7 Timing of Rail Access

0.7.1 Empirical Specification for the Timing of Rail Access

The primary approach in this analysis has assumed that all geographic units with rail access were affected in the same way, without accounting for how the timing of access might lead to different outcomes. For instance, early north-south lines, often driven by U.S. interests, may have shaped local economies differently than later east-west branch lines. Localities and regions that gained rail access earlier may have also experienced different effects due to greater exposure to markets compared to those connected decades later. My series of digitized rail maps documents the progression of expansion from 1874 to 1919 and allows me to incorporate the timing of rail access into the analysis.

By interacting timing with access, I test whether certain extensions of the network were more influential than others. I modify specification (1) to include an interaction term for timing with the rail access:

$$\ln(Population_{g,d}+1) = \alpha+\sum_{k=1}^{10} \beta_k \left(\mathbb{I}_{Rail}[g, 1910] \times PeriodGainedRail_{g,k}\right)+X_g'\delta+\epsilon_{g,d}, \quad (2)$$

where $PeriodGainedRail_g$ indicates the time period during in which geography g gained

rail access (e.g., was within 10km of the rail network in 1910). My panel of maps captures ten time periods ($K = 10$): before 1874, 1874 to 1878, 1878 to 1882, 1882 to 1884, 1884 to 1886, 1886 to 1892, 1892 to 1896, 1896 to 1900, 1900 to 1904, and 1904 to 1910.

Although I observe rail expansion across 10 time periods, there was little change in the network between 1874 and 1878 and insufficient variation to appropriately measure distinctions between the two years.[19] Consequently, the first two periods, rail before 1874 and rail gained between 1874 and 1878, are consolidated into one period.

As before, X_g is a vector of control variables, which includes geographic characteristics, pre-colonial population levels, and population in 1900. When measuring rail access using categorized distances, I continue to primarily focus on geographies within 10 km of the network, while also including additional distance indicators to account for spatial spillovers. These indicators capture any effects of being, for instance, 10 to 20 km or 20 to 30 km away from the network.

As discussed in Section 0.6.1, access to transportation infrastructure is often endogenous, meaning its placement can be influenced by political factors, investor interests, and local economic conditions. Consequently, estimates associated with rail access may be biased. If rail access is endogenous, then any interactions involving rail access in specification (2) are also endogenous. Additionally, the variable $PeriodGainedRail_g$, may itself be endogenous if the timing of expansions were not random.

As before, rail access can be measured using either categorized or continuous distance to the network. To address endogeneity, a 2SLS approach was previously applied, using distance to a least-cost path network as an instrument for distance to the actual rail network. Extending this approach to specification (2) would require nine separate instruments—one for each interaction between $PeriodGainedRail_g$ and distance to the nearest rail line. These instruments could be created by interacting $PeriodGainedRail_g$ with the

[19]See Appendix A.4.2 for large maps that illustrate the lack of expansion early on.

original instrument, distance to the least-cost path network. However, this approach was not feasible because the resulting instruments were collectively weak. In other words, the instruments did not sufficiently correlate with the endogenous variables and thus failed to satisfy the relevance condition required for the 2SLS approach. The weak instruments were driven by poorly correlated interactions during the period from 1874 to 1882, when the network was relatively sparse.

To strengthen the first stage and better capture the sequential nature of rail expansion, timing interactions with distance to nodes or historically significant locations used to construct the least-cost path network were also considered. These additional instruments account for the likelihood that areas near nodes were connected earlier during the network's expansion. For example, the main line of the Mexican Railway, connecting Mexico City to Veracruz, was built simultaneously from both ends (Donly, 1920). Similarly, construction of the Mexican Central began simultaneously in 1881, with work progressing southward from Chihuahua and northward from Mexico City, meeting and completing the line in 1884 (Powell, 1921). This construction pattern highlights the relevance of timing and node proximity.

Including these timing interactions ensured that all instruments collectively met the relevance and exogeneity conditions—they were sufficiently correlated with timing and actual rail access and weakly correlated with population, given that all localities within 30 km of any node were excluded from the estimation sample. However, including interactions between distance to the least-cost path network and timing, along with interactions between distance to the nearest node and timing, resulted in a total of 18 instruments. Such an approach, relying on the continuous measure of rail access and incorporating so many instruments, raised concerns about over-identification and excessive first-stage variation, which could undermine the reliability of second-stage estimates.

Other studies, such as Bogart et al. (2022), have avoided such challenges by ex-

plicitly incorporating a time dimension into the construction of their instrument. The authors create a measure of market potential prior to rail expansion and use the measure to prioritize connections across England and Wales. In creating their instrumental variable and least-cost path network, Bogart et al. first connect areas with the highest market potential and sequentially add least-cost path lines depending on the total kilometers of rail built in each year. Replicating this approach would be difficult without pre-1900 population data, which to my knowledge are unavailable at the needed geographic level.

Given these considerations, I use categorized distance to the network and an indicator for rail access, defined as being within 10 km of the network. As discussed in Section 0.6.1, applying a 2SLS approach for the endogenous binary variable of rail access would result in a mispecified first stage, where instruments may appear weak. To address this, a control function approach (Wooldridge, 2010; 2015) is again used to account for the endogeneity of the rail access indicator and its interactions with timing.

In my control function approach, the first stage is a probit model estimating the probability of being within 10 km of the actual rail network. These estimated probabilities are based on geographic characteristics, timing, distance to a least-cost path network connecting historically relevant locations (i.e., nodes), and distance to these locations. As before, the least-cost path network is an optimized hypothetical network connecting historically significant nodes, ensuring that distance from this network reflects optimization rather than any endogenous or unobserved determinants of rail expansion. The additional instruments for timing and node proximity, account for the progressive nature of the network's expansion.

Overall, using categorized distance and primarily focusing on the rail access indicator makes the first stage more estimable, less susceptible to outliers along the distance continuum, and better suited to capturing the local effects of rail proximity rather than the average effect across all distances. From the first stage probit model estimating rail

access, the generalized residuals are computed as $\hat{r} = \frac{\mathbb{I}_{Rail}[g,1910]-\hat{p}}{\hat{p}(1-\hat{p})}$, where $\hat{p}$ represents the predicted probabilities.

These generalized residuals, $\hat{r}$, are then included in the following second stage:

$$\begin{aligned} \ln(Population_{g,d}) = \alpha &+ \sum_{k=1}^{9} \beta_k \left(\mathbb{I}_{Rail}[g, 1910] \times PeriodGainedRail_{g,k}\right) \\ &+ \gamma\hat{r} + \sum_{k=1}^{10} \gamma_k \left(\hat{r} \times PeriodGainedRail_{g,k}\right) \\ &+ X_g'\delta + \epsilon_{g,d}, \end{aligned} \tag{3}$$

where the first two identifiable periods of expansion, having rail before 1874 and gained rail access between 1874 and 1878, are consolidated into one period. As a result, the index k for the number of periods represented by the categorical variable $PeriodGainedRail_{g,k}$ is nine, rather than the originally observed ten. The residuals, $\hat{r}$, account for variation in rail access that cannot be explained by geographic characteristics or instruments in the first stage. The additional inclusion of interactions between this unexplained variation in rail access and $PeriodGainedRail_{g,k}$ addresses potential bias associated with the timing of access. Together, $\hat{r}$ and $\hat{r} \times PeriodGainedRail_{g,k}$ help ensure that the interaction terms $\mathbb{I}_{Rail}[g, 1910] \times PeriodGainedRail_{g,k}$ appropriately isolate and capture the effect of the timing of rail access on population. This approach thus provides more reliable estimates of the effects of rail access during different periods of expansion.

Standard errors are bootstrapped and clustered at the state level to account for heteroskedasticity and within-state correlation. Bootstrapping adjusts confidence intervals to account for the additional uncertainty introduced by the two-stage process. This uncertainty mechanically originates from first-stage coefficient estimates, which add variability to the second stage via the residuals, $\hat{r}$. In other words, bootstrapping helps ensure reliable inference by addressing added variability within the control function approach.

Lastly, a multinomial probit or logit model may instinctively seem more appropriate

for the first stage. However, implementing such a model may be difficult given the nine categories associated with the timing of rail access. A multinomial model would, for example, produce a continuum of predicted probabilities that would need to be mapped to each possible value of $PeriodGainedRail_{g,k}$. The mapping of probabilities to distinct categorical outcomes produces a residual vector instead of a single residual value, which is needed in the second stage of the control function approach.

0.7.2 Timing of Rail Access and Population in the Long Run

Figures 0.8 and 0.9 illustrate the results of specification (3) on log population and highlight the estimated effect of gaining rail access during four selected time periods. A full table of results comparing the estimated effects of gaining rail access across the nine time periods of rail expansion, or the coefficient estimates associated with all interactions between $\mathbb{I}_{Rail}[g, 1910] \times PeriodGainedRail_{g,k}$, are in Tables A.24 and A.24 of the Table Appendix.

As shown in Figure 0.8, accounting for the timing of rail access reveals distinct relationships with population density. Without differentiating between phases of expansion, IV control function estimates suggest that grid cells with rail access by 1910 were approximately 67 percent denser in 1921 than cells farther from the network (Table 0.4). However, when the timing of rail access is considered, the estimated effects vary significantly by period. For instance, grid cells that gained rail access between 1904 and 1910, just before the Revolution, were similarly about 65 percent denser in 1921 than those farther from the network. In contrast, cells that gained rail access earlier, between 1882 and 1884 or 1884 and 1886, were more than twice as dense as those farther away. From 1921 onward, the effects of rail access during the four illustrated periods (1882–1884, 1884–1886, 1900–1904, and 1904–1910) are positive, significant, and often increase in magnitude. These findings suggest that the benefits of historical rail were persistent, sometimes cumulative, and strongly influenced by the timing of access. Prior analyses

that did not account for timing likely underestimated these effects due to omitted variable bias or the mischaracterization of rail access as a uniform "treatment."

Differentiating rail access by period provides a clearer understanding of how the effects of rail expansion evolved over time. While each phase of expansion has distinct effects on population density, these effects are not entirely independent of earlier infrastructure. This is especially true for later periods, such as 1900–1904 and 1904–1910, which primarily added branch lines to the existing network. The effectiveness of these lines reflects not only their own contribution but also the utility of the broader network built during earlier periods of expansion. For example, rail built between 1900 and 1904 likely facilitated local interactions between nearby communities, but more importantly, strengthened connections to other regions in Mexico and the global market. Thus, the effects of gaining rail during later periods cannot be fully disentangled from the foundational infrastructure of earlier expansions.

Nevertheless, the interdependence of extensions on the existing network does not imply that all later periods should have larger (or monotonically increasing or decreasing) effects on population. As discussed in Section 6, the marginal effect and persistence of any expansion may also depend on local economic characteristics, such as initial population density, complementary industries, and geographic constraints. This is evident in Figure 0.8, where by 2020, the effect of gaining rail access between 1882 and 1884 on log population is greater than that of rail access gained between 1900 and 1904.

Figure 0.9 presents the results of specification (3) for the same four periods of rail expansion but at the locality level. These disaggregated estimates are separated by rural and urban status, complementing the grid cell-level results that reflect aggregate changes in population density. Results at the locality level suggest that in the long run, locality population is sensitive to the timing of rail access, and that urban and rural localities were affected differently by the the historical rail network.

Previous locality-level IV results that do not differentiate rail access by timing suggest that, on average, in 1921 urban localities within 10 km of the rail network (by 1910) were more than twice as populated as those farther away (Table 0.6). However, when rail access is conditioned the timing of such access, it's evident that expansions beyond Mexico City and the port of Veracruz between 1878 and 1882 largely drove this result. Meanwhile, urban localities that gained rail access during later periods of expansion were no more populated in 1921 than those without access.

By 1930, urban localities that gained access to lines built after 1882 began to show relative advantages in population size. For instance, urban localities that gained rail access between 1882 and 1884 were approximately 60 percent more populated in 1930 than those farther from the network, while those gaining access between 1884 and 1886 were about 51 percent more populated (Figure 0.9, Panel A). Over time, the effect of gaining rail access between 1882 and 1884 remained statistically significant and progressively increased among urban localities. In contrast, IV estimates for rail access during all other periods of expansion were no longer positive or statistically significant after 1980. For instance, after 1980, urban localities that had gained access to the historical rail network between 1884 and 1886 were no more populated than other urban localities without access. While the coefficient estimates for this period appear negative after 1980 in Figure 0.9, they are not statistically significant.

For rural localities, accounting for the timing of rail access reveals a pattern similar to that observed for urban localities. During the first half of the 20th century, rural localities that gained access to the network between 1884 and 1886 were more populated than those without access and farther from the network. However, this effect diminished over time and was no longer evident after 1970, aligning with IV results that test the overall impact of historical rail access. Meanwhile, from 1921 onwards, rail access gained during most other periods of expansion did not have any significant effect on the population of

rural localities.

The lack of a persistent direct effect from historical rail access for most rural and many urban localities aligns with prior locality-level IV estimates that disregard the timing of access. However, prior analyses overlooked a subset of urban localities that continue to benefit directly from rail access gained between 1882 and 1884. This period of expansion is particularly significant for the growth of the overall network and the completion of the Mexican Central rail line, which connected Mexico City to the U.S. border, linked northern Mexico to the rest of the country, and initiated a mining boom in the region (Buffington and French, 2010).[20]

0.8 Substitution of New Infrastructure

As previously discussed, two possible mechanisms may be driving the persistent effect of historical rail: the coordination problem solved by old infrastructure and the substitution of new infrastructure. This section explicitly tests the latter mechanism in Mexico, where the modern interstate highway system was developed during the mid-20th century, decades after the rail boom of the *Porfiriato*.

Following World War I, increasing cooperation and trade between Mexico and the United States required reliable and modern interstate infrastructure (Waters, 2006). While state and federal governments were eager to invest in roads, the rail union, which played a significant role in the nationalized rail system, resisted such changes. Tensions between the government and union culminated into "ruinous competition," resulting in the disproportionate funding of roads over rail, the subsidization of road usage through taxes on rail, and the construction of highways parallel to rail (Fulwider, 2009). Roads inevitably prevailed; between 1940 and 1970, Mexico's road network expanded from ap-

[20]See Figure A.1 for the total increase in kilometers built and the Data Construction Appendix for maps showing the connections completed during this period.

proximately 10,000 km to 71,500 km (Díaz-Díaz and Garnica, 2007).

To study the connection between historical rail access and modern road access, a Probit model is used to predict modern road access in 1970, by which time most major roads had been built, and after which the estimated direct effect of historical rail diminishes (Table 0.5). Modern road access, defined as being within 10 km of a major road, is regressed on historical rail access, population in 1970, and the same baseline characteristics capturing geography and historical population proxies used throughout this analysis. Continuous variables are standardized to compare their relative predictive strength in determining the probability of having access to a major modern road. The model is run at both the grid cell and locality levels, with results using either categorized or continuous distance to historical rail presented in Table A.26 in the Table Appendix.

At the grid cell level, a one standard deviation increase in distance to the nearest historical rail line is associated with a 27 percentage point decrease in the probability of being within 10 km of a major modern road. Similarly, at the locality level, a one standard deviation increase in distance to historical rail is associated with a 29 percentage point decrease in the probability of having access to a major modern road.

The Probit model is reestimated using categorized distance as an alternative measure of rail access. Categorizing distance to historical rail into 10 km intervals between 0 and 30 km better captures the local association between historical rail and modern road access. For example, grid cells within 10 km of the historical rail network are associated with an 83 percentage point increase in the probability of being within 10 km of a major modern road. Likewise, localities within 10 km of the historical rail network are associated with a 98 percentage point increase in the probability of being within 10 km of a major modern road. These findings suggest that, when controlling for local characteristics, historical rail access is strongly linked to modern road access.

The strong association between old and new infrastructure suggests that roads were

likely built in parallel to or near the historical rail network. With this in mind, specification (1) is reestimated using 2SLS to examine the effect of historical rail access on population from 1960 onward, conditional on modern road access. The construction of roads was rolled out in multiple phases under different presidencies, but the network of major roads was largely completed by 1960. As a result, earlier decades are not reestimated using road access. In this analysis, rail access is defined using standardized distance to the nearest rail line or road to compare the relative influence of each type of transportation infrastructure. Results for all localities, as well as urban and rural localities, from 1960 to 2020 are presented in Table 0.8.

Previous IV results indicate that after 1970, historical rail access had little to no direct effect on locality population (Table 0.5 and Table A.23 in the Table Appendix). This finding is confirmed in Table 0.8, which examines the combined effects of historical and modern infrastructure. In 1960 and 1970, a one standard deviation increase in distance from historical rail is associated with a 0.8 to 0.7 standard deviation decrease in population among urban localities. By contrast, the same increase in distance from modern roads corresponds to only a 0.2 standard deviation decrease. This suggests that, prior to 1970, urban localities were less populated the further they were from both historical rail and modern roads, though they were more sensitive to distance from historical rail.

After 1970, the direct effect of historical rail access on urban population disappears, while distance to modern roads remains a strong predictor of population. Between 1980 and 2020, a one standard deviation increase in distance to the nearest modern road is associated with an approximate 0.2 to 0.3 standard deviation decrease in population. These findings confirm that with investments in modern interstate road infrastructure, the influence of historical rail access diminished. However, given the strong correlation between historical rail and road access, the persistent effect of modern road access reflects the path-dependent influence of 19th-century transportation infrastructure.

0.9 Conclusion

This study examines whether 19th century rail expansion had an enduring impact on the distribution of people across Mexico. Using harmonized historical census data and digitized archival maps, I analyze the relationship between population and historical infrastructure from 1900 onward. This analysis adopts a dual approach, focusing on the relationship between population and rail at both the grid cell and locality levels. This approach captures aggregate changes in density relative to the network while highlighting underlying patterns at the locality level. Together, these units provide a richer understanding of how Mexico's population distribution has shifted over time.

This research builds on the extensive economic history literature on railroads in Mexico, including the works of Kuntz-Ficker (1995; 2000; 2017; Kuntz-Ficker and Tena-Junguito, 2018), Coatsworth (1974; 1979; 1978; 1981; 1989), Haber (1989; 2011; Razo and Haber, 1998; Haber et al., 2004), and Herranz-Locán (2013). Kuntz-Ficker's work largely focuses on the political economy of rail expansion, particularly the export-oriented motivations and the dominance of foreign investment in shaping the rail network. Coatsworth's contributions were among the first to quantify the macroeconomic benefits of the rail network, highlighting its role in reducing transportation costs and fostering greater economic integration. This study builds upon and complements their contributions by applying modern data collection methods and econometric techniques, now prominent in the regional economics literature, to estimate the localized effects of rail expansion and resulting regional shifts. My results thus show how historical infrastructure shaped the broader, long-term spatial distribution of people in Mexico. Moreover, this analysis empirically tests theories regarding the conditions under which historical investments can lead to persistent or path-dependent outcomes (2022; 2012; Redding and Turner, 2015), while also underscoring the critical role of policy and foreign direct investment highlighted in Kuntz-Ficker's work.

The results of this analysis confirm that at the start of the 20th century, population was clustered near newly developed lines and density was associated with proximity to the network. Once the Díaz dicatorship and the railroad boom ended in 1910, areas and localities remained more populated near rail. During the first half of the 20th century, elasticity estimates confirm that population density and locality population levels were sensitive to distance from the rail network, but particularly pronounced within 20 kilometers of lines.

Although the rail network fell into relative obsolescence, its influence persisted and remained significant into the late 20th century. Long-run IV results, using the inconsequential units approach (Redding and Turner, 2015), reveal that up until 1970, historical rail access continued to play a deterministic role in locality population size. However, this effect diminishes throughout the 20th century, likely due to the gradual depreciation of rail's relative utility and associated infrastructure, such as stations and factories built near the network. Nevertheless, OLS results show a growing association between population density and historical rail access. A placebo analysis using records of planned but unbuilt rail lines further confirms that these findings are not driven by spurious correlations. Together, the OLS and IV results suggest that initial clustering near rail lines likely spurred economic activity, which in turn encouraged further clustering independent of direct proximity to the historical infrastructure.

I further investigate sources of heterogeneity driving overall changes in the spatial distribution of people relative to rail access over time. Analyzing localities separately by type reveals significant variation in the impact of historical rail on population, with urban localities and those in northern Mexico showing the strongest effects. Moreover, these persistent impacts were particularly pronounced among urban localities that gained rail access during a specific phase of expansion.

I used my portfolio of digitized archival maps to construct a panel of rail expansion

over 11 points in time from 1874 to 1910. Categorizing these phases of expansion and interacting them with rail access allows me to account for differences in the age of lines built during the rail boom. Without this adjustment, rail access would be characterized as a uniform "treatment", regardless of when the lines were built. By conditioning rail access by expansion phases, grid cell results show that lines constructed during specific periods have direct and persistent effects on population density even after 1970—a result previously obscured by omitted variable bias and the misspecification of rail access as a uniform "treatment". At the locality level, the persistent benefits of rail access are concentrated among urban localities in northern Mexico that were connected to the network via the Mexican Central Railway between 1882 and 1884.

Lastly, I test whether the observed persistence of historical rail access is driven by the expansion of modern road infrastructure. Probit results confirm that location of modern roads is highly correlated with the historical rail network, suggesting a substitution effect between these transportation modes. The cost co-location of old and new infrastructure is often done to help mitigate the sunk costs associated new investments and take advantage of residual non-transportation-related captial. Among urban localities, which continue to exhibit persistent effects of historical rail access, conditioning such access on proximity to modern roads reveals a significant shift. IV estimates show that the effect of historical rail access disappears after 1970 and is effectively replaced by modern road access from 1980 onward. This finding suggests that while historical railroads initially changed the spatial distribution of population, their long-run influence among urban localities was mediated and ultimately overtaken by the development of competing transportation infrastructure.

Locality level results show little to no persistent direct effect of rail among rural localities, or those that gained access during later periods of expansion outside of Northern Mexico. Where persistence is observed, there is relatively flat terrain with no inland

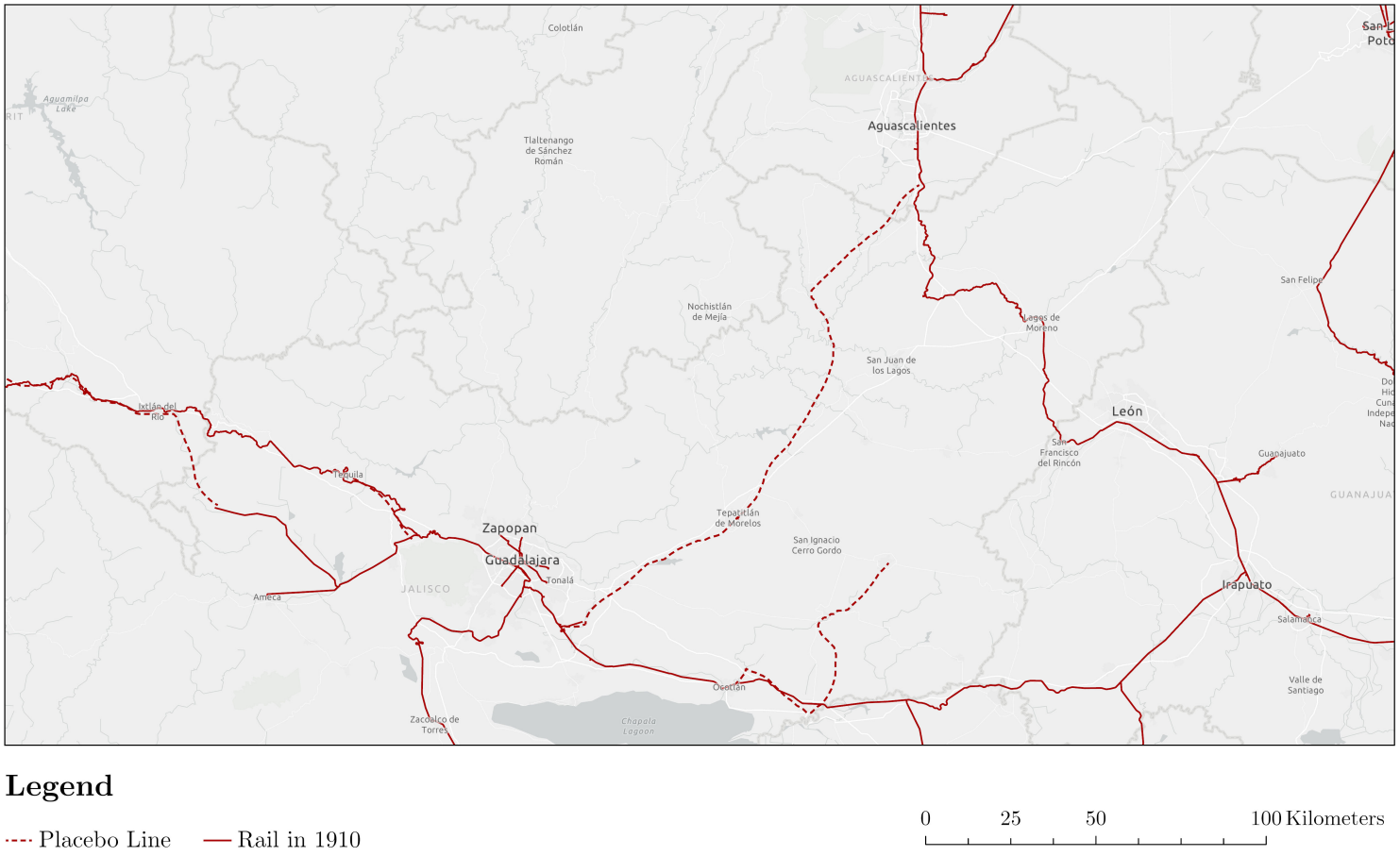
waterways that could have provided alternative forms of transportation. The lack of persistence in regions outside of Northern Mexico may stem from the high geographic costs of clustering in areas previously connected to historical infrastructure. These costs can arise unique geographies that can counter or limited the influence of the man-made or place based investments. In such cases, larger historical shocks are often required to permanently alter the distribution of economic activity. With the recent privatization of the rail network and efforts to revive passenger and freight transportation, rehabilitating the existing system could have unintended consequences if the goal is to reshape what may be seen as an undesirable distribution of economic activity (Lin, 2015; Lin and Rauch, 2022). If new rail investments or rehabilitation efforts are aimed at underdeveloped regions through existing infrastructure, the costs may be too high to overcome more deterministic geographic constraints.

This work highlights how Mexico's historical rail network has disproportionately benefited urban localities and the north, where favorable geography and co-location of modern infrastructure appears to have sustained the effects of rail access. These results provide valuable context for modern infrastructure policy, particularly in light of Mexico's recent efforts to reform its rail network.[21] Continued investments in place-based policies or preexisting infrastructure risk perpetuating the spatial distribution of people, which, as this study shows, has been shaped by history and is marked by significant regional inequalities in population density and economic activity. However, implementing new place-based policies to stimulate economic activity in rural or underdeveloped areas via new rail lines may prove either futile in the long run or cost prohibitive if natural geographic characteristics play a more decisive role than historical factors.

.

[21] As of October 2024, in efforts to revitalize passenger service, interstate connections, and domestic development, the Mexican government has reestablished control over the rail network, which was previously privatized in 1996.

Figure 0.7: Example Placebo Lines Alongside the Rail Network from 1910



See Figure A.10 in the Figure Appendix for an example of how an archival source was used to digitize projected lines as seen above.

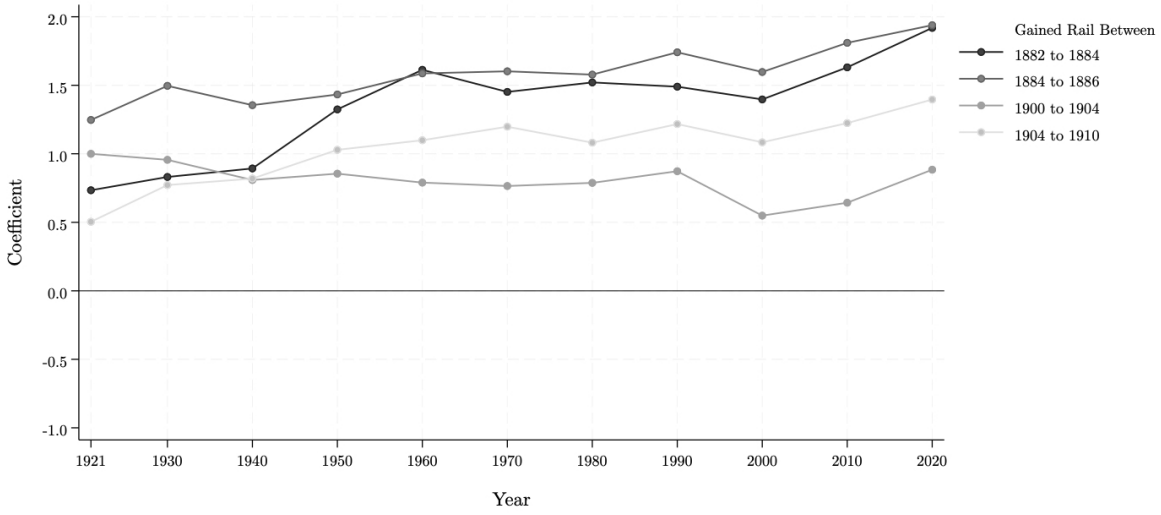
Table 0.7: Relationship Between Placebo Lines and Log Population From 1921 to 2020, OLS Results

Dependent Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Log Population in:	1921	1930	1940	1950	1960	1970	1980	1990	2000	2010	2020
Unit: Grid Cells											
OLS											
Within 10km from Placebo Line	0.193***	0.130	0.083	0.036	-0.001	-0.015	0.042	0.086	0.092	0.071	0.024
	(0.064)	(0.080)	(0.117)	(0.131)	(0.138)	(0.141)	(0.132)	(0.101)	(0.081)	(0.102)	(0.089)
Observations	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760
R-squared	0.721	0.710	0.676	0.653	0.625	0.601	0.598	0.613	0.624	0.621	0.621
OLS											
Log Distance to Placebo Line	-0.073**	-0.068*	-0.061	-0.053	-0.031	-0.062	-0.071	-0.086**	-0.091***	-0.090***	-0.058***
	(0.034)	(0.036)	(0.047)	(0.051)	(0.047)	(0.050)	(0.048)	(0.038)	(0.031)	(0.028)	(0.021)
Observations	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760
R-squared	0.721	0.711	0.676	0.653	0.625	0.602	0.599	0.614	0.624	0.622	0.621
Unit: All Localities Ever Inhabited											
OLS											
Within 10km from Placebo Line	0.032	0.012	0.081	0.080	0.069	0.038	-0.186	-0.193*	-0.187*	-0.139	-0.104
	(0.119)	(0.079)	(0.089)	(0.090)	(0.124)	(0.152)	(0.110)	(0.113)	(0.099)	(0.107)	(0.108)
Observations	55,716	63,181	74,215	82,956	93,599	118,116	155,407	218,756	290,292	320,681	344,469
R-squared	0.204	0.217	0.210	0.197	0.191	0.186	0.210	0.231	0.225	0.227	0.230
OLS											
Log Distance to Placebo Line	0.041	0.036	-0.009	-0.003	0.009	-0.002	0.079**	0.084**	0.074*	0.067*	0.058
	(0.045)	(0.029)	(0.032)	(0.032)	(0.042)	(0.045)	(0.037)	(0.041)	(0.038)	(0.037)	(0.037)
Observations	55,716	63,181	74,215	82,956	93,599	118,116	155,407	218,756	290,292	320,681	344,469
R-squared	0.204	0.217	0.210	0.197	0.191	0.186	0.210	0.232	0.225	0.227	0.231

This table presents OLS estimates of log population regressed on either categorized or log distance to placebo lines, controlling for latitude, geographic fundamentals, average air temperature and precipitation, and initial population proxies. Categorized distance to rail is defined in 10 km intervals from 0 to 40 km from the nearest rail line. Grid cells are 0.1×0.1 degrees (approximately 11×11 km along the equator). Standard errors at the grid cell level are adjusted for spatial dependence following Conley (1999) and using a 300 km cutoff. The number of localities in each decade includes those that were ever inhabited up until that point. Standard errors are clustered at the state level. Significance levels are denoted as * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Figure 0.8: Coefficient Estimates Associated with the Timing of Rail Access on Log Population

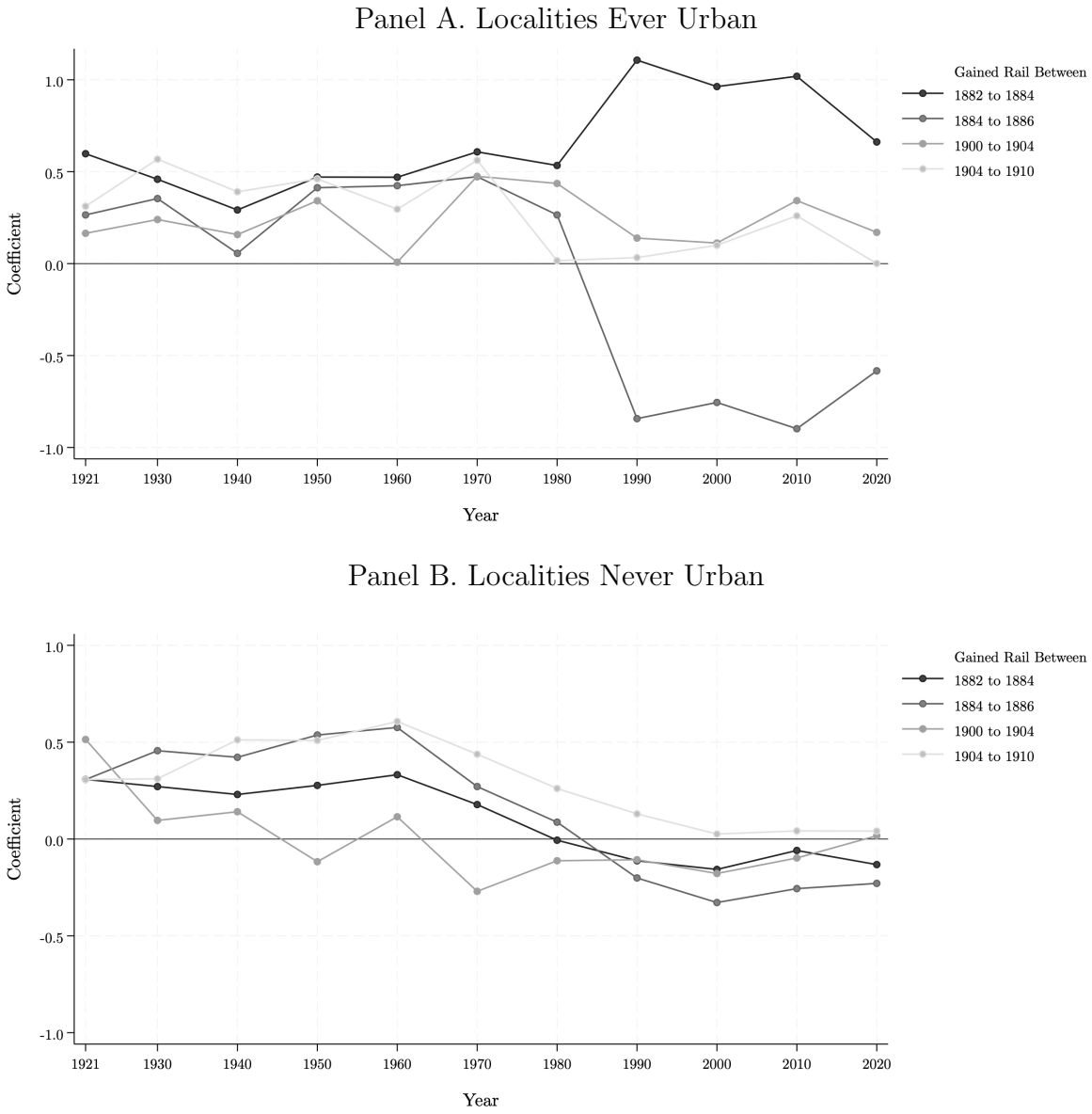
IV Control Function Estimates at the Grid Cell Level, 1921–2020



This figure plots the coefficient estimates for gaining rail access during selected periods on log population (1921–2020), based on specification (3) using an IV control function estimation approach. Exact estimates, coefficients for other periods, and bootstrapped standard errors are provided in Appendix Table A.24.

Figure 0.9: Coefficient Estimates Associated with the Timing of Rail Access on Log Population

IV Control Function Estimates for Urban and Rural Localities, 1921–2020



This figure plots the coefficient estimates for gaining rail access during selected periods on log population (1921–2020), based on specification (3) using an IV control function estimation approach. Panel A compares the effect and timing of rail access among localities that were ever urban during this period, while Panel B focuses on localities that remained consistently rural (i.e., were never urban). Exact estimates, coefficients for other periods, and bootstrapped standard errors are provided in Appendix Table A.24.

Table 0.8: Effect of Historical Rail and Modern Roads on Locality Population, Standardized 2SLS Estimates for Urban and Rural Localities, 1960–2020

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	All Localities						
	1960	1970	1980	1990	2000	2010	2020
Z Score of Distance to Rail in 1910 (km)	-0.065***	-0.039***	-0.015	-0.006	-0.004	-0.003	-0.004
	(0.020)	(0.013)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)
Z Score of Distance to Major Modern Road (km)	-0.002	-0.001	-0.000	-0.000	-0.001	-0.001	-0.001
	(0.006)	(0.004)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Z Score of Population in 1910	0.330***	0.266***	0.188***	0.237***	0.245***	0.257***	0.254***
	(0.024)	(0.025)	(0.019)	(0.028)	(0.031)	(0.032)	(0.030)
Observations	73,489	92,395	121,968	171,646	225,758	247,934	264,893
Kleibergen-Paap rk Wald F Statistic	18.48	19.75	15.09	12.63	13.91	13.67	14.40
	Never Urban						
	1960	1970	1980	1990	2000	2010	2020
Z Score of Distance to Rail in 1910 (km)	-0.022**	-0.012**	-0.005*	-0.003	-0.002	-0.002	-0.002*
	(0.009)	(0.005)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
Z Score of Distance to Major Modern Road (km)	-0.003	-0.002	-0.001	-0.001*	-0.000	-0.000	-0.000
	(0.002)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Z Score of Population in 1910	0.162***	0.102***	0.051***	0.050***	0.043***	0.040***	0.037***
	(0.007)	(0.005)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
Observations	70,835	89,633	119,118	168,733	222,801	244,938	261,878
Kleibergen-Paap rk Wald F Statistic	18.30	19.56	14.92	12.49	13.78	13.55	14.26
	Ever Urban						
	1960	1970	1980	1990	2000	2010	2020
Z Score of Distance to Rail in 1910 (km)	-0.821***	-0.672***	-0.273	-0.101	-0.079	-0.049	-0.130
	(0.292)	(0.259)	(0.221)	(0.297)	(0.307)	(0.338)	(0.322)
Z Score of Distance to Major Modern Road (km)	-0.196**	-0.151	-0.139**	-0.190**	-0.220**	-0.250***	-0.257***
	(0.094)	(0.093)	(0.067)	(0.089)	(0.090)	(0.096)	(0.092)
Z Score of Population in 1910	0.307***	0.260***	0.202***	0.270***	0.284***	0.298***	0.291***
	(0.028)	(0.034)	(0.025)	(0.038)	(0.041)	(0.042)	(0.039)
Observations	2,654	2,762	2,850	2,913	2,957	2,996	3,015
Kleibergen-Paap rk Wald F Statistic	31.44	34.49	36.39	32.27	36.56	35.65	37.21

Results are 2SLS estimates from regressing the z score of population on the z score of distance to historical rail and modern roads, while conditioning for all the same covariates used in prior analyses: latitude, geographic fundamentals, initial population, distance to locations associated with trade and distance to the nearest insurgent outbreak from 1910 to 1911. Distance to the nearest historical rail line is instrumented with distance to the terrain-based least cost path network constructed using the most populated localities in 1900 and custom stations. The sample of urban localities includes any locality that was urban or had at least 2,500 inhabitants at any point between 1921 and 2020. Localities that were never urban include all others that were inhabited but never had more than 2,500 inhabitants. Estimates are bounded by the 95 percent confidence interval using standard errors clustered at the state level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

APPENDIX A

CHUGGING ALONG: LONG RUN POPULATION DYNAMICS AFTER RAILROAD EXPANSION IN MEXICO

A.1 Table Appendix

Table A.1: OLS and 2SLS Estimates for Transportation Access on Population in the Literature

Paper	Table	Location/ Country	Measure of Transportation Access	Outcome of Interest	Description of Instrument		OLS Est.	IV Est.	Ratio (IV:OLS)
Bertazzini, 2021	4	The Horn of Africa	Within 10km of a road in 1940	Pop. density of grid cell (approximately 11×11 km) in 2000	Indicator if within 20km of a pre-1936 telegraphic or telephonic line		0.206	0.959	4.66
Bogart et al., 2022	3	England and Wales	Have a access to a rail station by 1851	Difference in log 1891 and log 1851 pop. of spatial unit (i.e., parishes, township, or hamlet)	Distance to least cost path connecting the most populous towns in 1801[1]		0.166	0.349	2.10
Fajgelbaum & Redding 2022	2	Argentina	The length of rail in each district in 1914	Log growth of district pop. from 1869-1914	Fraction of each districts surface area covered by least cost paths connecting their centroid to the top four trading ports		0.514	0.717	1.39
Marein, 2022	5	Puerto Rico	Log distance to roads in 1899	Log pop. density of cell (approximately 4.5×4.5 km) in 1920, 1940, and 2010	Log distance to straight line paths between the capitals of each department	Log pop. density in 1920:	-0.04	-0.12	3.00
						Log pop. density in 1940:	-0.07	-0.08 [2]	1.14
						Log pop. density in 2010:	-0.10	-0.26	2.60
Esteban-Oliver, 2023	4	Spain	Indicator for direct access to rail (i.e., within one hour of a line) after the first wave of construction from 1860-1877 (and the second wave 1900-1930)	Annual provincial pop. growth between 1860-1877 and 1877-1900 (and 1900-1930 associated with a second wage of rail construction)	Log distance to the least cost path network connecting provincial capitals and locations of economic importance[3] using cost estimates from Bogart et al. (2022) and Büchel and Kyburz (2018)	First wave of rail construction on pop. growth from 1877-1900:	0.119	0.932	0.78
						First wave of rail construction on pop. growth from 1877-1930:	0.219	0.346	1.58
						Second wave of rail construction on pop. growth from 1900-1930:	0.167	0.395	2.37
Fenske et al., 2023	2	India	Log distance to rail in 1881, 1891, 1901, 1911, 1921, and 1931	Log pop. in city	Interaction between year and log distance to the least cost path network constructed using historical cost data[4]		-0.017	-0.113	6.65

Coefficient estimates are significant unless otherwise noted. This table is an extension of the work in Table A.19 of Barsanetti (2023a), who finds that the ratio of IV and OLS coefficient estimate in similar studies to range between 0.60 and 2.33 with a median of 1.895. The ratios from more recent studies range between 0.78 - 6.65 with a median of 1.975.

[1] In Bogart et al. (2022) the least cost path network is uniquely constructed. The network is time varying and constrained by the total number of railway miles observed in a sample year. Due to this constraint, the least cost path is "rolled out" or sequentially connects the towns used as nodes.

[2] Estimate was not statistically significant.

[3] In Esteban-Oliver (2022), locations of economic importance used as nodes for the least cost path network include locations that fell at the intersection of either two truck roads, a truck road and a secondary road, or three or more secondary roads.

[4] In Fenske et al. 2023 the sequence of connections is determined using a gravity equation, where town connections with the highest value are prioritized until the expanse of the least cost path network equals the total distance of the true network.

Table A.2: Descriptive Statistics from Locality Population Panel

Panel A. Using Balanced Panel of Localities

	Percentile							
Census Decade	10^{th}	50^{th}	90^{th}	Mean	SD	Min.	Max.	N
1900	0	0	0	33	451	0	101,208	344,469
1910	0	0	18	36	503	0	119,468	344,469
1921	0	0	20	36	536	0	143,376	344,469
1930	0	0	42	42	662	0	179,556	344,469
1940	0	0	61	49	796	0	229,235	344,469
1950	0	0	82	65	1,262	0	377,016	344,469
1960	0	0	107	91	2,154	0	736,800	344,469
1970	0	0	152	132	3,577	0	1,193,601	344,469
1980	0	0	178	194	7,413	0	1,626,152	344,469
1990	0	0	190	236	8,247	0	1,650,042	344,469
2000	0	3	205	283	9,769	0	1,773,343	344,469
2010	0	2	226	327	10,610	0	1,815,786	344,469
2020	0	2	241	366	11,667	0	1,835,486	344,469

Panel B. Using Inhabited Localities Each Decade

	Percentile							
Census Decade	10^{th}	50^{th}	90^{th}	Mean	SD	Min.	Max.	N
1900	16	119	715	350	1,440	1	101,208	32,069
1910	13	98	628	313	1,444	1	119,468	40,177
1921	11	90	591	295	1,505	1	143,376	42,226
1930	11	82	547	281	1,686	1	179,556	51,848
1940	9	70	506	266	1,829	1	229,235	64,122
1950	10	86	598	335	2,840	1	377,016	67,259
1960	9	100	711	436	4,710	1	736,800	71,568
1970	6	61	692	474	6,757	1	1,193,601	96,187
1980	4	39	603	538	12,351	1	1,626,152	123,963
1990	3	24	525	529	12,355	1	1,650,042	153,318
2000	2	15	426	495	12,915	1	1,773,343	196,955
2010	2	15	494	591	14,270	1	1,815,786	190,279
2020	2	15	542	669	15,772	1	1,835,486	188,315

Table A.3: Estimates Capturing the Demographic Transition in Mexico, 1700 - 2000

	1700	1800	Pct. Change	Avg. Annual Growth Rate
Income Per Capita (2011 Prices)	1,476	1,305	-12%	0%
Population (millions)	4.5	5.1	13%	0%
	1800	1850	Pct. Change	Avg. Annual Growth Rate
Income Per Capita (2011 Prices)	1,305	1,054	-19%	0%
Population (millions)	5.1	7.6	49%	1%
	1850	1900	Pct. Change	Avg. Annual Growth Rate
Income Per Capita (2011 Prices)	1,054	1,822	73%	1.1%
Population (millions)	7.6	13.6	79%	1.2%
	1900	1950	Pct. Change	Avg. Annual Growth Rate
Income Per Capita (2011 Prices)	1,822	3,510	93%	1.3%
Population (millions)	13.6	27.6	103%	1.4%
Literacy	24	61	154%	1.9%
Life Expectancy	25	48	92%	1.3%
	1950	2000	Pct. Change	Avg. Annual Growth Rate
Income Per Capita (2011 Prices)	3,510	12,613	259%	2.6%
Population (millions)	27.6	98.6	257%	2.6%
Literacy	61	91	49%	0.8%
Life Expectancy	48	73	52%	0.8%

Income per capita data from 1700 to 2000 and population in 1700 are from the Maddison Project Database (2020); population data from 1800 to 2020 are from McCaaa (2000); and literacy and life expectancy estimates are from Astorga et al., 2005.

Table A.4: Percent of the Population in Urban and Rural Localities Over Time

Year t	Urban in 1900 but Not in t	Urban in 1900 and in t	Rural in 1900 but Urban in t	Rural in 1900 and t
1900	0.0%	31.0%	0.0%	69.0%
1910	1.3%	27.1%	3.4%	68.2%
1921	2.1%	25.9%	4.3%	67.7%
1930	1.7%	25.7%	5.7%	66.9%
1940	1.3%	24.8%	7.5%	66.4%
1950	0.8%	27.0%	12.1%	60.2%
1960	0.5%	28.6%	19.2%	51.7%
1970	0.2%	29.3%	27.8%	42.7%
1980	0.2%	35.1%	31.8%	33.0%
1990	0.1%	37.4%	34.1%	28.4%
2000	0.1%	38.2%	36.6%	25.2%
2010	0.1%	37.5%	39.5%	23.0%
2020	0.1%	36.7%	42.0%	21.2%

Table A.5: Descriptive Statistics from Grid Cell Population Panel

	Percentile							
Census Decade	10^{th}	50^{th}	90^{th}	Mean	SD	Min.	Max.	N
1900	0	0	1,837	669	2,464	0	106,029	16,787
1910	0	17	2,042	748	2,765	0	127,643	16,787
1921	0	18	2,041	743	2,876	0	151,439	16,787
1930	0	47	2,302	868	3,487	0	189,429	16,787
1940	0	76	2,570	1,014	4,172	0	241,853	16,787
1950	0	110	3,155	1,342	6,655	0	394,298	16,787
1960	0	149	3,897	1,858	12,287	0	777,073	16,787
1970	0	221	4,778	2,719	23,826	0	1,508,094	16,787
1980	0	281	5,841	3,972	43,256	0	2,313,813	16,787
1990	0	289	6,879	4,834	46,455	0	2,332,851	16,787
2000	0	302	7,781	5,807	54,140	0	2,594,455	16,787
2010	0	278	8,519	6,703	58,815	0	2,638,350	16,787
2020	0	243	9,057	7,510	64,190	0	2,643,168	16,787

Table A.6: Descriptive Statistics

	Using Localities				Using Grid Cells			
	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
Number of Pre Colonial Indigenous Communities	0.98	2.84	0.00	43.00	0.17	0.79	0.00	17.00
Elevation (km)	0.98	0.84	-0.01	3.53	1.00	0.81	-0.01	3.70
Average Daily Air Temperature (in Celsius)								
1901	20.92	3.62	11.96	27.61	20.51	3.55	11.96	27.61
2020	22.52	3.50	13.94	28.73	22.16	3.43	13.94	28.86
Average Monthly Precipitation (in Millimeters)								
1901	918.00	555.69	52.27	3,447.91	711.73	525.34	18.66	3,430.46
2020	953.74	535.99	53.89	3,276.03	753.27	534.63	51.10	3,276.03
Average Caloric Index (in Thousands)								
Bananas w/ High Inputs	6.33	3.32	0.00	10.37	5.81	3.88	0.00	10.43
Chickpeas w/ High Inputs	5.88	4.79	0.00	18.33	7.30	5.38	0.00	18.28
Coffee w/ High Inputs	17.64	11.76	0.00	31.61	13.74	12.91	0.00	31.93
Cotton w/ High Inputs	4.39	2.44	0.00	7.64	4.91	2.53	0.00	7.65
Maize w/ Low Inputs	8.29	3.83	0.00	17.04	6.71	4.69	0.00	17.07
Distance to the Nearest (in Hundreds)								
Intermittent River (km)	0.50	0.52	0.00	7.42	0.49	0.53	0.00	7.45
Perennial River (km)	0.37	0.77	0.00	9.31	0.57	0.86	0.00	9.36
Pacific Ocean (km)	2.82	2.30	0.00	9.35	3.02	2.42	0.00	9.34
Gulf of Mexico (km)	4.66	4.37	0.00	20.23	6.11	4.87	0.00	20.23
Coal Deposit (km)	3.75	2.44	0.00	13.09	3.59	2.58	0.01	13.08
Copper Deposit (km)	0.97	1.41	0.00	8.98	1.02	1.53	0.00	9.01
Iron Deposit (km)	1.33	1.40	0.00	10.57	1.40	1.51	0.01	10.58
Zinc Deposit (km)	1.13	1.42	0.00	10.27	1.16	1.50	0.00	10.33
Historical Gold Mine, Wide Distribution (km)	2.88	2.24	0.00	11.02	2.90	2.18	0.02	10.99
Historical Gold Mine, Moderate Distribution (km)	1.99	2.20	0.00	11.66	2.05	2.28	0.01	11.64
Historical Gold Mine, Limited Distribution (km)	3.23	1.90	0.00	11.14	2.89	2.06	0.00	11.14
Historical Silver Mine, Wide Distribution (km)	2.16	2.01	0.00	11.14	2.25	2.09	0.02	11.13
Historical Silver Mine, Moderate Distribution (km)	2.08	2.13	0.00	11.02	2.07	2.21	0.00	10.99
Historical Silver Mine, Limited Distribution (km)	6.12	3.81	0.00	16.21	5.72	3.76	0.02	16.28
U.S. Border (km)	6.51	3.39	0.00	14.85	5.44	3.57	0.00	14.91
Historical Customs Station (km)	1.40	0.94	0.00	7.35	1.51	0.94	0.01	7.39
Camino Real (km)	3.71	3.18	0.00	13.12	4.07	3.24	0.01	13.09
Insurgent Outbreak (km)	1.30	1.17	0.00	10.21	1.39	1.25	0.01	10.23

Summary statistics are computed using the balanced panel of 344,469 localities and 16,760 grid cells. Grid cells are 0.1×0.1 degrees (i.e., approximately 11×11 km at the equator) in size. For locality statistics, the number of pre-colonial indigenous communities corresponds to those within a 10km radius of each locality's coordinates. Average precipitation, air temperature and caloric suitability for each product is the average within the same size radius. For details on how characteristics are defined and on their data sources, see Appendix A.3.

Table A.7: Rail Access in 1892 and Log Population in 1900, OLS Results at the Grid Cell Level

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Within 10km of Rail in 1892	2.341***	1.758***	1.704***	1.276***	1.153***	1.286***	1.227***	1.256***
	(0.573)	(0.377)	(0.343)	(0.246)	(0.227)	(0.247)	(0.225)	(0.237)
Number of Pre Colonial Indigenous Communities		1.283***	0.992***	0.836***	0.668**	0.800***	0.798***	0.820***
		(0.295)	(0.284)	(0.305)	(0.278)	(0.302)	(0.308)	(0.306)
Latitude			-0.266***	-0.338	-0.549***	-0.386*	-0.867***	-0.461**
			(0.086)	(0.240)	(0.166)	(0.225)	(0.319)	(0.220)
Average Daily Air Temperature (C) in 1901				-0.116*	-0.044	-0.110*	-0.121*	-0.032
				(0.065)	(0.051)	(0.064)	(0.062)	(0.063)
Average Monthly Precipitation (mm) in 1901				-0.001	-0.000	-0.001	-0.000	-0.000
				(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Elevation				0.594***	0.314**	0.510***	0.440**	0.561***
				(0.180)	(0.127)	(0.164)	(0.176)	(0.192)
Average Caloric Index (in Thousands)								
Bananas w/ High Inputs				0.083***	0.046***	0.083***	0.067**	0.052***
				(0.023)	(0.016)	(0.029)	(0.027)	(0.014)
Chickpeas w/ High Inputs				0.029	0.031	0.022	0.023	0.016
				(0.033)	(0.025)	(0.030)	(0.032)	(0.031)
Coffee w/ High Inputs				0.007	0.002	0.013	0.004	0.012
				(0.014)	(0.012)	(0.016)	(0.014)	(0.015)
Cotton w/ High Inputs				-0.043	-0.043*	-0.052	-0.027	-0.037
				(0.048)	(0.025)	(0.048)	(0.038)	(0.046)
Maize w/ Low Inputs				0.068***	0.051***	0.062**	0.066***	0.041*
				(0.026)	(0.012)	(0.028)	(0.020)	(0.022)
Distance to the Nearest (in km)								
Intermittent River				0.528	0.285	0.606	0.654	0.537
				(0.510)	(0.501)	(0.514)	(0.513)	(0.516)
Perennial River				-0.960***	-0.208	-0.733***	-0.557**	-0.672***
				(0.227)	(0.205)	(0.240)	(0.275)	(0.230)
Pacific Ocean				0.078	0.459*	0.206	0.150	0.367
				(0.365)	(0.277)	(0.340)	(0.331)	(0.316)
Gulf of Mexico				0.077	0.252	0.119	0.286	0.325
				(0.245)	(0.173)	(0.254)	(0.244)	(0.221)
Historical Gold Mine, Wide Distribution					-0.428***			
					(0.114)			
Historical Gold Mine, Moderate Distribution					-0.396***			
					(0.102)			
Historical Gold Mine, Limited Distribution					0.288**			
					(0.129)			
Historical Silver Mine, Wide Distribution					0.009			
					(0.106)			
Historical Silver Mine, Moderate Distribution					-0.019			
					(0.132)			
Historical Silver Mine, Limited Distribution					0.027			
					(0.045)			
Coal Deposit						-0.167		
						(0.134)		
Copper Deposit						-0.381		
						(0.268)		
Iron Deposit						0.196		
						(0.188)		
Zinc Deposit						0.156		
						(0.247)		
U.S. Border							-0.563**	
							(0.234)	
Camino Real								-0.248***
								(0.085)
Constant	2.732***	2.547***	8.930***	11.579***	13.936***	12.278***	25.481***	11.180***
	(0.607)	(0.568)	(2.390)	(4.140)	(2.952)	(3.986)	(7.880)	(3.621)
Observations	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760
R-squared	0.040	0.139	0.272	0.362	0.421	0.369	0.376	0.382

This table presents estimates of log population in 1900 regressed on rail access in 1892 (i.e., being within 10 km of the nearest rail). Geographic characteristics include average monthly precipitation during each year, average air temperature during each year, elevation, distance to rivers and oceans, and land quality (Galor and Özak, 2016a). Distance to historical mines includes silver and gold mines documented in Panczer (1987) and distance to other mineral deposits are those documented by INEGI for copper, iron, zinc, and coal. Grid cells are 0.1×0.1 degrees (i.e., approximately 11 × 11 km at the equator) in size. Standard errors used to test for significance differences are adjusted for spatial dependence following Conley (1999) using a distance cutoff of 300 km. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.8: Rail Access in 1900 and Log Population in 1910, OLS Results at the Grid Cell Level

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Within 10km of Rail in 1900	2.607***	0.428***	0.467***	0.499***	0.518***	0.525***	0.498***	0.501***
	(0.449)	(0.096)	(0.097)	(0.120)	(0.118)	(0.114)	(0.121)	(0.119)
Number of Pre Colonial Indigenous Communities		0.090**	0.072**	0.065**	0.050*	0.072**	0.062**	0.066**
		(0.037)	(0.028)	(0.029)	(0.028)	(0.029)	(0.030)	(0.033)
Log Population in 1900		0.856***	0.828***	0.801***	0.782***	0.799***	0.798***	0.792***
		(0.023)	(0.026)	(0.023)	(0.025)	(0.023)	(0.022)	(0.024)
Latitude			-0.050**	-0.104	-0.080	-0.154**	-0.179*	-0.142**
			(0.022)	(0.066)	(0.053)	(0.075)	(0.104)	(0.057)
Average Daily Air Temperature (C) in 1910				-0.023**	-0.022	-0.015	-0.024**	0.001
				(0.011)	(0.014)	(0.010)	(0.010)	(0.012)
Average Monthly Precipitation (mm) in 1910				-0.000	0.000***	0.000	-0.000	0.000**
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Elevation				0.073	0.016	0.026	0.052	0.073
				(0.051)	(0.031)	(0.050)	(0.058)	(0.056)
Average Caloric Index (in Thousands)								
Bananas w/ High Inputs				-0.001	-0.005	-0.005	-0.002	-0.008*
				(0.003)	(0.004)	(0.005)	(0.004)	(0.005)
Chickpeas w/ High Inputs				0.012	0.008	0.010	0.012	0.008
				(0.008)	(0.007)	(0.007)	(0.008)	(0.007)
Coffee w/ High Inputs				0.001	-0.000	0.002	0.000	0.002
				(0.005)	(0.003)	(0.003)	(0.005)	(0.003)
Cotton w/ High Inputs				0.026**	0.027*	0.023**	0.028*	0.026*
				(0.012)	(0.015)	(0.011)	(0.014)	(0.014)
Maize w/ Low Inputs				0.006	-0.003	0.005	0.007	-0.001
				(0.012)	(0.010)	(0.011)	(0.012)	(0.010)
Distance to the Nearest (in km)								
Intermittent River				0.201	0.093	0.285**	0.220	0.197
				(0.142)	(0.150)	(0.143)	(0.148)	(0.142)
Perennial River				-0.348***	-0.089	-0.268***	-0.295**	-0.251***
				(0.111)	(0.090)	(0.090)	(0.118)	(0.083)
Pacific Ocean				0.040	-0.018	0.156	0.048	0.137
				(0.121)	(0.104)	(0.129)	(0.119)	(0.091)
Gulf of Mexico				0.043	0.020	0.112	0.072	0.131**
				(0.084)	(0.068)	(0.096)	(0.088)	(0.063)
Historical Gold Mine, Wide Distribution					-0.012			
					(0.027)			
Historical Gold Mine, Moderate Distribution					-0.090**			
					(0.046)			
Historical Gold Mine, Limited Distribution					0.151***			
					(0.040)			
Historical Silver Mine, Wide Distribution					-0.002			
					(0.059)			
Historical Silver Mine, Moderate Distribution					-0.040			
					(0.050)			
Historical Silver Mine, Limited Distribution					-0.050***			
					(0.016)			
Coal Deposit						0.004		
						(0.033)		
Copper Deposit						-0.264***		
						(0.064)		
Iron Deposit						0.219***		
						(0.038)		
Zinc Deposit						-0.114		
						(0.090)		
U.S. Border							-0.081	
							(0.093)	
Camino Real								-0.089***
								(0.023)
Constant	2.998***	0.720***	1.981***	3.226***	3.029***	3.490***	5.251**	3.039***
	(0.593)	(0.143)	(0.625)	(0.941)	(0.670)	(1.004)	(2.441)	(0.945)
Observations	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760
R-squared	0.065	0.777	0.781	0.786	0.790	0.789	0.787	0.789

This table presents estimates of log population in 1910 regressed on rail access in 1900 (i.e., being within 10 km of the nearest rail). Geographic characteristics include average monthly precipitation during each year, average air temperature during each year, elevation, distance to rivers and oceans, and land quality (Galor and Özak, 2016a). Distance to historical mines includes silver and gold mines documented in Panczer (1987) and distance to other mineral deposits are those documented by INEGI for copper, iron, zinc, and coal. Grid cells are 0.1×0.1 degrees (i.e., approximately 11 × 11 km at the equator) in size. Standard errors used to test for significance differences are adjusted for spatial dependence following Conley (1999) using a distance cutoff of 300 km. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.9: Rail Access in 1910 and Log Population in 1921, OLS Results at the Grid Cell Level

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Within 10km of Rail in 1910	2.472***	0.661***	0.698***	0.728***	0.756***	0.738***	0.741***	0.730***	0.709***
	(0.335)	(0.137)	(0.135)	(0.141)	(0.150)	(0.143)	(0.141)	(0.139)	(0.141)
Number of Pre Colonial Indigenous Communities		0.149**	0.129**	0.119**	0.089*	0.096**	0.114**	0.121**	0.108*
		(0.062)	(0.053)	(0.057)	(0.047)	(0.046)	(0.056)	(0.059)	(0.058)
Log Population in 1900		0.797***	0.765***	0.744***	0.724***	0.739***	0.739***	0.738***	0.741***
		(0.028)	(0.026)	(0.022)	(0.020)	(0.020)	(0.020)	(0.022)	(0.021)
Latitude			-0.055**	-0.114	-0.162**	-0.158**	-0.234*	-0.141*	-0.186**
			(0.022)	(0.080)	(0.075)	(0.075)	(0.123)	(0.076)	(0.085)
Average Daily Air Temperature (C) in 1921				-0.007	-0.001	-0.008	-0.009	0.008	0.001
				(0.021)	(0.022)	(0.019)	(0.019)	(0.018)	(0.022)
Average Monthly Precipitation (mm) in 1921				0.000	0.000***	0.000***	0.000***	0.000***	0.000
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Elevation				0.146**	0.115**	0.122*	0.116	0.149**	0.154**
				(0.069)	(0.057)	(0.074)	(0.082)	(0.076)	(0.071)
Average Caloric Index (in Thousands)									
Bananas w/ High Inputs				-0.010	-0.009	-0.003	-0.012	-0.015**	-0.010
				(0.007)	(0.009)	(0.009)	(0.009)	(0.007)	(0.008)
Chickpeas w/ High Inputs				0.018	0.013	0.009	0.017	0.015	0.016
				(0.011)	(0.011)	(0.009)	(0.011)	(0.011)	(0.011)
Coffee w/ High Inputs				-0.007	-0.007**	-0.001	-0.007*	-0.006	-0.006
				(0.004)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Cotton w/ High Inputs				0.022	0.016	0.009	0.025	0.022	0.023
				(0.017)	(0.019)	(0.015)	(0.019)	(0.018)	(0.018)
Maize w/ Low Inputs				-0.008	-0.013	-0.015	-0.008	-0.013	-0.009
				(0.012)	(0.013)	(0.013)	(0.012)	(0.012)	(0.011)
Distance to the Nearest (in km)									
Intermittent River				0.168	0.102	0.192	0.196	0.165	0.149
				(0.152)	(0.160)	(0.132)	(0.156)	(0.150)	(0.167)
Perennial River				-0.288**	-0.026	-0.171	-0.207	-0.232**	-0.142
				(0.122)	(0.115)	(0.114)	(0.139)	(0.100)	(0.108)
Pacific Ocean				0.037	0.131	0.139	0.054	0.104	0.158
				(0.118)	(0.144)	(0.114)	(0.112)	(0.106)	(0.126)
Gulf of Mexico				0.038	0.104	0.102	0.087	0.097	0.115
				(0.085)	(0.089)	(0.090)	(0.091)	(0.078)	(0.088)
Historical Gold Mine, Wide Distribution					0.012				
					(0.047)				
Historical Gold Mine, Moderate Distribution					-0.124***				
					(0.046)				
Historical Gold Mine, Limited Distribution					0.110**				
					(0.054)				
Historical Silver Mine, Wide Distribution					-0.162***				
					(0.050)				
Historical Silver Mine, Moderate Distribution					-0.003				
					(0.056)				
Historical Silver Mine, Limited Distribution					0.025				
					(0.026)				
Coal Deposit						-0.108***			
						(0.017)			
Copper Deposit						-0.013			
						(0.132)			
Iron Deposit						-0.010			
						(0.107)			
Zinc Deposit						0.006			
						(0.084)			
U.S. Border							-0.126		
							(0.095)		
Camino Real								-0.056**	
								(0.025)	
Insurgent Outbreak									-0.225***
									(0.061)
Constant	2.925***	0.838***	2.232***	3.264**	3.627***	3.922***	6.422**	3.177**	4.151***
	(0.568)	(0.141)	(0.591)	(1.317)	(0.975)	(1.215)	(2.725)	(1.284)	(1.360)
Observations	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760
R-squared	0.076	0.711	0.716	0.723	0.727	0.726	0.723	0.724	0.725

This table presents estimates of log population in 1921 regressed on rail access in 1910 (i.e., being within 10 km of the nearest rail). Geographic characteristics include average monthly precipitation during each year, average air temperature during each year, elevation, distance to rivers and oceans, and land quality (Galor and Özak, 2016a). Distance to historical mines includes silver and gold mines documented in Panczer (1987) and distance to other mineral deposits are those documented by INEGI for copper, iron, zinc, and coal. Grid cells are 0.1×0.1 degrees (i.e., approximately 11 × 11 km at the equator) in size. Standard errors used to test for significance differences are adjusted for spatial dependence following Conley (1999) using a distance cutoff of 300 km. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.10: Rail Access and Population Between 1900 and 1921, Tobit Average Marginal Effects at the Grid Cell Level

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
					Dependent Variable: Log Population in 1900				
Within 10km of Rail in 1892	2.594***	1.920***	1.835***	1.207***	1.083***	1.241***	1.169***	1.156***	
	(0.111)	(0.105)	(0.091)	(0.087)	(0.086)	(0.088)	(0.087)	(0.087)	
Pseudo R-squared	0.008	0.026	0.065	0.070	0.100	0.116	0.100	0.107	
					Dependent Variable: Log Population in 1910				
Within 10km of Rail in 1900	2.906***	0.322***	0.363***	0.410***	0.431***	0.454***	0.410***	0.407***	
	(0.090)	(0.041)	(0.042)	(0.043)	(0.043)	(0.043)	(0.043)	(0.042)	
Pseudo R-squared	0.013	0.271	0.274	0.282	0.286	0.285	0.282	0.284	
					Dependent Variable: Log Population in 1921				
Within 10km of Rail in 1910	2.739***	0.585***	0.626***	0.661***	0.689***	0.671***	0.667***	0.661***	0.640***
	(0.078)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)
Pseudo R-squared	0.0157	0.229	0.233	0.240	0.243	0.241	0.240	0.241	0.244
Initial Population Proxy		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Latitude			Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographic Characteristics				Yes	Yes	Yes	Yes	Yes	Yes
Distance to Historical Mines					Yes				
Distance to Other Mineral Deposits						Yes			
Distance to U.S. Border							Yes		
Distance to the Camino Real								Yes	
Insurgent Outbreaks									Yes
N	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760	16,760

This table presents Tobit estimates of log population regressed on rail access (i.e., being within 10 km of the nearest rail) from approximately ten years prior. Log population as a lower censored limit of zero. The regressions control for latitude, geographic fundamentals, average air temperature, precipitation, and initial population. Grid cells are 0.1×0.1 degrees (i.e., approximately 11×11 km at the equation) in size. Standard errors are robust. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.11: Coefficient Estimates and Standard Errors by Estimation Method, Relationship Between Rail and Population in 1921 at the Grid Cell Level

	Categorized Distance from Rail in	Log Distance from Rail in
Dependent Variable:	Log Pop. in 1921	Log Pop. in 1921
Rail Access in Year:	1910	1910
	(1)	(2)
Within 10km of Rail	0.862***	
Robust SE	(0.050)	
Conley SE		
w/ Distance Cutoff		
100 km	(0.122)	
200 km	(0.151)	
300 km	(0.157)	
10 - 20km from Rail	0.294***	‘
Robust SE	(0.047)	
Conley SE		
w/ Distance Cutoff		
100 km	(0.110)	
200 km	(0.095)	
300 km	(0.079)	
20 - 30km from Rail	0.191***	
Robust SE	(0.051)	
Conley SE		
w/ Distance Cutoff		
100 km	(0.110)	
200 km	(0.103)	
300 km	(0.070)	
30 - 40km from Rail	0.147***	
Robust SE	(0.053)	
Conley SE		
w/ Distance Cutoff		
100 km	(0.098)	
200 km	(0.081)	
300 km	(0.060)	
40 - 50km from Rail	0.121**	
Robust SE	(0.055)	
Conley SE		
w/ Distance Cutoff		
100 km	(0.102)	
200 km	(0.104)	
300 km	(0.076)	
50 - 60km from Rail	0.053	
Robust SE	(0.058)	
Conley SE		
w/ Distance Cutoff		
100 km	(0.088)	
200 km	(0.078)	
300 km	(0.060)	
Log Distance from Rail		-0.267***
Robust SE		(0.015)
Conley SE		
w/ Distance Cutoff		
100 km		(0.039)
200 km		(0.045)
300 km		(0.043)

This table presents estimates of log population in 1921 regressed on either categorized or log distance to historical rail in 1910 at the grid cell level. The regressions control for latitude, geographic fundamentals, average air temperature, precipitation, and initial population. For each coefficient estimate, robust standard errors and Conley standard errors (1999) are reported at various distance cutoffs. Grid cells are 0.1×0.1 degrees (i.e., approximately 11×11 km at the equation) in size. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.12: Distance to Rail and Log Population Between 1900 and 1921, OLS Results at the Grid Cell Level

Dependent Variable:	Log Pop. in 1900	Log Pop. in 1910	Log Pop. in 1921
Rail Access in Year:	1892	1900	1910
	(1)	(2)	(3)
Measured Using Categorized Distance			
Within 10km of Rail	1.282***	0.524***	0.862***
	(0.346)	(0.141)	(0.157)
10 - 20km from Rail	0.585*	0.087	0.294***
	(0.309)	(0.093)	(0.079)
20 - 30km from Rail	0.436	-0.002	0.191***
	(0.298)	(0.091)	(0.070)
30 - 40km from Rail	0.186	-0.049	0.147**
	(0.263)	(0.088)	(0.060)
40 - 50km from Rail	0.303	0.017	0.121
	(0.283)	(0.113)	(0.076)
50 - 60km from Rail	0.159	-0.130**	0.053
	(0.250)	(0.062)	(0.060)
R-squared	0.416	0.790	0.727
Measured Using Continuous Distance			
Log Distance from Rail	-0.425***	-0.153***	-0.267***
	(0.123)	(0.042)	(0.043)
R-squared	0.419	0.789	0.727
N	16,760	16,760	16,760

This table presents estimates of log population regressed on either categorized or log distance to historical rail (about) a decade prior, controlling for latitude, geographic fundamentals, average air temperature and precipitation, and initial population proxies. Categorized distance to rail is defined in 10 km intervals from 0 to 60 km from the nearest rail line. Grid cells are 0.1×0.1 degrees (i.e., approximately 11×11 km at the equation) in size. Standard errors used to test for significance differences are adjusted for spatial dependence following Conley (1999) using a distance cutoff of 300 km. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.13: Rail Access and Log Population Between 1900 and 1921, OLS Results at the Locality Level

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Dependent Variable: Log Population in 1900										
Within 10km of Rail in 1892	0.392***	0.234**	0.235**	0.177**	0.133**	0.103*	0.124**	0.131**	0.131**	
	(0.108)	(0.098)	(0.089)	(0.068)	(0.060)	(0.052)	(0.057)	(0.058)	(0.059)	
Number of Pre Colonial Indigenous Communities		0.077***	0.062***	0.033***	0.028***	0.027**	0.027**	0.028***	0.027***	
		(0.012)	(0.012)	(0.012)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	
R-squared	0.012	0.053	0.084	0.198	0.215	0.221	0.218	0.215	0.216	
N	32,069	32,069	32,069	32,069	32,069	32,069	32,069	32,069	32,069	
Dependent Variable: Log Population in 1910										
Within 10km of Rail in 1900	0.695***	0.326***	0.337***	0.323***	0.292***	0.260***	0.290***	0.291***	0.291***	
	(0.098)	(0.075)	(0.067)	(0.076)	(0.059)	(0.053)	(0.056)	(0.059)	(0.060)	
Number of Pre Colonial Indigenous Communities		0.071***	0.063***	0.032**	0.025**	0.025**	0.026**	0.025**	0.024**	
		(0.013)	(0.016)	(0.013)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)	
Log Population in 1900		0.266***	0.256***	0.254***	0.246***	0.244***	0.246***	0.246***	0.245***	
		(0.025)	(0.022)	(0.020)	(0.018)	(0.019)	(0.018)	(0.018)	(0.018)	
R-squared	0.020	0.147	0.152	0.216	0.223	0.228	0.224	0.223	0.223	
N	46,228	46,228	46,228	46,228	46,228	46,228	46,228	46,228	46,228	
Dependent Variable: Log Population in 1921										
Within 10km of Rail in 1910	0.715***	0.349***	0.377***	0.370***	0.360***	0.342***	0.345***	0.360***	0.360***	0.341***
	(0.085)	(0.061)	(0.052)	(0.069)	(0.058)	(0.051)	(0.058)	(0.058)	(0.058)	-0.057
Number of Pre Colonial Indigenous Communities		0.076***	0.063***	0.033***	0.028***	0.027***	0.027***	0.028***	0.028***	0.026***
		(0.011)	(0.010)	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	-0.009
Log Population in 1900		0.327***	0.315***	0.312***	0.308***	0.307***	0.308***	0.308***	0.308***	0.309***
		(0.020)	(0.017)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	-0.014
R-squared	0.018	0.174	0.182	0.229	0.234	0.237	0.238	0.235	0.234	0.236
N	55,716	55,716	55,716	55,716	55,716	55,716	55,716	55,716	55,716	55,716
Initial Population Proxy		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Latitude			Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects				Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographic Characteristics					Yes	Yes	Yes	Yes	Yes	Yes
Distance to Historical Mines						Yes				
Distance to Other Mineral Deposits							Yes			
Distance to U.S. Border								Yes		
Distance to the Camino Real									Yes	
Insurgent Outbreaks										Yes

This table presents estimates of log population regressed on rail access (i.e., being within 10 km of the nearest rail) from approximately ten years prior. Initial population includes the number of indigenous communities and log population in 1900. Geographic characteristics include average monthly precipitation during each year, average air temperature during each year, elevation, distance to rivers and oceans, land quality (Galor and Özak, 2016a). The number of pre-colonial indigenous communities corresponds to those within a 10km radius of each locality's coordinates. Average precipitation, air temperature and caloric suitability for each product is the average within the same size radius. Distance to historical mines are silver and gold mines documented in Panczer (1987) and distance to other mineral deposits are those documented by INEGI for copper, iron, zinc and coal. Population of localities in each decade includes those that were ever inhabited as of that year. Standard errors used to test for significance differences are adjusted clustered at the state level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.14: Relationship Between Rail and Population, Standardized OLS Results from 1900 to 1921

	Using Grid Cells			Using Localities		
Dependent Variable: Z Score of Population in	1900	1910	1921	1900	1910	1921
	(1)	(2)	(1)	(1)	(2)	(3)
Within 10km of Rail in 1892	0.550***			0.500***		
	(0.151)			(0.059)		
Within 10km of Rail in 1900		0.050***			0.064***	
		(0.012)			(0.017)	
Within 10km of Rail in 1910			0.061***			0.068***
			(0.010)			(0.020)
Z Score of						
Average Daily Air Temperature (C)	-0.093**	0.015***	0.025***	-0.110	0.020	0.010
	(0.044)	(0.005)	(0.007)	(0.073)	(0.014)	(0.023)
Average Monthly Precipitation (mm)	-0.015	0.013	0.026*	0.022	0.021	0.013
	(0.027)	(0.009)	(0.014)	(0.056)	(0.015)	(0.018)
Number of Pre Colonial Indigenous Communities	0.312***	-0.012	-0.025	0.110**	-0.006	-0.012
	(0.033)	(0.008)	(0.016)	(0.041)	(0.004)	(0.009)
Log Population in 1900		0.976***	0.946***		0.963***	0.925***
		(0.020)	(0.045)		(0.033)	(0.067)
Latitude	-0.361**	0.022	0.056**	-0.392*	0.051	0.038
	(0.149)	(0.014)	(0.028)	(0.198)	(0.049)	(0.061)
Elevation (km)	0.044***	-0.003	-0.006	-0.008	-0.011	-0.050**
	(0.012)	(0.004)	(0.007)	(0.061)	(0.011)	(0.021)
Bananas w/ High Inputs (Thousands)	0.028	0.001	-0.001	-0.083*	0.009	0.011
	(0.019)	(0.002)	(0.004)	(0.047)	(0.009)	(0.011)
Chickpeas w/ High Inputs (Thousands)	0.039	-0.003	-0.013**	0.081	0.004	-0.008
	(0.043)	(0.004)	(0.007)	(0.053)	(0.009)	(0.014)
Coffee w/ High Inputs (Thousands)	-0.009	-0.003	-0.008	-0.005	-0.008	-0.001
	(0.018)	(0.003)	(0.010)	(0.041)	(0.009)	(0.012)
Cotton w/ High Inputs (Thousands)	-0.029	0.000	-0.001	0.053	-0.000	-0.009
	(0.019)	(0.002)	(0.003)	(0.032)	(0.007)	(0.011)
Maize w/ Low Inputs (Thousands)	0.012	-0.001	-0.008	-0.115***	-0.027***	-0.026**
	(0.009)	(0.005)	(0.006)	(0.040)	(0.007)	(0.012)
Distance to						
Intermittent River (km)	0.008	0.005**	0.011***	0.076**	0.008	0.021**
	(0.022)	(0.002)	(0.002)	(0.035)	(0.007)	(0.008)
Perennial River (km)	-0.068**	-0.002	-0.008	-0.004	-0.035**	-0.063***
	(0.034)	(0.004)	(0.010)	(0.077)	(0.017)	(0.019)
Pacific Ocean (km)	0.131	-0.009	-0.027	0.053	-0.044	-0.026
	(0.101)	(0.012)	(0.026)	(0.157)	(0.036)	(0.048)
Gulf of Mexico (km)	0.134	-0.013	-0.031	0.009	-0.076	-0.042
	(0.166)	(0.018)	(0.038)	(0.244)	(0.056)	(0.071)
Historical Gold Mine, Wide Distribution (km)	-0.180***	-0.000	-0.000	0.153**	-0.002	-0.054**
	(0.039)	(0.006)	(0.021)	(0.066)	(0.017)	(0.025)
Historical Silver Mine, Wide Distribution (km)	0.043	0.001	0.020	-0.177*	0.005	0.049
	(0.080)	(0.006)	(0.017)	(0.101)	(0.018)	(0.031)
Coal Deposit (km)	-0.124***	-0.005	-0.029*	-0.115*	-0.015	-0.044**
	(0.034)	(0.007)	(0.016)	(0.062)	(0.016)	(0.020)
Copper Deposit (km)	0.053*	-0.011	-0.019	0.110	-0.052**	-0.030
	(0.031)	(0.007)	(0.014)	(0.107)	(0.020)	(0.037)
Iron Deposit (km)	-0.008	0.005	0.025**	0.038	0.005	0.034
	(0.040)	(0.007)	(0.010)	(0.081)	(0.017)	(0.024)
Zinc Deposit (km)	0.138***	0.006	0.008	-0.138	0.037	0.047
	(0.039)	(0.011)	(0.013)	(0.105)	(0.026)	(0.031)
Constant	-0.044***	-0.005**	-0.009**	0.417***	0.045*	0.040
	(0.015)	(0.002)	(0.004)	(0.070)	(0.026)	(0.029)
Observations	16,760	16,760	16,760	32,069	46,228	55,716
R-squared	0.243	0.951	0.884	0.020	0.929	0.852

This table presents coefficient estimates from regressing population in 1900 on rail access, or being within 10 km of the rail network in 1892; population in 1910 on rail access in 1900; and population in 1921 on rail access in 1921. The regressions control for latitude, geographic fundamentals, average air temperature, precipitation, and initial population. At the locality level, the number of pre-colonial indigenous communities corresponds to those within a 10km radius of each locality's coordinates. Average precipitation, air temperature and caloric suitability for each product is the average within the same size radius. All continuous covariates are standardized following Wooldridge (2012) by subtracting the mean and dividing by the standard deviation. Grid cells are 0.1×0.1 degrees (approximately 11×11 km along the equator). Standard errors are adjusted for spatial dependence following Conley (1999) and using a 300 km cutoff. Significance levels are denoted as * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.15: Shapely Decomposition
From Regressing Log Population in 1921 on Rail Access in 1910 at the Locality Level

	All Localities		Northern Localities		Southern Localities	
	Shapley Value	Percent	Shapley Value	Percent	Shapley Value	Percent
Within 10 km of Rail in 1910	0.00761	3.68	0.00981	6.33	0.00450	2.01
Other Distance to Rail Indicators	0.00236	1.14	0.00339	2.19	0.00295	1.32
Climate	0.01054	5.09	0.00922	5.95	0.01546	6.91
Geographic Characteristics	0.02751	13.29	0.01356	8.76	0.03887	17.39
Mineral Resources	0.01116	5.39	0.00705	4.55	0.01327	5.94
Soil Suitability	0.00940	4.54	0.00321	2.07	0.00888	3.97
Initial Populaiton Proxies	0.12801	61.86	0.09721	62.75	0.13012	58.21
Distance to Locations of Economic and Historical Importance	0.01036	5.01	0.01146	7.40	0.00948	4.24
Total	0.20695	100.00	0.15491	100.00	0.22353	100.00

Shapely decomposition results are from regressing log population in 1921 on rail access in 1910. Covariates include those for local geographic characteristics (latitude, distance to the nearest rivers and oceans, and elevation), climate (average air temperature and precipitation), soil suitability (using the average caloric indicies for bananas, chickpeas, coffee cotton and maize from Galor and Özak, 2016a), inital population proxies (the number of pre colonial indiegnous communities and population in 1900) and distance to other locations of economic and historical importance (such as the Camino Real and insurgent outbreaks from the Revolution). The number of pre-colonial indigenous communities corresponds to those within a 10km radius of each locality's coordinates. Average precipitation, air temperature and caloric suitability for each product is the average within the same size radius. Other distance rail indicators identify localities between 10 to 30 km away from the nearest rail line. Northern cells are those with a latitude greater than 22.5 degrees.

Table A.16: Rail Access and Population Between 1900 and 1921, Tobit Average Marginal Effects at the Locality Level

	Dependent Variable: Log Population in 1900									
Within 10km of Rail in 1892	0.392***	0.234**	0.234***	0.177***	0.133**	0.103**	0.124**	0.131**	0.131**	
	(0.108)	(0.098)	(0.089)	(0.068)	(0.060)	(0.052)	(0.057)	(0.058)	(0.059)	
Pseudo R-squared	0.003	0.015	0.025	0.062	0.068	0.070	0.070	0.070	0.070	
N	32,069	32,069	32,069	32,069	32,069	32,069	32,069	32,069	32,069	
	Dependent Variable: Log Population in 1910									
Within 10km of Rail in 1900	0.705***	0.338***	0.349***	0.345***	0.310***	0.273***	0.308***	0.308***	0.308***	
	(0.100)	(0.079)	(0.071)	(0.081)	(0.063)	(0.054)	(0.060)	(0.063)	(0.064)	
Pseudo R-squared	0.004	0.028	0.029	0.048	0.05	0.052	0.051	0.05	0.050	
N	46,228	46,228	46,228	46,228	46,228	46,228	46,228	46,228	46,228	
	Dependent Variable: Log Population in 1921									
Within 10km of Rail in 1910	0.738***	0.368***	0.397***	0.396***	0.383***	0.360***	0.365***	0.383***	0.383***	0.360***
	(0.086)	(0.063)	(0.055)	(0.080)	(0.069)	(0.062)	(0.069)	(0.069)	(0.068)	(0.067)
Pseudo R-squared	0.004	0.033	0.035	0.049	0.05	0.051	0.051	0.051	0.051	0.051
N	55,716	55,716	55,716	55,716	55,716	55,716	55,716	55,716	55,716	55,716
Initial Population Proxy		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Latitude			Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects				Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographic Characteristics					Yes	Yes	Yes	Yes	Yes	Yes
Distance to Historical Mines						Yes				
Distance to Other Mineral Deposits							Yes			
Distance to U.S. Border								Yes		
Distance to the Camino Real									Yes	
Insurgent Outbreaks										Yes

This table presents Tobit estimates of log population regressed on rail access (i.e., being within 10 km of the nearest rail) from approximately ten years prior. Log population as a lower censored limit of zero. The regressions control for latitude, geographic fundamentals, average air temperature, precipitation, and initial population. At the locality level, the number of pre-colonial indigenous communities corresponds to those within a 10km radius of each locality's coordinates. Average precipitation, air temperature and caloric suitability for each product is the average within the same size radius. Standard errors are clustered at the state level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.17: Short Run Results by Locality Type Using Categorized Distance to Rail

	All Localities	Never Urban	Ever Urban	Urban in 1900
	(1)	(2)	(3)	(4)
Dependent Variable: Log Population in 1900				
Within 10 km from Rail in 1892	0.240**	0.202**	0.382***	0.382***
	(0.091)	(0.094)	(0.077)	(0.077)
Observations	32,069	31,504	565	565
R-squared	0.220	0.226	0.140	0.140
Dependent Variable: Log Population in 1910				
Within 10 km from Rail in 1900	0.408***	0.349***	0.551***	0.231
	(0.080)	(0.082)	(0.162)	(0.144)
Observations	46,228	45,539	689	565
R-squared	0.227	0.206	0.284	0.573
Dependent Variable: Log Population in 1921				
Within 10 km from Rail in 1910	0.485***	0.440***	0.547***	0.678***
	(0.117)	(0.118)	(0.132)	(0.159)
Observations	55,716	54,933	783	565
R-squared	0.240	0.217	0.182	0.233

This table presents OLS estimates of log population regressed on categorized distance to historical rail (about) a decade prior, controlling for latitude, geographic fundamentals, average air temperature and precipitation, and initial population proxies. Categorized distance to rail is defined in 10 km intervals from 0 to 60 km from the nearest rail line. In each year, localities that were ever urban had at least 2,500 inhabitants at one point in time before or during that year. Standard errors clustered at the state level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.18: Short Run Results by Locality Type Using Continuous Distance to Rail

	All Localities	Never Urban	Ever Urban	Urban in 1900
	(1)	(2)	(3)	(4)
Dependent Variable: Log Population in 1900				
Log Distance from Rail in 1892	-0.040	-0.012	-0.153***	-0.153***
	(0.028)	(0.029)	(0.021)	(0.021)
Observations	32,069	31,504	565	565
R-squared	0.163	0.170	0.116	0.116
Dependent Variable: Log Population in 1910				
Log Distance from Rail in 1900	-0.089***	-0.066**	-0.183***	-0.089***
	(0.029)	(0.031)	(0.040)	(0.029)
Observations	46,228	45,539	689	46,228
R-squared	0.186	0.162	0.242	0.186
Dependent Variable: Log Population in 1921				
Within 10km from Rail in 1910	-0.108***	-0.088**	-0.193***	-0.249***
	(0.031)	(0.033)	(0.030)	(0.040)
Observations	55,716	54,933	783	565
R-squared	0.206	0.181	0.147	0.186

This table presents OLS estimates of log population regressed on log distance to historical rail (about) a decade prior, controlling for latitude, geographic fundamentals, average air temperature and precipitation, and initial population proxies. Categorized distance to rail is defined in 10 km intervals from 0 to 60 km from the nearest rail line. In each year, localities that were ever urban had at least 2,500 inhabitants at one point in time before or during that year. Standard errors clustered at the state level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.19: Kleibergen-Paap rk Wald F Stat. by Instrument and Unit Produced by Regressing Log Population in 1921 on Rail Access in 1910

Instrument Created Using:	Custom Stations and Top 1% of Localities (Sorted by Population)		Custom Stations, Cities with +10k Inhabitants and Prolific Mines		Custom Stations and Important Cities in the 1892 Census		Custom Stations, Important Cities in the 1892 Census, and Prolific Mines	
	Straight Line LCP	Terrain Based LCP	Straight Line LCP	Terrain Based LCP	Straight Line LCP	Terrain Based LCP	Straight Line LCP	Terrain Based LCP
Grid Cells								
Kleibergen-Paap rk Wald F Statistic	37.43	42.38	14.51	17.09	12.87	9.396	21.82	17.16
N	14,820	14,820	14,274	14,274	15,504	15,501	14,868	14,865
All Localities								
Kleibergen-Paap rk Wald F Statistic	43.78	103.6	16.23	54.15	12.56	11.79	20.37	11.46
N	264,893	264,893	252,575	252,575	297,594	297,594	281,171	281,171
Localities Ever Inhabited by 1921								
Kleibergen-Paap rk Wald F Statistic	61.02	102.9	16.05	31.61	7.210	11.97	2.365	10.83
N	43,333	43,333	40,769	40,769	50,067	50,067	46,392	46,392

Estimates are produced using the control function approach outlined in Heckman (1978) and Wooldridge (2010). The indicator for rail access, or being within 10 km of the network, is regressed on distance to a least cost path, along with covariates for latitude, geographic fundamentals, average air temperature, precipitation, and initial population. The resulting estimates from the Probit model are used to produced predicted probabilities for rail access. The predicted probabilities are used as an instrument for rail access in second stage, in which log population is regressed on rail access. This procedure is replicated for each of the eight possible instruments detailed above. Units within 30km of a node used to create the least cost paths are excluded from the sample. For localities, standard errors in the first and second stage are bootstrapped using 100 iterations and clustered at the state level. For grid cells, standard errors in the first and second stages are also bootstrapped using 100 iterations and spatial clusterd at with a 300 km cutoff following Conley (1999). The sample size and resulting Kleibergen-Paap rk Wald F statistic are reported. Grid cells are 0.1×0.1 degrees (i.e., approximately 11×11 km at the equator) in size.

Table A.20: Comparing 2SLS and IV Control Function Estimates at the Grid Cell Level

Dependent Variable: Log Population in 1921		
	2SLS	IV Control Function
	(1)	(2)
Within 10km from Rail in 1910	2.425	1.027***
	(1.887)	(0.345)
Log Population in 1900	0.711***	0.725***
	(0.034)	(0.009)
Observations	14,820	14,820
Kleinbergen-Papp rk Wald F Statistic	3.794	41.22

Column 1 presents 2SLS estimates from regressing log population in 1921 on categorized distance to rail network in 1910. The primary variable of interest is rail access, or being within 10 km of the network. Additional distance indicators capture being 10 to 20 and 20 to 30 km from the rail network. In the linear first stage, rail access is instrumented using log distance to the least cost path network, along with covariates for latitude, geographic fundamentals, average air temperature, precipitation, and initial population. Column 2 presents the same estimates but produced using the control function approach outlined in Heckman (1978) and Wooldridge (2010). This approach accounts for the non linear relationship between the instrument and the probability of having rail access in 1910. Standard error estimates from this approach are bootstrapped in the first and second stage using 100 iterations. Regardless of the approach, standard errors are adjusted for spatial dependence following Conley (1999) and using a 300 km distance cutoff. The sample size and resulting Kleibergen-Paap rk Wald F statistic are reported. Grid cells within 30 km of a node used to construct the least cost path network used to instrument rail access are excluded from the sample. Grid cells are 0.1×0.1 degrees (i.e., approximately 11×11 km at the equator) in size.

Table A.21: Kleibergen-Paap rk Wald F Stat. by Instrument and Unit Produced by Regressing Log Population in 1921 on Log Distance to Rail in 1910

Instrument Created Using:	Custom Stations and Top 1% of Localities (Sorted by Population)		Custom Stations, Cities with +10k Inhabitants and Prolific Mines		Custom Stations and Important Cities in the 1892 Census		Custom Stations, Important Cities in the 1892 Census, and Prolific Mines	
	Straight Line LCP	Terrain Based LCP	Straight Line LCP	Terrain Based LCP	Straight Line LCP	Terrain Based LCP	Straight Line LCP	Terrain Based LCP
Grid Cells								
Kleibergen-Paap rk Wald F Statistic	14.54	6.798	0.101	2.209	1.049	4.094	1.551	4.029
N	14,820	14,820	14,274	14,274	15,504	15,501	14,868	14,865
All Localities								
Kleibergen-Paap rk Wald F Statistic	37.04	132.8	10.06	52.65	3.912	3.892	1.182	3.368
N	264,893	264,893	252,575	252,575	297,594	297,594	281,171	281,171
Localities Ever Inhabited by 1921								
Kleibergen-Paap rk Wald F Statistic	65.78	113.4	12.28	36.39	7.403	8.523	1.665	7.828
N	43,333	43,333	40,769	40,769	50,067	50,067	46,392	46,392

Estimates are produced from using 2SLS and regressing log population in 1921 on log distance to the historical rail network in 1910, along with covariates for latitude, geographic fundamentals, average air temperature, precipitation, and initial population. In the first stage log distance to the least cost path network serves as an instrument for log distance to historical rail in 1910. Units within 30km of a node used to create the least cost paths are excluded from the sample. The Kleibergen-Paap rk Wald F statistic and resulting sample size are reported. Grid cells are 0.1×0.1 degrees (i.e., 11×11 km) in size.

Table A.22: Long Run Locality Estimates Using Categorical Distance to Rail in 1910

	1921	1930	1940	1950	1960	1970	1980	1990	2000	2010	2020
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Ever Inhabited											
OLS											
Within 10km from Rail in 1910	0.335***	0.257***	0.299***	0.303***	0.277**	0.183*	-0.045	-0.132	-0.171*	-0.120	-0.106
	(0.089)	(0.086)	(0.101)	(0.100)	(0.107)	(0.107)	(0.089)	(0.095)	(0.093)	(0.095)	(0.092)
N	55,716	63,181	74,215	82,956	93,599	118,116	155,407	218,756	290,292	320,681	344,469
R-squared	0.206	0.218	0.213	0.199	0.193	0.187	0.210	0.232	0.226	0.227	0.230
IV Control Function											
Within 10km from Rail in 1910	0.687***	0.520***	0.558***	0.747***	0.769***	0.650***	0.231	0.053	0.032	0.215	0.264
	(0.251)	(0.176)	(0.162)	(0.202)	(0.229)	(0.207)	(0.189)	(0.215)	(0.191)	(0.212)	(0.208)
N	43,333	49,253	58,079	64,957	73,489	92,395	121,968	171,646	225,758	247,934	264,893
Kleibergen-Paap rk Wald F Statistic	109.2	106.2	99.62	113	115.4	135.2	164.5	140.1	136.1	139.8	142.5
Never Urban											
OLS											
Within 10km from Rail in 1910	0.191*	0.122	0.174	0.169	0.133	0.042	-0.127	-0.189*	-0.220**	-0.174*	-0.162*
	(0.102)	(0.100)	(0.112)	(0.109)	(0.116)	(0.111)	(0.092)	(0.095)	(0.093)	(0.093)	(0.090)
N	52,041	59,337	70,146	78,719	89,129	113,345	150,486	213,696	285,124	315,322	339,041
R-squared	0.142	0.159	0.156	0.145	0.143	0.135	0.171	0.200	0.192	0.193	0.195
IV Control Function											
Within 10km from Rail in 1910	0.416*	0.268*	0.329**	0.494***	0.528**	0.440***	0.071	-0.062	-0.065	0.121	0.173
	(0.236)	(0.154)	(0.167)	(0.189)	(0.240)	(0.158)	(0.225)	(0.195)	(0.177)	(0.210)	(0.200)
N	41,072	46,888	55,570	62,383	70,835	89,633	119,118	168,733	222,801	244,938	261,878
Kleibergen-Paap rk Wald F Statistic	100.8	97.85	91.97	104.5	107.9	128.6	158.9	137.4	133.7	137.9	140.6
Ever Urban											
OLS											
Within 10km from Rail in 1910	0.287**	0.252***	0.299***	0.306***	0.332***	0.410***	0.077	-0.233	-0.227	-0.169	-0.078
	(0.114)	(0.088)	(0.090)	(0.101)	(0.113)	(0.111)	(0.224)	(0.253)	(0.271)	(0.294)	(0.277)
N	3,675	3,844	4,069	4,237	4,470	4,771	4,921	5,060	5,168	5,359	5,428
R-squared	0.244	0.272	0.277	0.188	0.130	0.110	0.362	0.379	0.355	0.314	0.276
IV Control Function											
Within 10km from Rail in 1910	0.767**	0.568**	0.275	0.800***	0.478	0.854***	0.225	-0.313	-0.281	-0.264	-0.371
	(0.362)	(0.282)	(0.310)	(0.253)	(0.317)	(0.295)	(0.464)	(0.577)	(0.569)	(0.597)	(0.567)
N	2,261	2,365	2,509	2,574	2,654	2,762	2,850	2,913	2,957	2,996	3,015
Kleibergen-Paap rk Wald F Statistic	104.0	106.7	102.6	104.1	104.7	113	116.8	103.6	113.8	105.4	117.7

This table presents estimates of log population regressed on rail access or being within 10 km of the historical rail network in 1910, controlling for latitude, geographic fundamentals, average air temperature and precipitation, and initial population proxies. Additional distance indicators up to 30 km are included to capture spillover effects. OLS and IV estimates are presented for comparison. For IV estimates, having rail access is instrumented using log distance to the least cost path network following the inconsequential units approach (Redding and Turner, 2015). IV estimates are produced using the control function approach outlined in Heckman (1978) and Wooldridge (2010). Localities within 30 km of nodes use to construct the least-cost path network are excluded from the sample. Standard errors clustered at the state level and IV control function estimates are bootstrapped using 100 iterations in both the first and second stages. Significance levels are denoted as * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.23: Long Run Locality Estimates Using Log Distance to Rail in 1910

	1921	1930	1940	1950	1960	1970	1980	1990	2000	2010	2020
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Ever Inhabited											
OLS											
Log Distance to Rail in 1910 (km)	-0.108***	-0.096***	-0.120***	-0.116***	-0.107***	-0.071*	0.024	0.058*	0.076**	0.063*	0.059*
	(0.031)	(0.029)	(0.033)	(0.034)	(0.038)	(0.038)	(0.033)	(0.033)	(0.032)	(0.033)	(0.032)
N	55,716	63,181	74,215	82,956	93,599	118,116	155,407	218,756	290,292	320,681	344,469
R-squared	0.206	0.219	0.214	0.200	0.194	0.187	0.209	0.231	0.226	0.228	0.231
2SLS											
Log Distance to Rail in 1910 (km)	-0.205***	-0.173***	-0.204***	-0.278***	-0.304***	-0.243***	-0.100	-0.028	-0.019	-0.072	-0.077
	(0.073)	(0.054)	(0.053)	(0.064)	(0.079)	(0.060)	(0.068)	(0.062)	(0.061)	(0.062)	(0.060)
N	43,333	49,253	58,079	64,957	73,489	92,395	121,968	171,646	225,758	247,934	264,893
Kleibergen-Paap rk Wald F Statistic	115.1	110.8	96.54	111.1	113.5	121.6	127.6	122.2	127.5	127.1	125.9
Never Urban											
OLS											
Log Distance to Rail in 1910 (km)	-0.054	-0.047	-0.075*	-0.066*	-0.052	-0.017	0.048	0.074**	0.090***	0.078**	0.075**
	(0.038)	(0.035)	(0.038)	(0.038)	(0.040)	(0.039)	(0.033)	(0.034)	(0.032)	(0.032)	(0.031)
N	52,041	59,337	70,146	78,719	89,129	113,345	150,486	213,696	285,124	315,322	339,041
R-squared	0.142	0.159	0.156	0.145	0.143	0.135	0.171	0.200	0.193	0.193	0.195
2SLS											
Log Distance to Rail in 1910 (km)	-0.131*	-0.103*	-0.141***	-0.206***	-0.235***	-0.188***	-0.056	0.001	0.005	-0.047	-0.053
	(0.072)	(0.055)	(0.053)	(0.064)	(0.077)	(0.056)	(0.066)	(0.061)	(0.060)	(0.061)	(0.060)
N	41,072	46,888	55,570	62,383	70,835	89,633	119,118	168,733	222,801	244,938	261,878
Kleibergen-Paap rk Wald F Statistic	110.0	106.2	92.55	106.8	110.1	119.4	126.3	121.2	126.4	126.0	124.9
Ever Urban											
OLS											
Log Distance to Rail in 1910 (km)	-0.107***	-0.094***	-0.129***	-0.157***	-0.180***	-0.220***	0.087	0.181	0.210	0.229	0.217
N	(0.031)	(0.025)	(0.025)	(0.023)	(0.025)	(0.027)	(0.110)	(0.125)	(0.134)	(0.145)	(0.151)
R-squared	3,675	3,844	4,069	4,237	4,470	4,771	4,921	5,060	5,168	5,359	5,428
2SLS											
	0.247	0.274	0.281	0.199	0.144	0.129	0.362	0.380	0.358	0.318	0.280
Log Distance to Rail in 1910 (km)	-0.280***	-0.212**	-0.147	-0.346***	-0.240**	-0.254**	-0.125	0.278	0.231	0.179	0.110
	(0.093)	(0.093)	(0.114)	(0.082)	(0.107)	(0.099)	(0.127)	(0.217)	(0.197)	(0.214)	(0.249)
N	2,261	2,365	2,509	2,574	2,654	2,762	2,850	2,913	2,957	2,996	3,015
Kleibergen-Paap rk Wald F Statistic	137.5	130.0	120.1	125.2	120.8	129.4	128.7	110.9	118.8	108.6	110.1

This table presents estimates of log population regressed on log distance to the historical rail network in 1910, controlling for latitude, geographic fundamentals, average air temperature and precipitation, and initial population proxies. Additional distance indicators up to 30 km are included to capture spillover effects. OLS and IV estimates are presented for comparison. For IV estimates, having rail access is instrumented using log distance to the least cost path network following the inconsequential units approach (Redding and Turner, 2015). IV estimates are produced using 2SLS. Localities within 30 km of nodes use to construct the least-cost path network are excluded from the sample. Standard errors clustered at the state level. Significance levels are denoted as * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A.24: Marginal Effects From the Timing of Rail Access on Log Population
IV Control Function Estimates at the Grid Cell Level, 1921 - 2020

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Gained Rail Access:	1921	1930	1940	1950	1960	1970	1980	1990	2000	2010	2020
Before 1878	-0.097	0.854**	0.336	1.007***	1.468***	2.146***	2.335***	2.747***	2.425***	2.514***	2.695***
	(0.446)	(0.354)	(0.346)	(0.387)	(0.424)	(0.488)	(0.766)	(0.635)	(0.748)	(0.780)	(0.761)
1878 to 1882	0.316	0.672*	0.532	1.135***	1.298***	1.446***	1.528***	1.560***	1.614***	1.719***	1.822***
	(0.419)	(0.343)	(0.394)	(0.352)	(0.380)	(0.402)	(0.324)	(0.419)	(0.375)	(0.445)	(0.451)
1882 to 1884	0.734***	0.831***	0.893***	1.324***	1.613***	1.452***	1.521***	1.490***	1.397***	1.631***	1.919***
	(0.247)	(0.246)	(0.234)	(0.235)	(0.236)	(0.217)	(0.271)	(0.252)	(0.258)	(0.304)	(0.331)
1884 to 1886	1.247***	1.496***	1.355***	1.433***	1.587***	1.602***	1.578***	1.741***	1.597***	1.810***	1.939***
	(0.227)	(0.224)	(0.229)	(0.246)	(0.248)	(0.276)	(0.268)	(0.266)	(0.306)	(0.329)	(0.346)
1886 to 1892	1.192***	1.300***	0.967***	1.087***	1.190***	0.985***	0.816***	1.016***	0.930***	1.171***	1.445***
	(0.219)	(0.224)	(0.205)	(0.210)	(0.229)	(0.193)	(0.211)	(0.198)	(0.215)	(0.225)	(0.281)
1892 to 1896	-0.207	-0.050	-0.056	0.222	0.294	0.429	0.612	0.732**	0.702**	0.992**	1.326***
	(0.488)	(0.416)	(0.469)	(0.405)	(0.437)	(0.351)	(0.390)	(0.352)	(0.324)	(0.388)	(0.367)
1896 to 1900	0.956***	1.293***	1.250***	1.429***	1.599***	1.556***	1.430***	1.486***	1.392***	1.582***	1.768***
	(0.243)	(0.190)	(0.188)	(0.224)	(0.240)	(0.185)	(0.210)	(0.203)	(0.211)	(0.231)	(0.259)
1900 to 1904	1.000***	0.956***	0.809**	0.855***	0.790**	0.765*	0.788**	0.873**	0.549	0.643*	0.884**
	(0.384)	(0.364)	(0.360)	(0.301)	(0.342)	(0.397)	(0.360)	(0.359)	(0.393)	(0.390)	(0.392)
1904 to 1910	0.504**	0.772***	0.819***	1.029***	1.099***	1.198***	1.081***	1.217***	1.084***	1.224***	1.396***
	(0.234)	(0.188)	(0.169)	(0.205)	(0.205)	(0.208)	(0.202)	(0.207)	(0.224)	(0.241)	(0.228)
Observations	14,820	14,820	14,820	14,820	14,820	14,820	14,820	14,820	14,820	14,820	14,820

The table presents coefficient estimates associated with gaining rail access or being within 10 km of the network during each time period. Gaining rail access is defined as being within 10 km of the rail network. Coefficients are from the IV control function model in which log population is regressed on categorized distance to rail, and controlling for latitude, geographic fundamentals, average air temperature and precipitation, and initial population proxies. The indicator for rail access is instrumented using log distance to the straight line least cost path network, constructed from the most populated localities in 1900 and custom stations, as well as log distance to the nearest of these two locations (i.e., nodes). Cells within 30 km of nodes used to construct the least-cost path network are excluded from the estimation sample. Standard errors are bootstrapped in the first and second stage using 100 iterations, and are adjusted for spatial dependence following Conley (1999) and using a 300 km distance cutoff. Grid cells are 0.1×0.1 degrees (i.e., approximately 11×11 km at the equator) in size. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$. The long run effects of gaining rail during select time periods is plotted in Figure 0.8.

Table A.25: Marginal Effects From the Timing of Rail Access on Log Population
IV Control Function Estimates for Urban and Rural Localities, 1921 - 2020

Panel A. Localities Ever Urban

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Gained Rail Access:	1921	1930	1940	1950	1960	1970	1980	1990	2000	2010	2020
Before 1878	-0.060	-0.315	-0.211	0.243	0.602	1.687***	-3.336	-3.779*	-3.603	-3.912	-4.946
	(0.426)	(0.771)	(0.415)	(0.417)	(0.465)	(0.598)	(2.242)	(1.963)	(2.334)	(13.027)	(3.341)
1878 to 1882	2.223**	1.354	0.597	0.316	0.348	0.241	0.330	0.860	0.796	0.309	-0.686
	(0.874)	(0.977)	(1.138)	(1.693)	(1.647)	(1.211)	(1.170)	(2.813)	(2.139)	(2.233)	(3.741)
1882 to 1884	0.598	0.459	0.292	0.471**	0.470**	0.608**	0.534**	1.107**	0.963**	1.019**	0.662*
	(0.406)	(0.320)	(0.322)	(0.237)	(0.237)	(0.286)	(0.229)	(0.458)	(0.440)	(0.457)	(0.363)
1884 to 1886	0.265	0.354	0.056	0.413**	0.424**	0.473**	0.265	-0.843	-0.755	-0.898	-0.583
	(0.348)	(0.270)	(0.231)	(0.200)	(0.213)	(0.214)	(0.247)	(0.809)	(0.788)	(0.716)	(0.553)
1886 to 1892	0.158	0.159	0.118	0.196	0.014	0.149	0.135	0.318	0.287	0.365	0.284
	(0.279)	(0.235)	(0.253)	(0.302)	(0.407)	(0.317)	(0.279)	(0.321)	(0.301)	(0.385)	(0.369)
1892 to 1896	-0.653	-0.475	-0.394	0.106	0.113	-0.399	-0.205	-0.698	-0.828	-0.517	-0.577
	(0.510)	(0.492)	(0.674)	(0.364)	(0.817)	(0.971)	(0.959)	(1.101)	(1.362)	(1.405)	(1.578)
1896 to 1900	0.145	0.417	0.228	0.431	0.408	0.463	0.524	-0.774	-0.913	-0.990	-1.263
	(0.322)	(0.268)	(0.282)	(0.412)	(0.352)	(0.295)	(0.359)	(0.911)	(0.969)	(0.918)	(1.073)
1900 to 1904	0.165	0.240	0.158	0.342	0.008	0.475*	0.436*	0.139	0.112	0.343	0.170
	(0.634)	(0.551)	(0.316)	(0.299)	(0.488)	(0.248)	(0.240)	(0.469)	(0.460)	(0.548)	(0.583)
1904 to 1910.	0.312	0.568**	0.391*	0.461	0.296**	0.562***	0.016	0.033	0.100	0.260	0.001
	(0.434)	(0.226)	(0.235)	(0.300)	(0.131)	(0.206)	(0.239)	(0.277)	(0.294)	(0.281)	(0.252)
Log Population in 1900	0.249***	0.228***	0.203***	0.169***	0.147***	0.133***	0.150***	0.230***	0.218***	0.192***	0.168***
	(0.021)	(0.017)	(0.023)	(0.016)	(0.016)	(0.011)	(0.020)	(0.032)	(0.031)	(0.026)	(0.025)
Observations	2,261	2,365	2,509	2,574	2,654	2,762	2,850	2,913	2,957	2,996	3,015

Panel B. Localities Never Urban

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Gained Rail Access:	1921	1930	1940	1950	1960	1970	1980	1990	2000	2010	2020
Before 1878	-0.362	-0.404	-0.683	-0.480	-0.598	-0.223	-1.344***	-1.931***	-1.772	-1.675***	-1.661***
	(0.397)	(0.391)	(0.648)	(0.549)	(0.381)	(0.896)	(0.434)	(0.708)	(1.550)	(0.589)	(0.579)
1878 to 1882	0.709	0.400	0.222	0.650	0.700	0.275	-0.254	-0.447	-0.598	-0.504	-0.427
	(0.438)	(0.574)	(0.900)	(0.729)	(1.138)	(1.376)	(1.148)	(1.748)	(1.111)	(1.033)	(1.113)
1882 to 1884	0.308	0.271	0.230	0.277	0.332	0.178	-0.006	-0.112	-0.157	-0.059	-0.132
	(0.247)	(0.251)	(0.231)	(0.293)	(0.304)	(0.191)	(0.242)	(0.186)	(0.211)	(0.196)	(0.221)
1884 to 1886	0.307	0.456*	0.422	0.537*	0.576**	0.271	0.087	-0.201	-0.328	-0.256	-0.229
	(0.304)	(0.273)	(0.314)	(0.283)	(0.276)	(0.299)	(0.299)	(0.324)	(0.328)	(0.298)	(0.319)
1886 to 1892	0.187	0.305	0.193	0.242	0.149	0.227	-0.001	-0.120	-0.029	0.074	0.099
	(0.212)	(0.187)	(0.201)	(0.196)	(0.214)	(0.188)	(0.187)	(0.169)	(0.154)	(0.153)	(0.149)
1892 to 1896	-0.089	0.097	0.171	0.339	0.515	-0.071	0.325	-0.298	-0.130	0.191	0.430
	(3.238)	(0.459)	(1.792)	(0.416)	(0.562)	(0.578)	(0.771)	(0.545)	(0.468)	(0.463)	(0.419)
1896 to 1900	0.295	0.298	0.310	0.478**	0.515**	0.355	0.303	0.041	-0.057	-0.078	-0.111
	(0.189)	(0.221)	(0.233)	(0.226)	(0.241)	(0.274)	(0.189)	(0.177)	(0.156)	(0.166)	(0.170)
1900 to 1904	0.514*	0.096	0.141	-0.117	0.115	-0.270	-0.112	-0.107	-0.178	-0.098	0.017
	(0.267)	(0.391)	(0.441)	(0.371)	(0.340)	(0.423)	(0.332)	(0.396)	(0.293)	(0.281)	(0.334)
1904 to 1910.	0.307*	0.311**	0.512***	0.509***	0.607***	0.438***	0.261	0.130	0.026	0.042	0.041
	(0.175)	(0.152)	(0.172)	(0.192)	(0.181)	(0.166)	(0.164)	(0.186)	(0.183)	(0.244)	(0.235)
Log Population in 1900	0.234***	0.231***	0.218***	0.242***	0.252***	0.251***	0.291***	0.343***	0.378***	0.407***	0.420***
	(0.018)	(0.012)	(0.011)	(0.012)	(0.011)	(0.012)	(0.011)	(0.014)	(0.020)	(0.024)	(0.025)
Observations	41,072	46,888	55,570	62,383	70,835	89,633	119,118	168,733	222,801	244,938	261,878

The table presents coefficient estimates associated with gaining rail access or being within 10 km of the network during each time period. Gaining rail access is defined as being within 10 km of the rail network. Coefficients are from the IV control function model in which log population is regressed on categorized distance to rail, using data at the locality level, or subsets of localities by urban status, and controlling for latitude, geographic fundamentals, average air temperature and precipitation, and initial population proxies. The indicator for rail access is instrumented using log distance to the terrain-based least cost path network, constructed from the most populated localities in 1900 and custom stations, as well as log distance to the nearest of these two locations (i.e., nodes). Localities within 30 km of nodes used to construct the least-cost path network are excluded from the estimation sample. Localities that were ever urban had at least 2,500 inhabitants at one point in time between 1921 and 2020. Standard errors are bootstrapped in the first and second stage using 100 iterations, and clustered at the state level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$. The long run effects of gaining rail during select time periods is plotted in Figure 0.9.

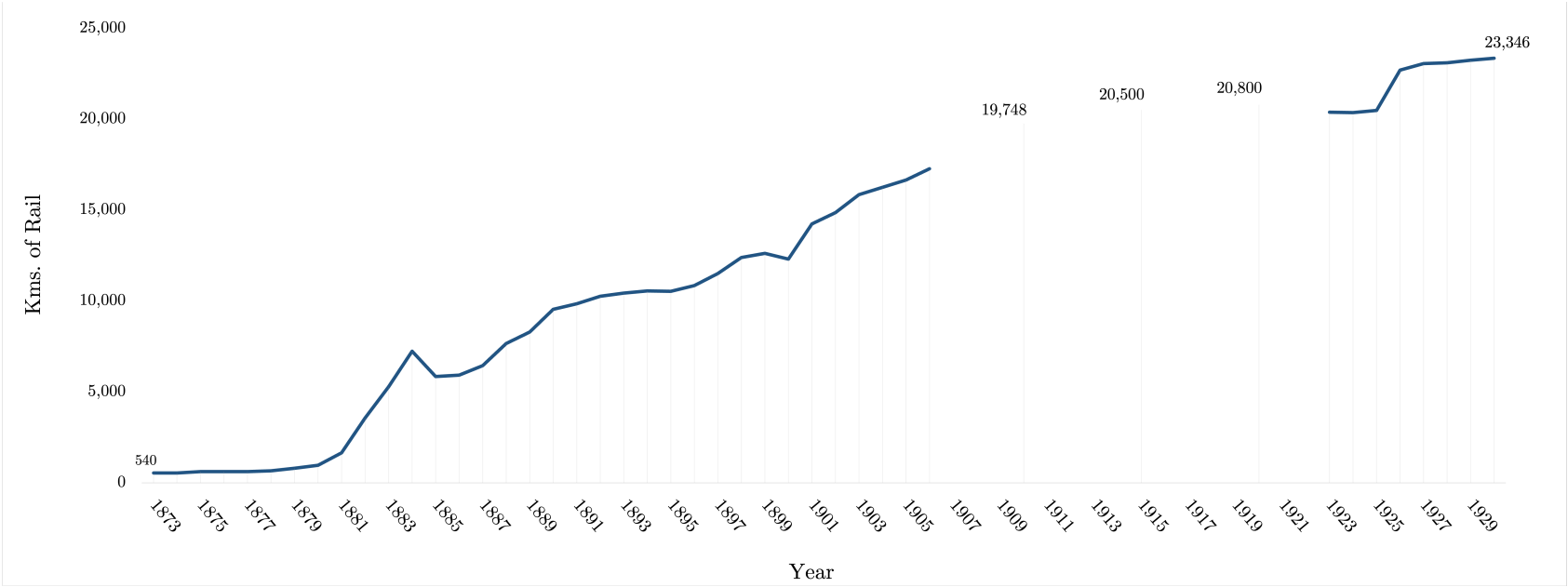
Table A.26: Standardized Probit Estimates for Access to Modern Roads in 1970
Marginal Effects at the Grid Cell and Locality Levels

	Using Grid Cells		Using All Localities	
	(1)	(2)	(3)	(4)
Within 10km of Rail in 1910		0.833***		0.987***
		(0.033)		(0.006)
10 - 20km from Rail in 1910		0.281***		0.302***
		(0.035)		(0.007)
20 - 30km from Rail in 1910		0.043		0.075***
		(0.039)		(0.008)
Z Score of				
Distance to Rail in 1910	-0.267***		-0.294***	
	(0.020)		(0.003)	
Population in 1970	1.004***	0.735***	0.147***	0.106***
	(0.208)	(0.175)	(0.032)	(0.026)
Population in 1900	0.166***	0.142***	-0.019**	-0.013*
	(0.048)	(0.043)	(0.008)	(0.007)
Number of Pre Colonial Indigenous Communities	-0.018	-0.020	0.159***	0.106***
	(0.015)	(0.015)	(0.003)	(0.003)
Latitude	-0.551***	-0.576***	-1.267***	-1.164***
	(0.061)	(0.060)	(0.014)	(0.014)
Average Daily Air Temperature (C) in 1970	0.176***	0.169***	-0.485***	-0.364***
	(0.015)	(0.016)	(0.008)	(0.008)
Average Monthly Precipitation (mm) in 1970	-0.121***	-0.081***	-0.394***	-0.314***
	(0.023)	(0.023)	(0.005)	(0.005)
Z Score of Distance to the Nearest (in km)				
Elevation (km)	-0.007	-0.009	-0.798***	-0.640***
	(0.014)	(0.014)	(0.008)	(0.008)
Intermittent River (km)	0.007	0.028*	0.102***	0.130***
	(0.017)	(0.017)	(0.004)	(0.004)
Perennial River (km)	-0.129***	-0.120***	-0.227***	-0.173***
	(0.023)	(0.023)	(0.005)	(0.005)
Pacific Ocean (km)	0.496***	0.512***	0.588***	0.635***
	(0.054)	(0.054)	(0.011)	(0.011)
Gulf of Mexico (km)	0.583***	0.554***	0.648***	0.650***
	(0.080)	(0.077)	(0.016)	(0.016)
Historical Gold Mine, Wide Distribution (km)	-0.036	-0.037	-0.136***	-0.091***
	(0.028)	(0.029)	(0.006)	(0.006)
Historical Silver Mine, Wide Distribution (km)	-0.025	-0.059	-0.067***	-0.097***
	(0.038)	(0.037)	(0.008)	(0.009)
Coal Deposit (km)	0.100***	0.050**	0.248***	0.223***
	(0.024)	(0.025)	(0.005)	(0.005)
Copper Deposit (km)	-0.287***	-0.203***	-0.095***	-0.157***
	(0.042)	(0.041)	(0.008)	(0.008)
Iron Deposit (km)	-0.286***	-0.181***	-0.284***	-0.239***
	(0.032)	(0.031)	(0.006)	(0.006)
Zinc Deposit (km)	0.513***	0.317***	0.379***	0.295***
	(0.036)	(0.032)	(0.007)	(0.007)
Z Score of Average Caloric Index				
Bananas w/ High Inputs (Thousands)	0.004	0.000	-0.122***	-0.136***
	(0.015)	(0.016)	(0.004)	(0.004)
Chickpeas w/ High Inputs (Thousands)	0.001	0.027	0.054***	0.080***
	(0.019)	(0.019)	(0.004)	(0.005)
Coffee w/ High Inputs (Thousands)	0.062***	0.071***	-0.012***	0.037***
	(0.020)	(0.020)	(0.004)	(0.005)
Cotton w/ High Inputs (Thousands)	-0.072***	-0.053***	0.065***	0.064***
	(0.015)	(0.015)	(0.004)	(0.004)
Maize w/ Low Inputs (Thousands)	-0.010	-0.000	-0.231***	-0.141***
	(0.016)	(0.016)	(0.004)	(0.003)
Constant	-0.591***	-0.772***	-0.093***	-0.411***
	(0.013)	(0.016)	(0.002)	(0.003)
N	16,787	16,787	344,469	344,469
Conditioning for Historical Rail Access Using				
Continous Distance	X		X	
Categorized Distance		X		X

This table presents marginal effects from a Probit model estimating modern road access or being within 10 km of a major modern road relative to historical rail access and local characteristics. Continuous covariates were standardized following Wooldridge (2012) by subtracting the mean and dividing by the standard deviation. Covariates include rail access, defined by categorical distance to the nearest rail line (0–30 km in columns 2 and 4) or continuous distance (columns 1 and 3). Results are reported at the grid cell level in columns (1)–(2) and at the locality level in columns (3)–(4). Additional covariates include 1970 population, the year by which most of major modern roads had been completed, and historical proxies for inital population such as 1900 population and the number of pre-colonial indigenous communities. Grid cells are 0.1×0.1 degrees (approximately 11×11 km along the equator). For grid cell level results, standard errors are adjusted for spatial dependence following Conley (1999) and using a 300 km cutoff. For locality-level results, standard errors are clustered at the state level. At the locality level, the number of pre-colonial indigenous communities corresponds to those within a 10 km radius of each locality's coordinates. Average precipitation, air temperature and caloric suitability for each product is the average within the same size radius. Significance levels are denoted as * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

A.2 Figure Appendix

Figure A.1: Length of Mexico's Rail Network Between 1873 to 1929



The figure illustrates annual estimates of total rail built (in kilometers) from Table 2.3 of Coatsworth (1981).

Figure A.2: Process for Harmonizing and Augmenting Locality Level Population Data

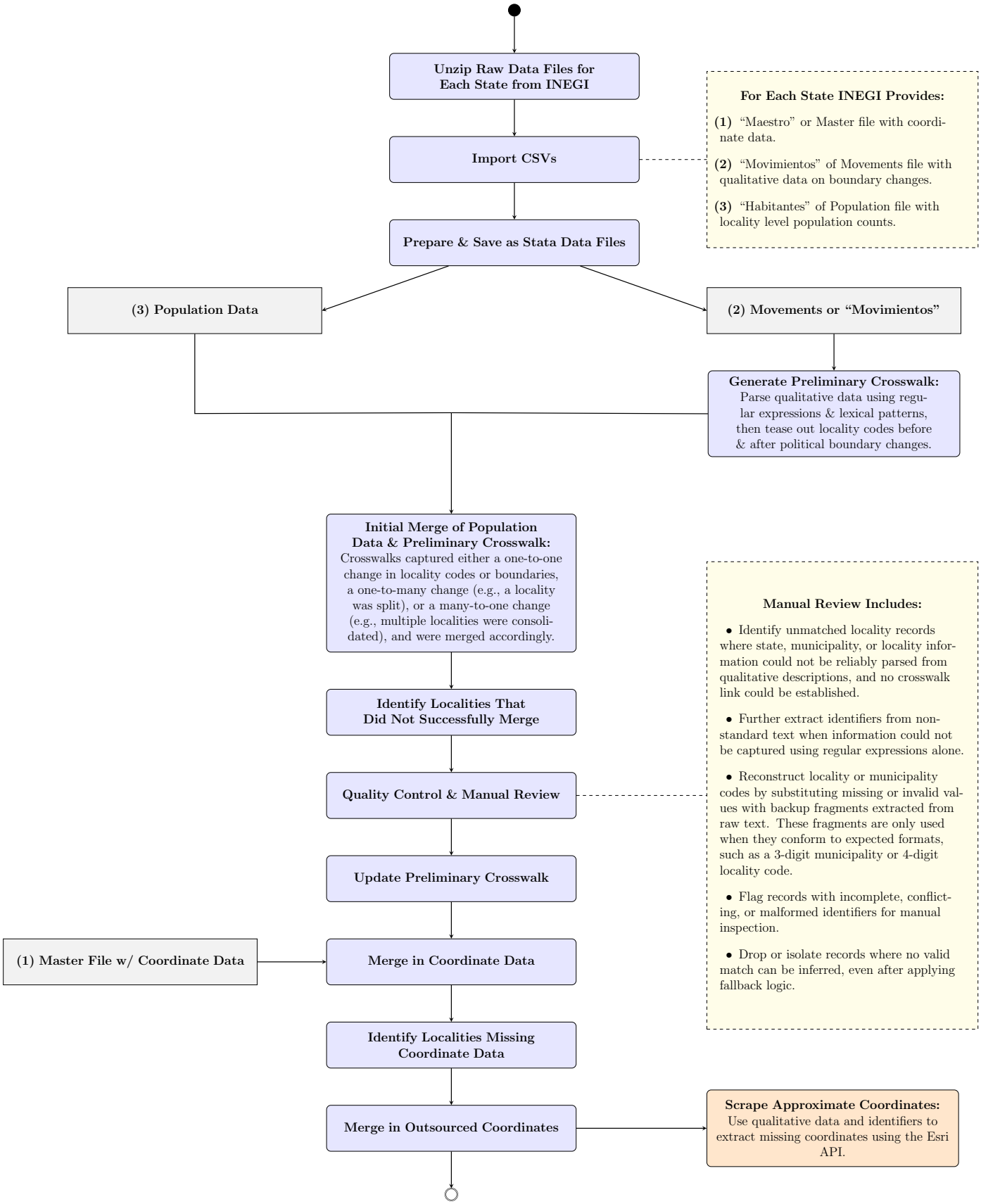
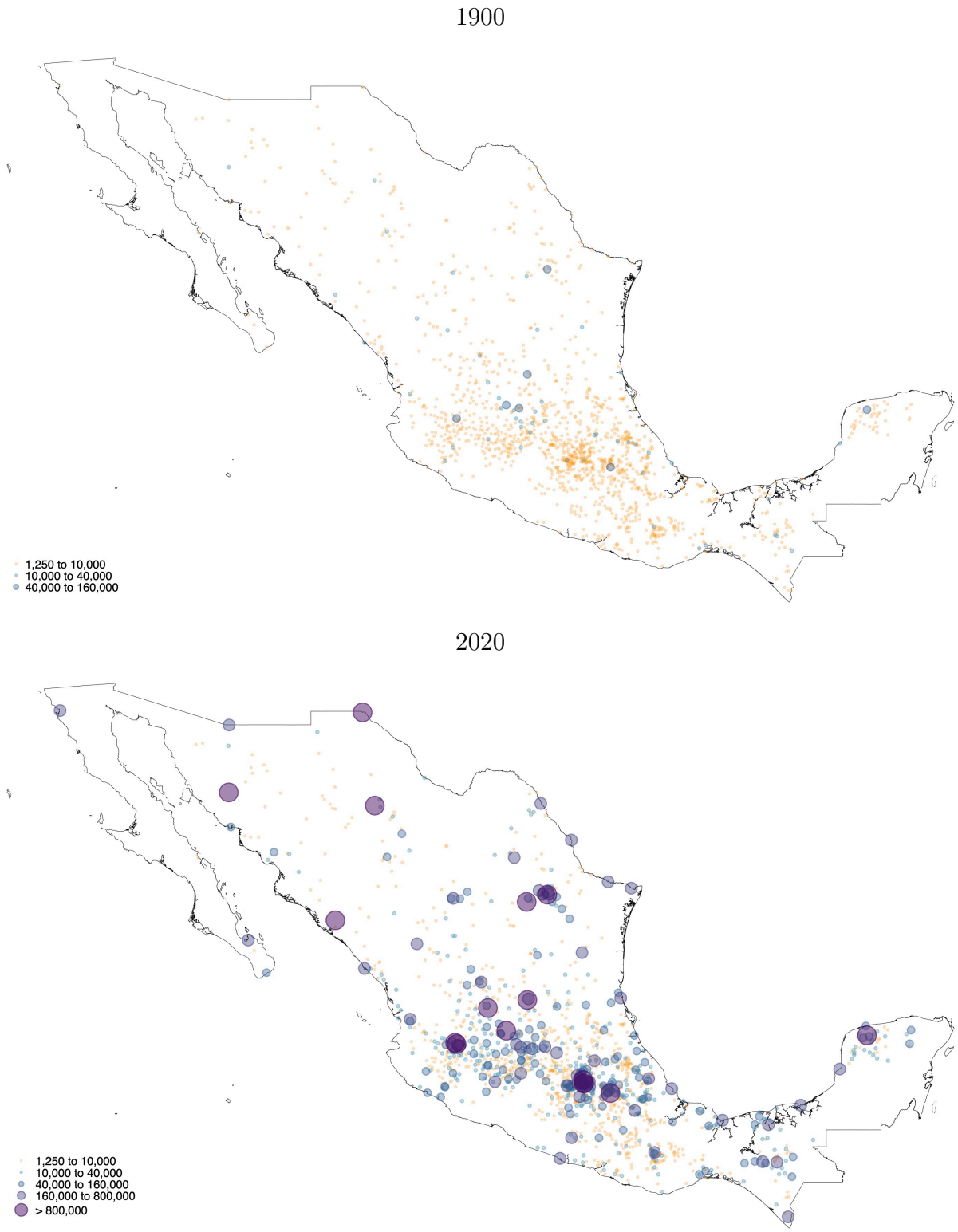


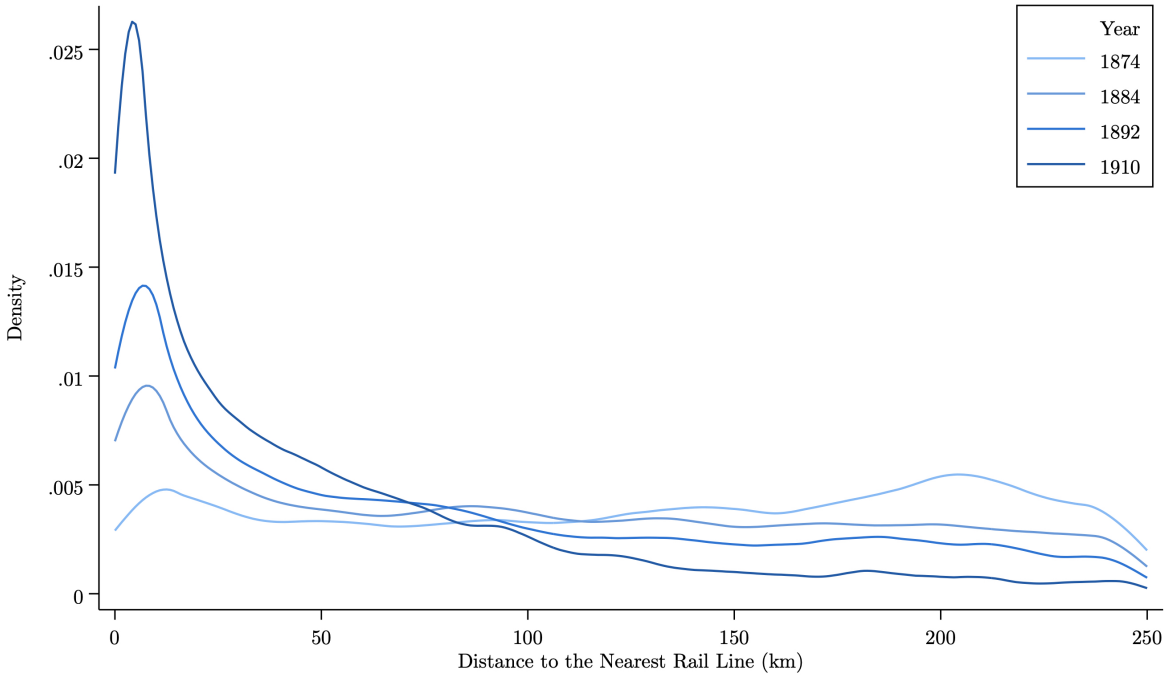
Figure A.3: Distribution of Urban Localities in 1900 and 2020



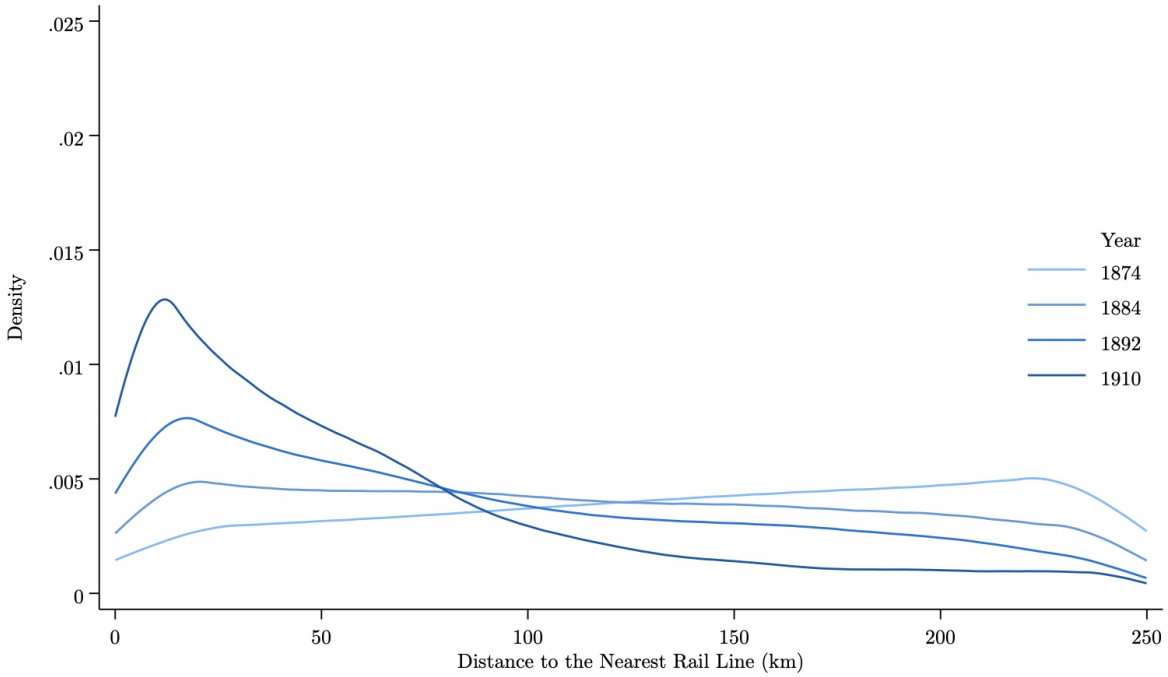
Each map presents the location and size of localities with at least 1,250 inhabitants, a cut off chosen solely for visualization purposes. Summary statistics are provided in Table A.2 of the Table Appendix.

Figure A.4: Kernel Density of Distance to the Nearest Rail Line

Panel A. Locality Level

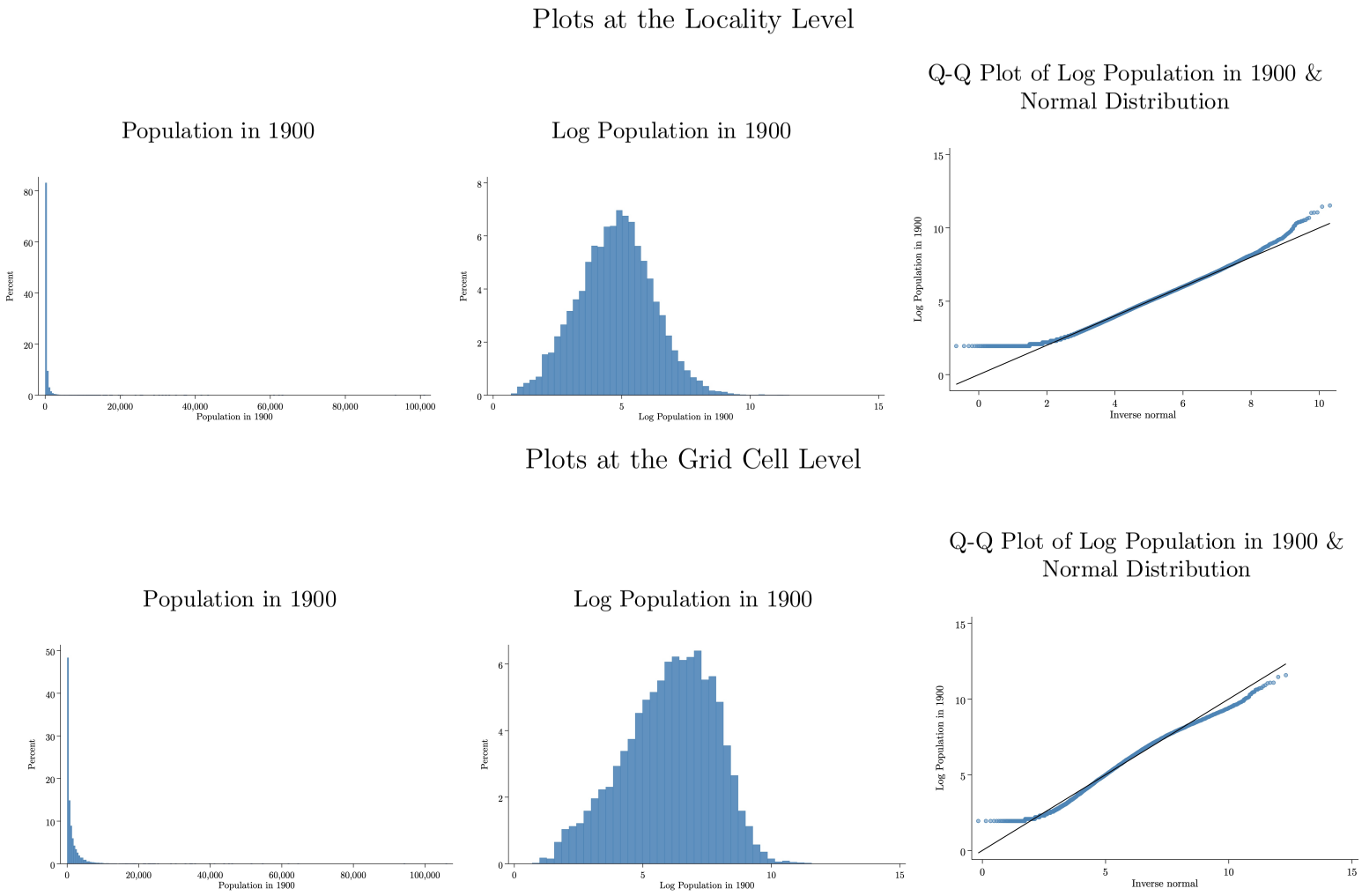


Panel B. Grid Cell Level



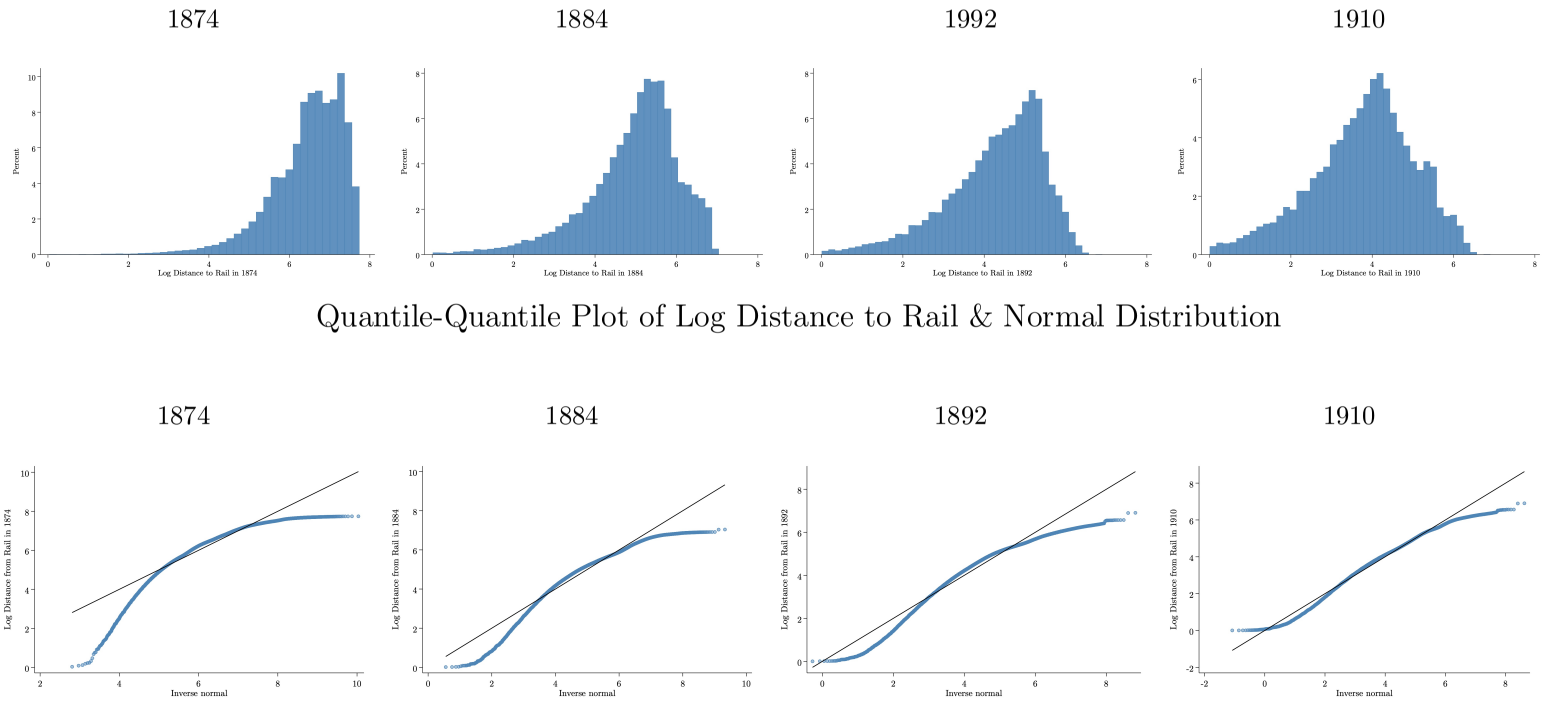
Distributions are produced using either the sample of all localities (Panel A) or the sample of grid cells spanning Mexico (Panel B), the Epanechnikov or parabolic kernel, and the half-width that minimizes the mean integrated squared error. Grid cells are 0.1×0.1 degrees (i.e., approximately 11×11 km at the equation) in size.

Figure A.5: Distribution and Transformation of Population Data



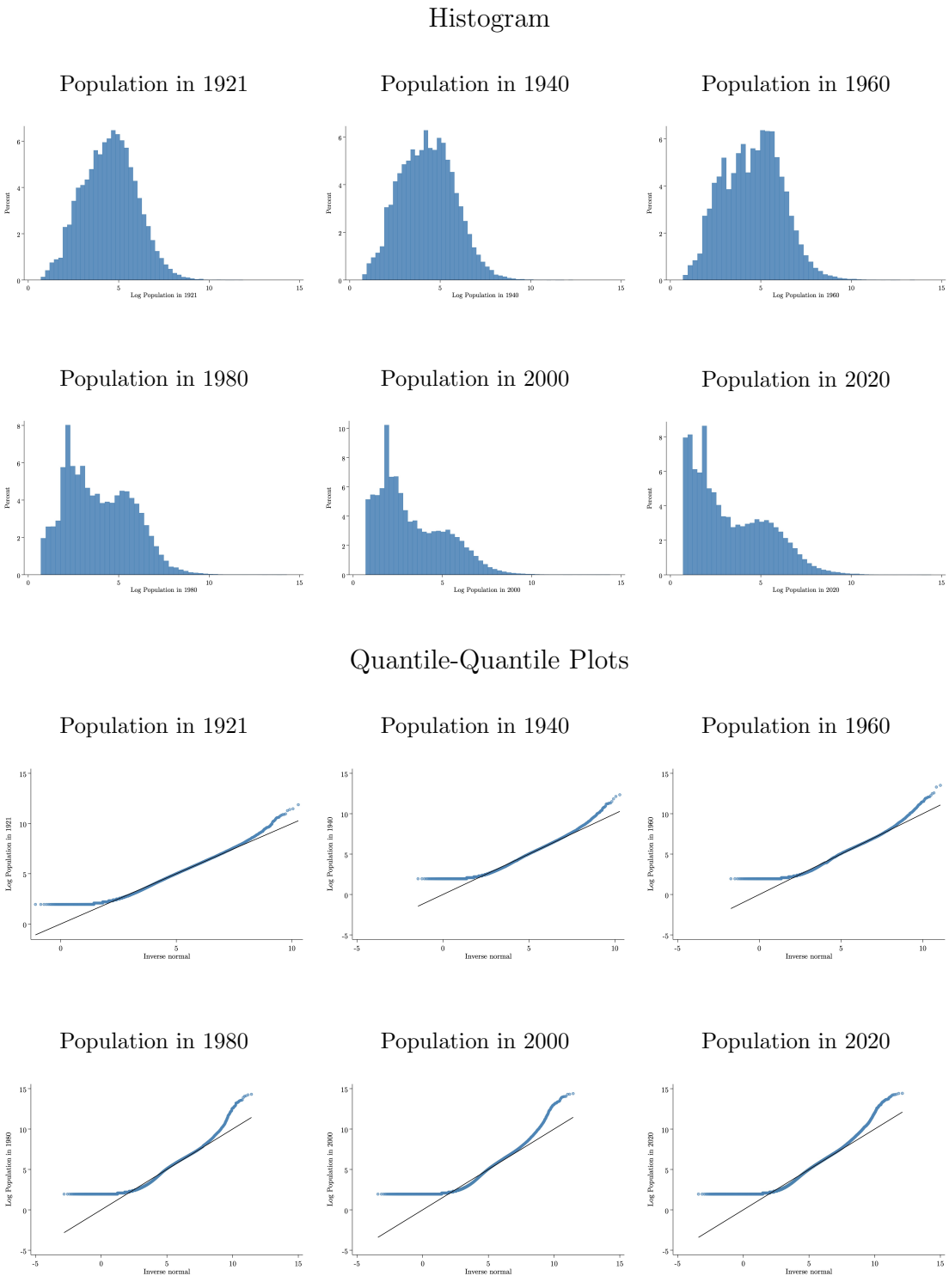
Using the inhabited sample of localities in each year or the full sample of grid cells. Grid cells are 0.1×0.1 degrees (i.e., approximately 11×11 km at the equation) in size.

Figure A.6: Distribution of Log Distance to Rail, 1874 - 1910



Using the inhabited sample of localities in each year.

Figure A.7: Distribution and Transformation of Locality Level Population Data



Plots produced using the inhabited sample of localities in each year.

Figure A.8: Least Cost Paths Across Chihuahua, Mexico

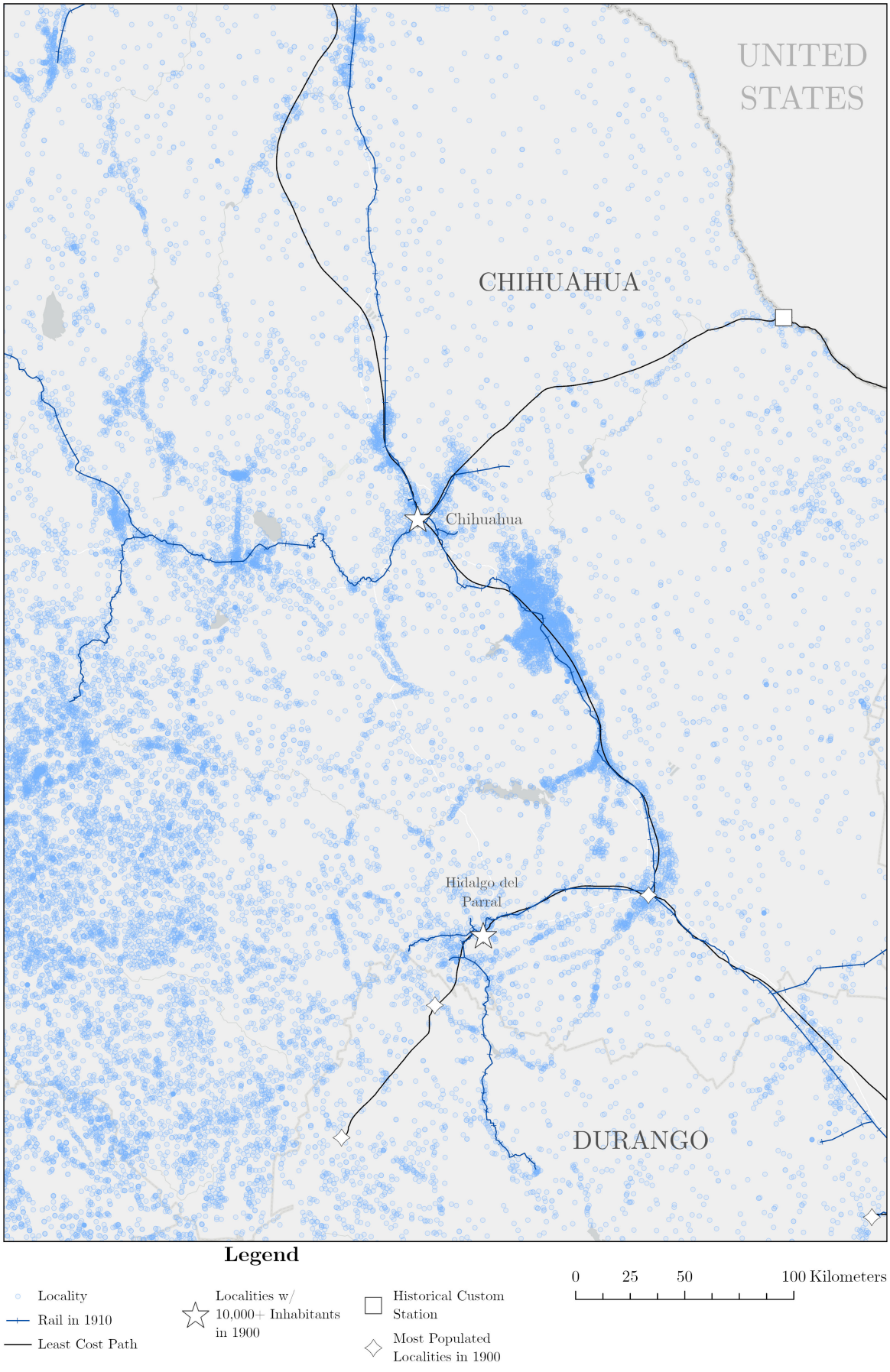


Figure A.9: Least Cost Path Networks

Panel A. Straight Line Least Cost Paths



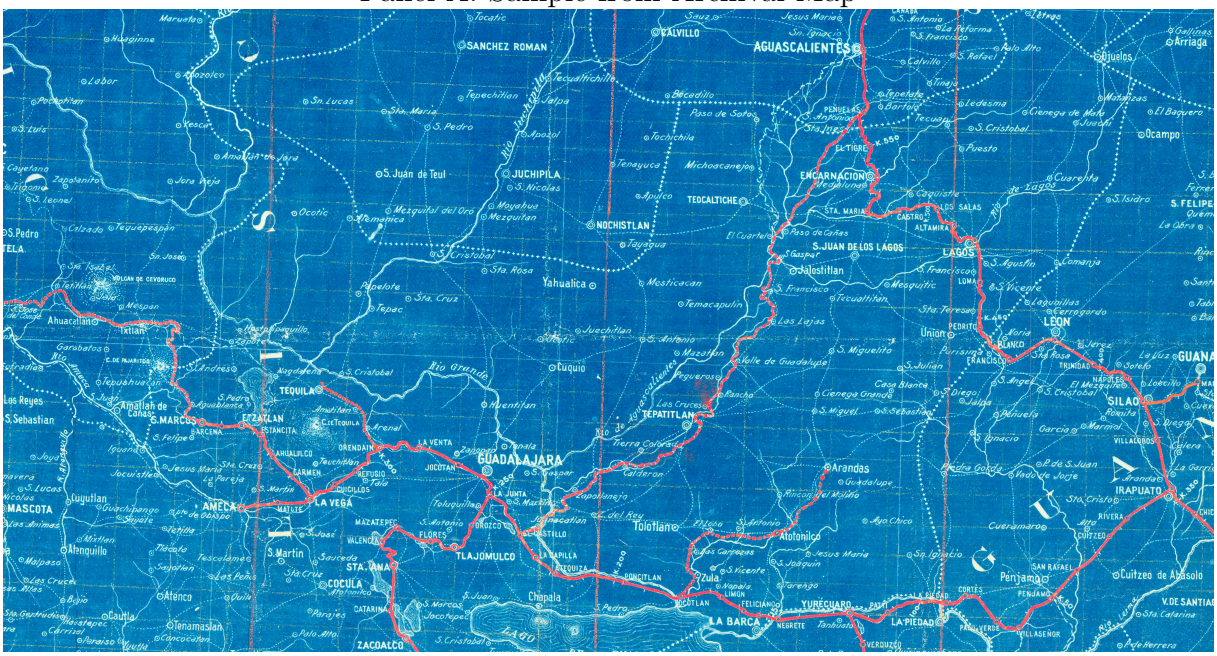
Panel B. Terrain Based Least Cost Paths



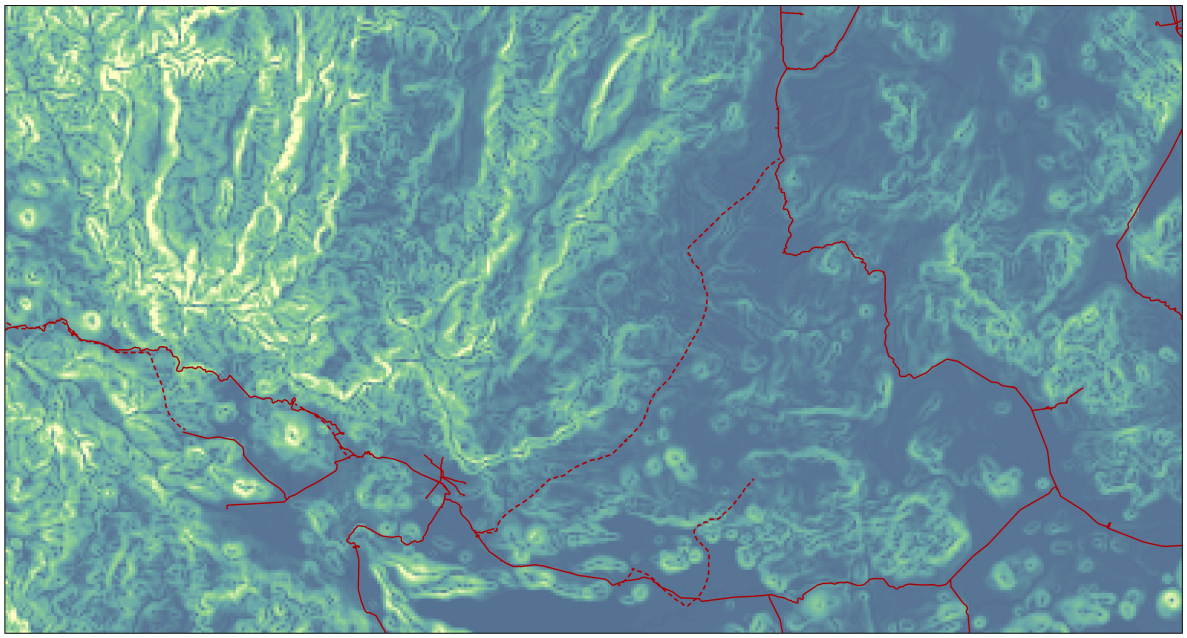
Nodes used to construct these least cost paths include the most populated cities in 1900 and historical customs stations from 1892. The least costs paths in Panel B. consider elevation, slope and geographic barriers, such as major lakes and oceans, to minimize effective distance through terrain. The resulting Kleibergen-Paap rk Wald F Statistics can be found in Appendix Tables A.19 and A.21.

Figure A.10: Placebo Line Construction

Panel A. Sample from Archival Map



Panel B. Cost Raster Used to Construct Placebo Lines



Legend

Placebo Line — Rail in 1910

0 25 50 100 Kilometers

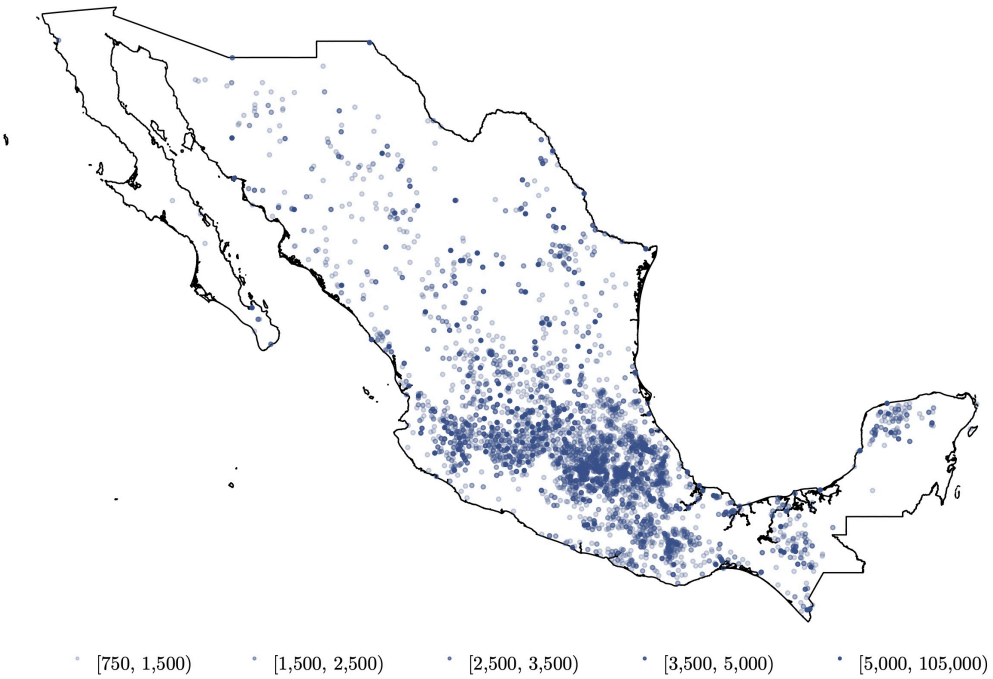
Insert text here.

Figure A.11: Visual Comparison of the Pre 1900 Population Proxy to the 1900 Census

Panel A. Density of Indigenous Communities from Tanck de Estrada et al. (2005) Used as a Pre 1900 Population Proxy



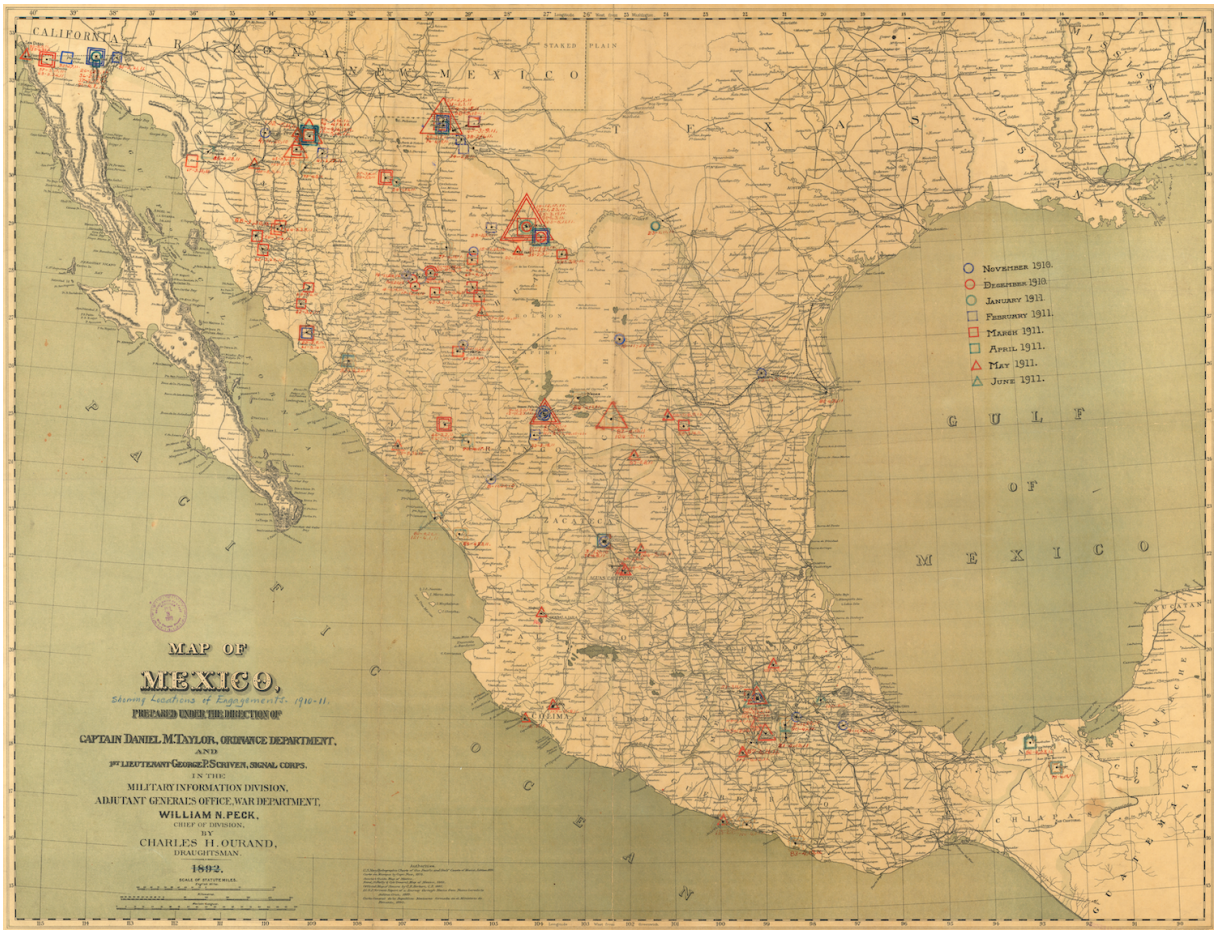
Panel B. Density of Localities with 750 or More Individuals in 1900



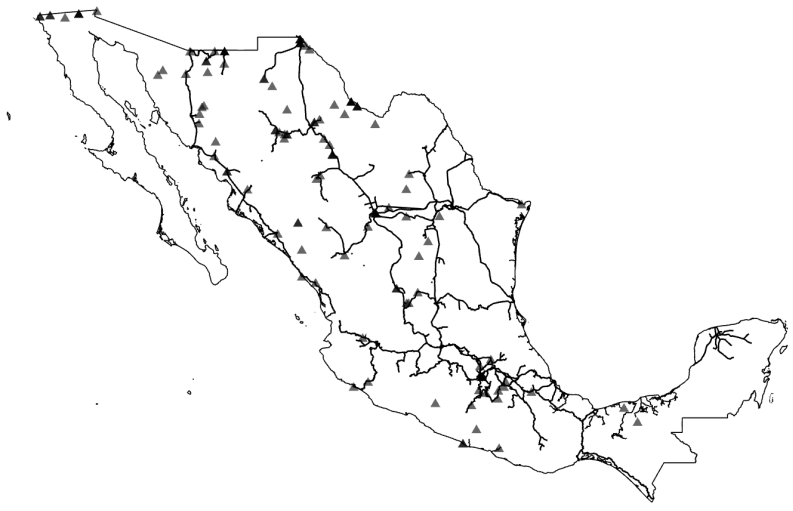
In 1900, approximately ten percent of all inhabited localities had a population of at least 750 inhabitants.

Figure A.12: Insurgent Outbreaks During the First Year of the Mexican Revolution, November 1910 to June 1911

Panel A. Archival Map of Insurgent Outbreaks



Panel B. Insurgent Outbreaks Relative to Rail in 1910



Archival map from Hartman, Hartman.

A.3 Control Variables

A.3.1 Population Data Before 1900

As previously noted, construction of the rail network initially began in the 1870s yet the first published national census is that of 1895 and the first national census with locality-level data is that of 1900. Data from the 1895 census are older, but still 20 years after initial rail expansion and are only available at aggregate geographies such as the district, municipal, and "partido" level.

In lieu of "initial" population before rail expansion, I use the number of pre-colonial and colonial indigenous communities either within a grid cell or 10km of each locality as a proxy. These communities were extensively documented and georeferenced by Tanck de Estrada et al. (2005) using archival records. A limitation of this proxy is that the exact period these documented communities existed during is largely unknown. Their existence was often documented by the Spanish during the colonial era, but their history prior documentation is largely unspecified.[1] Following Artega (2018), I use the number of indigenous communities either within a grid cell or certain distance from each locality. Figure A.11 in the Figure Appendix maps the density of pre-colonial indigenous communities and locality population in 1900. Although locality population in 1900 was more dispersed, particularly within the northwest region of Mexico, spatial patterns between the two data sources exist. By 1900, most localities in the top 10th percentile were clustered in Central Mexico, the center-west region along the Pacific coast, the Yucatán Peninsula, and northern Chiapas - all locations with a large documented number of indigenous communities.

[1]When possible, Tanck de Estrada et al. (2005) provide dates associated with each community using archival sources, but this is relatively uncommon.

A.3.2 Cultural and Economic Centers

I account for points along the Camino Real (The United Nations Educational, Scientific and Cultural Organization, 2010) and 19th century customs stations (INEGI, 1897), both of which served as coordinating points for domestic and international trade. The Camino Real was a colonial trading route with indigenous origins that facilitated long-distance inland trade, mineral exploration and religious expansion (Rees, 1975; Dep, 1997). The route primarily ran north to south within the Central Mexican Plateau, between the Sierra Mountains and into the United States. Its existence during the Colonial period, and before the United States Civil War in the 1860s, centralized trade along and near the route. Yet, to my knowledge, it remains unknown whether communities along the route developed and persisted more so than those farther away. However, conditional path dependence has also been observed along pre-colonial trading routes in Brazil (Barsanetti, 2021). It is therefore possible that communities along the Camino Real may have benefited from initial advantages provided by the route either independently of or in combination with rail access in the 20th century.

Customs stations, especially by ports, facilitated connections with international markets. As shown in Alvarez-Villa and Guardado (2020), colonial ports played a significant role in coordinating economic activity, and areas surrounding ports "took off" during the trade liberalization era in the 19th century. I thus use distance to historical locations that once served, and often continue to serve, as coordinating points for trade.

A.3.3 Resource and Environmental Data

I construct two sets of variables capturing the prevalence of mines. The first uses a historical encyclopedia of minerals in Mexico by Panczer (1987). The encyclopedia is an extensive catalogue of minerals historically mined in Mexico and at the time was considered the most complete synthesis of Mexican geology. I digitize the catalogue of

gold and silver mines, and in the process capture their location and relative richness. The specificity of a mine's location varies by entry but is never more aggregate than at the municipal level.[2] The richness of each mine is categorized by the author as either of limited, moderate or wide distribution. The second set of variables is sourced from INEGI (2002). The data are spatial and contain the point locations of copper, iron, zinc and coal mines. No consistent record of the richness and age of the mines is included, but these data do reflect the location of mineral deposits that are known to exist.

I use the Galor and Özak average caloric suitability index, which measures calories per hectare per year, to account for natural agricultural advantages (2016b). The average suitability of maize is assumed under the condition of low inputs and rainwater, which capture the subsistence quality of land before industrialization (Galor and Özak, 2016a). Meanwhile, the suitability of cash crops is assumed under the condition of high inputs and irrigation to reflect production potential.

Lastly, I use historical geospatial data on precipitation and average air temperature provided by the Climate Research Unit at the University of East Anglia (2024; Harris et al., 2020). Climate is considered to account for natural agglomeration near amenable climates and migration-driven changes over time (Hoffmann et al., 2024). Historically, climate, particularly drought, has played a significant role in agrarian crises throughout the 20th century, which have, in turn, affected population (Endfield and Tejedo, 2006; Dell, 2012; Wolfe, 2019). Historical data span from 1901 onward and are annualized. Data for 1901 is used to approximate conditions in 1900.

A.3.4 Revolutionary Insurgency

During the Revolution, significant migration, particularly to the U.S., was driven by insurgent violence (McCaa, 2003). Escamilla-Guerrero et al. (2023) find that this mi-

[2]Whenever possible, the locations of mines in Panczer (1987) were cross-referenced with the online mining database `mindat.org`.

gration was more common in areas with existing migration patterns, historical routes, and fewer haciendas, especially in central Mexico, despite the relatively high cost compared to regions near the U.S. border.

To account for the impact of revolutionary violence on population, two sources are used to geolocate insurgent activity. The first source is qualitative data from Dell's (2012) data appendix, which lists municipalities that experienced any insurgent activity throughout the war. Information on outbreaks was sourced from entries in the *Encyclopedia of Mexican Municipalities*. The second source is from NARA's United States War Department Map Collection. This source contains both a map and table documenting the location and outcome of 125 insurgent outbreaks during the first year of the Revolution from November 1910 to June 1911. Outbreaks and their relative position near the rail network can be seen in Figure A.12 of the Figure Appendix.

A.4 Data Construction Appendix

A.4.1 Constructing the Rail Shapefiles

Below is a list of maps used to determine the existence and extent of rail in Mexico before and during the *Porfiriato*. I used modern shapefiles as a base for creating the historical rail network, drawing from sources such as MEXLIST, USGS, INEGI, OpenStreetMap, and Esri. In cases where specific branches or connections were clearly and consistently labeled in historical maps or archival sources but did not appear in any modern shapefiles, I manually reconstructed them (i.e., hand drew the lines) as approximations. To do so, I overlaid the scanned historical maps onto a modern map of Mexico in ArcGIS. I then georeferenced each map by identifying latitude and longitude control points, adjusting the size and alignment of the historical image so that key features matched those in the modern base map.

Using the adjusted map image, I traced the likely location of historical rail lines using

visible infrastructure and labeled roads, many of which included references suggestive of prior rail presence. When such infrastructure was not available, I used modern roads and highways as rough proxies. In some cases, I chose not to hand draw a historically documented line if there was no supporting infrastructure, no consistent documentation, no historical discussion of the connection, and no geographically reasonable route linking the labeled points.

The maps listed below vary in geographic coverage, and only some depict the entire network at a given point in time. Many maps show only a portion of the system, rail lines belonging to a particular company, or rail within a specific state or region.[3] The sources are ordered below by date.

Table A.27: Summary of Historical Rail Maps Used

Citation	**Notes**
Colton and Fisher, 1874	Depicts rail lines across the country, with detailed coverage of the Tehuantepec region in southern Mexico. Used to refine the connection between Orizaba and Esperanza as shown in Salgado (1888).
Colton and Gray, 1878	Shows national rail alongside portions of the West Indies and Central America. Illustrates a connection from Jalapa to Tlaxcala via Perote. Later maps (McNally and Company, 1882; Blackie, 1882; Cubas, 1886; Salgado, 1888) indicate this line was under construction.
McNally and Company, 1882	Contains detailed station labels, though completed and projected lines are hard to distinguish. Used Blackie, 1882 for corroboration. Hand drew: • A small extension from the Silao branch to Guanajuato via San Josê de la Luz.

[3] All mistakes are my own. Any corrections, suggestions, or documentation detailing rail expansion, development, and use during this period are welcome.

Citation	**Notes**
Hartman, 1884	A detailed national rail map sourced from the National Archives and Records Administration (NARA), drawn by U.S. Army General Staff member H. Hartman.
Cubas, 1886	Atlas by Antonio García Cubas with both a national map and detailed state-level maps. The primary map of the entire country illustrates existing lines and those under construction. However, state-specific maps show in greater detail which parts of the network were under construction. Sometimes, country- and state-specific maps were not consistent: • Whether segments of the Yucatán line were operational or under construction. • Whether connections from Ticul to Peto and Campeche to Mérida were in use. Referenced additional maps before and after this period to confirm which branches were under construction.
Salgado, 1888; Sánchez, 1889	Local maps from CEDIF documenting *Interoceánico* lines, including under-construction segments. Used maps to refine sections of the 1886 network that I believe should have been noted as under construction but were drawn as completed, including connections between Puebla and San Martín Texmelucan, and Puebla and Izúcar.
Cubas, 1892	Another national map by García Cubas. Though less detailed, it labels main and terminal stations. Shows a completed Campeche–Mérida line, though earlier (Salgado, 1888) and later (Johnston, 1893; 1894; 1899; Cubas, 1899) maps depict it as incomplete. Treated accordingly.
Johnston, 1893	Atlas by Alexander Keith with detailed rail and station data. Used to confirm locations and expansions across years.

Citation	Notes
1894	First book of a two-volume series from CEDIF. Each volume includes a discussion of rail line extensions and multiple local maps, which were used to refine the digitized rail network for 1892 and 1896. For example, a region-specific map of Yucatán confirmed that the connection between Mérida and Campeche had not yet been completed by 1894. Based on the station labels shown in this 1894 map, I hand drew the corresponding segments as they would have appeared in 1886: • Salamanca to Jaral de Progreso via Valle de Santiago, Guanajuato. • Pachuca to Tizayuca, Hidalgo, verified with maps from 1878 (Caso, 1894) and 1903 (1903a). • Mérida–Conkal–Tixkokob branch in Yucatán (missing from 1896). These connections were only hand-drawn if they could be consistently verified in maps from subsequent years. Some local maps in this volume depict connections and branches that could not be corroborated using later maps or modern shapefiles. These were therefore excluded from the digitized network. Their omission does not imply that the lines never existed or were never under construction, but rather that I could not find consistent documentation confirming their completion. One additional map (1894) depicts a connection between San Blas–Huaristemba link via Zapotillo, Nayarit, but this is not confirmed in other years. ASimilarly, a 1904 map shows a nearby San Blas–Navarrete connection, also uncorroborated. These lines were not included due to insufficient verification. Given the uncertainty and lack of consistent evidence for this small branch line—if it existed at all—I chose not to include it in the digitized network.
von Glumer, 1896	A national map produced by Bodo von Glumer from the Lithographic Institute of Berlin under the director of Mexico's Secretary of State and the Office of Communications and Public Works. This is a detailed map of the entire country illustrating rail lines and postal routes by carriage, horse, foot, and canoe. Missing a previously noted Tizayuca–Zumpango link (north of Mexico City).

Citation	Notes
1899	Second book part of a two-volume series from the CEDIF (first book was published in 1894). Local maps were used to confirm 1900 rail network details.
Cubas, 1899	Second García Cubas atlas, again containing both national and state-specific maps. Used to verify 1900 rail extensions.

Citation	Notes
1901	Source is a book with large, detailed state-level maps from CEDIF. Map contain labeled government offices, narrow/wide-gauge rail with labeled stations, and postal routes. Hand drew the following branches and lines: • Chihuahua–Santa Eulalia, Chihuahua. • San Pedro to parallel line from Matamoros. • Parras–Saltillo, Coahuila. • Torres–Minas Prietas, Sonora. • Mexico City–El Guarda. • Ixtapalapa–Chalco, State of Mexico. • Tlalnepantla–San Ildefonso, State of Mexico. • Rincón–San Luis de la Paz, Guanajuato. • Tlacotepec–Tepeji area, Oaxaca (westbound route unclear). • Xalapa–Teocelo, Veracruz. • Conkal–Baca, Yucatán. • Tixkokob–Temax–Quintana Roo extension. • San Andrés–Chalchicomula, Puebla. • Ocotlán–Ejutla, Puebla. • Mérida–Muna via Molas, Temozón, and Yuncu, Yucatán. Some connections and branch lines that were clearly illustrated in these maps but were unidentifiable using satellite imagery and modern shapefiles include: • Minas Nuevas–Ojitos branch near Parral, Chihuahua. • Sabinas–San Felipe segment, Coahuila. • South of Mapimí, Durango. • Line paralleling Monterrey–Reta, Nuevo León. • Matehuala–Cerrito Blanco–La Viuda, San Luis Potosí. • Zumpango–San Cristóbal, Mexico. • Above Zempoala–Tulancingo segment, Mexico. • Chalchicomula–Tlachichuca extension, Puebla (unconfirmed by 1904; 1910; 1914; 1914).
190, 1903a	Detailed map from NARA of the Ferrocarril Central Mexicano. Covers Lechería–Pachuca–Tampico. Used to verify 1890s expansion.

Citation	Notes
1904	NARA map by the 2nd Division General Staff of the Military Information Division. Includes rail, trails, and place names. Station names varied, so locations were verified using earlier (1899; Cubas, 1899; 1901) and later (1914) maps. Hand drew: • San Quintín northward branch, Baja California. • Hidalgo de Parral–Minas Nuevas (Villa Escobedo). • Extension toward Durango via Biquillo. • Small branch east of Sabinas, Coahuila.
Alvarez, 1909	Detailed local map of Mexico City's outskirts. Used to refine southern and southwestern extensions.
1910	newline Map of the entire country from NARA which I believe was produced by the *Ferrocarriles Nacionales de México*. A very similar archival map from the company was also found at the CEDIF but dated 1914 with corrections. Referenced both when identifying expansions and hand drew: • El Oro–Yondese extension, Mexico. • Corral–Picachos branch, Sonora. • Conejos–Jaralito, Durango. Map shows an additional stop in Descubridora but was unable to identify where this connection could have been built. • Jarita–Colombia near the U.S. border, Nuevo León. Was unable to locate or identify the stations branching east from Linares to San José via Río de Conchos were unidentifiable. This map also showed the Chihuahua–Sinaloa connection as completed, though later maps (1914; 1914) still listed it as under construction.

Citation	Notes
1914	Source is a detailed map of the entire country from the National Oceanic and Atmospheric Administration's Historical Map and Chart Collection. Hand drew: • Tuxpan–Furbero, Veracruz. • Tres Valles–Cosamaloapan, Veracruz. • El Burro–San Andrés Tuxtla, Veracruz. • Apizaco–Tlaxo, Puebla. • Munoz–Chignahuapan, Puebla. • Corral–Onavas extension, Sonora. Could not identify the Conduacán–Villahermosa (San Juan Bautista), Tabasco line.

A.4.2 Rail Shapefiles by Year

Figure A.1: Rail in 1874



Figure A.2: Rail in 1878



Figure A.3: Rail in 1882



Figure A.4: Rail in 1884



Figure A.5: Rail in 1886



Figure A.6: Rail in 1892



Figure A.7: Rail in 1896



Figure A.8: Rail in 1900



Figure A.9: Rail in 1904



Figure A.10: Rail 1910



A.4.3 Placebo Lines

This section documents the sources used to create placebo lines. These lines were constructed as least-cost paths based on preliminary or planned rail routes that were ultimately never built. The projected lines were referenced in books, articles, and archival maps, often as reconnaissance or speculative plans. Many were never realized due to insufficient funding or disruptions caused by the Mexican Revolution.

All placebo lines were constructed in ArcGIS using elevation, slope, and physical barriers (e.g., lakes, oceans) to estimate the least-cost path across terrain. For more detail, see Section A.4.4. In several cases, placebo paths run close to or parallel with rail lines that were actually completed by 1910. The sources for these lines are grouped below.

From Secondary Sources

- A proposed line from Mexico City to Acapulco that "never crossed the coastal Sierra" and only reached Puente de Ixtla (Goldfrank, 1976).
- A planned route from the Gulf port of Antón Lizardo to the Pacific port of Puerto Ángel (Chassen-López, 2005).
- A projected line from Tehuantepec to Veracruz that went bankrupt (Goldfrank, 1976).
- A proposed line from Pachuca to Tampico via Actopan, Ixmiquilpan, El Higo, and Pánuco, which only reached Ixmiquilpan (Long, 1925).
- A Yucatán-bound line branching from the Guadalajara–Manzanillo line that was abandoned during the Revolution (Goldfrank, 1976).

Maps

Citation	Notes
1884	• A line beginning in northeastern Mexico and running along the Gulf through Tamaulipas: Matamoros, San Fernando, Santander, Aldama, and Coestocilas. • A route beginning in Mier, Nuevo León and continuing south through Villagrán, Hidalgo, Victoria, Huejutla, Pachuca, Mexico City, Morelos, Tixtla, and Acapulco. • A projected line from El Salado, San Luis Potosí, south through Cedral, Venado, Hiatono, San Luis Potosí (capital), León, and San Miguel. The segment between San Luis Potosí and León was never completed. • A line beginning in Alata, Sinaloa and running through Fresnillo, Zacatecas via Durango. Segments from Culiacán to Tamazula and from Durango to Fresnillo were uncompleted. • A route from San Blas to Guadalajara and San Juan de los Lagos. • A projected line from Manzanillo through Colima to Morelia, Michoacán. The Colima–Morelia segment was not completed. • A branch extending from Morelia southwest to Zihuatanejo, Guerrero. Only partial construction was completed; the segment east of Lázaro Cárdenas to Zihuatanejo remained unfinished. • A line from Orizaba to the Gulf of Tehuantepec via Oaxaca City. The segment from Ejutla to the Gulf was not built.

Citation	Notes
1901[4]	• Tepic to Compostela (Xalisco), San José del Conde, Tetitlán, Ahuacatlán, Ixtlán, and San Marcos. • From Colima northeast to Carpintero (west of Pihuamo), Calabozo, Colomos, and Tuxpan. • From Orendain to Tequila via Arenal, Jalisco. • From El Castillo through Juanacatlán, Zapotlanejo, Calderón, Tepatitlán, Valle de Guadalupe, San Gaspar de los Reyes, El Cuartel, and Peñuelas. • From Ocotlán (near Lake Chapala) to Zula, Atotonilco el Alto, and Arandas. • From Encarnación, Hidalgo, south past Zimapán, Tasquillo (opposite bank), to Ixmiquilpan, connecting to existing rail near Tetepango and Rosal. • From Tampico south through Chontla, passing east of Huautla (Veracruz), ending near Apulco, Hidalgo.

A.4.4 Constructing a Least Cost Path Network

To construct the least-cost path (LCP) instrument, I first built a geographic cost surface in ArcGIS that reflects variation in terrain. This cost raster combines elevation and slope to proxy for relative difficulty in constructing a rail line across Mexico. The following steps outline how the cost raster was created:

1. **Elevation base layer:** I began with a clipped elevation raster covering all of Mexico (U.S. Defense Mapping Agency, 2007).

2. **Masking water bodies:** Small inland bodies of water (CIAT and the Trustees of Columbia University, 2017) were excluded by applying a mask, generating a new raster that only includes land areas.

[4]Greater detail for the area around Tampico, Tamaulipas was provided in 190, 1903b, which includes reconnaissance routes from 1891 and 1892.

3. **Slope-based cost calculation:** Using the cleaned or clipped elevation raster, I computed slope in terms of percent rise over a 1,000-meter neighborhood. This was done using a quadratic surface type to better capture details in terrain. The resulting raster serves as the cost raster, assigning higher traversal costs to steeper areas.

4. **Distance accumulation:** This slope-based cost raster was then used as input to the `DistanceAccumulation` tool in ArcGIS. I computed the least-cost distance connecting a set of nodes (i.e., cities with at least 10,000 inhabitants in 1900, historical customs stations, and municipal centers). The calculation accounts for geodesic distance (Earth's curvature) and imposes a barrier to prevent paths from crossing country boundaries (Geospatial Management Office of the U.S. DHS, 2017; Flanders Marine Institute, 2017; 2014).

REFERENCES

Archival Sources

1894. Reseña Histórica y Estadística de los Ferrocarriles de Jurisdicción Federal: Desde Agosto de 1837 hasta Diciembre de 1894. Centro de Documentación e Investigación Ferroviarias (CEDIF), Secretaría de Comunicaciones y Obras Públicas, Puebla, MéxicoVol. 1 of 2.

1899. Reseña Histórica y Estadística de los Ferrocarriles de Jurisdicción Federal: Desde 1. de Enero de 1895 hasta 31. de Diciembre de 1899. Centro de Documentación e Investigación Ferroviarias (CEDIF), Secretaría de Comunicaciones y Obras Públicas, Puebla, MéxicoVol. 2 of 2.

1901. Cartas Postales de los Estados Unidos Mexicanos. Centro de Documentación e Investigación Ferroviarias (CEDIF), George F. Cram Co., Dirección General de Caminos, Sección de Transportes, Puebla, México

1901. Líneas de la Cía. Limitada del Ferrocarril Central Mexicano entre el Atlántico y el Pacífico. National Archives and Records Administration (NARA), College Park, College Park, MarylandNAID 207454648, local id 177-Mexico.

1903. Cía. Limitada del F.C. Central Mexicano, División del Pánuco: Línea de Lechería a Pachuca y Tampico. National Archives and Records Administration (NARA), College Park, College Park, MarylandPrepared under Lewis Kingman, chief engineer; NAID 207454656, local id 205-Mexico.

1903. Map of Mexican Central Railway, Tampico Short Line. National Archives and Records Administration (NARA), College Park, College Park, MarylandIncludes reconnaissance of 1891–1892; NAID 207454658, local id 209-Mexico.

1904. Map of Mexico. National Archives and Records Administration (NARA), College Park, War Department, Office of the Chief of Engineers, College Park, Maryland2nd Division General Staff, Military Information; NAID 207454619, local id 61-3-Mexico.

1910. Railways of Mexico. National Archives and Records Administration (NARA), College Park, Ferrocarriles Nacionales de México, College Park, Maryland

1914. Mapa de los Ferrocarriles Nacionales de Mexico. Centro de Documentación e Investigación Ferroviarias (CEDIF), Ferrocarriles Nacionales de México, Puebla, MéxicoIncludes lines of Ferrocarril Interoceánico de México, Ferrocarril Mexicano del Sur, and Ferrocarril Panamericano.

1914. Railroad Map of Mexico. Historical Map and Chart Collection, Office of Coast Survey, NOAA, War College Division, General Staff, War Department, Washington, DC Available at: https://www.historicalcharts.noaa.gov/image.php?filename=14827-00-1914.

Alvarez, Cristóbal T.. 1909. Carta de la República Mexicana a escala 1:100 000. Centro de Documentación e Investigación Ferroviarias (CEDIF), Secretaría de Fomento, Puebla, MéxicoCEDIF record no. 5100; sheet covering 30 × 30 around Mexico City.

Blackie, W. G.. 1882. The Comprehensive Atlas & Geography of the World: Mexico. David Rumsey Historical Map Collection, Blackie & Son, LondonPublication list no. 1589.000. Available at: https://www.davidrumsey.com/rumsey/download.pl?image=/162/1589077.jp2.

Caso, Antonio. 1894. Ferrocarriles de Hidalgo y del Nordeste: Plan General de la Vía entre México y Tuxpan con los Ramales a Zumpango, Irolo y Pachuca. National Archives and Records Administration (NARA), College Park, War Department, Office of the Chief of Engineers, College Park, MarylandWar Department Map Collection, NAID 207454623, local id 91-Mexico.

Colton, G. W. and R. Swainson Fisher. 1874. Colton's General Atlas. David Rumsey Historical Map Collection, G. W. & C. B. Colton Co., New YorkMap collection list no. 4587.068. Available at: https://www.davidrumsey.com/rumsey/download.pl?image=/D5005/4587068.sid.

Colton, G. Woolworth and Ormando Willis Gray. 1878. The National Atlas. David Rumsey Historical Map Collection, O. W. Gray & Son, PhiladelphiaMap collection list no. 3888.052. Available at: https://www.davidrumsey.com/rumsey/download.pl?image=/D5005/3888052.sid.

Cubas, Antonio García. 1886. Atlas Geográfico y Estadístico de los Estados Unidos Mexicanos. David Rumsey Historical Map Collection, Debray Sucesores, MéxicoPublication list no. 5758.000. Available at: https://www.davidrumsey.com/luna/servlet/detail/RUMSEY 8 1 33838 1171554.

Cubas, Antonio García. 1892. Ferrocarriles en la República Mexicana. David Rumsey Historical Map Collection, Oficina Impresora de Estampillas, MéxicoPublication list no. 9859.004. Available at: https://www.davidrumsey.com/rumsey/download.pl?image=/205/9859004.jp2.

Cubas, Antonio García. 1899. Atlas Metódico para la Enseñanza de la Geografía de la República Mexicana. David Rumsey Historical Map Collection, Eduardo Murguía, MéxicoPublication list no. 2850.000. Available at: https://www.davidrumsey.com/rumsey/download.pl?image=/D0121/3287056.sid.

Hartman, H. Map of Mexico Showing Locations of Engagements, 1910 - 1911. National Archives and Records Administration College Park (NARA), War Department, Office of the Chief of Engineers, College Park, MarylandPart of the War Department Map Collection. National Archives Identifier: 207454671. Local Identifier: 52-2 Mexico. The archival record is a set of two documents: a map and a table of insurgent events containing the date of each event, the number of deaths, commanding officers, victors and additional remarks.

Hartman, H. 1884. Railroad Map of the Republic of Mexico. National Archives and Records Administration (NARA), College Park, War Department, Office of the Chief of Engineers, College Park, MarylandWar Department Map Collection, NAID 207454620, local id 76-Mexico.

Johnston, Alexander Keith. 1893. Mexico in Keith Johnston's General Atlas. David Rumsey Historical Map Collection, W. & A. K. Johnston, EdinburghCollection list no. 3287.056. Available at: https://www.davidrumsey.com/rumsey/download.pl?image=/D0121/3287056.sid.

McNally, Rand and Company. 1882. Rand, McNally & Co.'s General Map of the Republic of Mexico. David Rumsey Historical Map Collection, Rand McNally & Co., ChicagoPublication list no. 3347.000. Available at: http://www.davidrumsey.com/maps5296.html.

Salgado, Cosme M. 1888. Ferrocarril Interoceánico: Carta General de la Línea México–Veracruz. Centro de Documentación e Investigación Ferroviarias (CEDIF), Puebla, MéxicoCEDIF map no. 3288.

Sánchez, Delfín. 1889. Carta General de las Líneas del Ferrocarril Interoceánico y Conexiones. Centro de Documentación e Investigación Ferroviarias (CEDIF), Puebla, MéxicoDrawn by Cosme M. Salgado under the supervision of Guillermo Hay; CEDIF map no. 3289.

von Glumer, Bodo. 1896. Carta de Comunicaciones de los Estados Unidos Mexicanos. David Rumsey Historical Map Collection, Lithographic Institute of Berlin; Secretaría de Estado y del Despacho de Comunicaciones y Obras Públicas, MéxicoPublication list no. 8388.000.

Data Sources

1897. Anuario Estadístico de la República Mexicana. Instituto Nacional de Estadística, Geografía e Informática Available at: https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825140250.

1900. Censo General de la República Mexicana. Instituto Nacional de Estadística, Geografía e InformáticaGeneral census publications are available at the state level. Census publications for the year 1900 were published years after by each state individually. Available at: https://en.www.inegi.org.mx/programas/ccpv/1900/.

2002. Conjunto de datos vectoriales Geológicos. Continuo Nacional. Minas otras ubicaciones geológicas: Escala 1:1,000,000. Instituto Nacional de Estadística, Geografía e Informática Available at: https://en.www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825267612.

2007. North America Elevation 1-Kilometer Resolution GRID. U.S. Defense Mapping Agency (now National Geospatial-Intelligence Agency) Available at: https://www.sciencebase.gov/catalog/item/4fb5495ee4b04cb937751d6d.

2014. Marine Gazetteer Placedetails: Polygon shapefile of the Pacific Ocean. Marine Regions, Flanders Marine InstituteBased on IHO 23-3rd: Limits of Oceans and Seas, Special Publication 23, 3rd Edition 1953, published by the International Hydrographic Organization. Available at: http://marineregions.org/mrgid/1903.

2017. Homeland Infrasturcutre Foundation-Level Data (HIFLD): Mexico and US Border. Geospatial Management Office of the United States Department of Homeland SecurityLine shapefile of border. Data hosted in conjunction with ESRI. Available at: https://hifld-geoplatform.opendata.arcgis.com/datasets/geoplatform::mexico-and-us-border/about.

2017. Marine Gazetteer Place Details: Polygon shapefile of the Gulf of Mexico. Marine Regions, Flanders Marine InstituteBased on IHO 23-3rd: Limits of Oceans and Seas, Special Publication 23, 3rd Edition 1953, published by the International Hydrographic Organization. Available at: http://marineregions.org/mrgid/4288.

2017. Mexican Rivers, Streams, Aqueducts and Irrigation Canals. The International Center for Tropical Agriculture (CIAT) and the Trustees of Columbia University, City of New York Available at: https://hub.arcgis.com/datasets/eabd5284fd764f0699998eec2f306928_0/about.

2017. Mexican Water Bodies. The International Center for Tropical Agriculture (CIAT) and the Trustees of Columbia University, City of New York Available at: https://hub.arcgis.com/datasets/eabd5284fd764f0699998eec2f306928$_2$/*about*.

2020. XII Censo General de Población y Vivienda 2020. Instituto Nacional de Estadística, Geografía e InformáticaGeographic identification and population variables. Available at: https://www.inegi.org.mx/app/scitel/Default?ev=9.

2024. CRU TS v. 4.08. Climatic Research Unit at the Univeristy of East Anglia Available at: https://crudata.uea.ac.uk/cru/data/hrg/current.

Alvarez, Estefanía, Constantino de León, Michael Bess, Mariana Ramos, and Fernando Marín. 2021. Extensión de Vías y Volumen de Servicios Ferroviarios 1873–1996. Based on data from *Estadísticas Históricas de México* by INEGI. Original table downloaded from the CD-ROM edition corresponding to volumes published in 2000. Contains information on rail line length (in kilometers), passengers, and freight volumes by year. Available at: http://hdl.handle.net/10089/17748.

Dell, Melissa. 2012. Path dependence in development: Evidence from the Mexican Revolution. See the Data Construction Appendix. Available at: https://dell-research-harvard.github.io/wp/396mxrev.

Galor, Oded and Ömer Özak (2016a, October). "The agricultural origins of time preference." *American Economic Review* 106(10), 3064–3103.

Galor, Oded and Ömer Özak. 2016. Caloric Suitability Index. Available at: https://ozak.github.io/Caloric-Suitability-Index/.

Harris, Ian, Timothy J. Osborn, Phil Jones, and David Lister (2020). "Version 4 of the cru ts monthly high-resolution gridded multivariate climate dataset." *Scientific Data - Nature* 7(109).

Panczner, William D. (1987). New York: Springer Science+Buisness Media.

Ruggles, Steven, Lara Cleveland, Rodrigo Lovaton, Sula Sarkar, Matthew Sobek, Derek Burk, Dan Ehrlich, Quinn Heimann, Jane Lee, and Nate Merrill. 2025. Integrated Public Use Microdata Series, International: Version 7.6 [dataset]. IPUMS, Minneapolis, MNSamples used: Mexico 1960, 1970, 1990, 2000, 2010, and 2020. Available at: https://doi.org/10.18128/D020.V7.6.

Tanck de Estrada, Dorthy, Jorge Luis Miranda Garcia, and Tania Lilia Chavez Soto (2005). *Atlas Ilustrado de los Pueblos de Indios: Nueva Espana, 1800*. Fomento Cultural Banamex.

The United Nations Educational, Scientific and Cultural Organization. 2010. Nomination File for the Camino Real de la Tierra. Sourced from the United Nations Educational, Scientific and Cultural Organization (UNESCO).

Nomination files are produced by state parties and are published by the World Heritage Centre at its website to ensure transparency. Available at: https://whc.unesco.org/en/list/1351/documents/.

Valle-Jones, Diego. Compiled census data from INEGI. Final shapefiles contain the location of amenities including, but not limited to, markets, plazas, schools and medical centers. Data are restricted to registered user. Available at: https://www.diegovalle.net/about/.

References

(1997). "National historic trail feasibility study: El camino real de tierra adentro." Technical document, United States Department of Interior, National Park Service.

(2019). "Manual de carogragía geoestadística: Encuesta nacional agropecuaria." Technical document, Instituto Nacional de Estadística y Geografía (INEGI).

Agredano, Hector (2019). *Rails to Revolution: Railroads, Railroad Workers and the Geographies of the Mexican Revolution*. Ph.D. thesis, The Graduate Center, City University of New York (CUNY).

Aguilar-Retureta, José (2015). "The GDP per capita of the Mexcian regions (1895 - 1930): New estimates." *Revista De Historia Económica / Journal of Iberian and Latin American Economic History* 33(3), 387–423.

Allen, Treb and Dave Donaldson (2022). "Persistence and path dependence: A primer." *Regional Science and Urban Economics* 94.

Alvarez-Villa, Daphne and Jenny Guardado (2020). "The long-run influence of institutions governing trade: Evidence from smuggling ports in colonial mexico." *Journal of Development Economics* 144, 102453.

Américo, Pedro (2022). "The industrialization paths: Railroads and structural transformation in Brazil 1872-1950." Working paper.

Arteaga, Fernando (2018). "The historical legacy of (pre?)colonial indigenous settlements in Mexico." Working Paper Working Paper in Economics No. 18-38, George Mason University.

Astorga, Pablo, Ame R. Berges, and Valpy Fitzgerald (2005). "The standard of living in Latin America during the twentieth century." *The Economic History Review* 58(4), 765–796.

Banerjee, Abhijit, Esther Duflo, and Nancy Qian (2020). "On the road: Access to transportation infrastructure and economic growth in China." *Journal of Development Economics* 145.

Barsanetti, Bruno (2021). "Cities on pre-columbian paths." *Journal of Urban Economics* 122, 103317.

Barsanetti, Bruno (2023a). "Road endpoints and city sizes." Working paper.

Barsanetti, Bruno (2023b, 09). "Road Endpoints and City Sizes." *The Review of Economics and Statistics*, 1–45.

Berger, Thor and Kerstin Enflo (2017). "Locomotives of local growth: The short- and long-term impact of railroads in Sweden." *Journal of Urban Economics* 98.

Bertazzini, Mattia C (2021, 05). "The long-term impact of Italian colonial roads in the Horn of Africa, 1935–2015." *Journal of Economic Geography* 22(1), 181–214.

Blackman, Allen, Emilio Leguízamo, and Laura Villalobos (2024, 5). "Points, cells, or polygons? on the choice of spatial units in forest conservation policy impact evaluation." *Environmental Research Letters* 19(5), 054046.

Bleakley, Hoyt and Jeffrey Lin (2012). "Portage and path dependence." *The Quarterly Journal of Economics* 127(2), 587–644.

Bogart, Dan, Xuesheng You, Eduard J. Alvarez-Palau, Max Satchell, and Leigh Shaw-Taylor (2022). "Railways, divergence, and structural change in 19th century England and Wales." *Journal of Urban Economics* 128.

Bolt, Jutta and Jan Luiten van Zanden. 2020. Maddison style estimates of the evolution of the world economy. A new 2020 update. Published by the Groningen Growth and Development Centre, University of Groningen. Available at: https://www.rug.nl/ggdc/historicaldevelopment/maddison/.

Buffington, Robert M. and William E. French (2010). "The culture of modernity." In William Beezley and Michael Meyer, eds., *The Oxford History of Mexcio* (1 ed.)., Chapter 13. Oxford University Press.

Büchel, Konstantin and Stephan Kyburz (2018, 12). "Fast track to growth? Railway access, population growth and local displacement in 19th century Switzerland." *Journal of Economic Geography* 20(1), 155–195.

Campos-Vazquez, Raymundo, Gerardo Esquivel, and Nora Lustig (2014). "The rise and fall or income inequality in Mexico, 1989-2010." In Giovanni Andrea Cornia, ed., *Falling Inequality in Latin America: Policy Changes and Lessons*, pp. 140–161. Oxford: Oxford University Press.

Campos-Vazquez, Raymundo M., Cristóbal Domínguez Flores, and Graciela Márquez (2017). *Long-Run Human Development in Mexico: 1895–2010*, pp. 89–112. Springer Cham.

Caruana-Galizia, Paul, Tomoko Hashino, and Max-Stephan Schulze (2021). *Underlying Sources of Growth: First and Second Nature Geography*, pp. 339–368. The Cambridge Economic History of the Modern World. Cambridge University Press.

Central Intelligence Agency. 2024. The World Factbook: Mexico - Transportation.

Chassen-López, Francie R. (2005). *From Liberal to Revolutionary Oaxaca: The View from the South, Mexico 1867–1911*. University Park, Pennsylvania: Penn State University Press.

Chiquiar, Daniel (2005). "Why Mexico's regional income convergence broke down." *Journal of Development Economics* 77(1), 257–275.

Coatsworth, John H. (1974). "Railroads, landholding, and agrarian protest in the early Porfiriato." *The Hispanic American Historical Review* 54(1), 48–71.

Coatsworth, John H. (1978). "Obstacles to economic growth in nineteenth-century mexico." *The American Historical Review* 83(1), 80–100.

Coatsworth, John H. (1979). "Indispensable railroads in a backward economy: The case of Mexico." *The Journal of Economic History* 39(4), 939–960.

Coatsworth, John H. (1981). *Growth Against Development: The Economic Impact of Railroads in Porfirian Mexico*. DeKalb: Northern Illinois University Press.

Coatsworth, John H. (1989). "The decline of the Mexican economy: 1800 - 1860." In Reinhard Liehr, ed., *América Latina en la època de Simón Bolívar: La formacion de las economías nacionales y los intereses económicoes Europeos 1800 - 1850*, pp. 27–53. Berlin: Bibliotheca Ibero-Americana.

Conley, T.G. (1999). "Gmm estimation with cross sectional dependence." *Journal of Econometrics* 92(1), 1–45.

Constitución Política de los Estados Unidos Mexicanos. 2024. Artículo 28: Prohibición de Monopolios y Exenciones de Impuestos.

Davis, Donald R. and David E. Weinstein (2002). "Bones, bombs, and break points: The geography of economic activity." *The American Economic Review* 92(5), 1269–1289.

Dell, Melissa (2010). "The persistent effects of Peru's minig mita." *Econometrica* 78(6), 1863–1903.

Dell, Melissa (2012). "Path dependence in development: Evidence from the Mexican Revolution." Working paper.

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

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

Donly, A. W (1920). "The railroad situation in Mexico." *The Journal of International Relations* 11(2), 234–251.

Díaz-Díaz, D. and P. Garnica. 2007. Mexico: Roads and Development. Available at: https://proceedings-paris2007.piarc.org/ressources/files/5/HS037-Diaz-E.pdf.

Endfield, Georgina H. and Isabel Fernández Tejedo (2006). "Decades of drought, years of hunger: Archival investigations of multiple year droughts in late colonial chihuahua." *Climate Change* 75, 391–419.

Escamilla-Guerrero, David, Edward Kosack, and Zachary Ward (2023, 8). "The impact of violence during the Mexican Revolution on migration to the United States." Working Paper Working Paper No. 31531, NBER.

Esteban-Oliver, Guillermo (2023). "On the right track? railways and population dynamics in Spain, 1860–1930." *European review of economic history* 27(4), 606–633.

Faber, Benjamin (2014). "Trade integration, market size, and industrialization: Evidence from china's national trunk highway system." *The Review of Economic Studies* 81(3 (288)), 1046–1070.

Fajgelbaum, Pablo and Stephen J. Redding (2022). "Trade, structural transformation, and development: Evidence from Argentina 1869–1914." *Journal of Political Economy* 130(5), 155–188.

Fenske, James, Namrata Kala, and Jinlin Wei (2023). "Railways and cities in India." *Journal of Development Economics* 161.

Fogel, Robert (1962). "A quantitative approach to the study of railroads in american economic growth: A report of some preliminary findings." *The Journal of Economic History* 22(2), 163–197.

Forero, Andrés, Francisco A. Gallego, Felipe González, and Matias Tapia (2021). "Railroads, specialization, and population growth: Evidence from the first globalization." *Journal of Population Economics* 34(3), 195–215.

Fujita, Masahisa, Paul Krugman, and Anthony J. Venables (1999). *The Spatial Economy: Cities, Regions and International Trade.* Cambridge, Massachusetts: The MIT Press.

Fulwider, Benjamin (2009). *Driving the nation: Road transportation and the postrevolutionary Mexican state, 1925–1960.* Ph.D. thesis, Georgetown University.

Galor, Oded (2005). "Chapter 4 from stagnation to growth: Unified growth theory. Volume 1 of *Handbook of Economic Growth*, pp. 171–293. Elsevier.

García, Laura Cházaro (2022). *Portraits for an Exhibition: The Making of a Statistical Culture for Public Life in Mexico During the Time of the Dirección General de Estadistica, 1882-1922*, Chapter 4, pp. 89–122. The Cambridge Economic History of the Modern World. Palgrave Macmillan Cham.

Garner, Paul (2011). "The extension of empire: The tehuantepec national railway, 1896–1918." In Giovanni Andrea Cornia, ed., *British Lions and Mexican Eagles: Business, Politics, and Empire in the Career of Weetman Pearson in Mexico, 1889–1919*, pp. 140–161. Stanford, California: Stanford University Press.

Gilly, Adolfo (2005). *The Mexican Revolution*. New York: The New Press.

Goldfrank, Walter (1976). "The ambiguity of infrastructure: Railroads in prerevolutionary Mexico." *Studies in Comparative International Development* 11(3), 3–24.

Gonzalez-Navarro, Marco, Roman David Zarate, Remi Jedwab, and Nick Tsivanidis (2023). "Land transport infrastructure." *VoxDevLit* 9(1).

Haber, Stephen (1989). *Industry and Underdevelopment: The Industrialization of Mexico, 1890 - 1940*. Standford, California: Standford University Press.

Haber, Stephen (2011). "The political economy of industrialization." In Victor Bulmer-Thomas, John Coatsworth, and Roberto Cortes-Conde, eds., *The Cambridge Economic History of Latin America*, pp. 537 – 584. Cambridge University Press.

Haber, Stephen, Armando Razo, and Noel Maurer (2004). *The Political Economy of Industrialization*. Cambridge University Press.

Hanlon, W.Walker and Stephan Heblich (2022). "History and urban economics." *Regional Science and Urban Economics* 94, 103751. Urban Economics and History.

Hardy, Osgood (1934). "The Revolution and the railroads of Mexico." *Pacific Historical Review* 3(3), 249–269.

Heath, Hilarie J. (2011). "El malogrado proyecto del ferrocarril peninsular del Distrito Norte de la Baja California, 1887-1892." *Estudios Fronterizos* 12(24), 185–221.

Heckman, James J. (1978). "Dummy endogenous variables in a simultaneous equation system." *Econometrica* 46(4), 931–959.

Herranz-Locán, Alfonso (2006). "Railroad impact in backward economies: Spain, 1850 - 1913." *The Journal of Economic History* 66(4), 853–881.

Herranz-Locán, Alfonso (2013). "Transport technology and economic expansion: The growth contribution of railways in Latin American before 1914." *Revista de Historia Económica* 32(1), 13–45.

Hoffmann, Roman, Guy Abel, Maurizio Malpede, Raya Muttarak, and Marco Percoco (2024). "Drought and aridity influence internal migration worldwide." *Nature Climate Change*.

i Font, Joan Costa and Eduardo Rodriguez-Oreggia (2005). "Is the impact of public investment neutral across the regional income distribution? evidence from Mexico." *Economic Geography* 81(3), 305–322.

International Transport Forum, Organisation for Economic Co-operation and Development (OECD). 2016. Establishing Mexico's Regulatory Agency for Rail Transport: Peer Review of Regulatory Capacity.

Jedwab, Rémi, Edward Kerby, and Alexander Moradi (2017). "History, path dependence and development: Evidence from colonial railways, settlers and cities in Kenya." *The Economic Journal* 127(603), 1467–1494.

Jedwab, Rémi and Alexander Moradi (2016). "The permanent effects of transportation revolutions in poor countries: Evidence from Africa." *Review of Economics and Statistics* 98(2), 268–284.

Jedwab, Rémi and Adam Storeygard (2021). "The average and heterogeneous effects of transportation investments: Evidence from Sub-Saharan Africa 1960–2010." *Journal of the European Economic Association* 20(1), 1–38.

Juhász, Réka (2018). "Temporary protection and technology adoption: Evidence from the Napoleonic blockade." *American Economic Review* 108(11), 3339–3376.

Kelly, Morgan (2022). "Understanding persistence." Working paper.

Kemper, Robert V. and Anya Peterson Royce (1979). "Mexican urbanization since 1821: A macro-historical approach." *Urban Anthropology* 8(3/4), 267–289.

Knapp, Frank A. (1952). "Precursors of american investment in mexican railroads." *Pacific Historical Review* 21(1), 43–64.

Knight, Alan (1986). "The revolution in power: 1 the great schism." In *The Mexican Revolution: Counter Revolution and Recconstruction*, Volume 2, Chapter 2, pp. 263–320. Cambridge University Press.

Krugman, Paul (1991). "Economic Backwardness and Firm Strategy: An American Railroad Corporation in Nineteenth-Century Mexico." *Journal of Political Economy* 99(31), 483–499.

Kuntz-Ficker, Sandra (1995). *Empresa Extranjera y Mercado Interno: El Ferrocarril Central Mexicano (1880-1907)*. El Colegio de Mexico.

Kuntz-Ficker, Sandra (2000, 05). "Economic Backwardness and Firm Strategy: An American Railroad Corporation in Nineteenth-Century Mexico." *Hispanic American Historical Review* 80(2), 267–298.

Kuntz-Ficker, Sandra (2017). "Mexico in the export era (1870s - 1929): Export boom, economic modernization, and industrialization." In *The First Export Era Revisited: Reassessing its Contribution to Latin American Economies*, pp. 235–278. Palgrave Studies in Economic History.

Kuntz-Ficker, Sandra and Antonio Tena-Junguito (2018). "Mexico's foreign trade in a turbulent era (1821-1870): A reconstruction." *Revista de Historia Económica / Journal of Iberian and Latin American Economic History* 36(1), 149–182.

Lewis, Daniel (2017). *Iron Horse Imperialism: The Southern Pacific of Mexico, 1880 -1951.* Tucson, Arizona: The University of Arizona Press.

Lin, Jeffrey (2015). "The puzzling persistence of place." *Business Review* (Q2).

Lin, Jeffrey and Ferdinand Rauch (2022). "What future for history dependence in spatial economics?" *Regional Science and Urban Economics* 94.

Long, William Rodney (1925). *Railways of Mexico.* Trade promotion series by the Department of Commerce, Bureau of Foreign and Domestic Commerce. U.S. Government Printing Office.

MacCallum, Robert C., Shaobo Zhang, Kristopher J. Preacher, and Derek D. Rucker (2002). "On the practice of dichotomization of quantitative variables." *Psychological Methods* 7(1).

Marein, Brian (2022). "Colonial roads and regional inequality." *Journal of Urban Economics* 131, 716–738.

Matthews, Matthew (2014). *The Civilizing Machine: A Cultural History of Mexican Railroads, 1876-1910.* Lincoln, Nebraska: University of Nebraska Press.

Matthews, Michael. 2017. Railroads and the Mexican Imagination during the Porfiriato and Revolution. Oxford University Press Available at: https://oxfordre.com/latinamericanhistory/display/10.1093/acrefore/9780199366439.001.0001/a 9780199366439-e-455.

McCaa, Robert (2000). "The peopling of Mexico from origins to Revolution." In Michael R. Haines and Richard H. Steckel, eds., *A Population History of North America*, Chapter 7. Oxford: Oxford University Press.

McCaa, Robert (2003). "Missing millions: The demographic costs of the Mexican Revolution." *Mexican Studies/Estudios Mexicanos* 19(2), 367–400.

McNeely, John H. (1964). "The railways of Mexico: A study in nationalization." *Southwestern Studies* 2(1), 3–56.

Medeles, Ana Maria (2016). "La ley del 26 del mayo de 1882 que constituyó a la dirección general de estadística." *Estatística e Sociedade* 8(4).

Mendoza-Velázquez, Alfonso, Vicente German-Soto, Mercedes Monfort, and Javier Ordóñez (2020). "Club convergence and inter-regional inequality in Mexico, 1940-2015." *Applied Economics* 52(6), 598–608.

Meyer, Michael C. and William L. Sherman (1991). *A Course of Mexican History*. New York: Oxford University Press.

Michaels, Albert L. (1970). "The crisis of Cardenismo." *Journal of Latin American Studies* 2(1), 51–79.

Miguel, Edward and Gérard Roland (2011). "The long-run impact of bombing Vietnam." *Journal of Development Economics* 96, 1–15.

Nunn, Nathan and Diego Puga (2012). "Ruggedness: The blessing of bad geography in Africa." *Review of Economics and Statistics* 94(1), 20–36.

Pindell, Terry and Lourdes Ramírez Mallis (1997). *Yesterday's Train: A Rail Odyssey Through Mexican History*. New York: Henry Holt & Co.

Pletcher, David M. (1950). "The building of the Mexican railway." *The Hispanic American Historical Review* 30(1), 26–62.

Powell, Fred Wilbur (1921). *The Mexican Central*, Chapter 15, pp. 127–131. New York: Stratford Company.

Pérez, Santiago (2018). "Railroads and the rural to urban transition: Evidence from 19th-century Argentina." Working paper.

Razo, Armando and Stephen Haber (1998). "The rate of growth of productivity in Mexico, 1850-1933: Evidence from the cotton textile industry." *Journal of Latin American Studies* 30(3), 481–517.

Redding, Stephen J. and Matthew A. Turner (2015). "Transportation costs and the spatial organization of economic activity." In Gilles Duranton, J. Vernon Henderson, and William C. Strange, eds., *Handbook of Regional and Urban Economics (Volume 5B)*, pp. 1339–1398. Elsevier Science & Technology.

Rees, Peter W. (1975). "Origins of colonial transportation in Mexico." *Geographical Review* 65(3), 323–334.

Roback, Jennifer (1982). "Wages, rents, and the quality of life." *Journal of Political Economy* 90(6), 1257–1278.

Rosen, Sherwin (1981). "The economics of superstars." *The American Economic Review* 71(5), 845–858.

Sánchez, Enrique Cárdenas (2015). pp. 87–187. México, D.F.': Fondo de Cultura Econńmica: El Colegio de México.

Serra, Maria Isabel, Maria Fernanda Pazmino, Genevieve M Lindow, Bennett W Sutton, and Gustavo Ramirez (2006). "Regional convergence in Latin America." Working Paper No. 2006/125, IMF.

Sokoloff, Kenneth L. and Stanley L. Engerman (2000). "Institutions, factor endowments, and paths of development in the New World." *Journal of Economic Perspectives* 14(3), 217–232.

Van Hoy, Teresa Miriam (2008). *A Social History of Mexico's Railroads: Peons, Prisoners, and Priests*. Rowman & Littlefield.

Waldinger, Maria (2017). "The long-run effects of missionary orders in mexico." *Journal of Development Economics* 127, 355–378.

Wasserman, Mark (1979). "Foreign investment in mexico, 1876-1910: A case study of the role of regional elites." *The Americas* 36(1), 3–21.

Waters, Wendy (2006). *Remapping Identities: Road Construction and Nation Building in Postrevolutionary Mexico*, pp. 221–242. New York, USA: Duke University Press.

Wells, Allen (1992). "All in the family: Railroads and henequen monoculture in Porfirian Yucatán." *Hispanic American Historical Review* 72(2), 159–209.

Wolfe, Mikael D. (2019, 08). "The Climate of Conflict: Politico-environmental Press Coverage and the Eruption of the Mexican Revolution, 1907–1911." *Hispanic American Historical Review* 99(3), 467–499.

Wooldridge, Jeffrey M. (2010). "Estimating average treatment effects." In *Econometric Analysis of Cross Section and Panel Data* (2 ed.)., Chapter 21. The MIT Press.

Wooldridge, Jeffrey M. (2012). "Limited dependent variable models and sample selection corrections." In *Introductory Econometrics: A Modern Approach* (5 ed.)., Chapter 17. Cengage Learning.

Wooldridge, Jeffrey M. (2015). "Control function methods in applied econometrics." *The Journal of Human Resources* 50(2), 420–445.

Xu, Ruonan (2021). "On the instrument functional form with a binary endogenous explanatory variable." *Economics Letters* 206, 109993.