Pontifícia Universidade Católica
do Rio de Janeiro

Pedro Américo de Almeida Ferreira

The Historical Origins of Development: Railways, Agrarian Elites, and Economic Growth in Brazil

Tese de Doutorado

Thesis presented to the Programa de Pós–graduação em Economia, do Departamento de Economia da PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Economia.

Advisor: Prof. Claudio Ferraz

Rio de Janeiro
June 2020

Pontifícia Universidade Católica
do Rio de Janeiro

Pedro Américo de Almeida Ferreira

The Historical Origins of Development: Railways, Agrarian Elites, and Economic Growth in Brazil

Thesis presented to the Programa de Pós–graduação em Economia da PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Economia. Approved by the Examination Committee:

Prof. Claudio Ferraz
Advisor
Departamento de Economia – PUC-Rio

Prof. Juliano Junqueira Assunção
Departamento de Economia – PUC-Rio

Prof. Renato Perim Colistete
Faculdade de Economia, Administração e Contabilidade – USP

Prof. William Roderick Summerhill
Department of History – UCLA

Prof. Leonardo Monteiro Monasterio
Escola Nacional de Administração Pública – Enap

Rio de Janeiro, June the 18th, 2020

Pedro Américo de Almeida Ferreira

Completed his Bachelor of Arts degree in Economics from Universidade Federal do Rio de Janeiro (UFRJ) in 2011 and obtained his Master of Science degree in Economics from Universidade Federal do Rio de Janeiro (UFRJ) in 2015. Now holds a PhD degree in Economics from PUC-Rio. During his Ph.D. studies, he was a Fulbright visiting researcher of the Anderson School of Management, UCLA.

Bibliographic data

Américo, Pedro

The Historical Origins of Development: Railways, Agrarian Elites, and Economic Growth in Brazil / Pedro Américo de Almeida Ferreira; advisor: Claudio Ferraz. – 2020.

183 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Economia, 2020.

Inclui bibliografia

1. Economia – Teses. 2. Ferrovias. 3. Mudança Estrutural. 4. Persistência. 5. Aglomeração. 6. Elites. 7. Educação. I. Ferraz, Claudio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Economia. III. Título.

CDD: 330

Aos meus pais.

Acknowledgments

First of all, I'm grateful for my family's unconditional support over all these years; without dedication and encouragement from my parents, Carlos Américo and Maria Lúcia, this project would be impossible. And for them, this thesis is dedicated. I'm also grateful for the guidance and affection from my grandparents João Alexandre (in memory), Sônia Maria, Samuel José (in memory), and Adir Machado. I'm grateful for the love and affection throughout this stressful period from my wife Giovana.

I am thankful for the teaching and dedication from the professors of the Department of Economics of PUC-Rio, especially my advisor, Claudio Ferraz, for the generosity, and patience. His guidance was essential to the result of this thesis. I'm also thankful to professors Marcelo de Paiva Abreu, Gustavo Gonzaga, and Juliano Assunção for their comments, suggestions, and conversations. Even before my Ph.D., I was encouraged by many professors from my undergraduate and master's studies, therefore I'm grateful to Jorge Chami, Antonio Licha, Getúlio Borges, Rolando Otero, and Rudi Rocha. I'm also thankful to the professors Nico Voigtländer and William Summerhill for having me at UCLA during my research fellowship visit. The conversations in California's pleasant climate contributed a lot to my work.

A special thanks to my friends from PUC-Rio, UFRJ, and CEFET for providing me moments of joy in my trajectory.

Finally, I would like to thank the financial support from CAPES, CNPq, and Fulbright Brazil.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Américo, Pedro; Ferraz, Claudio (Advisor). **The Historical Origins of Development: Railways, Agrarian Elites, and Economic Growth in Brazil**. Rio de Janeiro, 2020. 183p. Tese de Doutorado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

This dissertation explores the impact of transportation infrastructure on economic development and the effects of agrarian elites' political power on investments in education. The first chapter documents the impact of the Brazilian railway network on structural transformation between the late nineteenth and middle twentieth century. We exploit variation induced by geographic location, where municipalities near the least-cost routes were more likely to be connected to railway system, to identify the causal effects of railroads on structural transformation. We show that the expansion of the transportation infrastructure shifts the workers from agriculture to manufacturing. We provide evidence that market integration and the adoption of new technologies by manufacturing firms were two important mechanisms that explain the change in the occupational structure of the economy. The second chapter explores the persistence of the impact of Brazil's railway network on long-run economic development between 1950 and 2010. We exploit variation induced by geographic location, where municipalities near the least-cost routes were more likely to be connected to railway system, to identify the causal effects of railroads on economic activity. We show that, despite the decline of the railway network post-1950s, the effect on economic development persisted. Our findings suggest that agglomeration and urbanization were key drivers of economic activity persistence. In the third chapter, we study the relationship between the political power of agrarian elites and the spread of mass schooling in the early 20th century in Brazil. We use a novel dataset on the occupational structure of the voting elites in 1905 and historical censuses to test whether places, where more voters belonged to the agriculture elite, invested less in schooling. We find that municipalities with a higher share of agriculture voters have a lower literacy rate in 1920 and these effects persist until the 1970s. In the long-run, municipalities that had a higher share of agriculture voters have fewer years of schooling and lower income per capita. We provide evidence that the supply of educational inputs is the main mechanism that explains the long-term persistence.

Keywords

Railroads; Structural Transformation; Persistence; Agglomeration; Elites; Education.

Resumo

Américo, Pedro; Ferraz, Claudio. **As Origens Históricas do Desenvolvimento: Ferrovias, Elites Agrárias, e Crescimento Econômico no Brasil**. Rio de Janeiro, 2020. 183p. Tese de Doutorado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

Esta tese explora o impacto da infraestrutura de transporte sobre desenvolvimento econômico e os efeitos do poder político das elites agrárias no investimento em educação. O primeiro capítulo documenta o impacto da rede ferroviária brasileira na transformação estrutural da economia entre o final do século XIX e meados do século XX. Exploramos a variação induzida pela localização geográfica, onde municípios próximos às rotas de menor custo estavam mais propensos a serem conectados ao sistema ferroviário, para identificar os efeitos causais das ferrovias sobre transformação estrutural. Mostramos que a expansão da infraestrutura de transporte realoca os trabalhadores da agricultura para a manufatura. Fornecemos evidências de que a integração do mercado e a adoção de novas tecnologias pelas empresas manufatureiras são dois mecanismos importantes que explicam a mudança na estrutura ocupacional da economia. O segundo capítulo explora a persistência do impacto da rede ferroviária brasileira no desenvolvimento econômico de longo prazo entre 1950 e 2010. Exploramos a variação induzida pela localização geográfica, onde municípios próximos às rotas de menor custo tinham maior probabilidade de estarem conectados ao sistema ferroviário, para identificar os efeitos causais das ferrovias na atividade econômica. Mostramos que, apesar do declínio da rede ferroviária pós-1950, o efeito sobre desenvolvimento econômico persistiu ao longo do tempo. Nossos resultados sugerem que aglomeração e urbanização foram os principais impulsionadores da persistência da atividade econômica. No terceiro capítulo, estudamos a relação entre o poder político das elites agrárias e a disseminação da educação em massa no início do século XX no Brasil. Utilizamos um novo conjunto de dados sobre a estrutura ocupacional das elites votantes em 1905 e censos históricos para testar se locais onde mais eleitores pertenciam à elite agrícola investiam menos em educação. Constatamos que os municípios com maior participação de eleitores na agricultura têm uma menor taxa de alfabetização em 1920 e esses efeitos persistem até a década de 1970. A longo prazo, os municípios com maior participação de eleitores na agricultura têm menos anos de escolaridade e menor renda per capita. Apresentamos evidências de que o fornecimento de insumos educacionais é o principal mecanismo que explica a persistência a longo prazo.

Palavras-chave

Ferrovias; Mudança Estrutural; Persistência; Aglomeração; Elites; Educação.

Table of contents

List of figures

List of tables

1
On the Industrialization Track: Railroads and Structural Transformation in Brazil 1872-1950

1.1
Introduction

The railways played a fundamental role in the integration of international and national markets between the nineteenth and twentieth centuries. By connecting interior producers to ports and large cities, railways accelerated the convergence of commodity prices, both internationally and nationally (Findlay and O'rourke, 2009, p. 405). However, the impacts of this transportation technology revolution are not limited to its direct effects on market integration. The railroad expansion has increased the real income levels (Donaldson, 2018) and changed the occupational structure of the economy (Pérez, 2017; Yamasaki, 2017; Berger, 2019). Despite the expansion of the Brazilian railway network coincides with a substantial growth of manufacturing in the country, the existing estimates of the direct benefits of the railways to the industrial sector are relatively small (Summerhill, 2003).

This paper documents the impact of transportation infrastructure improvements on structural transformation between the end of the nineteenth century and the beginning of the twentieth century in Brazil and shows evidence on the mechanisms behind this effect. We show that by integrating markets and facilitating the import of new technologies, the railroads generated significant indirect economic gains for Brazilian industrial growth. Although the first Brazilian railway line was opened in 1854, the expansion of the railroad system did not occur until the 1870s. Faced with extremely high transportation costs and a country of continental size, railways became the best solution for connecting interior fertile lands to export coastal ports. Between 1870 and 1950, the Brazilian railroad network expanded by more than 36,000 kilometers, substantially reducing the transportation costs. Summerhill (2005) estimates a reduction in freight costs per ton-kilometer by about 86% due to the railway network expansion, generating an economy in freight services that represents 18% of the Brazilian GDP in 1913.

To estimate the impact of the railroads on structural transformation, we assemble a novel data set that combines the digitalization of the railway's historical maps and national censuses. The area under study covers the municipalities of the three most populous regions of Brazil: northeast, southeast, and south, where approximately 94% of the railroads' lines were located in 1950. The data allow us to follow almost 80 years of railroad expansion in Brazil and changes in the occupational structures of the economy. We use data on goods' prices, manufacturing plants and workers, and imported machinery to explore the mechanisms through which declines in transportation costs can shift the labor force out of agriculture.

The empirical analysis is based on the instrumental variables approach to address the problem of the endogenous placement of railroads. The purpose of the railroad expansion in Brazil was to connect the inland municipalities to the ports and to integrate the independent regional railways' systems. As a consequence, the propensity of a municipality to be connected to the transportation network varies according to its proximity to the least-costs routes. Similar to the strategy used by Fajgelbaum and Redding (2018), we instrument the railroad connection with the percentage of grid cells within each municipality that lie along the least-cost paths between the railroads' expansion targets interacted with the total extension of the railway network in the country. Conditional on the municipality fixed effects, year fixed effects, and controls we expect the process of railroad expansion to be faster in localities near the least-cost routes. A series of balance tests support the validity of our identification strategy.

We begin by showing that the railroad expansion reduced the share of agriculture workers in almost 20% between 1872 and 1950. Our estimates suggest that most shift of workers out of the agriculture occurred in the 1940s when the Brazilian industrialization took off. Also, we show that the workers' shift out from agriculture was channeled into manufacturing. The railroad expansion increased the share of manufacturing workers, by almost 14% between 1872 and 1950. The manufacturing growth took place, in both the extensive and intensive margins, with an increase in the number of industrial factories and in the number of workers by plants. We find heterogeneity in the impact of the railroads: the effect of the transportation costs is proportional to the distance to the railway line, the year of connection to the system, and the number of railroad stations. The impact of the railways is higher in municipalities within 10 kilometers of a railroad line, in those connected to the system until the 1920s, and also in those with more railway stations. All results are robust to the inclusion of geographic, baseline socioeconomic

characteristics, and pre-railway transportation infrastructure controls, as well as the municipality, and year fixed effects. In the same way, the results remain significant when we use different measures of railroad connection.

Despite the effects of the railroads on shift the workers from agriculture to manufacturing, we do not find impact of the railways on population and population density between 1872 and 1950, not even a spatial reallocation from the population. Contrary to other evidence found in the literature (see, for example, Pérez (2017)), we find that structural change occurs only by modifying the occupational structure of local economies and not by an increase in internal migration from rural areas to industrialized urban cities. The evidence suggests that immigration and agglomeration effects are not the mechanisms behind our results. Therefore, we analyze other mechanisms to shed light on how the expansion of the railway network changed the employment sectoral composition. First, the railroad reduced the price dispersion for corn, bean, liquor, and flour. Connected municipalities pairs have, on average, a price dispersion 22% to 30% (depending on the good) lower than those not connected. By integrating markets, railways potentially allowed municipalities to specialize in what they were more productive, changing the occupational structure.

Second, we show that the railroad's expansion increased the adoption of new technologies in textile manufacturing. Our estimates suggest that municipalities connected to the railway network had a higher probability to import textile machinery spindles from British exporters. The railroads allowed new technologies to be distributed to factories located in municipalities in the interior of the country, locations, in general, sought because of their energy potential in water power (Birchal, 1999). Given the importance of the textile sector for Brazilian industrialization (Stein, 1957; Dean, 1969), the adoption of new technologies seems to be a fundamental mechanism for the growth of the manufacturing sector.

Our paper contributes to three sets of literatures. First, it contributes to the work that measures the economic impact of transportation infrastructure. The literature is inaugurated by the seminal works of Fogel (1964) and Fishlow (1965) that, using the "social saving" methodology, analyze the impact of the railroads on the American economy. More recently, with the abundance of data and the new estimation methods, much of the literature has started to focus on the historical impact of the railroads on market integration (Keller and Shiue, 2008; Andrabi and Kuehlwein, 2010; Donaldson, 2018), urban growth (Atack et al., 2010; Hornung, 2015; Berger and Enflo, 2017), agricultural land values (Donaldson and Hornbeck, 2016; Donaldson, 2018), innovation (Andersson et al., 2020), and the spread of factories (Atack et al., 2008; Tang, 2014;

Hornbeck and Rotemberg, 2019). Other articles analyze the persistent effects of the railroad (Jedwab and Moradi, 2016; Jedwab et al., 2017; Okoye et al., 2019). More related to our work, Yamasaki (2017), Pérez (2017), Fajgelbaum and Redding (2018), and Berger (2019) study the effects of the railroads on structural transformation. Distinctly from these last articles, we show that for Brazil the impact of railroads on structural change was primarily due to the change in the local occupational structure, and not due to internal migration from rural to urban areas. We provide evidence that the immigration and agglomeration effects were limited between 1872 and 1950.

Second, our work also contributes to the structural transformation literature[1]. While some authors point the importance of productivity growth in the industrial sector for structural change (Alvarez-Cuadrado and Poschke, 2011; Yamasaki, 2017), others emphasize the role of agricultural labor-saving technological changes to shift the workers from agriculture to manufacturing (Bustos et al., 2016). In our article, we bring evidence that supports the thesis that the productivity growth of the manufacturing sector is decisive for structural transformation. More specifically, we show that the import of new technologies partly explains the growth of the industrial sector.

Third, the paper contributes to the classical debate about Brazilian industrialization and the importance of the railroads in this process. Although the importance of railways for Brazilian industrialization is minimized in the canonical works of Furtado (1959), and Dean (1969) we find significant effects of the railroads on the spread of the industrialization. Also, our article complements the qualitative historical or social saving approach evidences on the role of railways in the Brazilian economy from Mattoon Jr (1977), Saes (1981), Matos (1990), Lamounier (2012), Grandi (2007), Grandi (2013), Summerhill (1998), Summerhill (2003), providing robust evidence of the causal impact of railways on structural transformation. In related research, Summerhill (2005) calculates the direct effects of railways on the industrial sector due to the savings generated by the drop in transportation costs. Our results show that the indirect effects of the railways can be substantial due to the integration of markets, increased imports of new technologies, and structural changes.

The rest of the article is organized as follows. Section 1.2 outlines the historical background of the railroad network expansion in Brazil between 1854 and 1950. Section 1.3 presents our data and how it was built. Section 1.4 describes the empirical strategy. Section 1.5 presents the main results, and discuss the heterogeneity of the effects. Section 1.6 documents the mechanisms

[1]For a review of this literature see, for example Foster and Rosenzweig (2007), and Herrendorf et al. (2014).

underlying the impact of the railroads on structural transformation. The final section concludes.

1.2 Historical Background

1.2.1 Railroad Expansion in Brazil, 1854-1950

The first Brazilian railroad line was opened in 1854, connecting the Mauá port to the nearby district of Vila Inhomirim, in the city of Magé. Over a 14.5 km length, the railroad's goal was to reduce the costs to transport agricultural goods from the state of Minas Gerais to the country's capital, Rio de Janeiro. Despite this first initiative, the expansion of the railroad network in Brazil was slow between 1854 and 1870. Faced with a high-risk investment, and legislation that limited the maximum rates for freight and passenger services, investors had no incentive to fund the construction of new railways (Summerhill, 1998).

From the late 1860s to the early 1870s, the speed in the concession and construction of new railroads started to change. The government modified the concession rules, especially with the 1873 railroad law, offering guaranteed minimum dividends to investors[2]. The introduction of the government subsidy, beyond the fall in railroads' inputs prices and the increase in the amount of Britsh investment in Brazil, explain the growth of the Brazilian railroad network in the late nineteenth century (Summerhill, 2003; Mattoon Jr, 1977). The introduction of operating monopoly within 30 km of the railroad[3], in other words, no other railway could be built within 30 km of each side of the railway, increased incentives for investment in railroads too.

By 1860, the railroad network was limited, covering just a few cities nearby Rio de Janeiro and in the Northeast of the country, extending for less than 250 kilometers. By the beginning of the twentieth century, the length of the railway system had already increased to over 20,000 kilometers, an average growth of 478 kilometers of railway per year between 1860 and 1914. Figure 1.1(a) presents the extension of the railroad system between 1860 and 1950. Although it continued to grow, the speed of expansion of the railway system declined in the years during the First World War, with trade and capital flow disruption (Summerhill, 2005). After the war, railroads began to feel the first consequences

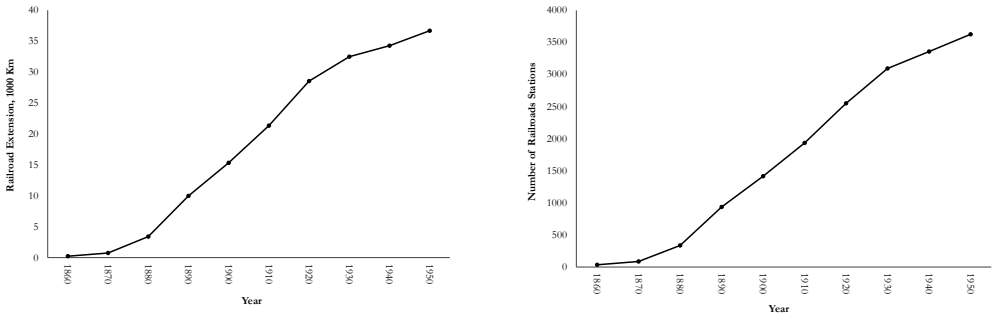
[2]Most of the railways built between 1870 and the early twentieth century had a guaranteed minimum dividend of 7% (Villela and Suzigan, 1975).

[3]The monopoly area could vary depending on the concession. See, for example, (Matos, 1990).

of car competition. Between 1920 and 1940, the growth of the railway network was 285 kilometers per year, over 40% reduction in expansion compared to 1860-1914. From the 1950s onwards, in the face of fierce car competition and a steady worsening of financial results, railroads have entered a long phase of decline[4]. A similar pattern is found when we analyze the expansion of the number of railways stations and not their extension, as we can see in Figure 1.1(b).

Figure 1.1: Railroad Network Expansion, 1860-1950

(1.1(a)) Railroad Extension, in 1,000 Km. (1.1(b)) Number of Railroads Stations



Notes: The figures show the evolution of the extension and number of stations from the railroads' network between 1860 and 1950. *Source:* Data on railroads extension in 1,000 kilometers from IBGE (2003), available at https://seculoxx.ibge.gov.br/. Number of railroads stations collected using web scraping from http://www.estacoesferroviarias.com.br/.

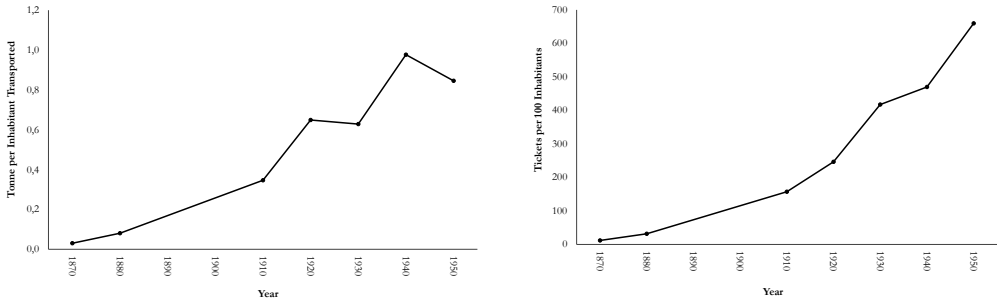
Figure 1.2 shows the expansion in manufacturing, agricultural, and mail goods transported by railroads, as well as the number of tickets sold for passengers between 1870 and 1950. The increase in tonnes *per capita* transported and passengers carried by railroads is consequence of the expansion of the railroad network and its integration of the market in the first half of the twentieth century. The fastest expanding period is between 1870 and 1920, with a reduction in 1920-1950. The evolution of the ton and passengers transported seems to follow the pattern of the expansion of the lines.

[4]In this context, the government started to encourage investments in roads and not more in railroads (Grandi, 2013).

Figure 1.2: Products and Passengers Carried by Railroads , 1870-1950

(1.2(a)) Tonne Per Capita Transported by Rail

(1.2(b)) Passengers Carried by Rail



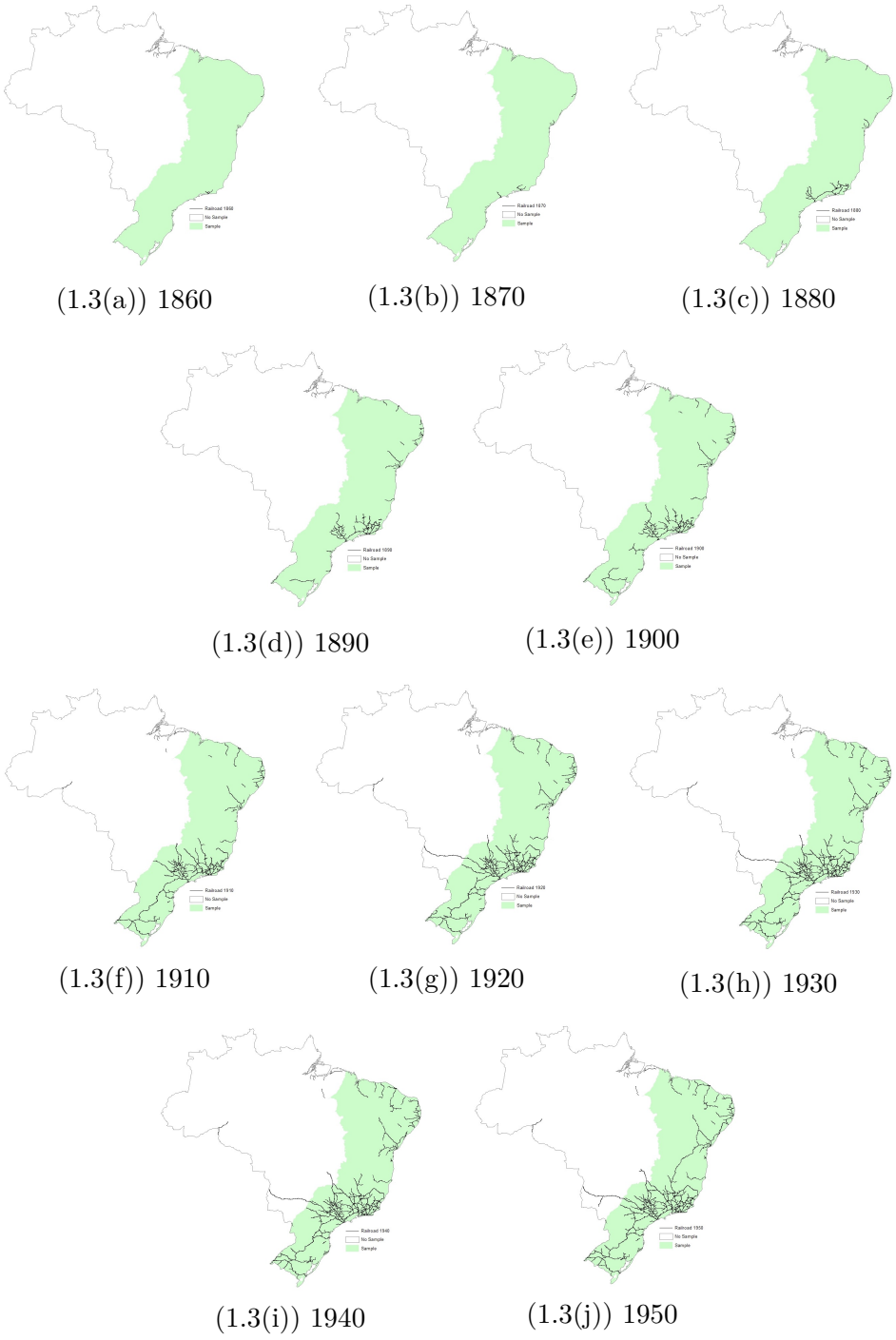
Notes: The figures show the evolution of the products and passengers transported by railroads between 1870 and 1950. *Source:* Data on tonnes of products transported by railway and passengers from Ipeadata, available at http://www.ipeadata.gov.br/. Original data from statistical yearbooks of transport for several years. Population data from IBGE.

Figure 1.3 shows the railroad network expansion from 1860 to 1950. The expansion of the Brazilian railroad network between the second half of the nineteenth century and 1930 aimed to connect the productive lands of the interior with the main ports to facilitate the exportation of agricultural goods. This initial expansion was largely financed by private capital from foreign investors and landowners, with government subsidies. Thus, the railway network created was characterized by an independent set of railways linking agricultural production areas to the nearest export port[5]. From the 1930s, with the diminishing profitability of the railways[6], the government participation in the sector began to increase. In 1929, more than 50% of railroads were privately run, falling to only 6% in the 1950s (Villela and Suzigan, 1975). As a consequence of the greater state participation, railroad expansion between 1930 and 1950 had the goal to integrate the independents' regional railways systems.

[5]In the Northeast, the railway lines followed the sugar and tobacco plantations. In the Southeast, on the other hand, they facilitated the connection with coffee-producing areas (Lamounier, 2012).

[6]By the 1940s, expenses already accounted for more than 90 percent of railroad revenues (Villela and Suzigan, 1975).

Figure 1.3: Railroads Network Expansion in Brazil, 1860-1950



(1.3(a)) 1860 (1.3(b)) 1870 (1.3(c)) 1880

(1.3(d)) 1890 (1.3(e)) 1900

(1.3(f)) 1910 (1.3(g)) 1920 (1.3(h)) 1930

(1.3(i)) 1940 (1.3(j)) 1950

Notes: This figure displays the expansion of the railroad network between 1860 and 1950. See main text and Appendix A for data sources and details on the construction of the railway network.

Before the railroads' era, animal-drawn carts and mules wagons were the main options for land transportation. All trade between the coast and the interior of the country was done by animals crossing roads in terrible conditions since the most populous regions had no navigable rivers. On the other hand, the connection between the cities of the coast was made by ships (Summerhill,

1998). The animal-drawn carts were slower than the trains, a 130 kilometers trip that took at least 7 days with the mules, reduced to less than a day with the railroads since mules wagons only reached a maximum speed of 3-4 leagues (14-19 Km.) per day (Silva, 1949). The expansion of the railway network reduced freight costs per ton-kilometer by about 86% (Summerhill, 2005). Summerhill (2003) estimates that this economy generated by railroad freight service represents 18% of the Brazilian GDP in 1913, more than double of the estimates found for the United States. Unlike the US, Brazil had a precarious pre-railway transportation structure. For example, in the richest regions of the country little was done to improve the navigability of rivers. Also, the existing roads were in poor condition. No major improvements over those roads used by indigenous people for centuries were done during the nineteenth century[7]. The poor pre-rail transportation infrastructure, besides a severe topography and regular rainstorms, caused a large drop in transportation costs with the arrival of the first railways.

If the roads connecting the coast to the interior of the country were in a terrible condition, the trade and people transportation between coastal cities were made by ships. Coastal shipping was a relatively efficient and inexpensive way to connect the major cities of the country. Thus, the railways did not compete with this mode of transport. Indeed, the railroad expansion took place at the East-West connection, replacing mules wagons and increasing the efficiency to export agricultural goods.

1.2.2 Railroads and Structural Transformation

The first consequence of the expansion of the railways was the expansion of the agricultural frontier. Distant areas, without agricultural production due to prohibitive trade costs, started to attract people creating new population centers and cities. The railroad stations, a place of great circulation of passengers, attracted new retail stores and commerce. Therefore, there was an increase in population density of the cities connected to the railroad network (Grandi, 2007). Furthermore, because of the maintenance of the trains, the railroads originated the first maintenance shops and small repair factories[8].

[7]There are many reports describing the terrible conditions of Brazilian roads in the 19th century. For example, Thomas Davatz, an immigrant visiting São Paulo in the 1850s, observed the conditions of the main roads of the state: "In the dry season they develop pot holes in many places due to the heavy and crisscrossing mule traffic, and in the rainy season these fill with water and mud so that the animals sink in up to the belly and while trying to walk on three feet seek out solid ground with the fourth" (Mattoon Jr, 1977).

[8]Despite the establishment of these repair factories, Summerhill (2005) conclude that Brazil's domestic backward linkages were limited: "No iron and steel industry arose to meet

By reducing the market integration costs, the railroads affected the location of the factories, inducing manufacturing centers to form near railroads[9]. Locating factories near railroads was a strategic decision to be able to sell products to a larger market, as well as to facilitate the importation of machinery and equipment (Birchal, 1999), since, until the mid-twentieth century, Brazilian industry was heavily dependent on machine imported from the US and Europe. In addition to affecting the expansion of factories, the literature emphasizes that the railroads freed up a considerable contingent of farm workers who were dedicated to transporting goods by roads with mules. Such employees start to work in other functions on the farms, or are moved to other productive sectors, such as manufacturing.[10]

The period of expansion of the railways is also a period of great modification of the Brazilian economy. There was a rapid growth of industrial production, especially for consumption goods, between the beginning and the middle of the twenty century. In 1907, the year of the first manufacturing census, there were 3,258 factories and more than 150 thousand industrial workers. Between 1907 and 1950, Brazil had a great expansion of its industry reaching more than 90 thousand factories and about one million workers.[11] Although infrastructure conditions have been a barrier to Brazilian industrial development (Villela and Suzigan, 1975), there is little quantitative evidence for the impact that the expansion of the railroad network has had on the structural change of the economy. An exception is Summerhill (2003), who shows that the railroads generated a direct saving of 5.8% over its total production for the Brazilian textile industry in 1913. In other words, without the railways, about 6% of the production value of these industries would be reverted to the payment of transport costs. The author also argues that the railroads helped to spread new and more productive manufacturing industries in the interior of the country.

most of the railroads' derived demand, and the bulk of inputs came from abroad."

[9] Jundiaí, Campinas, Americana, Limeira are an example of cities in the state of São Paulo, where industrial centers were formed near railroads (Matos, 1990).

[10] For a critical analysis of the role of railways on the release of workers from the former mule wagons transportation sector, see Lamounier (2012).

[11] Although the 1907 census is not exactly a census, but an incomplete survey of Brazilian industries (Dean, 1969), it gives us the first estimates for the size of the Brazilian industrial sector. Anyway, the comparison between the 1920 and 1950 censuses confirms the high growth of manufacturing in the first half of the 20th-century (Villela and Suzigan, 1975).

1.3 Data

This study combines four sets of historical data. First, we use censuses data to calculate, by the municipality, the share of workers that are employed in the agricultural, manufacturing, and service sector, as well as the total population. Second, we complement these data with industrial censuses, price data, and textile machinery imports information. Third, we merge the employment, population, and industrial data with railroad network data drawn from historical maps. Fourth, we achieve these data with socioeconomic, geographic, and transportation variables. The main data set used in this paper is a panel covering 547 municipalities from three regions of Brazil between 1872 and 1950, a total of 2,188 municipality by census year observations.

1.3.1 Census Data

The historical population information used in this paper comes from 1872, 1920, 1940, and 1950 Population Censuses. For each municipality, theses census records contain demographic and socioeconomics information. The employment data from theses censuses is used to calculate our main outcome variable: the share of workers employed in the agricultural, manufacturing, and service sectors. Also, the demographic information is used to calculate the total population and the population density by municipality. Information from the 1920 census on the number of factories, and its average number of workers, at the municipality level, complement our data set. Finally, the 1872 census allows us to build socioeconomic characteristics variables just at the beginning of the railroads' expansion. These variables are used to control for pre-existing social and economic aspects.

The census data set covers all the municipalities in the five regions of the country. However, our sample covers just three regions: Northeast, Southeast, and South, where around 94% of the existing Brazilian railway network and about 93% of the population of the country in 1950 were located (IBGE, 2003). Because of the large differences in the characteristics of the municipalities in the other regions from those in our sample, and due to the large concentration of railways in the regions analyzed, we decided to focus the sample only on the three most important regions of the country. By doing this, we mitigate the selection concerns.

The sample used in our main exercises maintains the 1872[12] border

[12]We can build some of the manufacturing outcome variables, like the number of factories,

definition. Therefore, we merge the data of the municipalities from 1950, 1940, and 1920 censuses to match the 1872 census boundaries[13]. As a result, in the main data set, we follow the 547[14] municipalities that existed in 1872 in our sample between 1872 and 1950, a total of 2,188 municipality by census observations.

1.3.2 Manufacturing, Price Dispersion, and Machinery Imports Data

We complement our outcome variables with additional manufacturing, price, and machinery imports information. The 1920 manufacturing census from Brazil allows us to examine the impact of the railroads on manufacturing in the short-term. These census records contain municipality-level information about factories' plants and the number of workers. To shed light on the impact of the railroads on the manufacturing technology adoption, we construct explanatory variables on the number of spindles imported by Brazilian textile firms between 1878 and 1933. The focus on the textile industry is due to the limitations of our database, which contains only information on machinery exports for this productive sector.[15] Although the textile industry is not representative of the entire Brazilian manufacturing sector due to significant differences in workers' skills, spatial distribution, and economies of scale, the textile firms played a decisive role in Brazilian industrial growth, mainly in the initial period of Brazilian industrialization (Stein, 1957; Dean, 1969). Therefore, given the data constraint and aware of the exercise limitations, we take the impact of the railways on the import of technologies for the textile industry as a proxy for the manufacturing sector in general. The historical data was collected by Saxonhouse and Wright (2010) and contains information on the number of spindles for textile exported to firms around the world by British machine producers, which were the leader exporters of textile machinery at the time. To identify the municipality in which the importing firms were located,

just for the 1920 period. Thus, in some of the empirical analysis, we maintain the 1920 border definition.

[13]A similar procedure has been used for the United States (Hornbeck, 2012) and Brazil (Rocha et al., 2017). The Brazilian administrative division can be found at https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial.html. See Appendix A for details.

[14]More precisely, for the Northeast, Southeast, and South regions there were 568 municipalities in 1872. But, for 6 municipalities we don't have census data, and 15 municipalities were incorporated by other cities before 1950.

[15]Although Suzigan (1986) collected data on the export of machinery and equipment from the United Kingdom, United States, Germany, and France to Brazil between 1855 and 1939 for all industrial sectors, it is not possible with these data to identify the municipality where the machine importing firm was located. Therefore, despite focusing only on the textile industry, we use the data from Saxonhouse and Wright (2010).

we combine these data with the list of textile industries from the industrial censuses of 1907 and 1920. In this way, we can build, by the municipality, outcome variables for importing textile machines.

Price dispersion data allows us to analyze the impact of railways on market integration, thus exploring a possible mechanism of the impact of railways on structural transformation and development. Therefore, based on historical records from the *Questionários sobre as Condições da Agricultura dos Municípios do Brasil*[16], we measure the price dispersion between municipalities for important goods (corn, bean, liquor, and flour) in 1910. The final data set contains price difference per product and pair of municipalities.

1.3.3 Brazilian Railway Network, 1860-1950

The purpose of this paper is to analyze the impact of the railroads on structural transformation. Our treatment variables are measures of the closeness to the railroad network from the municipality downtown (the main urban settlement of the municipality)[17]. To construct these variables, we created decade-by-decade digitized maps of the Brazilian railway network, using modern and historical maps of the railroads, in four steps. First, we started from a geo-referenced map of the modern railroad system created by the Ministery of Infrastructure[18]. Second, we digitized the railway network using historical maps following the same methodology from Atack (2013). We obtain the historical maps of the expansion of the railroad network from multiple sources, including *Plano Nacional de Aviação*, *Estatísticas das Estradas de Ferro do Brasil* and other historical reports[19]. Figure 1.4 reports an example of a historical map from the *Leopoldina* railway used in the construction of our database. Third, with the digitized railroad network, we calculated the planar distance from the municipality's main urban settlement to the nearest railway line in each decade. Finally, we certified the consistency of the railroad network by collecting data on the opening date of each railway station[20]. See Appendix

[16]The price data are not available for all Brazilian states, so municipalities in the states of Rio Grande do Sul, Bahia, Pernambuco, Piaui, and Maranhão will not be included in our price dispersion sample.

[17]Similar to Melander (2018), we think that to measure railway access, the distance from the main settlement is a better proxy than, for example, the distance from the centroid of the municipality.

[18]The modern railroad network shapefile is available at http://www.infraestrutura.gov.br/bit.html.

[19]Most of the maps and reports can be found at https://biblioteca.ibge.gov.br/ and http://memoria.org.br/index.php.

[20]The data are available at http://www.estacoesferroviarias.com.br/, and were collected using web scraping.

A for details on the construction of the dataset.

Figure 1.4: *Leopoldina* Railway in 1954

Notes: The map shows the lines from *Leopoldina* railways in 1954. *Source:* I Centenário das Ferrovias Brasileiras, 1954.

In addition to the distance to the railway line, we built a dummy to determine if the railway is less than 10 kilometers from the municipality. Since the digitizing of historical maps has a margin of error, dummy seems to us to be the most appropriate treatment variable. Similar connection measures are standard in the literature (see, for example, Jedwab et al. (2017)). Finally, we also use the number of railway stations as treatment variable.

1.3.4 Additional Data

To characterize pre-existing conditions at the municipality level, we make use of important sets of socioeconomic baseline controls, as well as geographic and transportation infrastructure controls. The 1872 census allows us to control for social and economic characteristics of the municipalities at the beginning of the expansion of the railroad network, since in the 1870s the railway extension was very limited and concentrated in few municipalities. Thus, we construct

municipality-level baseline controls: (i) share of literate individuals aged 6 or more; (ii) share of foreign-born people; (iii) share of slaves; (iv) share of workers in public administration; (v) share of workers in legal professions.

As we already described the expansion of the Brazilian railroad network aimed to connect the plantations in the interior to the ports at the coast. As the agricultural producing regions and the coast have very specific geographical characteristics, consequently it is expected that the expansion of the railways to be correlated with the geographic characteristic of the municipalities. Also, the expansion's cost of the railways depends on the geographic characteristics of the territory[21]. Consequently, we include geographic controls in our analysis: longitude, latitude, altitude, distance to the nearest coast, distance to the nearest state's capital, soil types, slope, and area.

The effects of railroads also depend on the pre-rail transportation infrastructure. Thus, we include in our analysis controls for the presence of transportations options just before the expansion of the railways. These controls are the distance to the nearest port, distance to the nearest road, and distance to the nearest river.

Finally, to construct our instrumental variable we use two geographic information: the average slope in each 0.5 x 0.5 Km^2 cell, and the presence of river at the same cell. Using high-resolution spatial data from CGIAR-CSI[22], and river's data from ANA[23] we build our measure of propensity to connect to the rail network. The least-cost paths are built assuming a cost function directly proportional to the average slope of the region, adding a penalty for the presence of a river. After defining the cost function, we calculate the cost minimization routes based on the algorithm created by Dijkstra et al. (1959). In Appendix A we present detailed information on the construction of the least-cost paths, as well as, definition of the variables used in this paper.

1.3.5 Descriptive Statistics

Summary statistics for key variables can be found in Table 1.1. The share of workers in the agricultural sector was 71%. The non-agricultural labor force represents 29%, which 10% in manufacturing and 19% in the service sector. Basically, until 1950 the Brazilian economy was dependent on the export of agricultural products to the international market. The average log of the

[21]In fact, much of the cost of building the railroads depended on the ground conditions and their slope (Lamounier, 2012).

[22]The data can be found at http://srtm.csi.cgiar.org/srtmdata/

[23]The data can be found at http://dadosabertos.ana.gov.br/

population at the 547 municipalities between 1872 and 1950 is around 10.32. The average distance to the nearest railroad line is about 105 kilometers, and almost 40% of the municipalities were connected to the Brazilian railway system until the 1950s.

Table 1.1: Summary Statistics

	Observations	Mean	S.D.	Min.	Max.
Panel A: Structural Transformation and Population, 1872 - 1950					
% Emp. agriculture	2,188	0.71	0.20	0.01	0.99
% Emp. manufacturing	2,188	0.10	0.08	0.00	0.55
% Emp. service	2,188	0.19	0.15	0.01	0.98
Log population	2,188	10.32	1.04	7.19	14.73
Log density	2,188	2.50	1.25	-2.66	7.62
Panel B: Manufacturing, 1920					
Dummy [Factories > 0]	1,076	0.62	0.49	0.00	1.00
Log number of factories	1,076	0.38	2.32	-2.30	7.34
Log workers by factories	665	4.24	1.17	0.11	7.83
Panel C: Log Price Dispersion, 1910 - 1913					
Corn	115,921	0.50	0.43	0.00	3.22
Bean	125,751	0.57	0.44	0.00	3.15
Liquor	169,071	0.52	0.41	0.00	3.22
Flour	152,628	0.53	0.44	0.00	4.27
Panel D: Imported Textile Machines, 1878 - 1933					
Dummy [spindles > 0]	1,076	0.05	0.22	0.00	1.00
Log number of imported spindles	1,076	0.45	1.95	0.00	12.58
Panel E: Railroads, 1872 - 1950					
Dummy [RR $\leq 10Km$]	2,188	0.39	0.49	0.00	1.00
Distance to Railroad, in Km	2,188	105.41	186.04	0.00	1,228.91
Number Railroads Stations	2,188	4.40	10.02	0.00	146.00

Continues in the next page...

Table 1.1: Summary Statistics (cont.)

	Observations	Mean	S.D.	Min.	Max.
Panel F: Geography					
Longitude	547	-42.51	4.79	-57.09	-34.86
Latitude	547	-15.14	8.21	-32.56	-1.66
Log altitude	547	5.60	1.10	1.93	7.23
Log distance to Coast	547	4.08	1.74	0.02	6.54
Log distance to state's capital	547	4.73	1.13	0.00	6.54
Log slope	547	1.53	0.66	-0.79	2.89
Log area	547	7.84	1.20	4.60	11.70
% Cambisol	547	10.07	20.50	0.00	99.87
% Latosol	547	24.59	29.73	0.00	100.00
% Argisol	547	29.62	29.43	0.00	100.00
% Spondosol	547	12.57	19.75	0.00	85.72
% Others soils types	547	23.14	27.98	0.00	100.00
Panel G: Socioeconomic Characteristics, 1872					
% Literate (aged 6+)	547	0.18	0.10	0.02	0.69
% Foreigners	547	0.02	0.04	0.00	0.50
% Slaves	547	0.15	0.10	0.01	0.57
Public Administration (in 1,000)	547	0.96	1.83	0.00	31.63
Legal professions (in 1,000)	547	0.75	0.70	0.00	5.26
Panel H: Transportation Controls					
Log distance to port, 1850	547	4.83	1.17	-1.24	6.61
Log distance to road, 1867	547	2.16	1.53	0.00	5.27
Log distance to river	547	3.10	1.42	0.02	5.26

Notes: Descriptive statistics for the variables used in the paper. Occupational and population data from economic and population censuses. Manufacturing data in 1920 from the economic census. Data for price dispersion from historical reports. Imported textile machines variables built with data originally from Saxonhouse and Wright (2010). Indicator and distance for railway built from historical maps. Geographic controls created using ArcGIS with data originally from IBGE, CGIAR-CSI, and Embrapa. Controls for socioeconomic characteristics in 1872 from population census. Transportation controls created using ArcGIS with data originally from IBGE, ANA, and historical reports. In all panels the sample is based on the 1872, 1910, or 1920 municipality boundaries. For data specific descriptions and sources, see the Appendix A.

1.4 Empirical Model

1.4.1 Empirical Strategy

The objective of the article is to examine the effects of the expansion of the railway system on structural transformation and population between 1872 and 1950. In order to do so, and similar to Jedwab et al. (2017) and Fajgelbaum and Redding (2018), we explore a variation in the expansion of the railroad network across municipalities due to the proximity from cost-minimizing routes. More specifically, we estimate the following equation:

$$Y_{irt} = \alpha_i + \delta_{rt} + \beta Railroad_{irt} + X'_{ir}\eta_t + \epsilon_{irt} \tag{1-1}$$

Where Y_{irt} is a economic outcome in municipality i, in a region r, and year t. Our key variable of interest is $Railroad_{irt}$, an indicator variable that takes a value of 1 if the linear distance from the municipality i and region r, in year t, to the nearest railroad line is equal or less than 10 kilometers[24]. Since train stations are not necessarily located in downtown city and there is a margin of error in building the railroad network data, we choose to use the indicator variable than the continuous distance[25]. The term α_i represents municipality fixed effects, which absorb time-invariant municipality characteristics and conditions, such as climate. We also include region-by-year fixed effects, δ_{rt}, to control for regional common time trends, like economic and political cycles. Furthermore, we include a set of time-invariant geographic, baseline socioeconomic characteristics, and transport controls, X_{ir}, described in detail below, all of which interacted with time effects, η_t, to capture differential trends across municipalities.

Our parameter of interest is β. If the location of the railway lines were random, this parameter would capture the effect of the railway on the economic outcomes. However, as discussed in section 1.2, much of the railroad's expansion had the intention to connect inland productive areas to the coast's ports before the 1930s. Consequently, although we consider a series of controls in equation 1-1, β can be biased by the influence of unobservable confounding

[24]Our results are robust for a continuous specification of the railroad access. See section 1.5.3 for the results using the natural logarithm of the distance to the nearest railroad as interest variable and the number of railroad stations.

[25]Most of the railway papers use distance dummies as a variable of interest, see, for example, Pérez (2017), and Jedwab et al. (2017).

trends. For example, if railroad expansion responds to an unobservable increase in agriculture productivity, we should expect attenuation bias in our estimates due to reverse causality. Alternatively, after the 1930s most of the expansion of the railway network was conducted by the state to integrate the independents' regional systems and the states' capitals. Thus, after the 1930s we should expect a positive bias in our estimates due to endogeneity in the choice of the targeted cities.

To address these potential concerns, we complement the analysis with an instrumental variable strategy that exploits the features of the railway system expansion to generate exogenous variation in the probability that a municipality gets connected.

1.4.2 Identification using the Least-Cost Path

The expansion of the Brazilian rail network had two main objectives: (i) to connect productive land from the interior to the ports; (ii) connect local railway systems and state's capitals. Therefore, our instrument exploits the fact that some municipalities are likely to be connected to the railroad network just because they are located along the least-cost routes between the interior and the ports, and between the state's capitals. In other words, our instrumental variable predicts the propensity to be connected to the railroad network based on the construction of least-cost paths between each municipality and the ports that existed before the introduction of the railway in Brazil, and between the state's capitals.

To build the instrument, we implemented a procedure similar to that one outlined in Fajgelbaum and Redding (2018). In particular, we discretize Brazil into a raster of grid cells (0.5 Km x 0.5 Km) and calculate the least-cost paths between each municipality and the existing ports in 1850[26], and between the state's capitals[27]. We use data on slope degrees and rivers to calculate the costs associated with each cell, assuming a cost function increasing monotonically with slope, and a river crossing penalty[28]. Then, for each municipality, we

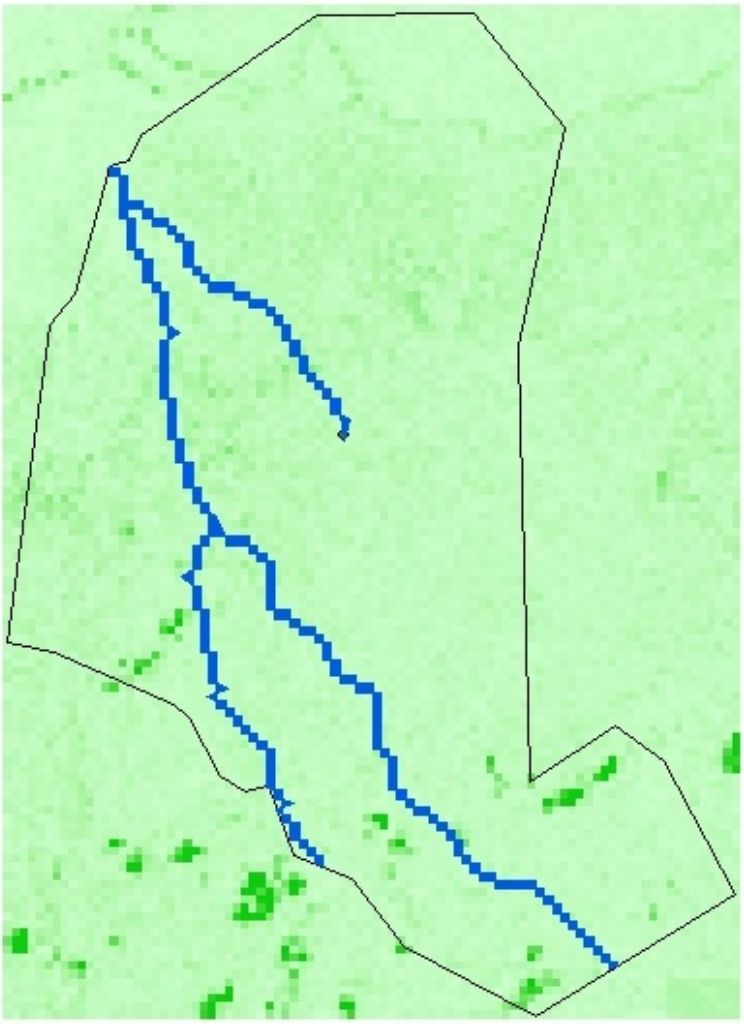
[26]The list of the most important ports in Brazil before the introduction of railways in the country is from Brasil (1850) and can be found at http://ddsnext.crl.edu/brazil. The list includes the following ports: Rio de Janeiro, Salvador, Recife, São Luiz, Rio Grande, Maceió, Santos, São José do Norte, Paranaguá, Aracajú, Fortaleza, Desterro, Porto Alegre, Aracaty, Parnayha, Vitória, Parnahyba, and Natal.

[27]More precisely, the integration of the regional railway systems and capitals occurred through the connection of Rio de Janeiro to the capitals of the other states from the southeast and northeast, and the connection of São Paulo to the southern system, crossing Itararé, Imbituva, Cruz Alta, reaching Santa Maria at Rio Grande do Sul (acts 10,432 from 1889, and 24,497 from 1934). Thus, we use these connections to build the least-cost routes.

[28]The penalty corresponds to an average slope of 1.7°, or 3%, the maximum slope allowed

compute the percentage of grid cells within its boundaries that lie along at least one of these least-cost paths. Figure 1.5 illustrates the construction of the percentage of grid cells situated along least-cost paths for the municipality of Caraúbas in the state of Rio Grande do Norte. The municipality corresponds to the area within the black border. The blue lines represent the least-cost paths that cross the municipality.

Figure 1.5: An Example of the Empirical Strategy: Municipality of Caraúbas, Rio Grande do Norte



Notes: The figure displays the municipality of Caraúbas (within the black border) in the state of Rio Grande do Norte. The blue cells represent the least-cost paths. The instrumental variable is the percentage of blue cells over the total number of green cells within the municipality interacted with the extension of the Brazilian railway network. See Appendix A for detailed information on the construction of the least-cost paths.

In this case, the percentage of blue cells over the total number of green cells within the border of the municipality represents the first component of our

in many late 19th-century railroad projects in Brazil.

instrument. Finally, since these percentages only vary over the municipality, we use the interaction between the propensity to receive railroads in the municipality i, given by the percentage of cells with the least-cost path, with the extension of the Brazilian railway network at year t as our instrumental variable. See Appendix A for detailed information on the construction of the least-cost paths.

Formally, the first-stage relationship for our second-stage equation 1-1 takes the following form:

$$Railroad_{irt} = \alpha_i + \delta_{rt} + \gamma LCP_{ir} * Extension_t + X'_{ir}\lambda_t + \nu_{irt} \tag{1-2}$$

Where LCP_{ir} is the percentage of grid cells within municipality i in the region r that lie along at least-cost paths between each municipality that existed in 1872 and the existing ports in 1850, and between the state's capitals. We interact the least-cost path with the total extension of the Brazilian railroad system, in kilometers, in year t, the term $Extension_t$. Therefore, the instrument exploits two sources of variation: (i) cross-sectional variation in the share of grid cells within the municipality that lie along least-cost paths (LCP_{ir}); (ii) time-series variation induced by changes in the total extension of the Brazilian railway system ($Extension_t$).

Since slope and river cover are crucial for calculating the least-cost routes, we control in all specifications for log municipality average slope and log distance to the nearest river. Additionally, to address the concern that larger municipalities are other things equal more likely to be along these least-cost paths, we control for log municipality land area. Finally, since ports and state capitals are targets for the construction of the least-cost paths, we also control in all specifications for log distance to the nearest port and state capital. The variable X_{ir} also includes controls to other geographic features, like, longitude, latitude, log altitude, log distance to coast, and percentage of the municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. To control for potential heterogeneity in initial levels of economic development, we control for baseline socioeconomic characteristics in 1872: (i) share of literate individuals aged 6 or more; (ii) share of foreign-born people; (iii) share of slaves; (iv) share of workers in public administration; (v) share of workers in legal professions. Since the impact of railways depends on pre-rail transport infrastructure, we also control for log distance to the nearest road in 1867, a proxy for the baseline transportation infrastructure.

Conditional on the municipality and region-by-year fixed effects, α_i and

δ_{rt}, and the controls X_{ir}, we expect that the expansion of the railroad network to be faster in localities crossed by more least-cost paths. In particular, municipality fixed-effects should absorb the confounding effects of the cross-sectional variation in the least-cost paths exposure, LCP_{ir}. Our instrument can be understood as a modified version of the shif-share instrument, where LCP_{ir} are the "shares" and $Extension_t$ is the "shift". Therefore, as shown by Goldsmith-Pinkham et al. (2018), the main identification assumption behind this type of instrument is that the shares are not correlated with other potential confounders in the baseline. In the next subsection we bring evidence that the shares of our instrument (the percentage of grid cells within municipality that lie along at least-cost paths) are not correlated with a set of social and economic characteristics in the baseline.

1.4.3
Least-Cost Paths and Expansion of Railways

Table 1.2 presents the first-stage results. We estimate standard errors clustered at the municipality level, to account for serial correlation within municipalities. All columns include controls for slope, distance to the nearest river, land area, distance to the nearest port, and states' capital. In column (1) we observe a positive and significant effect of the interaction term, $LCP_{ir} * Extension_t$, on the expansion of the railway with a KP F-statistic of 55.3. In columns (2)-(5) we add, gradually, the municipality and region-by-year fixed effects, the other geographic controls (longitude, latitude, log altitude, log distance to coast, and types of soil), the baseline characteristics, and the transport controls. In our most complete specification (column (5)), the impact of the interaction between the least-cost path and the length of the railway network on the likelihood of the municipality being connected to the railways remains positive and significant, with a KP F-statistic of 16.8. The magnitude and robustness of the coefficient do not vary much according to specifications. This indicates that the percentage of grid cells within each municipality that lie along at the least-cost paths is a strong predictor of the railway expansion over time, conditional upon municipality and year fixed-effects as well as on our full set of controls.

Table 1.2: First Stage Results, Panel 1872 - 1950

	Dummy for Railroad $\leq$ 10 Km.				
	(1)	(2)	(3)	(4)	(5)
$LCP_i * Extension_t$	0.289***	0.232***	0.206***	0.204***	0.204***
	[0.039]	[0.042]	[0.050]	[0.050]	[0.050]
Mean Dep. Var.	0.388	0.388	0.388	0.388	0.388
Observations	2,188	2,188	2,188	2,188	2,188
Municipalities	547	547	547	547	547
R^2	0.279	0.519	0.550	0.556	0.556
KP F-stat	55.318	30.813	17.264	16.823	16.874
Municipality FE	No	Yes	Yes	Yes	Yes
Region x Year FE	No	Yes	Yes	Yes	Yes
Geography x Year FE	No	No	Yes	Yes	Yes
Characteristics 1872 x Year FE	No	No	No	Yes	Yes
Transport x Year FE	No	No	No	No	Yes

Notes: This table reports first-stage results. All columns report the results from OLS regressions where the dependent variable is a dummy for the presence of a railroad line near the municipality, and the variable of interest is the percentage of grid points within each municipality i that lie on the least-cost paths interacted with the Brazilian railroad network extension in year t. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 547 municipalities based on the 1872 census boundaries. Standard errors are given in brackets and are clustered at the municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

As previously discussed, our main identification hypothesis is that the percentage of grid cells within the municipality that lie along at least-cost paths, or the term LCP_{ir} from our instrument, is not correlated with other characteristics of the municipalities in the baseline. Table 1.3 presents evidences that support our main identification assumption. In this table, we test whether the initial characteristics are related to the least-cost paths for the municipalities not yet connected to the railroad network in 1872. To do so, we regress at the cross-section the percentage of cells that lie on the least-cost routes, LCP_{ir}, on the baseline socioeconomic and transportation characteristics. The regressions are estimated for the 525 municipalities not connected to the railroad network based on the 1872 census boundaries. Each column is a separate regression where the LCP_{ir} is the dependent variable. All columns include controls for

slope, distance to the river, land area, distance to the port, states' capital, and region fixed effect. While in columns (1)-(9) we regress the LCP_{ir} in each variable separately, in column (10) we regress the least-cost path in all variables together. In all columns, we find results that are not statistically significant, even when we add all the variables simultaneously. The initial characteristics do not seem to be systematically related to the least-cost paths. Overall, the "shares" of our instrument are not correlated with the baseline economic and social structure, as well as the pre-railroad transportation infrastructure.

Table 1.3: Least-Cost Paths and Baseline Characteristics, 1872

	Least-Cost Path									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Log Population	0.005									0.018
	[0.071]									[0.075]
% Emp. agriculture		-0.365								-0.111
		[0.263]								[0.316]
% Emp. manufacturing			0.767							0.534
			[0.551]							[0.626]
% Literate rate				0.533						0.444
				[0.408]						[0.455]
% Foreigners					0.123					-0.381
					[1.262]					[1.369]
% Slaves						0.519				0.430
						[0.491]				[0.505]
Public Administration							0.034			0.022
							[0.028]			[0.029]
Legal professions								0.117		0.074
								[0.078]		[0.087]
Distance to Road									-0.030	-0.026
									[0.031]	[0.032]
Observations	525	525	525	525	525	525	525	525	525	525
Adjusted R^2	0.426	0.428	0.428	0.428	0.426	0.428	0.428	0.430	0.428	0.426

Notes: This table reports whether the baseline socioeconomic characteristics are systematically related to the least-cost paths for municipalities without railroad in 1872. All columns report the results from cross-section OLS regressions. Each column is a separate regression where we regress the percentage of grid points within each municipality that lie on the least-cost paths on the initial characteristics. In column (10) we regress the least-cost path on all baseline variables. All columns include controls for distance to the port, distance to states' capital, land area, slope, distance to the river, and region fixed effect. All regressions estimated for the 525 municipalities not connected to the railroad network based on the 1872 census boundaries. Standard errors are given in brackets and are clustered at the municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The evidence presented in this section indicates that our instrumental variable predicts the expansion of the railroad system between 1872 and 1950 and that the least-cost routes were balanced in terms of social, economic, and transportation characteristics in the baseline. This supports the idea that our instrumental variable can capture the causal effect of the railroad on employment structure and industrialization outcomes.

1.5 Main Results

In this section, we report the impacts of railroads on structural transformation, industrialization, and population dynamics. We also show that the results are robust to different connections to the railroad network measures. We complement the analysis examing the heterogeneity of the effects by distance to the nearest railroad line, year of connection to the railway network, and the number of railroad stations.

1.5.1 The Effects of Railways on Structural Transformation

We start the analysis by examining the effects of railroads' expansion on the structure of employment - the share of occupied workers in agriculture, manufacturing, and service sectors. First, we focus on the impact of the railroads just on the percentage of workers in the agriculture sector. Table 1.4, Panel A, reports the results from estimating the OLS regressions, based on equation 1-1. Panel B, otherwise, presents analogous results for our 2SLS specification, based on first-stage as in equation 1-2. All columns include controls for the determinants of the least-cost paths - slope, distance to the nearest river, land area, distance to the nearest port, and distance to nearest states' capitals - interacted with a year fixed effect. Throughout both panels, the second column adds municipality and region-by-year fixed effects. In column (3) we also control for other geographic conditions (longitude, latitude, log altitude, distance to the nearest coast, and types of soil) interacted with a year fixed effect. In columns (4) and (5) we add, respectively, controls for baseline socioeconomic characteristics and pre-railroad transportation infrastructure, both interacted with the year fixed effects. All regressions are estimated for the 547 municipalities based on the 1872 census boundaries, with standard errors clustered at the municipality level, to account for serial correlation within municipalities.

Table 1.4: Railroads and Structural Transformation, Panel 1872 - 1950

	Percentage of Workers in Agriculture				
	(1)	(2)	(3)	(4)	(5)
Panel A: OLS					
Dummy [RR $\leq$ 10 km]	-0.109***	-0.061***	-0.042***	-0.037**	-0.037**
	[0.011]	[0.015]	[0.015]	[0.015]	[0.015]
Panel B: 2SLS					
Dummy [RR $\leq$ 10 km]	-0.290***	-0.362***	-0.171*	-0.193**	-0.198**
	[0.055]	[0.094]	[0.101]	[0.095]	[0.095]
Mean Dep. Var.	0.707	0.707	0.707	0.707	0.707
Observations	2,188	2,188	2,188	2,188	2,188
Municipalities	547	547	547	547	547
Municipality FE	No	Yes	Yes	Yes	Yes
Region x Year FE	No	Yes	Yes	Yes	Yes
Geography x Year FE	No	No	Yes	Yes	Yes
Characteristics 1872 x Year FE	No	No	No	Yes	Yes
Transport x Year FE	No	No	No	No	Yes

Notes: This table reports the effects of railroads on the share of workers in the agriculture sector. All columns report the results from OLS (panel A) and 2SLS (panel B) regressions where the dependent variable is the share of workers in agriculture over the total number of occupied workers, and the variable of interest is a dummy for the presence of a railroad line near the municipality. For panel B, the instrumental variable is the percentage of grid points within each municipality i that lie on the least-cost paths interacted with the Brazilian railroad network extension in year t. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 547 municipalities based on the 1872 census boundaries. Standard errors are given in brackets and are clustered at the municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Panel A of Table 1.4 reports negative and significant OLS estimates, indicating a negative relationship between the railroad expansion and the share of occupied workers in agriculture. On Panel B, the 2SLS estimation, we observe that the coefficients remain negative, although larger point estimates. In all columns, the 2SLS estimated coefficients are significant at least at 10%. Overall, the comparison of OLS and 2SLS results suggests attenuation bias in OLS specifications. This is expected should the railroad expansion

respond to a non-observable increase in agriculture productivity, for example. In our most complete specification, column (5), the 2SLS estimated coefficient reports that the railroad impacts the employment structure decreasing the share of agriculture workers in almost 20 percentage points. This represents approximately 28% of the average share of agriculture workers between 1872 and 1950 (see Table 1.1).

Table 1.5 reports the 2SLS impact of the railroad on the percentage of workers in agriculture over time. In this alternative specification, we interact our main variable of interest with dummies of year. Thus, our coefficients capture the variation of the effects of railroads over time. Just like in the previous table, we just started controlling by the determinants of the least-cost paths, and gradually we added other controls. All columns present the 2SLS estimation, where we interact our instrument with year dummies. In the first two columns, we can see that the effect of railroads increases over time. In column (2), the impact of the railroads on the share of workers in agriculture is negative in 13 percentage points in 1920 relative 1872, 30 percentage points in 1940, and 38 percentage points in 1950. When we control for fixed effects, geography, baseline characteristics, and pre railway transportation infrastructure, column (5), there is a decrease in the point estimate, and the effect in 1920 became not statistically significant. However, the coefficients for 1940 and 1950 keep significant, ranging from -0.18 in 1940 and -0.22 in 1950. Summarising, Table 1.5 shows that the magnitude of the impact of the railroad on the employment structure increase over time, and becomes relevant, especially in the period when the industrialization took off in Brazil.

Table 1.5: The 2SLS Effects of Railroads on Structural Transformation By Year, Panel 1872 - 1950

	Percentage of Workers in Agriculture				
	(1)	(2)	(3)	(4)	(5)
Dummy [RR $\leq$ 10 km] X 1920	-0.134***	-0.114**	0.050	0.032	0.026
	[0.044]	[0.057]	[0.082]	[0.072]	[0.072]
Dummy [RR $\leq$ 10 km] X 1940	-0.307***	-0.300***	-0.151*	-0.175**	-0.180**
	[0.057]	[0.070]	[0.089]	[0.085]	[0.085]
Dummy [RR $\leq$ 10 km] X 1950	-0.383***	-0.364***	-0.206**	-0.226**	-0.229**
	[0.065]	[0.077]	[0.091]	[0.090]	[0.090]
Mean Dep. Var.	0.707	0.707	0.707	0.707	0.707
Observations	2,188	2,188	2,188	2,188	2,188
Municipalities	547	547	547	547	547
R^2	0.346	0.516	0.617	0.633	0.633
KP F-stat	17.715	15.672	7.960	7.957	8.059
Municipality FE	No	Yes	Yes	Yes	Yes
Region x Year FE	No	Yes	Yes	Yes	Yes
Geography x Year FE	No	No	Yes	Yes	Yes
Characteristics 1872 x Year FE	No	No	No	Yes	Yes
Transport x Year FE	No	No	No	No	Yes

Notes: This table reports the 2SLS effects of railroads on the share of workers in the agriculture sector by census year, using 1872 as the baseline year. All columns report the results from 2SLS regressions where the dependent variable is the share of workers in agriculture over the total number of occupied workers, and the variable of interest is a dummy for the presence of a railroad line near the municipality interacted with a dummy for year. The instrumental variable is the percentage of grid points within each municipality i that lie on the least-cost paths interacted with the Brazilian railroad network extension in year t and a dummy for year. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 547 municipalities based on the 1872 census boundaries. Standard errors are given in brackets and are clustered at the municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

As we have already shown, the railroad is associated with a decrease in the share of workers in the agriculture sector. However, did this reduction revert to growth for which sector? Table 1.6 reports OLS and 2SLS estimations based on equation 1-1 with the percentage of workers of agriculture (columns (1)-(2)), manufacturing (columns (3)-(4)), and service (columns (5)-(6)) as dependent

variables. In all specifications, we include all controls. The impact of the railroads on the agriculture is a decrease in almost 20 percentage points (p.p.) fo the share of workers, which 14 p.p. reverts in growth of the manufacturing sector (column(4)), and 6 p.p. in the service sector (column(6)), although for the latter the result is not statistically significant. Much of the reduction of the agriculture sector, therefore, was offset by the development of manufacturing.

Table 1.6: Railroads and Structural Transformation, Panel 1872 - 1950: Agriculture, Manufacturing, and Service

	Agriculture		Manufacturing		Service	
	(OLS)	(2SLS)	(OLS)	(2SLS)	(OLS)	(2SLS)
	(1)	(2)	(3)	(4)	(5)	(6)
Dummy [RR $\leq$ 10 km]	-0.037**	-0.198**	0.011	0.137**	0.026**	0.061
	[0.015]	[0.095]	[0.007]	[0.061]	[0.012]	[0.075]
Mean Dep. Var.	0.707	0.707	0.100	0.100	0.193	0.193
Observations	2,188	2,188	2,188	2,188	2,188	2,188
Municipalities	547	547	547	547	547	547
R^2	0.690	0.636	0.425	0.230	0.659	0.655
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Region x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Geography x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics 1872 x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Transport x Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the effects of railroads on the share of workers in the agriculture, manufacturing and service sector. All columns report the results from OLS or 2SLS regressions where the dependent variable is the share of workers in agriculture, manufacturing or service over the total number of occupied workers, and the variable of interest is a dummy for the presence of a railroad line near the municipality. The instrumental variable is the percentage of grid points within each municipality i that lie on the least-cost paths interacted with the Brazilian railroad network extension in year t. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 547 municipalities based on the 1872 census boundaries. Standard errors are given in brackets and are clustered at the municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The growth of the industrial sector may have occurred in three distinct ways: First, an expansion of the manufacturing at the extensive margin, with the increase in the number of the factories. Second, an expansion at the intensive margin, with the expansion of the manufacturing's firm sizes.

Third, a combination of the two previous effects. In Table 1.7 we show that the expansion of the manufacturing sector was in the extensive and intensive margins. Table 1.7 reports the cross-section OLS (Panel A) and 2SLS (Panel B) regressions, controlling for region fixed effects, geography, baseline characteristics, and transportation infrastructure. As we do not have manufacturing data from 1872 to 1950, we focus just in 1920. Here the instrumental variable is the log of the percentage of grid points within each municipality that lie along the least-cost paths, and the regressions are estimated for 1,076 municipalities based on the 1920 census boundaries. Looking at the effect on the probability to have at least one factory plant in 1920, column (1), we see that the impact of the railroad is positive and statistically significant. Thus, the expansion of the railroad system is associated with the increase in the probability of a municipality have factories. In column (2) we still analyze the extensive margin, but now with the log of the number of factories: municipalities connected to the railway system have a growth of more than 250% in the number of factories plants than those not connected. Finally, on column (3), we analyze the impact on the intensive margin. We conclude that the railroad's impact on the average number of workers in the manufacturing plants is positive. Municipalities connected to the railway system have a growth of more than 90% in the number of workers by factories than those not connected.

Table 1.7: Railroads and Manufacturing, 1920

Dependent variable:	Dummy [Factories>0]	Log factories (+1)	Log factories' size
	(1)	(2)	(3)
Panel A: OLS			
Dummy [RR $\leq$ 10 km]	0.244***	1.402***	0.098
	[0.036]	[0.167]	[0.101]
Panel B: 2SLS			
Dummy [RR $\leq$ 10 km]	0.407*	2.557**	0.905*
	[0.232]	[1.086]	[0.483]
Mean Dep. Var.	0.618	0.379	4.243
Observations	1,076	1,076	665
KP F-stat	19.124	19.124	29.213
Region FE	Yes	Yes	Yes
Geography	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes
Transportation	Yes	Yes	Yes

Notes: This table reports the effects of railroads on the number of manufacturing plants and its average number of workers in 1920. All columns report the results from OLS (panel A) and 2SLS (panel B) regressions where the dependent variable is defined at the top of each column. The variable of interest is a dummy for the presence of a railroad line near the municipality in 1920. For panel B, the instrumental variable is the log of the percentage of grid points within each municipality that lie on the least-cost paths. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the log of population, the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture and manufacturing computed over the total number of occupied workers. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 1,076 municipalities based on the 1920 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality boundaries. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In conclusion, our results show that the effect of the railroad expansion between 1872 and 1950 were a decrease in the share of workers in the agriculture sector and an increase in the manufacturing sector. The industrial sector growth has given both at the extensive and intensive margins. However, an issue still remains open. The structural transformation resulting from the expansion of the railway network can occur both by changes in the local occupational structure and by the migration from rural to urban areas (Pérez,

2017). In the next section, we analyze the impact of railways on population dynamics to understand whether the results on industrialization found so far are due to population displacement.

1.5.2 The Effects of Railways on Population

As we do not have migration data for the entire period analyzed, we use measures of population and population density to find out to what extent the observed structural change is a consequence of population agglomeration in urbanized areas. Table 1.8 presents the impact of the railroad expansion on the log of the number of people living in a municipality, and on the log of population density between 1872 and 1950. Panel A, reports the results from estimating the OLS regressions, based on equation 1-1, while Panel B presents analogous results for our 2SLS specification, based on first-stage as in equation 1-2. All columns include controls for the determinants of the least-cost paths-slope, distance to the nearest river, land area, distance to the nearest port, and distance to the nearest states' capitals. Throughout both panels, the columns (2) and (7) add municipality and region-by-year fixed effects. In column (3) and (8) we also control for other geographic conditions interacted with a year fixed effect. In columns (4) and (9) we add controls for baseline socioeconomic characteristics interacted with the year fixed effects. Finally, in columns (5) and (10) we include pre-railroad transportation infrastructure interacted with the year fixed effects. All regressions are estimated for the 547 municipalities based on the 1872 census boundaries, with standard errors clustered at the municipality level, to account for serial correlation within municipalities.

Despite detecting statistically significant effects in regressions where we do not include controls, in our preferred specification, columns (5) and (10), the estimated coefficients for log population and log population density are not statistically significant, for both the OLS and 2SLS estimates. Therefore, although the point estimates are positive, we cannot say that they are different from zero. We can conclude that between 1872 and 1950, the municipalities with railroads do not have a population growth compared to those not connected to the transportation system. The results indicate, therefore, that the internal movement of the population does not explain the decrease in importance of agriculture, the structural transformation seems to be due to the local change in the employment structure.

Table 1.8: Railroads and Population, Panel 1872 - 1950

Dependent variable:	Log Population					Log Population Density				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: OLS										
Dummy [RR $\leq$ 10 km]	0.402***	0.088*	0.061	0.062	0.062	0.885***	0.086*	0.069	0.071	0.071
	[0.063]	[0.048]	[0.050]	[0.049]	[0.049]	[0.066]	[0.049]	[0.050]	[0.049]	[0.049]
Panel B: 2SLS										
Dummy [RR $\leq$ 10 km]	-1.316***	-0.308	0.124	0.168	0.172	3.099***	-0.460	0.006	0.039	0.052
	[0.338]	[0.286]	[0.344]	[0.347]	[0.347]	[0.425]	[0.292]	[0.352]	[0.355]	[0.356]
Mean Dep. Var.	10.321	10.321	10.321	10.321	10.321	2.504	2.504	2.504	2.504	2.504
Observations	2,188	2,188	2,188	2,188	2,188	2,188	2,188	2,188	2,188	2,188
Municipalities	547	547	547	547	547	547	547	547	547	547
Municipality FE	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Region x Year FE	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Geography x Year FE	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Characteristics 1872 x Year FE	No	No	No	Yes	Yes	No	No	No	Yes	Yes
Transport x Year FE	No	No	No	No	Yes	No	No	No	No	Yes

Notes: This table reports the effects of railroads on the population and population density. All columns report the results from OLS (panel A) and 2SLS (panel B) regressions where the dependent variable is the log of the population (columns (1)-(5)) or the log of the population density (columns (6)-(10)). The variable of interest is a dummy for the presence of a railroad line near the municipality. For panel B, the instrumental variable is the percentage of grid points within each municipality i that lie on the least-cost paths interacted with the Brazilian railroad network extension in year t. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 547 municipalities based on the 1872 census boundaries. Standard errors are given in brackets and are clustered at the municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

We find no effects on the population even in the short term. Table 1.9 reports the 2SLS impact of the railroad on population and population density over time. In this specification, we interact our main variable of interest, as well as our instrumental variable, with dummies of year. Thus, our coefficients capture the variation of the effects of railroads over the census year. In the specifications where we do not include controls, we find negative effects on population and, however, positive effects on population density. In both cases, the estimated coefficients are statistically significant. In our preferred specification, columns (5) and (10), although the positive and significant OLS coefficients (Panel A), the 2SLS estimates are not statistically significant. Therefore, we find no effects on the population even in the short term, 1920, as well as for the medium term, 1950.

Table 1.9: The 2SLS Effects of Railroads on Population
By Year, Panel 1872 - 1950

Dependent variable:	Log Population					Log Population Density				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Dummy [RR ≤ 10 km] X 1920	-1.235***	-0.308*	-0.238	-0.180	-0.167	2.910***	-0.401*	-0.416	-0.379	-0.350
	[0.313]	[0.186]	[0.269]	[0.249]	[0.249]	[0.391]	[0.218]	[0.319]	[0.307]	[0.307]
Dummy [RR ≤ 10 km] X 1940	-1.362***	-0.326	0.055	0.087	0.089	2.936***	-0.483*	-0.067	-0.037	-0.030
	[0.343]	[0.244]	[0.315]	[0.321]	[0.320]	[0.420]	[0.253]	[0.325]	[0.334]	[0.333]
Dummy [RR ≤ 10 km] X 1950	-1.255***	-0.185	0.250	0.289	0.293	3.154***	-0.299	0.182	0.210	0.220
	[0.343]	[0.257]	[0.322]	[0.330]	[0.331]	[0.438]	[0.260]	[0.329]	[0.337]	[0.339]
Mean Dep. Var.	10.321	10.321	10.321	10.321	10.321	2.504	2.504	2.504	2.504	2.504
Observations	2,188	2,188	2,188	2,188	2,188	2,188	2,188	2,188	2,188	2,188
Municipalities	547	547	547	547	547	547	547	547	547	547
KP F-stat	17.715	15.672	7.960	7.957	8.059	17.715	15.672	7.960	7.957	8.059
Municipality FE	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Region x Year FE	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Geography x Year FE	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Characteristics 1872 x Year FE	No	No	No	Yes	Yes	No	No	No	Yes	Yes
Transport x Year FE	No	No	No	No	Yes	No	No	No	No	Yes

Notes: This table reports the 2SLS effects of railroads on the population and population density by census year. All columns report the results from 2SLS regressions where the dependent variable is the log of the population (columns (1)-(5)) or the log of the population density (columns (6)-(10)). The variable of interest is a dummy for the presence of a railroad line near the municipality. The instrumental variable is the percentage of grid points within each municipality i that lie on the least-cost paths interacted with the Brazilian railroad network extension in year t. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 547 municipalities based on the 1872 census boundaries. Standard errors are given in brackets and are clustered at the municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Summarising, the results show that the expansion of railroads in Brazil between the 1870s and 1950s is associated with the shift of workers from agriculture to manufacturing. In particular, this effect seems to be due to a local structural change, and not due to population agglomeration in urban regions. The connected municipalities do not become poles of attraction for internal migrants, at least until the 1950s. The migration and agglomeration effects seem to be limited to explain the process of structural change, in this regard. In section 1.6 we explore two mechanisms that can be the drivers of this economic transition process.

1.5.3
Alternative Measures

Our main variable of interest is an indicator that takes value 1 if the municipality downtown is within 10 kilometers of a railroad line. However, as we can see in Table 1.10, our results are robust using the continuous distance and number of railway stations as interest variables. The table reports the 2SLS effects of the railways on the share of agriculture workers, log population and

log population density. In all columns, we report the results for the specification with all controls. In column (1), an increase in 100% in the distance to the railroad line is associated with an increase in 0.06 percentage points on the share of agriculture workers over the total number of occupied workers. Also, in column (2), we show that an increase in the number of railroad stations is associated with a decrease in the share of agriculture workers. However, we do not find statistically significant effects of the distance of the railway and the number of railroad stations on population and population density. Overall, our results are robust to these different measures of railway connection.

Table 1.10: Railroads, Structural Transformation, and Population Alternative Measures (2SLS), Panel 1872 - 1950

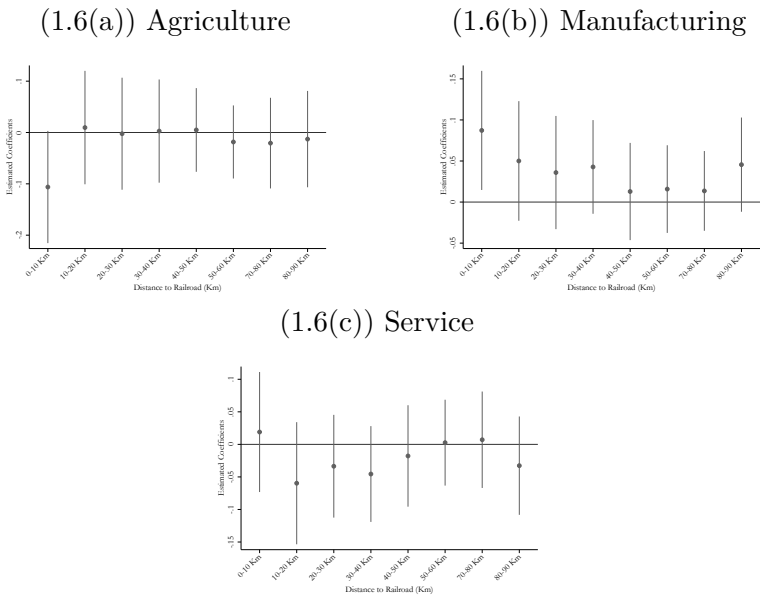
	% Emp. Agr.		Log Population		Log Population Density	
	(1)	(2)	(3)	(4)	(5)	(6)
Log Distance RR	0.059**		-0.051		-0.016	
	[0.030]		[0.103]		[0.106]	
Log RR stations		-0.109**		0.094		0.029
		[0.050]		[0.187]		[0.195]
Mean Dep. Var.	0.707	0.707	10.321	10.321	2.504	2.504
Observations	2,188	2,188	2,188	2,188	2,188	2,188
Municipalities	547	547	547	547	547	547
R^2	0.605	0.668	0.802	0.812	0.767	0.769
KP F-stat	14.152	14.984	14.152	14.984	14.152	14.984
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Region x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Geography x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics 1872 x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Transport x Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the effects of railroads on structural transformation, and population. All columns report the results from 2SLS regressions where the dependent variable is the share of workers in agriculture over the total number of occupied workers (columns (1)-(2)), log of the population (columns (3)-(4)), or the log of the population density (columns (5)-(6)). In columns (1), (3), and (5) the log of the distance from the municipality to the nearest railroad line. In columns (2), (4), and (6) the variable of interest is the log of the number of railways stations. The instrumental variable is the percentage of grid points within each municipality i that lie on the least-cost paths interacted with the Brazilian railroad network extension in year t. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 547 municipalities based on the 1872 census boundaries. Standard errors are given in brackets and are clustered at the municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

1.5.4
Heterogeneity by Distance, Time, and Number of Stations

In the previous section, we show that railways are associated with the reduction of the agricultural sector and the growth of the industrial sector. However, does the impact reverberate in the cities around the municipalities connected to the rail system? In other words, is there any heterogeneity by distance from railway lines? Figure 1.6 reports the 2SLS estimation where we include dummy variables for 10 km bins of distance from the railroad network as interest variables and interact our instrument with the same dummies, in the specification where we include all controls. Therefore, each coefficient represents the effect of the railroad for the municipalities at a certain distance from the railway line about municipalities distant 90 km or more from the same line.

Figure 1.6: The 2SLS Effects of Railroads on Structural Transformation, Panel 1872-1950:
Heterogeneity by Distance



Notes: The figures show the heterogeneity of the impact of railroads on structural transformation by distance to the nearest railroad line. The figures display a modified version of our 2SLS estimation where we include separate dummy variables for 10 km bins of distance to the railroad network as interest variables and interact our instrument with dummies of distance to the railroad, in the specification where we include all controls. The dependent variable is the share of workers in (a) agriculture/(b) manufacturing/(c) service over the total number of occupied workers. All regressions estimated for the 547 municipalities based on the 1872 census boundaries. Standard errors are clustered at the municipality level.

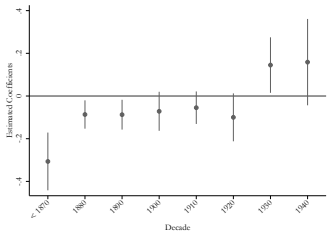
The effect of railways on structural change is concentrated only in munic-

ipalities within 10 km from a railway line. There is no significant impact on municipalities nearby connected areas.

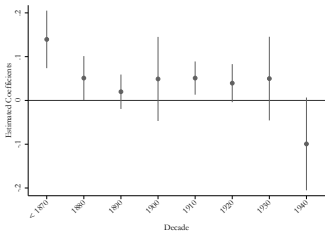
Figure 1.7 reports the heterogeneous effect of railways on structural change by connection year at the railroad network. The figures display a modified version of the 2SLS estimation where we interact the variable of interest and the instrument with dummies of connection year to the railroad system. As we can see, the impact of the railroad is higher the sooner the municipality was connected to the railway system. The effect is higher for municipalities connected before the 1920s and is not statistically significant for those connected just in the 1940s.

Figure 1.7: The 2SLS Effects of Railroads on Structural Transformation , Panel 1872-1950:
Heterogeneity by Time

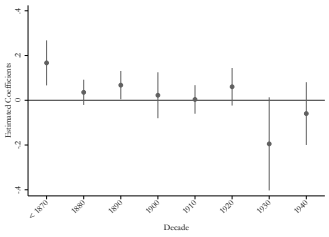
(1.7(a)) Agriculture



(1.7(b)) Manufacturing



(1.7(c)) Service



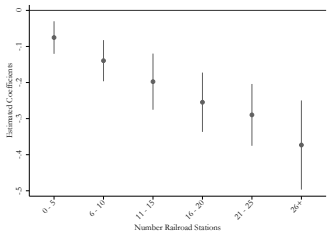
Notes: The figures show the heterogeneity of the impact of railroads on structural transformation by the time of connection to the railroad network. The figures display a modified version of our 2SLS estimation where we include the interaction between our dummy of railroad nearby 10 Km with dummy variables for the year of connection to the railroad network as interest variables and interact our instrument with dummies of year, in the specification where we include all controls. The dependent variable is the share of workers in (a) agriculture/(b) manufacturing/(c) service over the total number of occupied workers. All regressions estimated for the 547 municipalities based on the 1872 census boundaries. Standard errors are clustered at the municipality level.

Figure 1.8 reports the effect of railways on structural change in the intensive margin. Again, the figures display a modified version of the 2SLS estimation where we include dummy variables for the number of railroad stations as interest variables and interact our instrument with the same

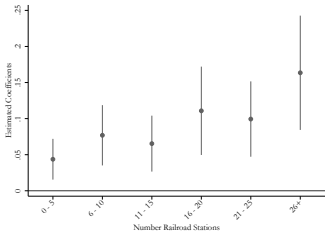
dummies. The results indicate that the magnitude of the impact is a linear function of the number of stations. Therefore, there is variability in the magnitude of the impact in the intensive margin. Although on average, we do not find significant results of the railroads on the service sector, these effects are significant for municipalities with more than six railway stations.

Figure 1.8: The 2SLS Effects of Railroads on Structural Transformation , Panel 1872-1950:
Heterogeneity by Stations

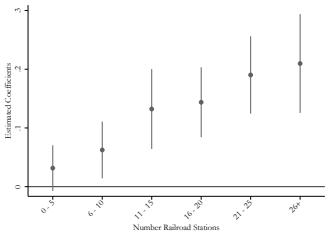
(1.8(a)) Agriculture



(1.8(b)) Manufacturing



(1.8(c)) Service



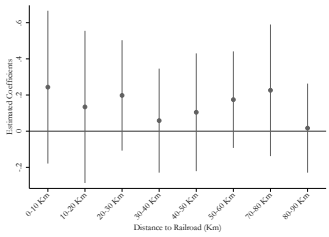
Notes: The figures show the heterogeneity of the impact of railroads on structural transformation by the number of railroad stations in each municipality. The figures display a modified version of our 2SLS estimation where we include separate dummy variables for the number of railroad stations as interest variables and interact our instrument with dummies number of railroad stations, in the specification where we include all controls. The dependent variable is the share of workers in (a) agriculture/(b) manufacturing/(c) service over the total number of occupied workers. All regressions estimated for the 547 municipalities based on the 1872 census boundaries. Standard errors are clustered at the municipality level.

In Figure 1.9 we do the same exercise as the previous figures, however, for the population. Our objective is to verify if the agglomeration effects are present for the municipalities that were more distant from railways, were connected at the beginning of the railway expansion, or for those that have more stations. In the three graphs, the dependent variable is the log of the population. We find no heterogeneity in the effect by distance or year of connection. In both cases, railways have no impact on population size, regardless of the distance and the time of connection to the railroad system. However, we find heterogeneity by the number of railroad stations. Although

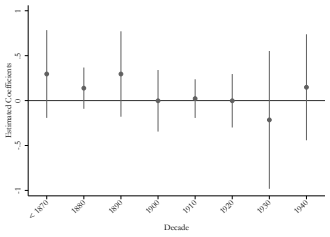
we do not find the effects of railways on the population, this impact becomes statistically significant for municipalities with more than six railroad stations, and its magnitude grows monotonically with the number of stations.

Figure 1.9: The 2SLS Effects of Railroads on Population, Panel 1872-1950: Heterogeneity by Distance, Time, and Stations

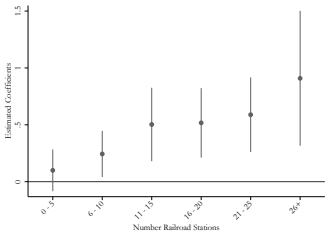
(1.9(a)) Distance



(1.9(b)) Time



(1.9(c)) Stations



Notes: The figures show the heterogeneity of the impact of railroads on log population by distance to the nearest railroad line (a), time to connection to the railroad network (b), and number of railroad stations in each municipality (c). The figures display a modified version of our 2SLS estimation where we include separate dummy variables for 10 km bins of distance to the railroad network (a), year of connection to the railroad network (b), or number of railroad stations (c) as interest variables, and interact our instrument with these dummies, in the specification where we include all controls. The dependent variable is log of the population. All regressions estimated for the 547 municipalities based on the 1872 census boundaries. Standard errors are clustered at the municipality level.

The results presented above show that there are considerable heterogeneities in the impact of railways, especially on structural transformation. Moreover, the heterogeneities on population effects indicate that the immigration and agglomeration effects only occur for municipalities with more than six railway stations.

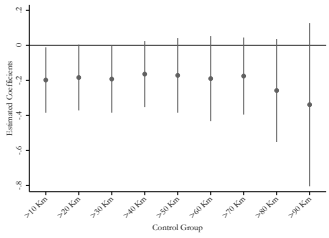
1.5.5 Spatial Reallocation

Are our results driven by a real change in the employment structure or its consequence of the reallocation of the economic activity across locations? In Figures 1.10 and 1.11 we test whether the spatial reallocation is important.

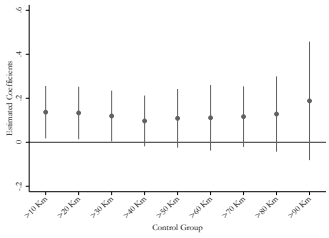
Both figures display a modified version of our 2SLS estimation where we gradually shift the control group in 10 km, in the specification where we include all controls. While in Figure 1.10 the dependent variables are employment structure, in Figure 1.11 are the population measures.

Figure 1.10: The 2SLS Effects of Railroads on Structural Transformation , Panel 1872-1950:
Spatial Reallocation?

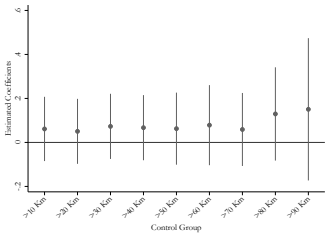
(1.10(a)) Agriculture



(1.10(b)) Manufacturing



(1.10(c)) Service



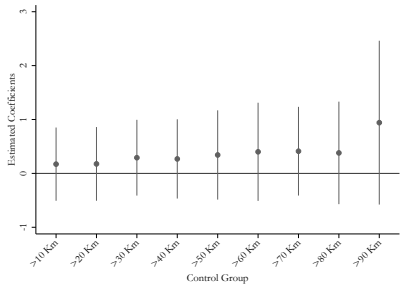
Notes: The figures show the impact of railroads on structural transformation changing the control group. The figures display a modified version of our 2SLS estimation where we gradually shift the control group in 10 Km, in the specification where we include all controls. from the nearest railroad line. The dependent variable is the share of workers in (a) agriculture/(b) manufacturing/(c) service over the total number of occupied workers. All regressions estimated for the 547 municipalities based on the 1872 census boundaries. Standard errors are clustered at the municipality level.

If spatial reallocation matters, the magnitude of the effects should decrease as the control group is shifted to cities farther from the railway network (Berger, 2019; Berger and Enflo, 2017; Redding and Turner, 2015). For example, if the impact of the railways is higher when the control group is the municipalities that are more than 10 km from a line than when the control group is the municipalities that are more than 20 km from the same line, this means that there is a reallocation of the economic activity from the municipalities that are between 10 and 20 km from the railway to those less than 10 km from the same railway. For both occupational structure and population, we find no evidence of spatial relocation. Although we change control groups, variations in the effect magnitudes are at most limited. Thus the reduction in the agricultural sector and the increase in manufacturing that

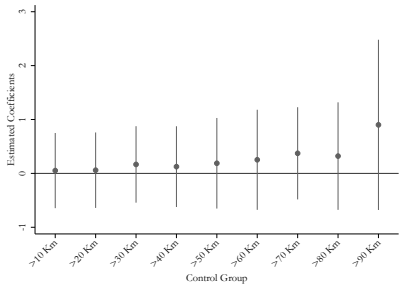
we find seem not to be coming from the spatial reorganization of economic activity but the establishment of new economic activities. In the same way, the null effects found for the population do not come from a spatial relocation of the people to areas surrounding the municipalities connected by the railways.

Figure 1.11: The 2SLS Effects of Railroads on Population , Panel 1872-1950: Spatial Reallocation?

(1.11(a)) Log Population



(1.11(b)) Log Population Density



Notes: The figures show the impact of railroads on population changing the control group. The figures display a modified version of our 2SLS estimation where we gradually shift the control group in 10 Km. from the nearest railroad line, in the specification where we include all controls. The dependent variable is the log of population (a) or log of the population density (b). All regressions estimated for the 547 municipalities based on the 1872 census boundaries. Standard errors are clustered at the municipality level.

1.6 Mechanisms

Municipalities connected to the railroad network had more workers in the manufacturing sector, and less in the agriculture than those municipalities not connected between 1872 and 1950. However, this change in occupational structure does not appear to have come from immigration and agglomeration effects. In this section, we discuss two mechanisms through which railroads might have resulted in this structural transformation: market integration and technology adoption.

1.6.1 Market Integration

One of the mechanisms by which railways can impact the shift from agriculture to manufacturing is market integration. The decrease in transportation costs might have allowed municipalities with low agricultural productivity to import such products from neighboring locations, decreasing the number of agricultural workers. Pérez (2017), for example, find that the railroads resulted in a reduction in the local demand for agricultural labor in connected districts from 19th-century Argentina, and most of the transition out of farming took place in agricultural unproductive districts. Therefore, in this subsection, we explore whether railways have contributed to market integration, and consequently, reduced price dispersion among Brazilian municipalities.[29]

Following Donaldson (2018) and Andrabi and Kuehlwein (2010), we use price dispersion, as a proxy for market integration. Therefore our dependent variable is the absolute value of the log price difference between municipalities: $P_{ij} = |log(P_i) - log(P_j)|$ for all $i \neq j$. Since our price information is a cross-section data covering some municipalities in 1910[30], our estimating equation is the following:

$$P_{ij,1910} = \alpha + \beta BothRR_{ij,1910} + \sigma D_{ij} + X'_{ij}\eta + \lambda_r + \epsilon_{ij} \tag{1-3}$$

Where $BothRR_{ij,1910}$ is an indicator variable that takes value 1 for both municipalities i and j within 10 kilometers from a railroad line, and 0 otherwise. The term D_{ij} is the log distance between the municipalities i and j, a proxy to

[29]The expansion of the railroads comes along with the convergence of prices in Brazil. Comparing the dispersion of coffee prices among the municipalities of the state of São Paulo, Summerhill (2003) finds a reduction of 48% between 1854 and 1906.

[30]The original data were collected between 1910 and 1913. See Appendix A for details

the trade costs. Since the distance impacts the price dispersion, we would like to compare a pair of municipalities with similar distances. Since we just have a cross-section data, we can't control for municipalities pair and time fixed effects. Thus, to control for unobservables confounding factors we include a set of geographic, socioeconomics, and transportation controls, X_{ij}. All controls are indicator variables that take values 1 if both municipalities i and j are in the first quartile of the control variable distribution. For example, to control for distance to the nearest port, we include a dummy variable that takes value 1 if both municipalities are in the first quartile of the distance to the port distribution. The set of controls is similar to those included in the specifications from Section 1.5. Finally, λ_r is a region fixed effect.

However, even controlling for this set of controls, there may be unobserved characteristics that affect the propensity of the pair of municipalities to be connected to the railroad network, and that also affect price dispersion. Therefore, we use the least-cost path to instrument $Railroad_{ij,1910}$. The first-stage for our equation 1-3 is the following equation:

$$BothRR_{ij,1910} = \alpha + \beta MinLCP_{ij,1910} + \psi D_{ij} + X'_{ij}\phi + \tau_r + \nu_{ij} \qquad (1\text{-}4)$$

Where $MinLCP_{ij,1910} = min\{LCP_i, LCP_j\}$ for LCP_i, LCP_j - the percentage of grid cells within each municipality that lie on the least-cost paths, for i and j. In other words, the instrumental variable is the minimum percentage of grid cells within each municipality that lie on the least-cost paths between the municipalities i and j. Our identification hypothesis is that conditional on the controls, the least-cost paths just affect the price dispersion by the expansion of the railway system. As shown in section 1.4, there is strong evidence that this hypothesis is valid.

First, we analyze the impact of the railroads on the price dispersion for corn. Table 1.11, Panel A, reports the results from estimating the OLS regressions, based on equation 1-3. Panel B, otherwise, presents analogous results for our 2SLS specification, based on first-stage as in equation 1-4. All columns include controls for the determinants of the least-cost paths - indicator variable for slope, distance to the nearest river, land area, distance to the nearest port, and distance to the nearest states' capitals, as well the log distance between the municipalities i and j.

Table 1.11: Mechanism - Railroads and Corn Market Integration, 1910

	Corn Price Dispersion				
	(1)	(2)	(3)	(4)	(5)
Panel A: OLS					
Both [RR $\leq$ 10 km]	-0.139***	-0.139***	-0.137***	-0.138***	-0.138***
	[0.010]	[0.010]	[0.010]	[0.010]	[0.010]
Panel B: 2SLS					
Both [RR $\leq$ 10 km]	-0.264***	-0.278***	-0.209***	-0.218***	-0.217***
	[0.069]	[0.067]	[0.064]	[0.064]	[0.065]
Mean Dep. Var.	0.497	0.497	0.497	0.497	0.497
Observations	115,921	115,921	115,921	115,921	115,921
Municipalities	482	482	482	482	482
KP F-stat	90.220	100.485	101.589	101.504	101.426
Geography	No	Yes	Yes	Yes	Yes
Characteristics 1872	No	No	Yes	Yes	Yes
Transport	No	No	No	Yes	Yes
Region F.E.	No	No	No	No	Yes

Notes: This table reports the effects of railroads on corn price dispersion. All columns report the results from OLS (panel A) and 2SLS (panel B) regressions where the dependent variable is the corn price difference, in log, between all combinations of the 481 municipalities of the sample: $|log(CornPrice)_i - log(CornPrice)_j|$ for all i, j with $i \neq j$. The variable of interest is a dummy for the presence of a railroad line near both municipalities i and j in 1910. For panel B, the instrumental variable is the minimum percentage of grid points within each municipality that lie on the least-cost paths between the pair of municipalities. All columns include controls for distance to the port, distance to states' capital, land area, slope, distance to the river, and distance between the municipalities. Geographic controls include altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the population, share of the literate population, percentage of foreigners, percentage of slaves, the total number of workers in public administration and legal professions relative to the total population. The transportation control is the distance to the nearest road in 1867. All controls are an indicator variables built for the pair of municipalities. All regression is estimated for the municipalities with price information in 1910, based on the 1910 boundaries. Standard errors are given in brackets and are clustered at the destination municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Throughout both panels, the second column adds other geographic controls - dummy variables for altitude, distance to the nearest coast, and soil for the pair of municipalities. In column (3), we also control for the baseline 1872 socioeconomic characteristics, dummies for population, literate rate, the

share of foreigners, the share of slaves, workers in public administration, and legal professions relative to the total population. In column (4), we add the transportation control, an indicator variable for the presence of a road. Finally, in column (5) we control for region fixed effect. All regression is estimated for the 482 municipalities with corn price information in 1910, based on the 1910 boundaries. The final dataset is a pair combination of the 482 municipalities. The standard errors are clustered at the destination municipality level.

In all specifications our KP F-statistics is greater than 10, indicating a strong instrument. Throughout the five columns, the OLS coefficients are negative and significant, the railroad is associated with a decrease in the price dispersion in almost 14%. For the 2SLS estimates, the magnitudes of the coefficients increase. In our preferred specification, column (5), the expansion of the railroad impacted on the decline of the corn price dispersion in almost 22%. Our estimates are larger than those found for XIX and XX centuries British India (Andrabi and Kuehlwein, 2010) and XIX century Europe (Keller and Shiue, 2008), which may be justified by the poor pre-railway transportation infrastructure in Brazil.

Table 1.12 reports the 2SLS results for other products, beyond corn: bean, liquor, and flour. In all specifications, we control for the geographic, baseline socioeconomic, transportation controls, and region fixed effects. Likewise corn, the railroad expansion impact on the decrease of the price dispersion for bean, liquor, and flour. Connected municipalities pairs have, on average, a price difference 26%, 37%, and 30% lower than those not connected for, respectively, bean, liquor, and flour.

Table 1.12: Mechanism - The 2SLS Effects of Railroads on Market Integration, 1910

	Price Dispersion			
	Corn	Bean	Liquor	Flour
Both [RR $\leq$ 10 km]	-0.217***	-0.257***	-0.369***	-0.297**
	[0.065]	[0.091]	[0.100]	[0.133]
Mean Dep. Var.	0.497	0.568	0.521	0.528
Observations	115,921	125,751	169,071	152,628
Municipalities	482	502	582	553
R^2	0.039	0.012	0.071	0.027
KP F-stat	101.426	95.558	87.308	84.992
Geography	Yes	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes	Yes
Transport	Yes	Yes	Yes	Yes
Region F.E.	Yes	Yes	Yes	Yes

Notes: This table reports the effects of railroads on price dispersion for differents products. All columns report the results from 2SLS regressions where the dependent variable is the price difference, in log, between all combinations of the municipalities of the sample: $|log(Price)_{ik} - log(Price)_{jk}|$ for all municipalities i, j with $i \neq j$, and product k. The variable of interest is a dummy for the presence of a railroad line near both municipalities i and j in 1910. The instrumental variable is the minimum percentage of grid points within each municipality that lie on the least-cost paths between the pair of municipalities. All columns include controls for distance to the port, distance to states' capital, land area, slope, distance to the river, and distance between the municipalities. Geographic controls include altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the population, share of the literate population, percentage of foreigners, percentage of slaves, the total number of workers in public administration and legal professions relative to the total population. The transportation control is the distance to the nearest road in 1867. All controls are an indicator variables built for the pair of municipalities. All regression is estimated for the municipalities with price information in 1910, based on the 1910 boundaries. Standard errors are given in brackets and are clustered at the destination municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

1.6.2 Technology Adoption

The adoption of new technologies and equipment by Brazilian industries in the late nineteenth and early twentieth depended heavily on transportation infrastructure. Birchal (1999, p. 166) describes how difficult was the transport of imported machines from the port of Rio de Janeiro to factories located in the state of Minas Gerais:

"The most affected industries were those that employed technologies embodied in the form of equipment and machinery. Due to the lack of roads, the transport of textile machinery from the port of Rio de Janeiro to the hinterlands of Minas Gerais was full of adventures and obstacles. Machines had to be carried on the back of beasts of burden. Roads were rough and sometimes bridges had to be built along the way."

The expansion of the railroad system has considerably reduced the difficulty of importing machinery by industries. According to Birchal (1999, p. 167):

"Towards the end of the nineteenth century, the problem of transporting machinery became less acute as the railway was penetrating farther and farther into the hinterlands of Minas Gerais."

We use textile machinery imports data between 1878 and 1933 from Saxonhouse and Wright (2010) to test the impact of the railroads on the imports of new textile machines, and technology by the Brazilian factories. The adoption of new technologies, and the consequent increase in productivity, can explain the impact of railways on the increased participation of the industrial sector in the economy between 1872 and 1950.

Table 1.13 reports the impact of the railroads on the accumulated number of imported textile machines between 1878 and 1933. Our specification is an adaptation to the model presented in section 1.4 for the cross-section data. We control for region fixed effects, geography, baseline characteristics, and transportation infrastructure. While the variable of interest is a dummy for railway connection in 1930, the instrumental variable is the log of the percentage of grid points within each municipality that lies along the least-cost paths, and the regressions are estimated for 1,076 municipalities based on the 1920 census boundaries. In the first two columns, the dependent variable is an indicator variable that takes value 1 if the municipality imported, at least, one textile machine spindle from the Britsh exporters. For the last two columns, the dependent variable is the log of the number of imported textile machines spindles. As we can see in the 2SLS estimated coefficients, the expansion of the railroad impact the probability of the connected municipality import

textile machines, as well the number of textile machinery spindles imported: Municipalities connected to the railway system imported, on average, 200% more spindles than those not connected. Both estimated coefficients are significant at 5%. The results confirm the historical evidence for the importance of railways on the adoption of new technologies.

Table 1.13: Mechanism - Railroads and Textile Machinery Imports Between 1878 and 1933

Dependent variable:	D. [imported spindles > 0]		Log imported spindles (+1)	
	(OLS)	(2SLS)	(OLS)	(2SLS)
Dummy [RR ≤ 10 km]	0.009	0.232**	0.062	1.951**
	[0.013]	[0.108]	[0.111]	[0.923]
Mean Dep. Var.	0.052	0.052	0.450	0.450
Observations	1,076	1,076	1,076	1,076
R^2	0.226	0.042	0.249	0.074
KP F-stat		24.347		24.347
Region FE	Yes	Yes	Yes	Yes
Geography	Yes	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes	Yes
Transport	Yes	Yes	Yes	Yes

Notes: This table reports the effects of railroads on the number of imported textile machines between 1878 and 1933. All columns report the results from OLS ((1) and (3)) or 2SLS ((2) and (4)) regressions where the dependent variable is defined at the top of each column. The variable of interest is a dummy for the presence of a railroad line near the municipality in 1930. The instrumental variable is the log of the percentage of grid points within each municipality that lie on the least-cost paths. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the log of population, the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture and manufacturing computed over the total number of occupied workers. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 1,076 municipalities based on the 1920 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality boundaries. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

1.7 Conclusion

Between the late nineteenth and middle twentieth centuries, the Brazilian economy underwent a major process of transformation, with the growth of industrial labor forces, and the reduction of the participation of agriculture in the economy. The shift of workers from agriculture to manufacturing is a common factor in modern economic growth. In this article, we analyze how the expansion of transportation infrastructure affects the occupational transition process in a country with a continental dimension and high pre-railroad transportation costs. We exploit a key feature of the Brazilian railway expansion to document the effects of transportation on structural transformation: The railroad system's expansion goal, which was to connect the ports to the interior and the state capitals, generated variability in the propensity of municipalities to connect to the system solely because of their geographical position.

We follow almost 80 years of Brazilian railroad expansion to identify the impact of the transportation infrastructure on the structural transformation between the 19th and 20th centuries. Also, we explore the mechanisms by which transportation costs reduction can impact the occupational structure of the economy. To do so, we used data from 1872, 1920, 1940, and 1950 Brazilian censuses, as well as prices and imported machinery data. We document that railway expansion decreased the share of agriculture workers in almost 20 percentage points between 1872 and 1950. Much of this impact occurred when Brazilian industrialization took off, in the 1940s. This reduction in the proportion of agricultural workers came with a growth in the percentage of manufacturing workers, in almost 14 percentage points. We do not find statistically significant impacts on the service sector. The industrial growth took place on the extensive margin, with an increase in the number of manufacturing factories, as well in the intensive margin, with the increment in the number of industrial workers by factories. Despite this change in the occupational structure of the economy, we find no effects of railways on population and population density, perhaps an evidence that the results are not a consequence of immigration and agglomeration effects.

We also show that there is heterogeneity on the railways' impacts: the magnitude of the impact increases with nearness, the time of connection, and the number of stations. Finally, we explore two mechanisms: market integration and technology adoption. The railway expansion reduced the price dispersion and increased the number of textile machinery imported from Britsh exporters. The results of this paper indicate that transportation costs are

important to the mobility out of farming from the workers. Moreover, market integration and technology adoption seem to be more relevant to explain the process of structural change than immigration and agglomeration effects. In particular, our article differs from those papers that show that the structural change was due both to the change in the local occupational structure and to the internal migration from agricultural areas to urban centers (Berger, 2019; Fajgelbaum and Redding, 2018; Yamasaki, 2017; Pérez, 2017), since we do not find agglomeration effects for Brazil between 1872 and 1950. The article also contributes to the classical discussion about the Brazilian historical industrialization, in which the role of railways has been minimized given its export-oriented character. We show that the impact of railways is greater than the simple connection between inland productive areas and the international market, the railways contributed to the integration of the Brazilian internal market and changed its structure.

2
The Persistence Paths: Railways and Economic Development in Brazil 1950-2010

2.1
Introduction

Although the importance of history on economic development is well established[1], a large debate remains about the influence of history on the spatial equilibrium of economic activity. Some articles bring evidence of the recovery capacity of cities in the face of adverse negative shocks, such as the bombing in war time (Davis and Weinstein, 2002, 2008; Miguel and Roland, 2011), which can be interpreted as evidence of the existence of only one spatial equilibrium. On the other hand, others show that temporary historical shocks can have persistent long-term effects on the spatial location of economic activity. Bleakley and Lin (2012), for example, document that, although the natural advantage of canoe portage sites across the U.S. was made obsolete due technology changes, these portage sites are likely to be agglomeration centers in modern times. The authors interpret the result as evidence of spatial multiple equilibria. Such spatial persistence was also found in other articles, like, Redding et al. (2011), Ahlfeldt et al. (2015), Jedwab and Moradi (2016), Jedwab et al. (2017), Hanlon (2017), Michaels and Rauch (2018), and Brooks and Lutz (2019).

This paper documents the persistent effects of the Brazilian railway network on economic development and economic activities' concentration and examines the mechanisms behind this persistence. Despite the rapid expansion of the Brazilian railway system between 1854 and the middle twentieth century, the railways began to decline from the 1950s onwards. The Brazilian government closed almost 10,000 kilometers of railroad lines between 1960 and 2000. The decrease in the number of railway stations is even more remarkable: From more than 3,600 railway stations in 1950 to less than 400 in 2010. The

[1]For a summary of the historical persistence literature see, for example, Nunn (2009, 2014), and Spolaore and Wacziarg (2013). In general, geographic, institutional, and human capital aspects are considered the main persistence mechanisms in this literature. See, for example, Engerman and Sokoloff (1997, 2002), Acemoglu et al. (2001, 2002), Dell (2010), Porta et al. (1998), and Glaeser et al. (2004).

objective of this article is to analyze the persistence effects of the railroads over time, despite its decline after 1950.

To estimate the persistent effects of the railroads on economic activity, we assemble a novel data set that combines the digitalization of the railway's historical maps and national censuses, besides a set of historical reports. The area under study covers the municipalities of the three most populous regions of Brazil: northeast, southeast, and south, where around 94% of the railroads' lines were located in 1950. The data allow us to follow almost 60 years of railroad decline in Brazil and its consequence on economic outcomes. Historical occupational and population data enable us to explore the mechanisms through which transportation infrastructure persists through time.

The empirical analysis is based on the instrumental variables method to address the problem of the endogenous placement of railroads. The goal of the railroad expansion in Brazil was to connect the inland municipalities to the ports and to integrate the independent regional railways' systems. As consequence of these expansion purposes, the propensity of a municipality to be connected to the transportation network varies according to the proximity to the least-costs routes. Based on the strategy used by Fajgelbaum and Redding (2018), we instrument the railroad connection by the percentage of grid cells within each municipality that lie along the least-cost paths between the railroads' expansion targets. Conditional on the region fixed effects, and controls we expect the process of railroad expansion to be faster in localities near the least-cost paths. A series of balance tests support the validity of our identification strategy.

Our results show a significant and positive long-run effect of the railroad in 1950 on economic development in 2010, despite the decline of this mode of transport. In this respect, the connection of a municipality to the railway system in 1950 increases the income per capita by about 34% in 2010. Also, we find that places within 10 kilometers from a railroad line in 1950 have today a GDP per capita that is 60% higher than those municipalities not connected to the railway system. All results are robust to the inclusion of geographic, baseline socioeconomic characteristics, and pre-railway transportation infrastructure controls, as well as region fixed effects. In the same way, the results remain significant when we use different measures of railroad connection. The persistence of the effect of the railways is not limited, however, to income and GDP per capita. Having been connected to the railway system in 1950 is associated with an increase in population density, as well as a decrease in the share of workers in the agriculture sector over the total occupied workers. We show that the effects of railways on structural change have persisted since the 1950s,

when municipalities connected to the system were already more industrialized than those not connected. Finally, we also show that our results are not driven by alternative transportation infrastructure, like roads. Indeed, the results are robust to the control of the presence of major roads between 1960 and 2010.

To shed light on the mechanism behind the persistence of the effects of the railroads, we analyze the impact of the railways on agglomeration, and urbanization. First, we show that, despite the decline of the railroads, municipalities that were connected to its network have become centers of attraction for residents of other municipalities, perhaps due to the beginning of the process of structural change and income growth. The effects of the 1950s railroads persisted on population density between the 1960s and 2000s. And this agglomeration effect is partly a consequence of the increase in internal migration, municipalities connected by railways have become poles of attraction for domestic migrants. In the 1980s, places that were already connected to railways in 1950 had the share of migrants over the total population 9 percentage points higher than places that were not connected to the railway system, an increase of about 32% in relation to the average proportion of migrants.

Finally, we examine the persistence of the effects of the railroads on urbanization. Along with the agglomeration effects, the expansion of the railway lines increased the proportion of people living in urban areas between 1950 and 1991. Therefore, we can conclude that the differences in per capita income in 2010 between municipalities connected and not connected to the 1950 railway system are driven by agglomeration and urbanization.

Our study is related to two sets of studies. First, our article is connected to the path dependence literature. As Bleakley and Lin (2012), Redding et al. (2011), Ahlfeldt et al. (2015), Hanlon (2017), Michaels and Rauch (2018), and Brooks and Lutz (2019) we find that a declining transport system can have permanent effects on the spatial equilibrium of economic activity. Distinctly from most articles, we provide evidence on the specific mechanisms behind this persistence. We find that agglomeration, and urbanization matters to explain the persistence on economic development. In particular, although the relevance of structural transformation for growth is well established[2], we show that occupational persistence can be important to explain the spatial equilibrium of economic activity.

Our paper is also related to the literature on the effect of railways. The literature is inaugurated by the seminal works of Fogel (1964) and Fishlow

[2]See, for example, Herrendorf et al. (2014).

(1965) that, using the "social saving" methodology, analyze the impact of the railroads on the American economy. More recently, with new estimation methods and the abundance of data, much of the literature has started to focus on the historical impact of the railroads on market integration (Keller and Shiue, 2008; Andrabi and Kuehlwein, 2010; Donaldson, 2018), urban growth (Atack et al., 2010; Hornung, 2015; Berger and Enflo, 2017), agricultural land values (Donaldson and Hornbeck, 2016; Donaldson, 2018), innovation (Andersson et al., 2020), spread of factories (Atack et al., 2008; Tang, 2014; Hornbeck and Rotemberg, 2019), and structural transformation (Yamasaki, 2017; Pérez, 2017; Fajgelbaum and Redding, 2018; Berger, 2019). More related to our work, other articles analyze the persistent effects of the railroad on economic development (Jedwab and Moradi, 2016; Jedwab et al., 2017; Okoye et al., 2019). Differently from these articles, we show that immigration and urbanization are important mechanisms to explain the persistence. Finally, there is no evidence on the long-term effects of railroads for South America. As far as we know, our article is the first to provide evidence on the impact of the railroads on long-run development for this region.

The rest of the article is organized as follows. Section 2.2 outlines the historical background of the railroad network expansion and decline in Brazil between 1854 and 2010. Section 2.3 presents our data and how it was built. Section 2.4 describes the empirical strategy. Section 2.5 presents the main results. Section 2.6 documents the mechanisms underlying the impact of the railroads in 1950 on modern time development. Section 2.7 presents the robustness checks. The final section concludes.

2.2
Historical Background

2.2.1
Railroad Expansion in Brazil, 1854-1950

Before the introduction of railroads, the transportation of goods and people in Brazil took place almost exclusively by roads and shipping routes on the coast. Given the country's continental dimensions, a difficult topography, and rainy weather, the pre-railway transportation was, to a large extent, extremely inefficient. All trade between the coast and the interior of the country was done by animal-drawn carts and mules wagons. Since the most populous regions had no navigable rivers, roads were the only option to connect the interior to the coast. The transport by animals was extremely slow, a 130 kilometers trip on the Brazilian best roads took 7 days since mules wagons

only reached a maximum speed of 3-4 leagues (14-19 Km.) per day (Silva, 1949). This speed could decrease considerably in periods of rain when the roads became practically impassable (Silva, 1949; Mattoon Jr, 1977). Throughout the 19th century, little was done to improve road conditions or make rivers navigable. As a consequence, transportation compromised a considerable part of the farmers' profits[3] (Summerhill, 2003).

Water transport was much more efficient, however, the navigable rivers were mostly located in the sparsely populated areas in the north of the country. The most populous areas, in the southeast and south regions, had a few kilometers of navigable rivers. The maritime navigation was dominant only to connect of coastal cities by steamships (Summerhill, 1998). Coastal shipping was a relatively efficient and inexpensive way to connect the major cities of the country, however, it was not an alternative to transport the agricultural goods from the interior to the exportation ports.

The introduction of railways significantly reduced transportation time and costs. A 130 kilometers trip that took at least 7 days with the mules, reduced to less than a day with the railroads. The trains traveled, on average, 134 kilometers per day, well above the 20 kilometers from the mules. The expansion of the railway network reduced freight costs per ton-kilometer by about 86%, generating an economy in freight services that represent 18% of the Brazilian GDP of 1913, more than double of the estimates found for the United States (Summerhill, 2005).

The railroads have made the connection between the agriculturally productive areas of the interior and the main ports more efficient[4]. In this sense, the expansion of the railroads in Brazil took place from the coast to the interior, mainly until the 1930s. From the 1930s, with the diminishing profitability of the railways, the government participation in the sector began to increase (Villela and Suzigan, 1975), starting a period of expansion to integrate the independent regional railways' systems. The expansion of the railway network continued until the 1950s when the railroad started to decline.

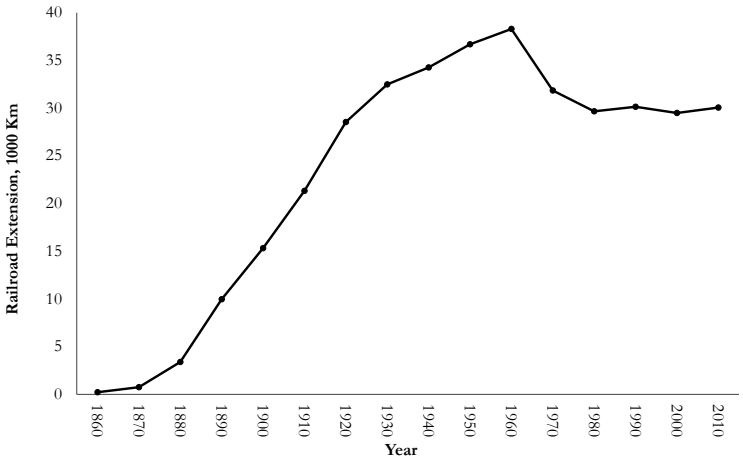
Figure 2.1 shows the expansion of the Brazilian railway network between 1860 and 2010. Between the 1860s and the 1950s, the railroad network grew more than 16 thousand percent, going from 223 kilometers to 36,681 kilometers, an average growth of 3.6 thousand kilometers per decade. The period of

[3]The poor condition of the roads was also an obstacle to the import of machinery by the factories. See, for example, Birchal (1999).

[4]The railway lines followed the sugar and tobacco plantations, in the Northeast. On the other hand, they facilitated the connection with coffee-producing areas in the Southeast (Lamounier, 2012). For the role of railroad expansion on coffee production, see Saes (1981), Matos (1990) and Grandi (2007).

greatest expansion occurred between the years 1880 and 1930, largely financed by private capital from foreign investors and landowners, with government subsidies. From the 1930s, with the decrease in the profitability of railway companies, many began to be nationalized. Thus, the railway expansion between 1930 and 1950 aimed to integrate the previously independent regional rail systems.

Figure 2.1: Railroad Network Expansion, 1860-2010



Notes: The figure shows the evolution of the extension from the railroads' network between 1860 and 2010. *Source:* Data on railroads extension in 1,000 kilometers from IBGE (2003), available at https://seculoxx.ibge.gov.br/, for 1860-1980. For 1990-2000, data from Ipeadata, available at http://www.ipeadata.gov.br/. For 2010, data from *Pesquisa CNT de Ferrovias, 2011*, available at https://cnt.org.br/pesquisa-cnt-ferrovias.

Figures 2.2(a)-2.2(d) present the railroad network in Brazil for 1860-1950 decades. The railway network connected the interior of the country to the coast. Most of the railway system was concentrated in the states of São Paulo, Minas Gerais, and Rio de Janeiro, in the southeastern region of the country, the most populous territory and with the highest agricultural production. As we will see below, the expansion of the railway network ends in the 1950s, and a great process of decline begins.

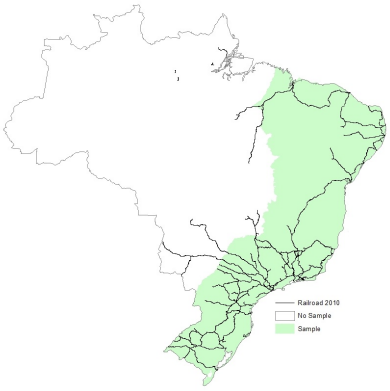
Figure 2.2: Railroads Network Expansion in Brazil, 1860-2010

(2.2(a)) 1860

(2.2(b)) 1890

(2.2(c)) 1920

(2.2(d)) 1950



(2.2(e)) 2010

Notes: This figure displays the expansion of the railroad network between 1860 and 2010. See main text and Appendix B for data sources and details on the construction of the railway network.

2.2.2
Railroad Decay in Brazil, 1950-2010

Beginning in the 1950s, the railway network began to decline. Competition with automobiles, which became more popular and cheaper, the terrible finan-

cial situation from railway companies, and changes in public transportation policies generated a decrease in the railways' system participation in Brazilian transportation. With the increase in the number of cars and trucks since the 1920s, the road transportation will gain more and more importance, since it became more efficient than railways. As the railways were poorly integrated and connected the interior to the ports, the highways offered greater flexibility to connect the domestic market (CMBEU, 1954).

In addition to the competition from the roads, the financial situation of railway companies has deteriorated since the 1930s. Villela and Suzigan (1975) show that since the 1930s, business expenses have represented over 90% of revenues from the railroads' companies. The increase in companies' deficits has several reasons, including over-staffing, political pressure, a non-integrated system with different track gauges, inadequate tariff policy, and outdated equipment (CMBEU, 1954; Paula, 2000). The variety of track gauges[5] and the non-integration of the railroad system, making traffic interruption constant, has always been a limiting factor in the expansion of Brazilian railways. Together with the non-readjustment of tariffs in the face of inflationary pressures, these are perhaps the two most important factors that explain the precarious financial situation of railway companies since the 1930s.

Given the financial problems of the railway companies, the public transportation policies start to change their focus. The expansion of the road system takes precedence over the railway system since the 1950s (Grandi, 2013). Therefore, in the 1950s we see an increase in the road network and stagnation in the railway network. This process will intensify in the 1960-1990s, with policies to close railway branches and reduce the length of the railway network. The closed branches were those that the government judged to have no economic potential (Paula, 2000). Therefore, from the 1950s onwards, we noticed a decrease in the Brazilian railway network, with the closure of branches and stations. Those railroads designed to transport agricultural products from the interior to export ports on the coast were no longer functional due to the change in the Brazilian economic structure and the and the bankruptcy of railway companies, leading the government to close them.

The decrease in the railway network is shown in Figure 2.1. In 1950, the Brazilian railroad system was just over 36 thousand kilometers. Despite the beginning of a process of restructuring the sector, between 1950 and 1960 the network still grew, reaching about 38 thousand kilometers in 1960. However, since the 1960s, the balance of opened and closed railroads is negative. In the

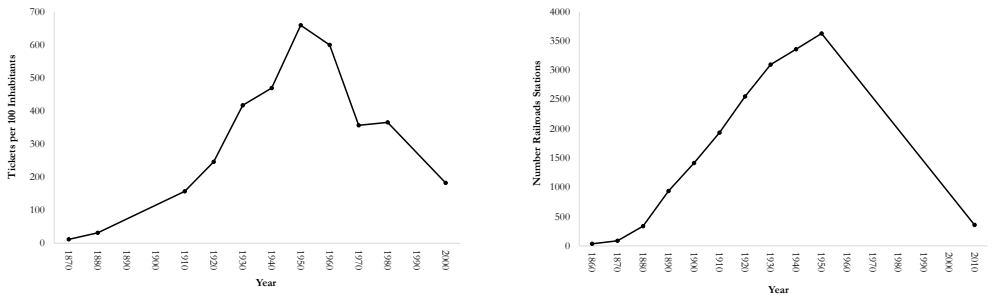
[5]There were five track gauges in Brazil: 1.60m., 1.00m., 0.76m., 0.66m. and 0.60m. (Villela and Suzigan, 1975).

1960s, the government closed almost 7,000 kilometers of the railway lines. The decline process continues over the years 1970-1990, reaching its lowest level in 2000, with a railway network of only 29 thousand kilometers, a reduction of 24% of the total network. In four decades, the Brazilian government closed almost 10,000 kilometers of railroad lines.

The decline in the importance of railways is even greater when we focus on passenger transport. The number of tickets sold to passengers fell by more than 70% between 1950 and 2000, from 660 per 100 inhabitants to just 183 (Figure 2.3(a)). The decrease in the number of passengers transported by railway is related to the closure of stations, as we can see in Figure 2.3(b). In 1950 there were more than 3,600 railway stations in Brazil. In 2010 there are less than 400 railway stations in the country, a reduction of almost 90% in 60 years.

Figure 2.3: Passengers and Stations, 1860-2010

(2.3(a)) Passengers Carried by Railroad (2.3(b)) Number of Railroad Stations



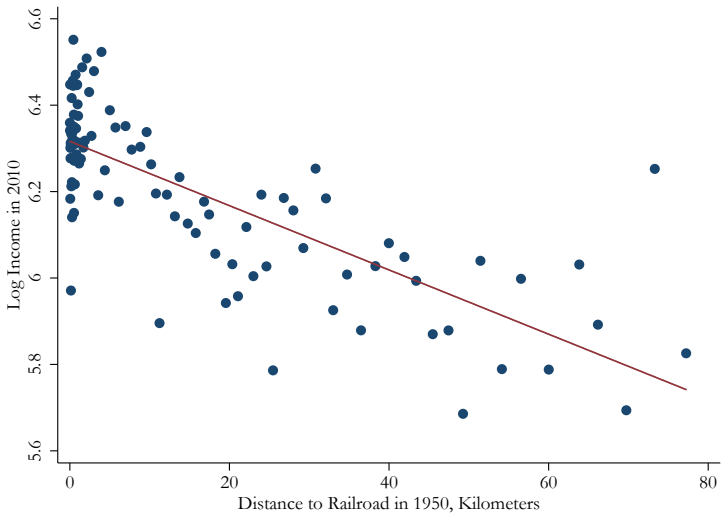
Notes: The figures show the evolution of the passengers transported by railroads between 1870 and 2000, and the number of railroads stations between 1860 and 2010. *Source:* Data on passengers carried by railway from Ipeadata, available at http://www.ipeadata.gov.br/. Original data from statistical yearbooks of transport for several years. Population data from IBGE. Number of railroads stations collected using web scraping from http://www.estacoesferroviarias.com.br/.

The Figures 2.2(d) and 2.2(e) present the railroad network in 1950 and 2010, respectively. Comparing the two figures, we see a drastic reduction in the density of the Brazilian railway network, especially in the richest region of the country, the southeast. This region concentrated most of the production of agricultural commodities for export throughout the late 19th century and the first half of the 20th century in Brazil. Therefore, the railroads built to export these products have become obsolete with the coffee production crisis, and the industrialization of these states. Consequently, most of these railways were closed after the 1950s.

2.2.3
Railroad and Economic Persistence

Despite the decrease in the importance of the railways, and their extension decline, the proximity of a railway line is still associated with an increase in income per capita in 2010, as we show in Figure 2.4. The graph presents the unconditional correlation between the log per capita income in 2010 and the distance to the nearest railway line in 1950, restricted to a 100 kilometers distance threshold. Observations are sorted into 100 bins of equal size and the dots indicate the mean value in each group. The unconditional relationship is negative and highly significant. Although almost 10,000 kilometers of railroad lines were closed between 1950 and 2010, the negative relationship remains strong. This negative correlation leads us to suspect that the impacts of the railways may have persisted over time. We show that this suspicion is true and explore the mechanisms of this persistence in the sections 2.5 and 2.6.

Figure 2.4: Modern Income and Railroads in 1950



Notes: The figure displays the unconditional non-parametric relationship between log per capita income in 2010 and the distance to the nearest railroad line in 1950, restricted to a 100 kilometers distance threshold. Observations are sorted into 100 bins of equal size and the dots indicate the mean value in each group. The red line presents the linear fit line. See main text and Appendix B for data sources and details on the construction of the railway network.

2.3
Data

This study combines four sets of historical and modern data. First, we use GDP and censuses data to calculate, by the municipality, the GDP per capita,

the income per capita, as well the population density for 2010. Historical and modern censuses allow us to build the share of workers that are employed in the agricultural, manufacturing, and service sector, besides the share of urban, and migrant populations. Second, we complement these data with industrial censuses. Third, we merge the employment, population, and industrial data with railroad network data drawn from historical maps. Fourth, we achieve these data with socioeconomic, geographic, and transportation variables. The main data set used in this paper is a cross-section covering 1,663 municipalities from three regions of Brazil between 1950 and 2010.

2.3.1 Census Data

The GDP data from IBGE allows us to build GDP per capita by the municipality in 2010. The modern and historical population information used in this paper comes from 1872, 1950, 1960, 1970, 1980, 1991, 2000, and 2010 Population Censuses. These census records contain, for each municipality, demographic and socioeconomic information. The income and population data from the 2010 census are used to calculate the income per capita for each municipality in 2010. The employment data from the 1950-2000 censuses are used to calculate some outcome variables: the share of workers that employed in the agricultural, manufacturing, and service sector. Also, the demographic information is used to calculate the total population, population density, as well as the share of urban and migrant population by municipality.

Information from the 1950 industrial census on the number of factories, its average number of workers, and production, at the municipality level, complement our data set. Finally, the 1872 census allows us to build socioeconomic characteristics variables just at the beginning of the railroads' expansion. These variables are used to control for pre-existing social and economic aspects.

The census data set covers all the municipalities in the five regions of the country. However, our sample covers just three regions: Northeast, Southeast, and South, where around 94% of the existing Brazilian railway network and about 93% of the population of the country were located in 1950 (IBGE, 2003). Because of the large differences in the socioeconomic characteristics of the municipalities in the other regions (North and Midwest) from those in our sample, and due to the large concentration of railways in the regions analyzed, we decided to focus the sample only on the three most important regions of the country. By doing this, we mitigate the selection concerns.

To follow the municipalities over time, we kept the 1950 border definition.

Therefore, we merge the data of the municipalities from the 2010-1960 censuses to match the 1950 census boundaries[6]. Finally, to control for the baseline socioeconomic characteristics, we match the municipalities from 1950 to those that existed in the 1872 census. As a result, in the main data set, we follow the 1,663 municipalities[7] that existed in 1950 in our sample between 1950 and 2010. See Appendix B for details in how we merge the 1950-2010 data censuses.

2.3.2
Brazilian Railway Network, 1860-2010

The objective of the paper is to analyze the persistence of the impact of railroads in 1950 on long-term development, thus our treatment variables are measures of the closeness to the railroad network from the municipality downtown (the main urban settlement of the municipality) in 1950[8]. To construct these variables, we created a digitized maps of the Brazilian railway network, using modern and historical maps of the railroads, in four steps. First, we started from a geo-referenced map of the modern railroad network created by the Ministery of Infrastructure[9]. Second, we digitized the railway network using historical maps following the same methodology from Atack (2013). We obtain the historical maps of the expansion of the railroad network from multiple sources, including *Plano Nacional de Aviação, Estatísticas das Estradas de Ferro do Brasil* and other historical reports[10]. Third, with the digitized railroad network, we calculated the linear distance from the municipality's main urban settlement to the nearest railway line in 1950. Finally, we certified the consistency of the railroad network by collecting data on the opening date of each railway station[11]. See Appendix B for details on the construction of the dataset.

In addition to the distance to the railway line, we built a dummy to determine if the railway is less than 10 kilometers from the municipality. Since

[6]A similar procedure has been used for the United States (Hornbeck, 2012) and Brazil (Rocha et al., 2017). The Brazilian administrative division can be found at https://www.ibge.gov.br/geociencias/organizacaodo-territorio/estrutura-territorial.html.

[7]More precisely, for the Northeast, Southeast, and South regions there were 1,678 municipalities in 1950. But, for 1 municipality we don't have census data in 1950 and for 14 municipalities we don't have census data in 1872.

[8]Similar to Melander (2018), we think that to measure railway access, the distance from the main settlement is a better proxy than, for example, the distance from the centroid of the municipality.

[9]The modern railroad network shapefile is available at http://www.infraestrutura.gov.br/bit.html.

[10]Most of the maps and reports can be found at https://biblioteca.ibge.gov.br/ and http://memoria.org.br/index.php.

[11]The data are available at http://www.estacoesferroviarias.com.br/, and were collected using web scraping.

the digitizing of historical maps has a margin of error, dummy seems to us to be the most appropriate treatment variable. Similar connection measures are standard in the literature (see, for example, Jedwab et al. (2017)). Finally, we also use the number of railway stations as treatment variable.

2.3.3 Additional Data

We make use of important sets of socioeconomic baseline controls, as well as geographic and transportation infrastructure controls. The 1872 census allows us to control for social and economic characteristics of the municipalities at the beginning of the expansion of the railroad network, since in the 1870s the railway extension was very limited and concentrated in few municipalities. Thus, we construct municipality-level baseline controls: (i) log of population; (ii) share of literate individuals aged 6 or more; (iii) share of foreign-born people; (iv) share of slaves; (v) share of workers in public administration; (vi) share of workers in legal professions; (vii) share of workers in agriculture over the total number of occupied workers; (viii) share of workers in manufacturing over the total number of occupied workers.

As we already described the objective of the Brazilian railroad network expansion was to connect the plantations in the interior to the ports in the coast. Therefore, it is expected that the expansion of the railways to be correlated with the geographic characteristic of the municipalities. In fact, much of the cost of building the railroads depended on the ground conditions and their slope (Lamounier, 2012). Consequently, we include geographic controls in our analysis: longitude, latitude, altitude, distance to the nearest coast, distance to the nearest state's capital, soil types, slope, and area.

The impact of railroads depends on the pre-railway transportation infrastructure. Thus, we include in our analysis controls for the presence of transportations options just before the expansion of the railways. These controls are the distance to the nearest port in 1850, distance to the nearest road in 1867, and distance to the nearest river. Also, to verify the robustness of our results to the expansion of roads in Brazil, we use data from roads between 1960 and 2010 from the Ministry of Infrastructure to create additional controls for the presence of roads over the time.

Finally, to construct our instrumental variable we use two geographic information: the average slope in each 0.5 x 0.5 Km^2 cell, and the presence of river at the same cell. Using high-resolution spatial data from CGIAR-CSI[12],

[12]The data can be found at http://srtm.csi.cgiar.org/srtmdata/

and river's data from ANA[13] we build our measure of propensity to connect to the railway network. The least-cost paths are built assuming a cost function directly proportional to the average slope of the region, adding a penalty for the presence of a river. After defining the cost function, we calculate the cost minimization routes based on the algorithm created by Dijkstra et al. (1959). In Appendix B we present detailed information on the construction of the least-cost paths.

Table 2.1 presents the summary statistics for key variables. The average log income per capita in the sample is slightly above 6.1. Also, the average distance to the nearest railroad line in 1950 was 28.7 kilometers. Around 50% of the municipalities were less than 10 kilometers from a railway. Finally, the average number of railroad stations in 1950 was 2.18. In Appendix B, we provide the definition of the variables used in this paper, and their source and units.

Table 2.1: Summary Statistics

	Observations	Mean	S.D.	Min.	Max.
Panel A: Vars in 2010					
Log income p.c.	1,663	6.15	0.49	4.79	7.60
Log GNP p.c.	1,663	9.25	0.70	7.75	12.30
Log population density	1,663	3.75	1.25	0.14	9.47
% Emp. agriculture	1,663	0.31	0.18	0.01	0.75
% Emp. manufacturing	1,663	0.20	0.10	0.03	0.58
% Emp. service	1,663	0.49	0.12	0.17	0.86
Panel B: Vars in 1950					
Share urban	1,663	0.24	0.17	0.00	1.00
% Emp. agriculture	1,663	0.76	0.19	0.01	0.97
% Emp. manufacturing	1,663	0.08	0.09	0.00	0.73
% Emp. service	1,663	0.16	0.12	0.02	0.85
Log number of factories	1,600	3.23	1.20	-2.30	8.91
Log workers by factories	1,578	2.48	0.92	-0.82	7.32
Log production by factories	1,578	5.61	1.33	-0.20	10.32
Panel C: Railroads in 1950					
Dummy [RR $\leq 10Km$]	1,663	0.52	0.50	0.00	1.00
Distance to Railroad, in Km	1,663	28.74	53.87	0.00	535.92
Number Railroads Stations	1,663	2.18	4.41	0.00	90.00

Continues in the next page...

[13]The data can be found at http://dadosabertos.ana.gov.br/

Table 2.1: Summary Statistics (cont.)

	Observations	Mean	S.D.	Min.	Max.
Panel D: Geography					
Longitude	1,663	-44.24	4.96	-57.09	-34.86
Latitude	1,663	-17.39	7.51	-33.52	-1.19
Log altitude	1,663	5.91	0.97	1.47	7.37
Log distance to Coast	1,663	4.76	1.42	0.02	6.66
Log distance to state's capital	1,663	5.05	0.93	0.00	6.57
Log slope	1,663	1.58	0.65	-0.84	2.89
Log Area	1,663	6.84	1.09	1.95	10.29
% Cambisol	1,663	10.02	22.85	0.00	100.00
% Latosol	1,663	31.02	35.16	0.00	100.00
% Argisol	1,663	31.02	34.68	0.00	100.00
% Spondosol	1,663	10.73	21.09	0.00	100.00
Panel E: Socioeconomic Characteristics, 1872					
Log population	1,663	9.67	0.74	7.19	12.52
% Literate (aged 6+)	1,663	0.19	0.11	0.02	0.69
% Foreigners	1,663	0.02	0.05	0.00	0.50
% Slaves	1,663	0.14	0.09	0.01	0.57
Public Administration (in 1,000)	1,663	0.96	1.42	0.00	31.63
Legal Professionals (in 1,000)	1,663	0.75	0.59	0.00	5.26
% Emp. agriculture	1,663	0.52	0.15	0.01	0.94
% Emp. manufacturing	1,663	0.14	0.07	0.00	0.47
Panel F: Transportation Controls					
Log distance to port, 1850	1,663	5.21	0.94	-1.24	6.72
Log distance to road, 1867	1,663	2.73	1.41	0.00	5.51
Log distance to river	1,663	3.07	1.26	0.01	5.26
Dummy [Road $\leq 10Km$], 1960	1,663	0.14	0.35	0.00	1.00
Dummy [Road $\leq 10Km$], 1970	1,663	0.39	0.49	0.00	1.00
Dummy [Road $\leq 10Km$], 1980	1,663	0.47	0.50	0.00	1.00
Dummy [Road $\leq 10Km$], 1990	1,663	0.56	0.50	0.00	1.00
Dummy [Road $\leq 10Km$], 2000	1,663	0.60	0.49	0.00	1.00
Dummy [Road $\leq 10Km$], 2010	1,663	0.62	0.49	0.00	1.00

Notes: Descriptive statistics for the main variables used in the paper. Occupational and population data from economic and population censuses. Manufacturing data in 1950 from the economic census. Indicator and distance for railway built from historical maps. Geographic controls created using ArcGIS with data originally from IBGE, CGIAR-CSI, and Embrapa. Controls for socioeconomic characteristics in 1872 from population census. Transportation controls created using ArcGIS with data originally from IBGE, ANA, DNIT and historical reports. In all panels the sample is based on the 1950 municipality boundaries. For data specific descriptions and sources, see the Appendix B.

2.4 Empirical Model

2.4.1 Empirical Strategy

Our goal is to examine the persistent effect of the expansion of the railway network on long-term economic development. In order to do so, we explore variation in the expansion of the railroad network across municipalities due to the closeness from cost-minimizing routes (similar to Jedwab et al. (2017) and

Fajgelbaum and Redding (2018)). More specifically, we estimate the following equation:

$$Y_{ir,2010} = \alpha + \delta_r + \beta Railroad_{ir,1950} + X'_{ir}\eta + \epsilon_{ir} \quad (2\text{-}1)$$

Where $Y_{ir,2010}$ is a economic outcome in municipality i, in a region r measured in 2010. Our key variable of interest is $Railroad_{ir,1950}$, an indicator variable that takes a value of 1 if the linear distance from the municipality i and region r in 1950 to the nearest railroad line is equal or less than 10 kilometers[14]. We follow Pérez (2017) and choose to use the indicator variable than the continuous distance for two reasons: (i) train stations are not necessarily located in downtown city; (ii) there is a margin of error in building the railroad network data. Thus, the 10 kilometers bin of distance to the railroad helps to contain these imprecisions. We also include region fixed effects, δ_r, to control for time-invariant regions characteristics and conditions. Furthermore, we include a set of time-invariant geographic, baseline socioeconomic characteristics, and transport controls, X_{ir}, described in detail below, to capture differential characteristics across municipalities before the introduction of the railways.

Our parameter of interest is β. If the location of the railway lines before 1950 were random, this parameter would capture the effect of the railway on the economic outcomes. However, as discussed in section 2.2, much of the railroad's expansion had the intention to connect inland productive areas to the ports before the 1930s. Consequently, although we consider a series of controls in equation 2-1, β can be biased by the influence of unobservable confounding trends. For example, if railroad expansion responds to an unobservable increase in agriculture productivity, we should expect attenuation bias in our estimates due to reverse causality. Alternatively, after the 1930s most of the expansion of the railway network was conducted by the state to integrate the independents' regional systems and the states' capitals. Thus, after the 1930s we should expect a positive bias in our estimates due to endogeneity in the choice of the targeted cities.

To address these potential concerns, we complement the analysis with an instrumental variable strategy that exploits the features of the railway system expansion to generate exogenous variation in the probability that a municipality gets connected.

[14]Our results are robust for alternative measures of railroad access. See section 2.5.4 for the results using the natural logarithm of the distance to the nearest railroad as interest variable and the number of railroad stations.

2.4.2
Instrumental Variable: Least-Cost Path

The expansion of the Brazilian railway network had two main objectives: (i) to connect productive land from the interior to the ports; (ii) connect local railway systems and state's capitals. Therefore, our instrument exploits the fact that some municipalities are more likely to be connected to the railroad network just because they are located along the least-cost routes between the interior and the ports, and between the state's capitals. In other words, our instrumental variable predicts the propensity to be connected to the railroad network based on the construction of least-cost paths between each municipality and the ports that existed before the introduction of the railway in Brazil, and between the state's capitals.

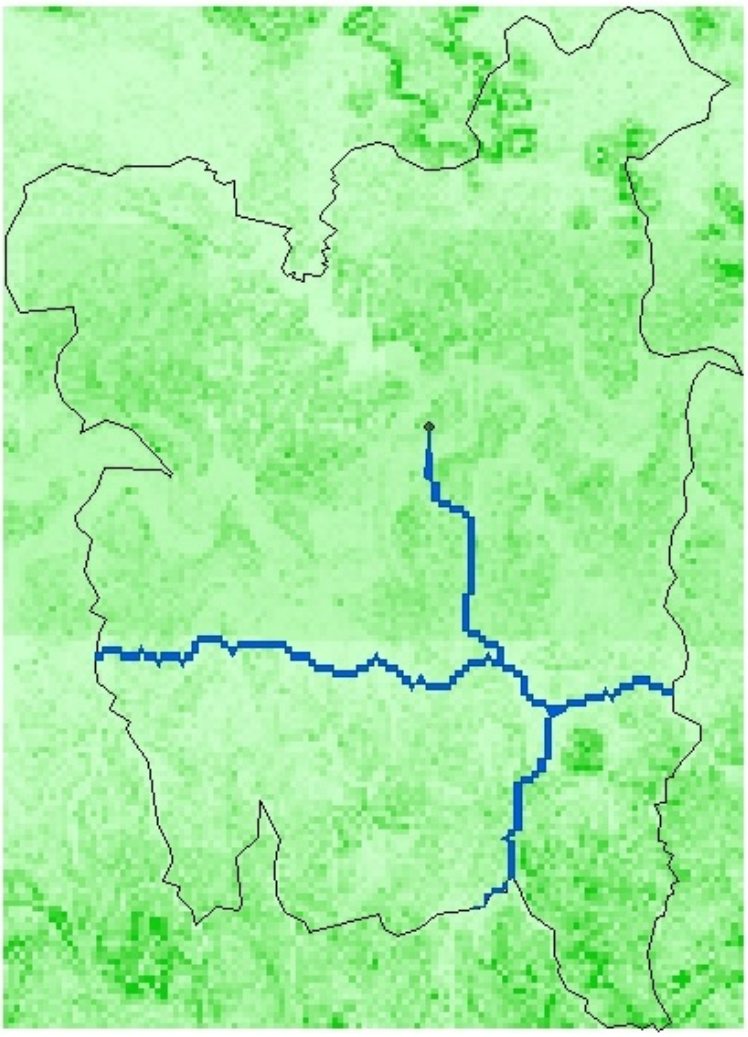
To build the instrument, we implemented a procedure similar to that one outlined in Fajgelbaum and Redding (2018). In particular, we discretize Brazil into a raster of grid cells (0.5 Km x 0.5 Km) and calculate the least-cost paths between each municipality that existed in 1872 and the existing ports in 1850[15] (municipalities and ports existing at the beginning of the expansion of railways in Brazil), and between the state's capitals[16]. We use data on slope degrees and rivers to calculate the costs associated with each cell, assuming a cost function increasing monotonically with slope, and a river crossing penalty[17]. Then, for each municipality, we compute the percentage of grid cells within its boundaries that lie along at least one of these least-cost paths. Figure 2.5 illustrates the construction of the instrumental variable for the municipality of Rio Pardo in the state of Rio Grande do Sul. The municipality corresponds to the area within the black border. The blue lines represent the least-cost paths that cross the municipality.

[15]The list of the most important ports in Brazil before the introduction of railways in the country is from Brasil (1850) and can be found at http://ddsnext.crl.edu/brazil. The list includes the following ports: Rio de Janeiro, Salvador, Recife, São Luiz, Rio Grande, Maceió, Santos, São José do Norte, Paranaguá, Aracajú, Fortaleza, Desterro, Porto Alegre, Aracaty, Parnayha, Vitória, Parnahyba, and Natal.

[16]More precisely, the integration of the regional railway systems and capitals occurred through the connection of Rio de Janeiro to the capitals of the other states from the southeast and northeast, and the connection of São Paulo to the southern system, crossing Itararé, Imbituva, Cruz Alta, reaching Santa Maria at Rio Grande do Sul (acts 10,432 from 1889, and 24,497 from 1934). Thus, we use these connections to build the least-cost routes.

[17]The penalty corresponds to an average slope of 1.7°, or 3%, the maximum slope allowed in many late 19th-century railroad projects in Brazil.

Figure 2.5: An Example of the Empirical Strategy: Municipality of Rio Pardo, Rio Grande do Sul



Notes: The figure displays the municipality of Rio Pardo (within the black border) in the state of Rio Grande do Sul. The blue cells represent the least-cost paths. The instrumental variable is the log of the percentage of blue cells over the total number of green cells within the municipality. See Appendix B for detailed information on the construction of the least-cost paths.

In this case, our instrument will be the log of the percentage of blue cells over the total number of green cells within the border of the municipality. See Appendix B for detailed information on the construction of the least-cost paths.

Formally, the first-stage relationship for our second-stage equation 2-1 takes the following form:

$$Railroad_{ir,1950} = \alpha + \delta_r + \gamma LCP_{ir} + X'_{ir}\lambda + \nu_{ir} \tag{2-2}$$

Where LCP_{ir} is the log of the percentage of grid cells within municipality i in the region r in 1950 that lie along least-cost paths. Since slope and river cover are crucial for calculating the least-cost routes, we control in all specifications for log municipality average slope and log distance to the nearest river. Additionally, to address the concern that larger municipalities are other things equal more likely to be along these least-cost paths, we control for log municipality land area. Finally, since ports and state capitals are targets for the construction of the least-cost paths, we also control in all specifications for log distance to the nearest port and state capital.

In addition to the controls mentioned above, X_{ir}, includes controls to other geographic features, like, longitude, latitude, log altitude, log distance to the nearest coast, and percentage of the municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. To control for potential heterogeneity in initial levels of economic development, we control for baseline socioeconomic characteristics in 1872: (i) log of population; (ii) share of literate individuals aged 6 or more; (iii) share of foreign-born people; (iv) share of slaves; (v) share of workers in public administration; (vi) share of workers in legal professions; (vii) share of workers in agriculture over the total number of occupied workers; (viii) share of workers in manufacturing over the total number of occupied workers. Since the impact of railways depends on pre-rail transport infrastructure, we also control for log distance to the nearest road in 1867, a proxy for the baseline transportation infrastructure.

Conditional on the region fixed effect, δ_r, and the controls X_{ir}, we expect that the expansion of the railroad network to be faster in localities crossed by more least-cost paths. Our identification assumption is that, conditional on fixed-effect and control variables, the percentage of grid cells within each municipality that lie along at least one of the least-cost paths, just impact the economic outcomes by the likelihood of connecting the municipality to the railroad system. In other words, there are no other shocks correlated with this differential exposure, LCP_{ir} that affect the economic outcomes. Although not directly testable, we show in the next section that our instrument is not correlated with municipalities' initial characteristics.

2.4.3
Least-Cost Paths and Expansion of Railways

Table 2.2 reports the first-stage results. We estimate standard errors clustered at the 1872 municipality border level to account for serial correlation within municipalities. All columns include controls for slope, distance to the

nearest river, land area, distance to the nearest port, and states' capital.

Table 2.2: First Stage Results, 1950

	Dummy for Railroad ≤ 10 Km.				
	(1)	(2)	(3)	(4)	(5)
Least-Cost Path	0.124***	0.135***	0.142***	0.137***	0.138***
	[0.027]	[0.024]	[0.024]	[0.024]	[0.024]
Mean Dep. Var.	0.518	0.518	0.518	0.518	0.518
Observations	1,663	1,663	1,663	1,663	1,663
R^2	0.123	0.193	0.214	0.215	0.215
KP F-stat	20.989	30.708	35.935	33.431	33.431
Geography	No	Yes	Yes	Yes	Yes
Characteristics 1872	No	No	Yes	Yes	Yes
Transportation	No	No	No	Yes	Yes
Region FE	No	No	No	No	Yes

Notes: This table reports first-stage results. All columns report the results from OLS regressions where the dependent variable is a dummy for the presence of a railroad line near the municipality in 1950, and the variable of interest is the log of the percentage of grid points within each municipality that lie on the least-cost paths. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the log of population, the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture and manufacturing computed over the total number of occupied workers. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 1,663 municipalities based on the 1950 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality boundaries. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In column (1) we observe a positive and significant effect of the percentage of cells that lie on the least-cost routes on the probability to be connected on the railroad, with a KP F-statistic of almost 21. In columns (2)-(5) we add, gradually, the other geographic controls (longitude, latitude, altitude, distance to coast, and types of soil), the baseline characteristics, the other transportation controls, and the region fixed effect. In our most complete specification (column (5)), the impact of the least-cost path on the likelihood of the municipality being connected to the railways remains positive and significant, with a KP F-statistic of 33.4. The magnitude and robustness of the coefficient do not vary much according to specifications. This indicates

that the percentage of grid cells within each municipality that lie along at least one of the least-cost paths is a strong predictor of the railway in 1950, conditional upon geographic, baseline socioeconomic, transportation controls, as well as region fixed effect.

Table 2.3 presents evidences that support our identification assumption. In this table, we test whether the initial characteristics are related to the least-cost paths for the municipalities not yet connected to the railroad network in 1872. To do so, we regress at the cross-section the log of the percentage of cells that lie on the least-cost routes on the baseline socioeconomic and transportation characteristics.

Table 2.3: Least-Cost Paths and Baseline Characteristics, 1872

	Least-Cost Path									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Log Population	0.027									0.032
	[0.023]									[0.024]
% Emp. agriculture		-0.087								0.015
		[0.097]								[0.117]
% Emp. manufacturing			0.253							0.251
			[0.198]							[0.226]
% Literate rate				0.143						0.147
				[0.153]						[0.166]
% Foreigners					0.215					0.048
					[0.536]					[0.568]
% Slaves						0.127				0.087
						[0.175]				[0.179]
Public Administration							0.005			0.002
							[0.008]			[0.008]
Legal professions								0.031		0.023
								[0.023]		[0.025]
Distance to Road									-0.007	-0.006
									[0.011]	[0.012]
Observations	525	525	525	525	525	525	525	525	525	525
Adjusted R^2	0.436	0.436	0.436	0.435	0.435	0.435	0.435	0.437	0.435	0.433

Notes: This table reports whether the baseline socioeconomic characteristics are systematically related to the least-cost paths for municipalities without railroad in 1872. All columns report the results from OLS regressions. Each column is a separate regression where we regress the log of the percentage of grid points within each municipality that lie on the least-cost paths on the initial characteristics. In column (10) we regress the least-cost path instrument on all baseline variables. All columns include controls for distance to the port, distance to states' capital, land area, slope, distance to the river, and region fixed effect. All regressions estimated for the 525 municipalities not connected to the railroad network based on the 1872 census boundaries. Standard errors are given in brackets and are clustered at the municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The regressions are estimated for the 525 municipalities not connected to the railroad network based on the 1872 census boundaries. Each column is a separate regression where the log of the percentage of cells that lie on the least-cost paths is the dependent variable. All columns include controls for slope, distance to the river, land area, distance to the port, states' capital,

and region fixed effect. While in columns (1)-(9) we regress the instrument in each variable separately, in column (10) we regress the least-cost path in all variables simultaneously. In all columns, we find results that are not statistically significant, even when we add all the variables simultaneously. The initial characteristics do not seem to be systematically related to the least-cost paths. Overall, our instrument is not correlated with the baseline economic and social structure, as well as the pre-railroad transportation infrastructure.

The evidence presented in this section indicates that the least-cost paths predict the connection to the railway system in 1950 and that they are not related with the social, economic, and transportation characteristics in the baseline. This supports the idea that our instrument can capture the causal effect of the railroad on long-term development.

2.5
Main Results

Despite the decline of railroads in Brazil since the 1950s, their effects are still present today. In this section, we report the persistent effects of railways on long-term development, as well as their impacts on agglomeration and structural change. We complement the analysis checking the robustness of the results using other measures of connection to the railroad system, and controlling for the presence of major roads.

2.5.1
The Long-Term Effects of Railways on Income

We start by reporting the effects of railway network connection in 1950 on economic development in 2010. First, we use the log of the per capita income to measure the level of development. Table 2.4, Panel A, reports the results from estimating the OLS regressions, based on equation 2-1. Panel B, otherwise, presents analogous results for our 2SLS specification, based on first-stage as in equation 2-2. Since our instrumental variable is built based on some geographic information, all columns include controls for the determinants of the least-cost paths - slope, distance to the nearest river, land area, distance to the nearest port, and distance to the nearest states' capitals. Throughout both panels, the second column adds controls for other geographic conditions (longitude, latitude, altitude, distance to the nearest coast, and types of soil). In columns (3) and (4) we add, respectively, controls for baseline socioeconomic characteristics and pre-railroad transportation infrastructure. Finally, in column (5) we include region fixed effect. All regressions are

estimated for the 1,663 municipalities based on the 1950 census boundaries, with standard errors clustered at the 1872 municipality level, to account for serial correlation within municipalities.

Table 2.4, Panel A, reports positive and significant OLS estimates, revealing a positive and significant impact of the railway transport structure in 1950 on per capita income in 2010. As we can see in column (5), being connected to the railroads in 1950 is associated with an increase of about 0.1 log points of the per capita income in 2010. On Panel B, the 2SLS estimation, we observe that the coefficients remain positive, although larger point estimates. Apart from the first column, all coefficients remain significant at 1%, and their point estimates do not vary much with the inclusion of new controls. The increase in 2SLS point estimates to the OLS indicates attenuation bias in OLS specifications. This is expected should the railroad expansion respond to a non-observable decrease in long-run economic growth, for example. In our most complete specification, column (5), the 2SLS estimated coefficient reports that being 10 kilometers closer to a railway line in 1950 gives about 0.3 log points of the per capita income in 2010. In other words, municipalities connected to the railroad system in 1950 have an income per capita around 34% higher than those not connected. Despite the diminishing importance of railways as a transportation system between 1950 and 2010, their effects still persist today on per capita income.

Table 2.4: Railroads and Development Persistence, Income in 2010

	Log Income per capita				
	(1)	(2)	(3)	(4)	(5)
Panel A: OLS					
Dummy [RR $\leq$ 10 km]	0.310***	0.145***	0.144***	0.144***	0.142***
	[0.029]	[0.016]	[0.015]	[0.015]	[0.015]
Panel B: 2SLS					
Dummy [RR $\leq$ 10 km]	-0.007	0.428***	0.370***	0.376***	0.341***
	[0.348]	[0.111]	[0.098]	[0.102]	[0.102]
Mean Dep. Var.	6.155	6.155	6.155	6.155	6.155
Observations	1,663	1,663	1,663	1,663	1,663
Geography	No	Yes	Yes	Yes	Yes
Characteristics 1872	No	No	Yes	Yes	Yes
Transportation	No	No	No	Yes	Yes
Region FE	No	No	No	No	Yes

Notes: This table reports the persistent effects of railroads on the log of income per capita in 2010. All columns report the results from OLS (panel A) and 2SLS (panel B) regressions where the dependent variable is the log of income per capita in 2010, and the variable of interest is a dummy for the presence of a railroad line near the municipality in 1950. For panel B, the instrumental variable is the log of the percentage of grid points within each municipality that lie on the least-cost paths.All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the log of population, the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture and manufacturing computed over the total number of occupied workers. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 1,663 municipalities based on the 1950 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality boundaries. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In addition to income per capita, GDP can be another interesting indicator of economic development. Therefore, in Table 2.5 we do the same exercises as in Table 2.4, however, using the log of GDP per capita in 2010 as dependent variable.

Table 2.5: Railroads and Development Persistence, GDP in 2010

	Log GDP per capita				
	(1)	(2)	(3)	(4)	(5)
Panel A: OLS					
Dummy [RR $\leq$ 10 km]	0.405***	0.203***	0.199***	0.198***	0.195***
	[0.042]	[0.029]	[0.029]	[0.029]	[0.029]
Panel B: 2SLS					
Dummy [RR $\leq$ 10 km]	0.411	0.726***	0.655***	0.647***	0.606***
	[0.433]	[0.212]	[0.196]	[0.205]	[0.205]
Mean Dep. Var.	9.251	9.251	9.251	9.251	9.251
Observations	1,663	1,663	1,663	1,663	1,663
Geography	No	Yes	Yes	Yes	Yes
Characteristics 1872	No	No	Yes	Yes	Yes
Transportation	No	No	No	Yes	Yes
Region FE	No	No	No	No	Yes

Notes: This table reports the persistent effects of railroads on the log of GNP per capita in 2010. All columns report the results from OLS (panel A) and 2SLS (panel B) regressions where the dependent variable is the log of GDP per capita in 2010, and the variable of interest is a dummy for the presence of a railroad line near the municipality in 1950. For panel B, the instrumental variable is the log of the percentage of grid points within each municipality that lie on the least-cost paths.All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the log of population, the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture and manufacturing computed over the total number of occupied workers. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 1,663 municipalities based on the 1950 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality boundaries. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The persistence of the railroads' impact on GDP per capita is similar to that for per capita income, despite an increase in point estimates. In the first column, we just control for the determinants of the least-cost paths - slope, distance to the nearest river, land area, distance to the nearest port, and distance to the nearest states' capitals. While the OLS estimate indicates a positive and significant impact of the railways on GDP per capita, the 2SLS estimate is not significant (column (1)). However, in column (2), where we control for other geographic characteristics (longitude, latitude, altitude,

distance to the nearest coast, and types of soil), the estimated effect of 2SLS becomes significant at 1%. Municipalities connected to the railroad system in 1950 have, on average, a GDP per capita 73% higher than those not connected. The introduction of other controls, despite reducing the point estimates, does not alter the statistical significance of the results. In our most complete specification, column (5), we find that being connected to the railway network in 1950 is associated with an increase of about 20% of the GDP per capita, for the OLS estimation. The impact increases for 61% in the 2SLS estimation.

2.5.2
The Long-Term Effects of Railways on Agglomeration

There is evidence in the literature about the persistence over time of the effects of railways on population agglomeration and city creation (Jedwab and Moradi, 2016; Jedwab et al., 2017). Therefore, we test whether this relationship also applies to the Brazilian case. Table 2.6 reports the impact of the railroads in 1950 on population density in 2010. As in the previous tables, all columns include controls for the determinants of the least-cost paths - slope, distance to the nearest river, land area, distance to the nearest port, and distance to the nearest states' capitals. Columns (2)-(5) gradually include controls for geography, baseline characteristics, transportation, and region fixed effect. On Panel A, the OLS estimation, we observe that to be within 10 kilometers from a railroad line in 1950 is associated with an increase in 0.5 log points on the population density. The point estimate for the effect of 2SLS increases to 1.9 (column (1)). This increase makes sense as a large part of the railroads aimed to connect the ports to agricultural producing regions, places in general with low demographic density. The inclusion of the other controls practically does not alter the point estimates, as well as their significance levels. In our most complete specification, the impact of railroads in 1950 on population density in 2010 is 1.8 log points when we use the log of the percentage of grid cells that lie on least-cost routes as the instrument from the connectivity dummy, Panel B column (5).

Table 2.6: Railroads and Development Persistence, Population Density in 2010

	Log Population Density				
	(1)	(2)	(3)	(4)	(5)
Panel A: OLS					
Dummy [RR $\leq$ 10 km]	0.559***	0.523***	0.507***	0.506***	0.505***
	[0.050]	[0.050]	[0.050]	[0.050]	[0.050]
Panel B: 2SLS					
Dummy [RR $\leq$ 10 km]	1.909***	1.738***	1.841***	1.883***	1.858***
	[0.496]	[0.409]	[0.395]	[0.417]	[0.416]
Mean Dep. Var.	3.750	3.750	3.750	3.750	3.750
Observations	1,663	1,663	1,663	1,663	1,663
Geography	No	Yes	Yes	Yes	Yes
Characteristics 1872	No	No	Yes	Yes	Yes
Transportation	No	No	No	Yes	Yes
Region FE	No	No	No	No	Yes

Notes: This table reports the persistent effects of railroads on the log population density in 2010. All columns report the results from OLS (panel A) and 2SLS (panel B) regressions where the dependent variable is the log population density in 2010, and the variable of interest is a dummy for the presence of a railroad line near the municipality in 1950. For panel B, the instrumental variable is the log of the percentage of grid points within each municipality that lie on the least-cost paths.All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the log of population, the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture and manufacturing computed over the total number of occupied workers. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 1,663 municipalities based on the 1950 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality boundaries. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In conclusion, the effects of railways on the persistence of population agglomerations are robust and strong, indicating that the effects on development may be associated with the formation of new spatial equilibriums. In section 2.6 we test to what extent urbanization and agglomeration are the mechanisms that generate persistent effects of railways on development.

2.5.3
The Long-Term Effects of Railways on Structural Transformation

The long-term effects of the railways are not limited to per capita income and population density. In Table 2.7 we show that being connected to the railroad system also impacts the economy's occupational structure.

Table 2.7: Railroads and Development Persistence, Structural Transformation in 2010

Dependent variable:	% Emp. Agriculture	% Emp. Manufacturing	% Emp. Service
	(1)	(2)	(3)
Panel A: OLS			
Dummy [RR $\leq$ 10 km]	-0.111***	0.036***	0.075***
	[0.009]	[0.005]	[0.006]
Panel B: 2SLS			
Dummy [RR $\leq$ 10 km]	-0.364***	0.126***	0.239***
	[0.073]	[0.038]	[0.053]
Mean Dep. Var.	0.314	0.198	0.489
Observations	1,663	1,663	1,663
Geography	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes
Transportation	Yes	Yes	Yes
Region FE	Yes	Yes	Yes

Notes: This table reports the persistent effects of railroads on the share of workers in the agriculture, manufacturing and service sector. All columns report the results from OLS (panel A) and 2SLS (panel B) regressions where the dependent variable is the share of workers in agriculture, manufacturing or service over the total number of occupied workers in 2010. The variable of interest is a dummy for the presence of a railroad line near the municipality in 1950. For panel B, the instrumental variable is the log of the percentage of grid points within each municipality that lie on the least-cost paths.All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the log of population, the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture and manufacturing computed over the total number of occupied workers. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 1,663 municipalities based on the 1950 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality boundaries. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In columns (1)-(3), we include all the controls. The dependent variable is the proportion of agricultural, manufacturing, or service workers over the total number of occupied workers in 2010. The variable of interest is a dummy for the presence of a railroad line near (10 kilometers or less) the municipality in 1950. In Panel A, the OLS estimation, the railway is associated with a decrease in 11 percentage points on the share of agricultural workers, and an increase in 4 and 7 percentage points on the share of manufacturing and service sectors, respectively. The magnitudes of the effects are higher for the 2SLS estimates, Panel B, being 10 kilometers closer to a railway line decreases the share of agriculture workers in 35 percentage points, and increases the share of manufacturing and service workers in 12 and 24 percentage points, respectively.

The impact of railroads on industrialization is not just for 2010. Indeed, this effect was already present in the 1950s. In Table 2.8 we explore the effects of railways on the industrial sector in 1950. The table reports the effect of railways on the number of factories, the average number of workers per plant, and the average production of these factories in 1950. The railways are associated with the increase, both in the extensive and intensive margins, of the Brazilian industry. Municipalities connected to the railroad network in 1950 had, on average, more factories (column (1), increase in 0.7 log points), and more workers and production per factory (columns (2) and (3), increase in 1.3 and 1.4 log points, respectively) in the same year.

Table 2.8: Railroads and Industrialization, 1950

Dependent variable:	Log factories (+1)	Log factories' size	Log factories' production
	(1)	(2)	(3)
Panel A: OLS			
Dummy [RR $\leq$ 10 km]	0.628***	0.349***	0.891***
	[0.060]	[0.052]	[0.074]
Panel B: 2SLS			
Dummy [RR $\leq$ 10 km]	0.770*	1.372***	1.424***
	[0.400]	[0.429]	[0.454]
Mean Dep. Var.	3.231	2.480	5.612
Observations	1,600	1,578	1,578
R^2	0.289	0.053	0.281
Geography	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes
Transportation	Yes	Yes	Yes
Region FE	Yes	Yes	Yes

Notes: This table reports the effects of railroads on the number of manufacturing plants, its average number of workers and production in 1950. All columns report the results from OLS (panel A) and 2SLS (panel B) regressions where the dependent variable is defined at the top of each column. The variable of interest is a dummy for the presence of a railroad line near the municipality in 1950. For panel B, the instrumental variable is the log of the percentage of grid points within each municipality that lie on the least-cost paths. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the log of population, the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture and manufacturing computed over the total number of occupied workers. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the municipalities based on the 1950 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality boundaries. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

If in the 1950s, the municipalities connected to the railway system were already more industrialized, then this effect persisted over time without major changes as we can see in Table 2.9. In this table we report the 2SLS estimates for the effects of the railroads in 1950 on the percentage of workers in the agriculture, manufacturing, and services over the total number of employed workers between 1950 and 2000.

Table 2.9: The 2SLS Effects of Railroads on Structural Transformation, 1950 - 2000

	Structural Transformation, 2SLS				
Census:	1950	1970	1980	1991	2000
Panel A: % Emp. Agriculture					
Dummy [RR $\leq$ 10 km]	-0.339***	-0.151***	-0.463***	-0.436***	-0.440***
	[0.072]	[0.030]	[0.086]	[0.085]	[0.083]
Panel B: % Emp. Manufacturing					
Dummy [RR $\leq$ 10 km]	0.156***	0.072***	0.272***	0.200***	0.167***
	[0.043]	[0.017]	[0.053]	[0.043]	[0.040]
Panel C: % Emp. Service					
Dummy [RR $\leq$ 10 km]	0.182***	0.079***	0.191***	0.236***	0.272***
	[0.043]	[0.022]	[0.051]	[0.057]	[0.058]
Observations	1,663	1,663	1,663	1,663	1,663
Geography	Yes	Yes	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes	Yes	Yes
Transportation	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the persistent effects of railroads on the share of workers in the agriculture, manufacturing and service sector. All columns report the results from 2SLS regressions where the dependent variable is the share of workers in agriculture, manufacturing or service over the total number of occupied workers in 1950 - 2000. The variable of interest is a dummy for the presence of a railroad line near the municipality in 1950. The instrumental variable is the log of the percentage of grid points within each municipality that lie on the least-cost paths.All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the log of population, the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture and manufacturing computed over the total number of occupied workers. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 1,663 municipalities based on the 1950 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality boundaries. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In all columns, we control by geographic characteristics, socioeconomic characteristics in 1872, as well as transportation controls and region fixed effects. Each panel presents the results for an economic sector[18]. In column (1)

[18]Unfortunately, the occupational structure data for the 1960 census has not yet been

we can see that the railways are associated with a decrease in the proportion of workers in the agricultural sector. Municipalities connected to the railway system in 1950 had, on average, 34 percentage points fewer on the share of agricultural workers in the same year (Panel A). In the same way, there is an increase in 16 and 18 percentage points, respectively, on the share of workers in the manufacturing and service sectors (Panels B and C, respectively).

Despite the reduction in the magnitude of the impact in the 1970s, the effect of railways on structural change persists over time. The 1960s was the biggest fall in the extension of the Brazilian railway network (Figure 2.1), thus the 1970 census captures a reduction in the magnitude of the railroads' effects on structural change, indicating a period of readaptation as the effects grow again in the subsequent decades. In 1980, municipalities connected to the railroads in 1950 had a 46 percentage point reduction in the proportion of agricultural workers, with an increase of 27 and 19 percentage points, respectively, for the manufacturing and services sectors. If during the 1950-1980s the structural change was concentrated on the growth of the industrial and service sectors, in the 1990s and 2000s the growth of the services sector became prevailing.

Overall, the results found indicate that, despite the decline of the railroads since the 1950s, their impact on income, population density, and occupational structure persisted over time. The evidence can be understood in light of the agglomeration literature. The railroads, by connecting the markets, determined the location of economic activity. And despite its decline, the effect persisted. The economic activity remains concentrated in the same locations. In section 2.6 we explore how this persistence happened.

2.5.4 Alternative Measures

The results found are robust to alternative measures as variable of interest. In Table 2.10 we report the 2SLS estimations using the log distance to the nearest railroad line in 1950 as interest variable, as well as the log of the number of railroad stations. In all columns, we control for geography, baseline characteristics, pre-railway transportation infrastructure, and region fixed effect. In all regressions, the KP F-statistics of the first stage is larger than 10 (17.5 for log distance, and 28.5 for log stations), indicating that the instrument is also strong in these cases.

fully digitized. However, the omission of that year does not compromise the interpretation of the results.

Table 2.10: Railroads and Development Persistence, Alternative Measures (2SLS)

	Log Income p.c.		Log Pop. Density		% Emp. Agr.	
	(1)	(2)	(3)	(4)	(5)	(6)
Log Distance RR	-0.092***		-0.502***		0.098***	
	[0.031]		[0.134]		[0.024]	
Log RR stations		0.211***		1.151***		-0.226***
		[0.063]		[0.266]		[0.047]
Mean Dep. Var.	6.155	6.155	3.750	3.750	0.314	0.314
Observations	1,663	1,663	1,663	1,663	1,663	1,663
R^2	0.667	0.737	0.085	0.423	0.499	0.032
KP F-stat	17.545	28.576	17.545	28.576	17.545	28.576
Geography	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes	Yes	Yes	Yes
Transportation	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the persistent effects of railroads on development in 2010. All columns report the results from 2SLS regressions where the dependent variable is defined at the top of each column. The variable of interest is the log of the distance from the municipality to the nearest railroad line in 1950 or the log number of railroads stations for the same year. The instrumental variable is the log of the percentage of grid points within each municipality that lie on the least-cost paths. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the log of population, the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture and manufacturing computed over the total number of occupied workers. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 1,663 municipalities based on the 1950 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality boundaries. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Both for the linear distance of a railway and the number of stations, we find persistence in the impact of the railway on economic development, population agglomeration, and occupational structure. All results are significant and maintain our conclusions so far.

2.6 Mechanisms

In the previous section, we show that the effects of railways persist over time. Despite the decline of the Brazilian railway system, municipalities connected to the railroads in 1950 have higher per capita income, GDP per capita, population density, and occupational structure concentrated in the manufacturing and services sectors in 2010, compared to unconnected municipalities. In this section, we explore three mechanisms that may explain the persistence of railroad effects in modern times.

Since the railroads increased the shift of workers from agricultural to industry and services, as well as expanded the population density in 2010, thus agglomeration, and urbanization can be potential mechanisms behind this persistence.

2.6.1 Agglomeration

Agglomeration effects are relevant to explain the results found in the previous section. Table 2.11 reports the OLS (panel A) and 2SLS (panel B) estimates for the effects of railroads in 1950 on the log population density between 1960 and 2000. In all columns we include geography, baseline socioeconomic characteristics, and transport controls, as well region fixed effect. The results from the 2SLS estimates indicate that municipalities connected to the railway network in 1950 have a higher population density even 50 years after the beginning of the transportation system's decline.

Table 2.11: Mechanism - Railroads and Population Density, 1960 - 2000

	Log Population Density				
Census:	1960	1970	1980	1991	2000
Panel A: OLS					
Dummy [RR $\leq$ 10 km]	0.329***	0.357***	0.407***	0.472***	0.493***
	[0.038]	[0.043]	[0.047]	[0.047]	[0.049]
Panel B: 2SLS					
Dummy [RR $\leq$ 10 km]	0.728**	1.111***	1.444***	1.595***	1.738***
	[0.283]	[0.324]	[0.366]	[0.385]	[0.401]
Mean Dep. Var.	3.195	3.347	3.444	3.572	3.677
Observations	1,663	1,663	1,663	1,663	1,663
R^2	0.615	0.499	0.418	0.414	0.397
Geography	Yes	Yes	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes	Yes	Yes
Transportation	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the effects of railroads on the log of the population density in 1960 - 2000. All columns report the results from OLS (panel A) and 2SLS (panel B) regressions where the dependent variable log of the population density for the census year defined at the top of each column. The variable of interest is a dummy for the presence of a railroad line near the municipality in 1950. For panel B, the instrumental variable is the log of the percentage of grid points within each municipality that lie on the least-cost paths. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the log of population, the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture and manufacturing computed over the total number of occupied workers. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the municipalities based on the 1950 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality boundaries. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The impact of railroads on the log population density is 0.7 points, for the 1960 census. The magnitude of the effect increases monotonically over time, reaching 1.7 log points in 2000. Despite the decline of the railway system, the municipalities connected by it continued to be locations of population

agglomeration. Much of the agglomeration effect is a consequence of increased internal migration.

Table 2.12: Mechanism - Railroads and Agglomeration, 1970 - 1991

	Share of Migrants		
Census:	1970	1980	1991
Panel A: OLS			
Dummy [RR $\leq$ 10 km]	0.038***	0.052***	0.038***
	[0.010]	[0.009]	[0.008]
Panel B: 2SLS			
Dummy [RR $\leq$ 10 km]	0.015	0.094*	0.091*
	[0.054]	[0.051]	[0.047]
Mean Dep. Var.	0.232	0.287	0.276
Observations	1,663	1,663	1,663
R^2	0.432	0.493	0.472
Geography	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes
Transportation	Yes	Yes	Yes
Region FE	Yes	Yes	Yes

Notes: This table reports the effects of railroads on the share of internal migrants over the total population in 1970 - 1991. All columns report the results from OLS (panel A) and 2SLS (panel B) regressions where the dependent variable is the share of migrants over the total population *1,000 for the census year defined at the top of each column. The variable of interest is a dummy for the presence of a railroad line near the municipality in 1950. For panel B, the instrumental variable is the log of the percentage of grid points within each municipality that lie on the least-cost paths. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the log of population, the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture and manufacturing computed over the total number of occupied workers. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the municipalities based on the 1950 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality boundaries. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.12 shows the effect of the railways in 1950 on the proportion of the population not born in the municipality, in other words, who migrated, for 1970-1991 census years. Although we do not find significant effects for 1970, municipalities connected to the railway system in 1950 have a higher share of the migrant population in 1980 and 1991, an increase of around 9 percentage points in both years, an increase of about 32% in relation to the average proportion of migrants.

To conclude, the agglomeration effects are relevant to explain how, despite the decline in the length of the railroad network, its impact has persisted over time. The municipalities connected to the railroads in 1950 continued to have higher population density, in part because they are places of attraction for people from other municipalities. From the 1980s, when the internal migration process in Brazil intensified, regions that received the railroads seem to have been the biggest recipients of these domestic migrants.

2.6.2 Urbanization

Just as we find an impact of railroads on population density, we find on urbanization. Table 2.13 reports the effect of being connected to the railway system in 1950 on the share of urban population between 1950 and 1991. In all columns, we include geography, socioeconomic characteristics in 1872, transport controls, and region fixed effect. Like the previous tables, panel A presents the OLS estimates, while panel B the 2SLS estimates using the log of the percentage of grid points that lie on the least-cost path as instrumental variable.

Table 2.13: Mechanism - Railroads and Urbanization, 1950 - 1991

	Share of Urban Population				
Census:	1950	1960	1970	1980	1991
Panel A: OLS					
Dummy [RR $\leq$ 10 km]	0.107***	0.019***	0.144***	0.144***	0.135***
	[0.008]	[0.005]	[0.009]	[0.010]	[0.010]
Panel B: 2SLS					
Dummy [RR $\leq$ 10 km]	0.269***	0.088**	0.296***	0.336***	0.274***
	[0.063]	[0.040]	[0.070]	[0.075]	[0.071]
Mean Dep. Var.	0.242	0.425	0.385	0.498	0.599
Observations	1,662	1,663	1,663	1,663	1,663
R^2	0.415	0.687	0.452	0.495	0.489
Geography	Yes	Yes	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes	Yes	Yes
Transportation	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the effects of railroads on the share of people living in urban areas over the total population in 1950 - 1991. All columns report the results from OLS (panel A) and 2SLS (panel B) regressions where the dependent variable is the share of urban population over the total population for the census year defined at the top of each column. The variable of interest is a dummy for the presence of a railroad line near the municipality in 1950. For panel B, the instrumental variable is the log of the percentage of grid points within each municipality that lie on the least-cost paths. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the log of population, the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture and manufacturing computed over the total number of occupied workers. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the municipalities based on the 1950 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality boundaries. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Analyzing panel B, we find a positive and significant impact of the railroads on urbanization between 1950 and 1991. In all cases, the effect is significant at least at 10%. Being connected to the railway system in 1950 increases, on average, between 9 and 34 percentage points the share of urban population, which represents an increase of about 22% and 68% on the

proportion of people living in urban areas. While the biggest impact is on the 1980 census, the smallest is on the 1960 census. In the 1950s and 1970s, the magnitude of the impact is around 27 and 30 percentage points, respectively.

Despite the decline of the railroads since the 1950s, the municipalities connected to the railroad system continued to attract migrants from other municipalities, increasing its population density. This process was accompanied by an increase in urbanization. These mechanisms help to explain why the effects of the railways have persisted over time resulting in differences in income per capita between connected and unconnected municipalities in modern times.

2.7
Are Persistent Effects Driven by Other Transportation Infrastructure?

One concern with the results found so far is that if the railways have been replaced by highways, then the effects found would not be of the persistence of the railways, but simply effects of the construction of new roads. To deal with this concern, in this section we present the main persistence results obtained controlling for the presence of roads. In particular, we include a dummy for the proximity of a road, the presence of a road within a maximum of 10 kilometers from the municipal headquarters, as an additional control. After the 1950s, we have road data available for 1960-2010. Therefore, although this variable is endogenous, it can help us to separate the persistence effects of railways from those of new roads.

In Table 2.14 we test if our results are driven by the presence of major roads. In the first column, we report the 2SLS effects of the railroads in 1950 on long-term economic development, controlling by the determinants of our instrument, the geographical characteristics, the initial socioeconomic characteristics, transportation options, and region fixed-effects. In columns (2)-(7), we report the same effects and, however, controlling for the presence of roads between 1960 and 2010. Each column presents the effect of railways on development, controlling for roads every decade. We present the effects on log income per capita in 2010, however, the results are similar using the other outcomes presented in the article. The inclusion of road controls, despite slightly lowering the point estimates, practically does not change the level of significance of the results. When we do not control by road, we find that railroads in 1950 are associated with a 34% increase in income per capita. However, when we add the new controls, the magnitude of the impact is reduced to between 27% and 32%. Summarising, despite the decline

of Brazilian railroads since the 1950s, their effect has persisted overtime on different measures of economic development. Also, the persistence effects are not driven by the construction of new highways to replace the old railroads.

Table 2.14: Robustness Checks - The 2SLS Effects of Railroads on Long-Term Development Controlling by Roads

	Log of Income Per Capita						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dummy [RR $\leq$ 10 km]	0.341***	0.325***	0.289***	0.275**	0.283**	0.288**	0.302***
	[0.102]	[0.105]	[0.109]	[0.112]	[0.112]	[0.112]	[0.111]
Dummy [Road $\leq$ 10 km] in 1960		0.050**					
		[0.024]					
Dummy [Road $\leq$ 10 km] in 1970			0.088***				
			[0.024]				
Dummy [Road $\leq$ 10 km] in 1980				0.092***			
				[0.024]			
Dummy [Road $\leq$ 10 km] in 1990					0.082***		
					[0.023]		
Dummy [Road $\leq$ 10 km] in 2000						0.074***	
						[0.022]	
Dummy [Road $\leq$ 10 km] in 2010							0.064***
							[0.022]
Mean Dep. Var.	6.155	6.155	6.155	6.155	6.155	6.155	6.155
Observations	1,663	1,663	1,663	1,663	1,663	1,663	1,663
R^2	0.731	0.737	0.753	0.758	0.754	0.752	0.747
KP F-stat	33.431	30.545	28.382	26.569	27.137	27.253	28.526
Geography	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Transportation	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the persistent effects of railroads on development in 2010 controlling for roads between 1970 and 2010. All columns report the results from 2SLS regressions where the dependent variable is the log of the per capita income. The variable of interest is a dummy for the presence of a railroad line near the municipality in 1950. The instrumental variable is the log of the percentage of grid points within each municipality that lie on the least-cost paths. In columns (2)-(7) we add controls for roads, a dummy for the presence of a road near the municipality between 1970 and 2010. All columns include controls for distance to the port, distance to states' capital, land area, slope, and distance to the river. Geographic controls include longitude, latitude, altitude, distance to coast, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The socioeconomics characteristics in 1872 include the log of population, the share of the literate population, percentage of foreigners, percentage of slaves, total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture and manufacturing computed over the total number of occupied workers. The transportation control is the distance to the nearest road in 1867. All regressions estimated for the 1,663 municipalities based on the 1950 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality boundaries. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2.8 Conclusion

Between the middle of the 19th century and the 20th century, there was a great expansion of the Brazilian railway network. However, from the 1950s, with the popularization of automobiles and in the face of the crisis in the railway system, rail transportation entered a long phase of decline, with a large decrease in the length of the network and in the number of stations. In this article, we analyze the persistence of the railroads' effects on economic outcomes, exploring the mechanisms behind this path dependence. We exploit a key feature of the Brazilian railway expansion to document the persistent effects of transportation on economic development: The railroad system's expansion goal, which was to connect the ports to the interior and the state capitals, generated variability in the propensity of municipalities to connect to the system solely because of their geographical position.

We followed almost 60 years of Brazilian railroad expansion and decline to identify the persistent impact of transportation infrastructure on development and agglomeration of the economic activity between the 19th and 21st centuries. We find that railway expansion impacted economic outcomes, even after 60 years of its decline: Municipalities connected to the railroad network before 1950 had an income per capita 34% higher, and a per capita GDP 60% higher in 2010 when compared to not connected municipalities. We also find that these same municipalities have a higher population density and more workers in the manufacturing and services sectors. We show that the effects of railways on structural change have persisted since the 1950s, when municipalities connected to the system were already more industrialized than those not connected. The results are robust to changes in how we measure the connection to the railways, using the linear distance or the number of stations does not alter our conclusions. We also show that the persistence effects on long-term development are not driven by road construction. Finally, we explore the mechanisms that can explain how, despite their decline, the effects of railways on economic outcomes persist until modern times. To do so, we used data from 1872 - 2010 Brazilian censuses, as well as novel historical railway network data.

We documented two mechanisms that can help explain the differences observed in income: First, the railroads had an impact on population agglomeration. Municipalities connected to the railway network before 1950 received migrants in the 1980s and 1990s. Third, despite the decline of the railways, they continued to impact the growth of urbanization in the connected municipalities. Overall, the agglomeration and migration effects are two important

mechanisms that help to explain the current differences in income, despite the decline of the railways.

The results presented by this paper indicate that the effects of market integration due to the expansion of transport infrastructure can persist over time, even when transport technology becomes obsolete. By showing the importance of the agglomeration and migration effects as mechanisms of persistence, we contribute to both the persistence and the urban agglomeration literature.

3
Agrarian Elites, Education, and Long-Term Development[a]

3.1
Introduction

Throughout the 19th and 20th centuries, the expansion of mass education was associated with sustained economic growth (Easterlin, 1981; Galor, 2005; Becker et al., 2011). Indeed, investment in human capital played a key role in the transition from Malthusian stagnation to modern economic growth (Galor and Weil, 2000; Galor, 2005). However, the diffusion of mass public education took place very unevenly between countries (Lindert, 2004; Gallego, 2010) and even between regions of the same country (Go and Lindert, 2010). Two different literatures emerged to explain this phenomenon by associating political power with investment in education. The first establishes the relationship between extending the franchise and the provision of public goods, such as education (Acemoglu and Robinson, 2000b; Engerman et al., 1999; Engerman and Sokoloff, 2002; Gallego, 2010). However, despite the extent of the franchise matters at the national level, it can not explain subnational variations. The second establishes that the identity of political elites matters for the provision of education, insofar as these elites have different incentives (Galor and Moav, 2006; Becker and Hornung, 2019; Corvalan et al., 2020).

In two seminal articles Galor and Moav (2006) and Galor et al. (2009) theoretically show, and provide empirical evidence, that due to the low degree of complementarity between human capital and land, landowners have no incentive to support public schooling institutions. On the other hand, due to the high complementarity in production between physical and human capital, there are incentives for industrialists to support mass education policies. However, the evidence on the relationship between the political power of agricultural elites and the spread of education are not conclusive yet. While some articles find evidence of a negative relationship between the political power of agrarian elites and investment in education[1], others show that in

[a]This chapter is co-authored with Claudio Ferraz (UBC and PUC-Rio).

[1]See, for example, Ramcharan (2010), Vollrath (2013), Cinnirella and Hornung (2016), Tapia and Martinez-Galarraga (2018) and Goñi (2018).

certain circumstances traditional elites can support mass schooling policies[2]. In addition to the mixed evidence, most articles that analyze the relationship between agricultural elites and the expansion of education use land inequality as a proxy for the political importance of these elites[3], which makes the distinction between the impacts of political and economic inequality very difficult.

In this paper, we show the effect of landowners' political power on the spread of mass education at the beginning XX century in the Brazilian state of São Paulo and examine the persistent impact on long-run economic development. More specifically, we digitize a novel dataset on the occupational structure of the voting elites to evaluate the impacts of the agrarian elites' political power in 1905, measured by the share of farmers' voters over the total number of voters, on educational outcomes and economic development. The state of São Paulo represents an excellent setting to document the persistent effects of the distribution of political power across economic elites on mass education and economic development for several reasons. First, at the beginning of the 20th century, the state's economy was concentrated in agriculture[4] and dependent on coffee exports, which made landowners important political players (Dean, 1969). Second, due to restrictive electoral rules, such as the one that only guaranteed the right to vote for literate men over 21 years of age, the control of state policies belonged to economic elites (Love, 1970)[5]. Third, since the municipalities from the state of São Paulo share common national institutions, we can rule out alternative institutional explanations. Fourth, since the industrialization process took place throughout the 20th century, we can analyze the role of the agricultural elite in a context of high complementarity between physical and human capital.

We show that municipalities where the agricultural elites have greater political power in 1905 have a lower literacy rate in the short-term, in 1920. The impact on educational variables persists over time, however, shifting from literacy to higher education throughout the 20th century. We also find this persistence to be associated with lower per capita income in the long-run. Since we exploit variation across municipalities within the same country, the results cannot be explained by differences in other political institutions. We also show that the results are not derived from alternative channels other than

[2]See, for example, Nafziger (2011), Andersson and Berger (2019), and Cvrcek and Zajicek (2019).

[3]Few are the articles that directly measure the political power of landowners. For some exceptions see, for example, Becker and Hornung (2019) and Andersson and Berger (2019).

[4]In 1920 almost 78% of the total occupied workers were in the agriculture sector at the state of São Paulo.

[5]In 1905 only 3.5% of the total population had the right to vote at the state of São Paulo.

the educational ones, such as blocking technological innovation, obstruction of the expansion of transport infrastructure, or labor coercion. Also, we show that the results obtained are due to the agricultural elites' political power, and not the concentration of political power within the elites. To explain the long-term differences in economic outcomes, we analyze the investment in educational inputs throughout the XX century as a mechanism. We find that the agrarian elites' political power is associated with less investment in educational inputs, like schools, teachers, and school attendance. Finally, robustness tests suggest that results are not driven by immigration.

To measure the effects of the farmers' political power on the spread of public education and long-term development, we assemble a novel data set that combines the 1872-2010 Brazilian censuses with political data from the state statistical yearbooks, as well as auxiliary information from historical and contemporaneous sources. To quantify the agrarian elites' political power we use the share of farmers voters over the total number of registered voters in the elections of 1905 as the main explanatory variable. We believe that the proportion of farmers' voters is a good proxy for the political power of this elite because the right to vote was restricted to only less than 4% of the population at the beginning of the 20th century, therefore, the farmers who participated in the political process were most likely those who belonged to the local economic elites. Our main empirical strategy compares educational and economic outcomes across municipalities with different levels of landowners' political power but with similar pre-existing characteristics. Because the political power of farmers is not exogenous, we control for a large set of observable characteristics, we construct control variables that measure the extent of the franchise, geographic, land inequality, transportation, and baseline socioeconomic characteristics. The inclusion of these controls do not change our results.

We start by showing that the political structure of municipalities in 1905 is associated with the literacy rate in 1920. More precisely, an increase in one standard deviation in the share of farmers' voters over the total number of voters is associated with a 6.11 percentage point (or almost 22%) decrease in the literacy rate. The effects on the literacy rate persisted until the 1970s. With the universalization of elementary education, the impact of economic elites on education is no longer reflected in literacy, but on completed years of schooling. Despite the decline in the importance of agricultural elites in the Brazilian political scenario throughout the 20th century, the impact of their decisions on educational public policies persisted over time: an increase in one standard deviation in the share of farmers' voters led to a decrease in 0.14 (or 2.5%) in the years of schooling in 2000. The consequence of the persistent

impact on educational variables is reflected in the current per capita income, an increase in one standard deviation in the landowners' political power in 1905 is associated with decrease in 5.23% in the average per capita income in 2010. We also show that the slow expansion of mass schooling is not associated with political power concentration within local economic elites, but just with landowners' political power.

To ensure that the results are not driven by other factors, we implement a set of robustness tests: (i) we show that the results are robust to controlling for the extent of the franchise, geographic, land inequality, transportation, and baseline socioeconomic characteristics; (ii) we present evidence that the results are not driven by alternative channels other than the educational ones, such as blocking technological innovation, obstruction of the expansion of transport infrastructure, or labor coercion; (iii) the results remain significant when we control for municipality and year fixed-effects; (iv) we present evidence that the results are not driven by immigration.

Lastly, to shed light on the mechanisms behind the persistence of the impacts of the agrarian elites' political power on the spread of mass education and development, we present empirical evidence of the importance of low levels of investment in school inputs as a transmission mechanism. We find that the agrarian elites' political power in 1905 is associated with lower enrollment rates, teachers, and schools per school-aged children. Also, we show that throughout the twentieth century the supply limitation shifted from the inputs of elementary education to the inputs of technical and higher education.

The results found in the article can be interpreted in the light of the theories proposed by the political economy and development literature. Due to the low complementarity of human capital and land at the beginning of the 20th century in an agrarian export economy like Brazil, landowners limited the expansion of educational spending through, among other things, to reduce the mobility of the rural workforce (Galor and Moav, 2006; Galor et al., 2009), to not lose their political power (Bourguignon and Verdier, 2000; Acemoglu and Robinson, 2000b), and to avoid land taxation (Vollrath, 2013; Colistete, 2016). As a consequence, regions, where this elite held greater political power, had lower rates of literacy. Despite the decrease in the political power of agrarian elites throughout the 20th century and the increase in the importance of the industry for the economy of the state of São Paulo (Dean, 1969; Love, 1970), the impact persisted over time, resulting in less economic development in the long-term.

The article is related to three sets of literature. First, it speaks to a

recent literature debating the role of elites in the spread of mass schooling[6]. In particular, this literature has grown a lot with the seminal work of Galor and Moav (2006) and Galor et al. (2009) that establish a theoretical framework for the negative relationship between the landed elites' political power and the spread of mass schooling found in several countries. Given the differences in the degree of complementarity between factors of production and human capital, landed and industrial elites would have opposite incentives to invest in mass education. While the former would use its political influence to block the spread of mass schooling due to the low complementarity between land and human capital[7], the latter would support investment in education because of the high complementarity between physical and human capital in the industrial production. We contribute to this literature in three ways: (i) Unlike most articles that use land inequality as a proxy for the political power of agricultural elites[8], we measure the political power of this elite directly from voter data. Therefore, we were able to measure the impact exclusively from the landowners' political power, and not of a mixture from political and economic inequality; (ii) The empirical evidence of the relationship between agrarian elites and investment in education is still mixed[9]. In this paper, we show that in the context of an agrarian-export and late industrialized economy[10], the political power of agrarian elites is associated with low levels of investment in mass schooling; (iii) We also provide evidence that the effects on education persisted over time, influencing the long-term development process[11].

Second, our results contribute to the more general literature that debates

[6]See, among others, Lindert (2004), Go and Lindert (2010), Ramcharan (2010), Vollrath (2013), Cinnirella and Hornung (2016), Tapia and Martinez-Galarraga (2018) and Goñi (2018).

[7]Or to restrict the mobility of agricultural workers (Galor et al., 2009).

[8]See, for example, Vollrath (2013), Cinnirella and Hornung (2016), Tapia and Martinez-Galarraga (2018), and Goñi (2018).

[9]Although much of the evidence points to a negative relationship between the agrarian elites' political power and the spread of mass schooling, there are some exceptions. See, for example, Nafziger (2011), Cvrcek and Zajicek (2019), and Andersson and Berger (2019).

[10]For an analysis of the impact of political power inequality on education in countries of late industrialization see, for example, Engerman et al. (1999), Engerman and Sokoloff (2002), Chaudhary et al. (2012), Acemoglu et al. (2009), and Musacchio et al. (2014). Distinctly from the first three papers, we analyze a state within Brazil, which makes the jure macro institutions constant. Also, we explore the mechanism of educational persistence throughout the twentieth century, complementing the analysis of Acemoglu et al. (2009), and Musacchio et al. (2014).

[11]For related literature that focuses on Brazil, see de Carvalho Filho and Colistete (2010), Summerhill (2010), and Funari (2017). Our paper differs from this literature by analyzing the specific educational mechanism behind the historical persistence of human capital and its impact on long-run development. Furthermore, unlike the results presented here, Summerhill (2010) finds no effect of agrarian elites' political power on long-term development. Also, Funari (2017) does not find impact of the extension of the franchise on long-term development for some Brazilian states.

the role of elites in the modernization process.[12] Squicciarini and Voigtländer (2015), for example, show that the enlightened elite was an important driver of city growth in Revolutionary France. Ashraf et al. (2018) present evidence that the capital-owning elites supported the end of labor coercion in industrializing Prussia. Tyrefors et al. (2019) analyzing Sweden in the 19th century find that a suffrage reform that shifted the de jure distribution of political power from landowners to industrialists impacted in the reduction of labor coercion and the increase of investment in railways and the adoption of new technologies[13]. Finally, Becker and Hornung (2019) find that higher vote inequality across the three franchise classes in the Prussian parliament is associated with more liberal orientated policies. We contribute to this literature by showing for the Brazilian case that the concentration of political power in the hands of landowners at the beginning of the 20th century affect the development trajectory of the municipalities by blocking the expansion of mass education, and not by other mechanisms such as labor coercion, adoption of new technologies or infrastructure transport projects. Also, we show that is the political power of the agrarian elites that explains the low investment in education, and not the inequality of political power within the economic elites.

Finally, our article is related to the growing literature that analyzes the persistence of human capital over time[14]. Valencia Caicedo (2019) finds a positive effect of Jesuit missions in South America on education and long-term development. The enduring impact is explained by occupational persistence and technology adoption in agriculture. Also, Huillery (2009) shows that early colonial investments in education had large and persistent effects on current outcomes in West Africa. We contribute to this literature, mainly, by showing that the consequences of economic elites' decisions on public educational policies can persist for more than a century. We also bring evidence on the importance of disseminating educational inputs to explain the persistence of human capital over time. In this sense, our paper is related to Rocha et al. (2017) that show the role of the supply of educational inputs to propagate the effects of a temporary human capital shock.

The rest of the paper is organized as follows. Section 3.2 provides the historical background. Section 3.3 presents the data and descriptive statistics.

[12]In addition to this debate, there is also a large literature associating elites' choices with social control (Bourguignon and Verdier, 2000) and the establishment of a national state via the development of civic values (Weber, 1976; Bandiera et al., 2019).

[13]Bogart (2018) also finds that interest groups affected the slow diffusion of infrastructure projects during Britain's first transportation revolution.

[14]See, for example, Wantchekon et al. (2015), and Droller (2018). For Brazil, see de Carvalho Filho and Monasterio (2012), de Carvalho Filho and Colistete (2010), and Rocha et al. (2017).

Section 3.4 describes the empirical strategy. Section 3.5 contains the main results on education and income, as well as a discussion on the mechanism of persistence. Section 3.6 presents the robustness checks. Finally, section 3.7 presents some concluding remarks.

3.2 Historical Background

3.2.1 Historical Context and The Political Elites

In 2017, the GDP of the state of São Paulo represented more than 30% of the total Brazilian GDP, being by far the richest state in the country. The dominance of São Paulo's economy was established in the second half of the 20th century with the intensification of the transition from an agricultural economy to a robust industrial one. The rise of the economy of the state occurs between the second half of the 19th century and the 20th century, with the production of coffee for export as its main factor of economic dynamism. At the beginning of the twentieth century, Brazilian coffee production represented almost 80% of the world's product supply, with São Paulo being responsible for almost 70% of national production (Luna and Klein, 2014). Together with the production of coffee for export, the increase in the flow of immigrants and the investment in the integration of markets with railways, allowed the state to transform its productive structure and emerge as the richest and most industrialized region in the country (Dean, 1969; Saes, 1981; Cano, 1977). By integrating the different regions of the state with the ports, consequently reducing transportation costs and travel time, the railroads generated growth in coffee exports and diversification of the productive structure of the economy[15] (Summerhill, 2003). Also, the growth of immigration solved the problem of reduced labor force supply after the end of slavery (Leff, 1972).

With the introduction of the railroads and the increase in the flow of immigrants, the production of coffee may expand to the West of São Paulo in search of more fertile lands. Landowners used up all the productive potential of the soil and migrated to other locations in search of better yields since lands were cheap relative to capital and labor (Dean, 1969). The expansion of coffee culture took place in a framework of great land inequality, and in which the landowners held great political power. The transition from the monarchic

[15]For the effects of the railways on coffee production and development in Brazil, see Mattoon Jr (1977); Matos (1990); Summerhill (1998, 2005); Grandi (2007); Lamounier (2012).

to the republican regime increased the political power decentralization, from the federal government to the states, expanding the political power of these local elites. The political influence of the landowners grew during much of the first half of the 20th century, only decreasing with the coffee price's crisis and the diversification of the economic structure of the state, with the decline of agriculture and an increase in the relative importance of industries and services (Luna and Klein, 2014; Love, 1970).

At the end of the 19th century, the monarchical regime that had existed until then entered a phase of strong political deterioration. In 1889, a military-political revolution inaugurated the Brazilian republican regime. Despite being peaceful, the transition from a monarchical to a republican regime at the end of the 19th century generated significant changes in the Brazilian political structure. There is an increase in the autonomy of the states and the establishment of regular elections (Luna and Klein, 2014). However, the political influence of regional oligarchies remained. According to Love (1970), the Brazilian constitution of 1891 guaranteed the right to vote for all literate Brazilian males aged twenty-one and older[16]. With the new constitution, there was an increase in the percentage of voters, however, the right to vote remained extremely restricted to a small elite (Love, 1970; Summerhill, 2010). Given the low literacy rate and the exclusion of women from the political process, in 1905 the percentage of voters over the total population was less than 4% in the state of São Paulo. Indeed, just a small elite voted. The franchise rate for Brazil as a whole at the beginning of the 20th century was slightly more than 2%, a rate much lower than that verified at the same time, for example, for the United States, which had 20% of the population voting. Even if we compare with other Latin American countries, the extent of voting in Brazil was concentrated in a small elite. For example, the franchise rate for Costa Rica, Argentina, and Uruguay was around 10% at the beginning of the 20th century, a figure around 5 times higher than the Brazilian one (Engerman and Sokoloff, 2002). In addition to the legal restrictions on voting, there were still several de facto restrictions, which together with the non-secret vote, made the electoral process undemocratic and the target of numerous frauds.

The electoral fraud was more common in landowner control regions, where patriarchal tradition and manipulated votes allowed the local elites to control the state policies (Love, 1970). There is evidence for Chile that elites' control over rural elections were greatly facilitated where balloting was open, implying greater control over the political behavior of workers by the

[16]Despite much debate, the female vote was not allowed. The right to vote for women will only occur in 1932 (Luna and Klein, 2014).

landowners (Baland and Robinson, 2008), which also seems to happen in Brazil (Love, 1970). Despite the restricted franchise and the oligarchic power of rural elites, there was great variation between the municipalities in the political participation of the different segments of society (Summerhill, 2010). And we will explore this variation to analyze the impact of the agrarian elites' political power on educational outcomes and long-term development.

3.2.2
Education and Educational Policies in São Paulo

At the beginning of the 20th century, the schooling rates of Brazilian population were very low, even when compared to other countries in Latin America. In the 1890 census, only 14.8 % of people over 4 years old knew how to read and write, an index comparable only to the poorest countries in Latin America (Luna and Klein, 2014). Although investment in primary education remains the responsibility of state governments, the transition to the republic and the 1891 constitution resulted in the decentralization of the collection of export taxes. Such modification in the tax collection allowed the states to have a considerable increase in their revenues at the turn of the 19th to the 20th century, enabling more effective initiatives for educational expansion. From 1892, municipalities also began to allocate part of their revenues to invest in elementary education (Colistete, 2019). At the beginning of the 20th century, Brazil had the highest growth in school enrollment among Latin American countries (Colistete, 2016). Despite these changes, the Brazilian educational backwardness persisted. In 1907 the enrollment rate of children in primary schools was 29 per 1,000 inhabitants, one of the lowest levels in the world. Countries like Argentina, Uruguay, Peru, and Chile, for example, had enrollment rates two or three times higher than Brazil (de Carvalho Filho and Colistete, 2010). In the 1920 census, the literacy rate in Brazil was still less than 30%, well below Argentina[17], for example. Despite the Brazilian educational backwardness, the state of São Paulo managed to improve its literacy and enrollment rates in the face of economic growth that arose from the coffee exportation boom [18].

Given the economic dynamism of the coffee export sector, and the decentralization of tax collection agreed in the 1891 Constitution, the government

[17]Although Argentina and Uruguay had the best educational indicators in Latin America at the beginning of the 20th century, such rates are still relatively low when taking into account the per capita income of these countries (Lindert, 2010).

[18]For an analysis of the educational statistics and policies in São Paulo between the mid-19th and 20th centuries, see de Carvalho Filho and Colistete (2010) and Colistete (2016, 2019).

of the state of São Paulo and the local councils made more investments in public goods, including in education (Colistete, 2016). Thus, compared to other states in the federation, São Paulo showed great progress in its educational indicators. If in 1889 the enrollment rate in primary education was 6.3% in São Paulo, this rate grows to 31.6% in 1933, an increase of more than 25 percentage points. For Brazil as a whole, the growth in the school enrollment rate was around 16.3 percentage points in the same period, from 7% to 23.3% (Musacchio et al., 2014). Indeed, São Paulo was advancing more than the rest of Brazil in the educational performance, despite still being debauched about Europe and some Latin American countries.

In addition to the increase in income from coffee exports, educational reforms were also responsible for improving the state's educational indicators (Colistete, 2016). In 1892 an educational reform was introduced in the state creating new schools and hiring trained teachers for graded schools (*grupos escolares*), for example. The passage from the 19th to the 20th centuries is also marked by an increase in the immigration flow in São Paulo. These, since they come from countries with better educational indexes, started to demand greater investments in education from the local political elites (Colistete, 2016; Rocha et al., 2017; Witzel de Souza, 2018).

Although the 1891 Constitution made primary education free and universal, it did not specify how the federal units would finance the expansion of the educational system nor did it impose a goal of expanding the educational system. This created a political dilemma for local elites. As demand for educational investments grew, the agrarian elites resisted to the introduction of income and wealth taxes (Colistete, 2016) and to the reallocation of resources to education. It is in the debate over the sources of taxation and the allocation of municipal and state funds that agrarian elites held greater bargaining power to block the advance of primary education since investment in education could only increase if part of their wealth was taxed by the state or if the resource already available were reallocated to education. Despite the educational indicators of the state of São Paulo improved more than the rest of the country due to the economic dynamism of the region, the movement of agricultural elites was in the sense of blocking the expansion of elementary education in order not to have their land and wealth taxed (Colistete, 2016) and not disturb the political structure giving more political voice to workers (Musacchio et al., 2014). The educational improvements in the state were largely due to the increase in immigration (Rocha et al., 2017) and the educational demands of segments of the society (Colistete, 2016; Witzel de Souza, 2018).

3.3 Data

This study combines three sets of historical and modern data. First, we use voters' data from 1905. The electoral data allows us to create measures of political power for elites of different occupations in the early 20th century. Since the right to vote was very restricted, the number of voters by profession gives us a good measure of the political importance of each economic elite. Second, we complement these data with educational and economic information to build human capital and development outcomes from 1920 to 2010.We use census data to calculate, by municipality, the income per capita for 2010. Modern and historical educational measures are built with population and educational censuses. Third, we achieve these data with socioeconomic, geographic, land inequality, and transportation variables. The main data set used in this paper is a cross-section covering 161 municipalities in the state of São Paulo between 1905 and 2010.

3.3.1 Educational and Income Data

Three sources of information are the basis for the construction of the main outcomes of the article. The population censuses from 1920 to 2010 are used to build educational variables both in the past and in the present, allowing us to analyze the impacts of the political power of the agrarian elites on education between 1920 and 2010 and on economic development in 2010 (per capita income). The educational censuses and reports complement the sources of information, which combined allow us to create a set of educational and economic variables for the period 1920-2010.

To follow the municipalities over time, we kept the 1905 border definition, since our variable of interest is only available for 1905. Therefore, we merge the data of the municipalities from the 2010-1920 censuses to match the 1905 census boundaries[19]. Finally, to control for the baseline socioeconomic characteristics, we match the municipalities from 1905 to those that existed in the 1872 census. As result, in the main data set, we follow the 161 municipalities from the state of São Paulo[20] that existed in 1905 in our sample between 1905 and 2010. See Appendix C for details in how we merge the 1872-2010 data

[19]A similar procedure has been used for the United States (Hornbeck, 2012) and Brazil (Rocha et al., 2017). The Brazilian administrative division can be found at https://www.ibge.gov.br/geociencias/organizacaodo-territorio/estrutura-territorial.html.

[20]More precisely, for in the state of São Paulo there were 171 municipalities in 1905. But, 10 municipalities were incorporated by other cities between 1920 and 1960.

censuses to the 1905 administrative division, as well as the definition of the variables used in this paper, and their source and units.

These census records contain, for each municipality, demographic and socioeconomic information. Besides, historical reports and educational censuses contain information on the number of schools and teachers per municipality. The income and population data from the 2010 census are used to calculate the income per capita for each municipality in 2010. The educational data from the 1920-2010 censuses and the educational censuses/reports are used to calculate the following outcome variables at the municipality level: literacy rate, the share of children attending school, number of teachers and schools per aged-school children, the share of people with completed elementary/middle school education, number of elementary/middle, high, technical school teachers and universities' professors over total population, percentage of people with middle, high school or bachelor degree, and the average number of years of schooling. Unfortunately, most of these variables cannot be created for the entire period, only the literacy rate is available in all censuses.

In addition to income per capita income, from the variables mentioned above the following are available for 2000 or 2010: the average years of schooling for people aged 25 years or more, the percentage of people aged 25 years or older with a school degree by educational level, number of teachers over total population aged 25 years or older by educational level, the share of people aged 25 years or older with completed elementary education, the number of schools over total children aged 7-14 years old, the share of children attending school, and the literacy rate for people aged 5 years old or more. From our historical educational variables, the share of people with completed elementary education is only available for 1940-1950, and the 1970-2010 censuses, however, they could only be built differently, the first for people over 10 years of age and the second for people over 25 years old. The number of schools is only available for the 1920 and 1940 censuses, and the number of teachers is available for the 1920 and 1970-2010 censuses. Finally, the share of children attending school is disposal for 1940, 1960, and 1970-2010 censuses. And as noted earlier, just the literacy rate is available for all censuses between 1920 and 2010.

To explore possible alternative channels of the impact of the concentration of power on agricultural elites on development, we additionally collect agricultural data from the 1920 census to compose the following variables: share of farmers with agricultural machinery, coffee production in tons, and wages in construction and agriculture. Also, for the robustness checks, we create the share of foreigners over the total population as an additional control, using data from the 1920 census.

3.3.2 The Voters Data from 1905

The objective of the paper is to analyze the impact of the agrarian elites' political power in 1905 on educational outcomes and long-term development, thus our treatment variables are measures of landowners' political power. The statistical yearbook of the state of São Paulo for 1905 provides data on registered voters by occupation and municipality. The list of occupations includes farmers, artists, clergy, traders, civil servants, industrialists, short-term contract workers, military personnel, and factory workers. As the percentage of the population registered to vote was very low, around 3.5% of the population, these data give us the occupational composition of the political elite of the municipalities. Therefore, in addition to the total percentage of people registered to vote over the total population a measure of the extent of the franchise, we created two variables that represent the political power of the agrarian elites and the political inequality of the municipalities: (i) the percentage of farmers (agricultures) voters over the total number of voters in 1905[21], our main variable of interest; (ii) the Herfindahl index of the share of the voters by occupation: $\sum_{i=1}^{n} S_i^2$, where S_i is the share of voters in the occupation i over the total number of voters.

The first variable captures the relative political importance of the agrarian elite related to the elites of other occupations, while the second variable captures the concentration of the political power across the elites. Higher values of the index correspond to higher political concentration. The closer the index gets to 1, the greater the concentration of political power in a few elites. While the share of farmers voters is our variable of interest, the political concentration index helps us to verify the extent to which the relationship between the political power structure in the early 20th century and the educational/economic outcomes is given by the political power of agricultural elites or by the concentration of political power among local economic elites. As we will show in the results section, the political concentration index helps rule out potential alternative political channels.

3.3.3 Additional Controls

We make use of important sets of socioeconomic baseline controls, as well as geographic and transportation infrastructure controls. The 1872 census

[21]For some municipalities the 1905 data is not available, in this case, we use the 1904 data.

allows us to control for social and economic characteristics of the municipalities at the end of the XIX century. Thus, we construct municipality-level baseline controls: (i) population density; (ii) share of literate individuals aged 6 or more; (iii) share of children attending school; (iv) number of teachers per school-aged child; (v) share of foreign-born people; (vi) share of slaves; (vii) share of workers in public administration; (viii) share of workers in legal professions; (ix) share of workers in agriculture over the total number of occupied workers; (x) share of workers in manufacturing over the total number of occupied workers. Since the occupational structure of voters may be related to the occupational structure of the economy itself, the inclusion of the economic characteristics of the municipalities in 1872 serves to mitigate this type of endogeneity. Also, the inclusion of educational and social controls in the baseline allows us to compare municipalities with similar social characteristics.

The expansion of the economy of the state of São Paulo between the end of the 19th century and the beginning of the 20th century is closely linked to coffee production (Dean, 1969), which depends on appropriate geographical conditions, therefore in our study, we include a set of geographical variables as control: latitude, longitude, altitude, slope, municipality area, soil types, distance to the nearest coast, and distance to the states' capital. To build these variables different sources of information were used, as IBGE, CGIAR-CSI, CPRM, and 1920 population census.

In addition to geographical conditions, land inequality is an important and determinant variable for the expansion of the political power of agrarian elites, which can impact investments in education (see, for example, Galor et al. (2009) and Cinnirella and Hornung (2016)). Therefore, in our analysis, we include the land Gini as a additional control. The land Gini measure is calculated using data on the number of farms and its average size by a given land size interval.[22] The variable are constructed with the 1920 census data and adapted for the 1905 administrative division. As we do not have agricultural data available for 1905, we take the data from 1920 as proxy for the agrarian structure of the municipalities in the early 20th century.

Since the expansion of São Paulo's agrarian elites and the increase in coffee production were associated with the expansion of the state's transportation system, especially the railways (Matos, 1990), we include in our analysis controls for the presence of transportations options in the baseline period. These controls are the distance to the nearest port in 1850, distance to the

[22]The 1920 Census divides the area of rural properties into ten intervals, reporting the number of properties and the average farm size for each of the land size intervals. With this information we can construct the Gini coefficient of the land for each municipality following the same formula as Nunn (2008) and Funari (2017). See Appendix C for details.

nearest road in 1867, distance to the nearest river, and distance to the nearest railway in 1870.

Table 3.1 presents the summary statistics for key variables. The literacy rate in 1920 was almost 30%. Also, just 3.5% of the population was registered to vote in the 1905 elections. The farmers represented more than 60% of the total voters.

Table 3.1: Summary Statistics

	Observations	Mean	S.D.	Min.	Max.
Panel A: Income in 2010					
Log income p.c.	161	6.59	0.24	5.96	7.23
Panel B: Education in 2010, 2000 and, 1920					
% Literate (aged 5+) in 2010	161	93.28	1.87	83.56	96.69
% Children attending school in 2010	161	97.64	0.95	93.77	100.00
% People with bachelor degree in 2010	161	10.09	3.32	2.98	20.54
Years of schooling in 2000 (aged 25+)	161	5.48	0.76	3.34	7.56
% Literate (aged 5+) in 1920	161	28.34	10.08	6.13	67.28
Schools/children (*1,000) in 1920	161	0.29	0.30	0.00	1.54
Teachers/children (*1,000) in 1920	161	5.15	3.63	0.39	22.71
Panel C: Voters, 1905					
% Voters	161	3.52	1.77	0.72	10.87
% Farmers voters	161	62.97	17.45	2.50	90.73
Voters Concentration	161	0.48	0.15	0.18	0.83

Continues in the next page...

Table 3.1: Summary Statistics (cont.)

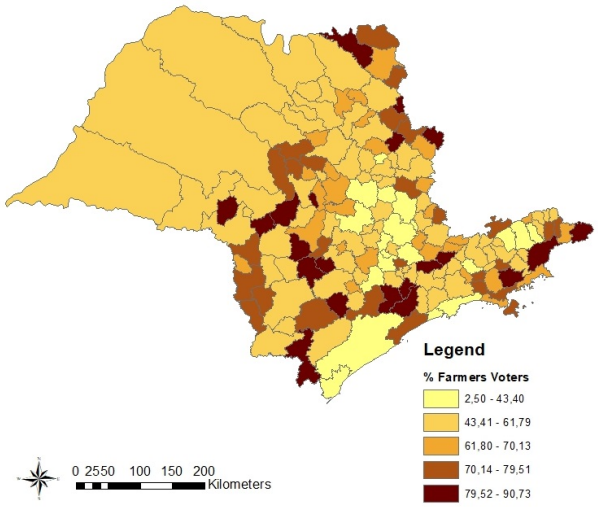
	Observations	Mean	S.D.	Min.	Max.
Panel D: Geography					
Longitude	161	-47.31	1.29	-50.28	-44.32
Latitude	161	-22.70	0.95	-25.01	-20.04
Log altitude	161	6.48	0.32	4.73	7.23
Log distance to coast	161	4.68	1.46	-3.73	6.11
Log distance to state's capital	161	5.01	0.76	0.00	6.01
Log slope	161	1.71	0.53	0.57	2.88
Log area	161	6.69	0.93	4.62	10.66
% Cambisol	161	10.84	22.06	0.00	89.31
% Latosol	161	46.81	37.48	0.00	100.00
% Argisol	161	36.95	35.48	0.00	100.00
% Spondosol	161	1.74	6.28	0.00	52.23
Panel E:Land Inequality					
Land Gini in 1920	161	0.66	0.11	0.31	0.90
Panel F: Transportation Controls					
Log distance to port, 1850	161	5.01	0.91	0.00	6.14
Log distance to road, 1867	161	2.71	1.50	0.00	5.00
Log distance to river	161	2.08	1.42	-2.18	4.19
Log distance railway, 1870	161	4.71	1.23	-3.72	5.94
Panel G: Socioeconomic Characteristics, 1872					
Density	161	7.36	6.17	0.15	28.61
% Literate (aged 6+)	161	20.02	10.23	4.88	46.55
% Children attending school	161	14.87	10.92	2.66	76.36
Teachers/children (*1000)	161	4.35	3.57	0.00	20.07
% Foreigners	161	0.01	0.02	0.00	0.08
% Slaves	161	0.16	0.09	0.04	0.53
Public Administration (in 1,000)	161	1.19	1.07	0.00	4.67
Legal Professionals (in 1,000)	161	0.92	0.74	0.00	5.24
% Emp. agriculture s	161	0.59	0.11	0.35	0.91
% Emp. manufacturing	161	0.11	0.05	0.02	0.24

Notes: Descriptive statistics for the main variables used in the paper. Population and education data from population and education censuses. GNP data from IBGE. Political data in 1905 from *Anuário Estatístico do Estado de São Paulo, 1904 and 1905*. Geographic controls created using ArcGIS with data originally from IBGE, CGIAR-CSI, and Embrapa. Land inequality control from the 1920 censues. Controls for socioeconomic characteristics in 1872 from population census. Transportation controls created using ArcGIS with data originally from IBGE, ANA, and historical reports. In all panels the sample is based on the 1905 municipality boundaries. For data specific descriptions and sources, see the Appendix C.

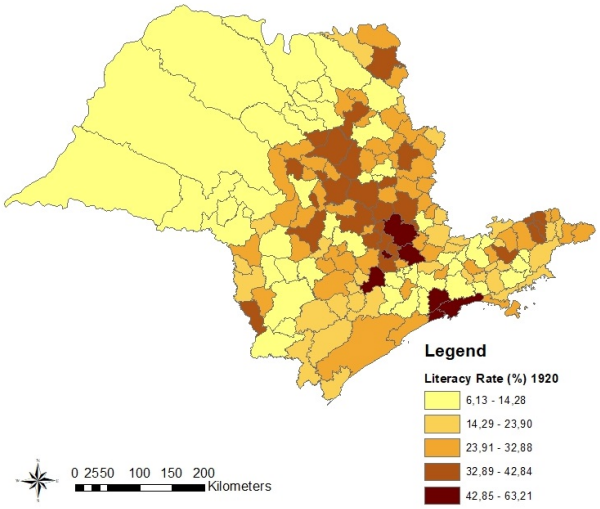
Figure 3.1 presents the spatial distribution of the agrarian political power in 1905 and the literacy rate in 1920, with darker colors reflecting the higher share of farmers voters (Figure 3.1(a)) or higher literacy rates (Figure 3.1(b)). When we compare these figures, we find that their distributions are almost opposite, municipalities where agricultural elites have greater political power are the places with a lower rate of literacy. We explore this difference in political inequality and the expansion of mass schooling across municipalities in our empirical strategy.

Figure 3.1: Agricultural Elites in 1905 and Literacy Rate in 1920

(3.1(a)) % Farmers Voters



(3.1(b)) Literacy Rate (%), 1920



Notes: The figures show the spatial distribution of the farmers' political power in 1905 and the literacy rate (%) in 1920 for the municipalities of the state of São Paulo. See the main text and Appendix C for data sources and details on the construction of the variables.

3.4 Empirical Strategy

Our main goal is to analyze the impact of the agrarian elites' political power on the evolution of educational outcomes over time and long-run development. Our conceptual framework suggests that differences in capital-skill complementarity across the economic sectors would induce variations in the support for education investment between the agrarian and industrial elites. While the accumulation of human capital by the working class at the industrialization process will favor the industrial elites, it will impair the landowners due to the increase of the rural labor force mobility. Therefore the political power of agrarian elites would be associated with the lower spread of elementary education (Galor and Moav, 2006; Galor et al., 2009). Additionally, in a scenario in which landowners have great political power, the concentration of political power may be associated with a reduction in economic development, because the state's capture by the interests of the agrarian elites.

To empirically test the effect of the agrarian elite's political power on education and economic development, we first exploit variations in a cross-section of São Paulo municipalities using data from the 1920-2010 censuses. Specifically, we compare municipalities with different levels of political power concentration, but with similar pre-1905 characteristics, and its relationship with educational and economic outcomes. We examine the short and long-term impact of the agrarian elite political power on education and development following the standard OLS framework:

$$Y_i = \alpha + \beta AgPower_{i1905} + \lambda Voters_{i1905} + G_i'\sigma + L_i'\pi + T_i'\mu + S_i'\gamma + \epsilon_i \quad (3\text{-}1)$$

where Y_i is the educational or economic outcome of interest in the municipality i measured in the censuses years between 1920 and 2010. $AgPower_{i1905}$ is our interest variable of elites' political power in 1905, which will be measured by the share of the agrarian elites' voters, the percentage of farmers (agriculturers) voters over the total number of voters.

Since the percentage of voters was very low at the beginning of the 20th century, the percentage of farmers (agriculturers) voters captures the political power of the agrarian elites. $Voters_{i1905}$ is the overall share of the population registered to vote in 1905, capturing the extent of the franchise by the municipality. Since the expansion of the economy of the state of São Paulo was highly dependent on coffee exports between the mid-19th and 20th centuries,

geographic characteristics are important covariates to be included in our model, and they are represented by the vector G_i in our equation 3-1. The geographic controls include latitude, longitude, altitude, slope, municipality area, presence of different types of soil, distance to the nearest coast, and distance to the states' capital. There is evidence that land inequality is an important factor in explaining the spread of elementary education in several countries (Engerman and Sokoloff, 2002; Galor et al., 2009), so we will control for L_i, the land Gini for 1920. The infrastructure conditions can also be important determinants of both the economic and educational expansion of the municipalities, so we control for the transport infrastructure vector T_i that it includes distances to the nearest port, road, river, and railway line in the middle XIX century. Finally, S_i is a vector of socioeconomic characteristics measured at the baseline in 1872, allowing us to control by pre-existing social and economic conditions that can impact both the concentration of political power and educational outcomes. The variables of this vector include population density, literacy rate, share of children attending school, number of teachers, share of foreign-born people, share of slaves, public administration workers, share of workers in legal professions, and share of workers in the agriculture and manufacturing sectors.

The regression is estimated with the 1905 administrative division, which contains 161 municipalities. Since these municipalities were originally unit of the less fragmented 1872 administrative division, used in the construction of our baseline socioeconomic characteristics, we report the standard errors clustered at the 1872 aggregated geographic division to account for possible interdependence of error terms across municipalities. The identification hypothesis is that conditional to the franchise, geographic, socioeconomic characteristics, transportation, and land inequality controls, the distribution of the political power of the agrarian elites is not correlated with any unobserved determinants of educational and development outcomes. In this case, β captures the causal impact of the political power concentration on education and long-term development. A discussion on the capacity of our estimator to recover the causal effect and possible alternative channels of the impact of the concentration of political power will be made in the next section. Also, we discuss further concerns on our identification hypothesis in the robustness section of the paper.

A potential concern about the specification from equation 3-1 is that there might be unobservable characteristics affecting both the agrarian elites' political power and the spread of mass schooling. In this case, our estimator would be capturing effects other than those associated with the impact of rural elites' political power on investment in education. One way to mitigate the endogeneity arising from the unobservable characteristics fixed over time

is to use panel data and control for municipality fixed-effect. Therefore, we build an alternative data set based on the 1872 municipalities' boundaries and pool censuses data from 1920-2010 to measure the impact of the agrarian elites' political power on education, controlling for municipality and year fixed-effects. Similar to Hornbeck (2012) and Rocha et al. (2017), our estimating equation is:

$$Y_{it} = \alpha + \beta_t AgPower_{i1905} + \lambda_t Voters_{i1905} + G_i'\sigma_t + L_i'\pi_t + T_i'\mu_t + S_i'\gamma_t + \eta_i + \phi_t + \epsilon_{it} \quad (3\text{-}2)$$

where Y_{it} is our educational outcome in the municipality i measured in the census year t. The variables η_i and ϕ_t represent the municipality and year fixed-effects. The variables $Voters_{i1905}$, G_i, L_i, T_i and S_i are, respectively, controls for franchise, geography, land inequality, transportation infrastructure and socioeconomic characteristics measured at the baseline in 1872. All controls are interacted with year fixed-effects to allow for differential trends across municipalities. $AgPower_{i1905}$ is our interest variable of agricultural elites' political power in 1905. ϵ_{it} is an error term. Finally, β_t is interpreted as the average difference in the outcome between municipalities with different levels of landowners' political power in a given year t relative to 1872. The regression is estimated with the 1872 administrative division, which contains 87 municipalities. The standard errors are clustered at the municipality level. For 1920-2010, the municipalities are aggregated into the original 87 municipalities from 1872. From our educational outcomes, just the literacy rates can be compared across all censuses between 1872 and 2010, thus the estimation is limited to this variable.

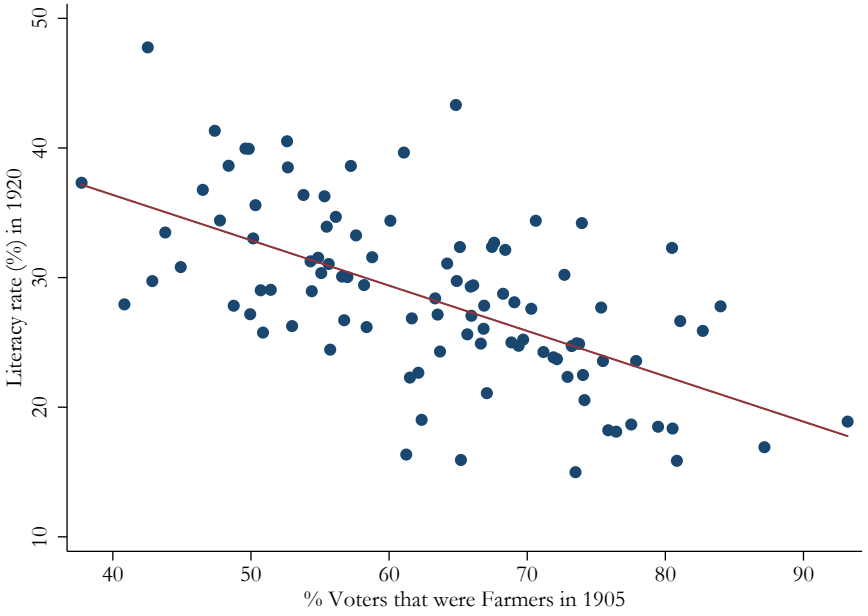
3.5 Main Results

In this section, we report the effects of the agrarian elites' political power in 1905 on educational outcomes in 1920 in municipalities from the state of São Paulo. We complement the analysis examing the medium and long-term effects of the elites' political power, showing the human capital persistence between 1920 and 2010, and the impacts on economic development in 2010.

3.5.1
Political Elites and Education in 1920

We start the analysis examining the effects of agrarian elites' political power in 1905 on education in 1920. We use the literacy rate, the percentage of literate people aged 5 years old or more over the total population aged 5 years or more, as the dependent variable. Before presenting the results, we illustrate the relationship between the landowners' political power and the literacy rate in 1920 through Figure 3.2. The graph plots the literacy rate in 1920 versus the percentage of voters that were farmers in 1905. As we can see, there is a strong negative relationship between the political power of agrarian elites and the educational outcome in 1920. As we will show in the regressions this relationship, besides being negative, is statistically significant.

Figure 3.2: Education in 1920 and Agricultural Elites' Political Power in 1905



Notes: The figure displays the non-parametric relationship between literacy rate (%) in 1920 and the percentage of farmers' voters over the total number of voters in the municipality in 1905, conditional on the % of voters, geography, land inequality, transportation and baseline economic controls. Observations are sorted into 100 bins of equal size and the dots indicate the mean value in each group. The red line presents the linear fit line.

Table 3.2 reports the results from estimated OLS regressions coefficients, based on equation 3-1 and using the percentage of farmers voters over the total number of registered voters as the variable of interest. In column (1), we present the results for the specification without any control. Our elites' political

power measure is negative and significant, showing that municipalities, where agrarian elites had more political power, have lower literacy rates.

Table 3.2: The Effects of Agricultural Elites' Political Power on Education, 1920

	Literacy rate (%), 1920					
	(1)	(2)	(3)	(4)	(5)	(6)
% Farmers Voters	-0.370***	-0.359***	-0.364***	-0.353***	-0.358***	-0.350***
	[0.047]	[0.051]	[0.049]	[0.049]	[0.050]	[0.054]
Mean Dep. Var.	28.344	28.344	28.344	28.344	28.344	28.344
Observations	161	161	161	161	161	161
R^2	0.411	0.412	0.526	0.532	0.559	0.621
% Voters	No	Yes	Yes	Yes	Yes	Yes
Geography	No	No	Yes	Yes	Yes	Yes
Land Inequality	No	No	No	Yes	Yes	Yes
Transport	No	No	No	No	Yes	Yes
Characteristics 1872	No	No	No	No	No	Yes

Notes: This table reports the effects of elites' political power on the literacy rate in 1920. All columns report the results from OLS regressions where the dependent variable is the municipality percentage of literate people aged 5+ years over total population aged 5+ years in 1920. The variable of interest is the percentage of farmers voters over the total number of voters in the municipality in 1905. In columns (2)-(6), we control for the overall share of the population registered to vote. Geographic controls include longitude, latitude, altitude, slope, land area, distance to the nearest coast, distance to the states' capital, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The land inequality control is the land Gini from the 1920 Census. The transportation control includes the distance to the nearest railroad in 1870, the distance to the nearest port in 1870, the distance to the nearest road in 1867, and the distance to the nearest river. The socioeconomics characteristics in 1872 include the share of the population density, literacy rate, share of children attending school, number of teachers per school-aged child, percentage of foreigners, percentage of slaves, the total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture, manufacturing over the total number of occupied workers. All regressions estimated for the 161 municipalities based on the 1905 administrative boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The following columns test the consistency of the results by adding, sequentially, a set of control variables. In column (2), we control for the franchise, adding the percentage of registered voters over the total population in 1905 as control. The inclusion of the new control does not alter the statistical significance of the estimated coefficient nor its magnitude, which practically does not change. Column (3) reports the effect including geographic characteristics as controls. Again, the inclusion of new controls does not change our results.

In column (4) we control for land inequality. The estimated coefficient hardly changes with the inclusion of the land Gini control. In columns (5) and (6) we add controls, respectively, for transportation infrastructure and socioeconomic characteristics in 1872. The inclusion of these controls does not significantly change the estimated coefficient.

The results showed in Table 3.2 indicate that the negative relationship between the political power of the agrarian elites is not derived from geographical, land inequality, transportation infrastructure, or socioeconomic conditions in the baseline. In our most complete specification, column (6), an increase in 1 percentage point in the share of farmers voters decreases the literacy rate in 0.35 percentage point, or an increase in 1 standard deviation in the share of farmers voters led to a 6.11 percentage point decrease in the literacy rate, almost 22% reduction given the average of 28.34% in 1920. The result indicates a negative impact of agricultural elites' political power in 1905 on the literacy rate in 1920.

In Table 3.3, we explore other potential effects of landed elites' political power concentration that has been featured in the literature. The idea is to analyze whether the results found in Table 3.2 are not the consequence of alternative channels. For example, Acemoglu and Robinson (2000a) argues that political elites, fearful of losing their power, can block technological innovations. Indeed, Tyrefors et al. (2019) find a negative impact on the political power of landed elites in the adoption of labor-saving technologies in agriculture and investments in railways for XIX century Sweden. Bogart (2018) also finds effects of political interest groups on the diffusion of transportation infrastructure for XVIII century Britain. Finally, the political power concentration can also be associated with control over workers' outside options and labor coercion, resulting in lower wages for workers (Tyrefors et al., 2019).

Table 3.3: Alternative Channels: Political Elites, Agriculture and, Infrastructure in 1920

	Alternative Channels				
	% Farms with machinery	Log coffee product.	Log wages construc.	Log wages agric.	Log distance railway
	(1)	(2)	(3)	(4)	(5)
% Farmers Voters	0.0002	-0.0016	-0.0016	-0.0007	-0.0053
	[0.001]	[0.003]	[0.001]	[0.001]	[0.012]
Mean Dep. Var.	0.223	-0.908	1.995	1.369	0.858
Observations	161	153	114	93	161
R^2	0.408	0.340	0.529	0.757	0.306
% Voters	Yes	Yes	Yes	Yes	Yes
Geography	Yes	Yes	Yes	Yes	Yes
Land Inequality	Yes	Yes	Yes	Yes	Yes
Transport	Yes	Yes	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the effects of elites' political power on agriculture and infrastructure characteristics in 1920. All columns report the results from OLS regressions where the dependent variable is listed in the top of the columns. The variable of interest is the percentage of farmers voters over the total number of voters in the municipality in 1905. In all columns, we control for the overall share of the population registered to vote. Geographic controls include longitude, latitude, altitude, slope, land area, distance to the nearest coast, distance to the states' capital, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The land inequality control is the land Gini from the 1920 Census. The transportation control includes the distance to the nearest railroad in 1870, the distance to the nearest port in 1870, the distance to the nearest road in 1867, and the distance to the nearest river. The socioeconomics characteristics in 1872 include the share of the population density, literacy rate, share of children attending school, number of teachers per school-aged child, percentage of foreigners, percentage of slaves, the total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture, manufacturing over the total number of occupied workers. All regressions estimated for the 161 municipalities based on the 1905 administrative boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

We find no evidence that the vote share of landed elites affected the adoption of labor-saving technologies in the agriculture, Table 3.3 column (1), the impacts of our elites' political power measures on the percentage of farms with machinery in 1920 are not significant. Also, the elites' political power is not associated with coffee productivity by hectare, column (2). Thus, the block on the diffusion of new technologies in agricultural production can't explain our results. We also find no evidence of the labor coercion hypothesis, there is no impact of the elites' political power on rural workers' wages in construction and agriculture (columns (3)-(4)). Finally, we find no evidence of agrarian

elites' political power on the distance to the nearest railway in 1920. Thus, the impact channel does not appear to be for blocking infrastructure projects. Overall, the impact of the elites' political power on literacy in 1920 can't be explained by these alternative channels.

While some articles point to a negative relationship between the political power of agricultural elites and the expansion of mass education (Galor et al., 2009; Cinnirella and Hornung, 2016; Tapia and Martinez-Galarraga, 2018), others suggest that it is the concentration of political power within elites, or the inequality of political power, that is responsible for the slow expansion of elementary education (Acemoglu et al., 2009). In Table 3.4 we present the impact of political power in the hands of agricultural elites on the literacy rate in 1920 by adding a Herfindahl index of political concentration among voting elites and the interaction between our main variable of interest and that index as control variables. Our measure of political concentration is the Herfindahl index of the share of the voters by occupation[23]. In all columns, we control for the extent of the franchise, geography, land inequality, transportation, and baseline socioeconomics characteristics. Additionally, in column (2) we include the political concentration index, and in column (3) we include this index and the interaction term between the political power of agrarian elites and the political power concentration index.

As we can see from Table 3.4, column (2), the inclusion of the index of political power concentration does not alter either the sign or the statistical significance of the coefficient of agricultural elites' political power. Indeed, the point estimate increases from 0.350 to 0.503. The estimated coefficient for political concentration is positive and significant at 10%, indicating that the concentration of political power within the elites is associated with an increase in the literacy rate. In column (3), the inclusion of the interaction term practically does not change the coefficient of landed elites. However, the political concentration index is no longer significant. The interaction term, although negative, is also not significant. Overall, the results indicate that the slow expansion of mass schooling is associated with the landowners' political power, and not with the political power concentration within economic elites. Therefore, the degree of competition among economic elites does not explain our results. The degree of concentration of political power in the hands of agricultural elites is the relevant variable to explain the different paths of expansion of elementary education.

[23]Or, $\sum_{i=1}^{n} S_i^2$, where S_i is the share of voters in the occupation i over the total number of voters.

Table 3.4: Alternative Channels: Political Power Concentration in 1905

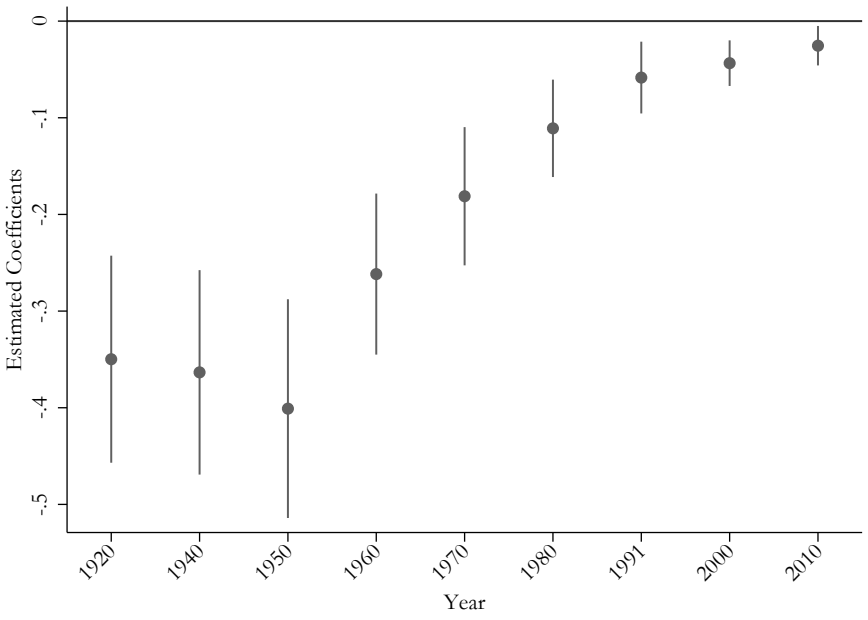
	Literacy rate (%), 1920		
	(1)	(2)	(3)
% Farmers Voters	-0.350***	-0.503***	-0.467***
	[0.054]	[0.082]	[0.125]
Political Concentration		18.621*	35.931
		[10.004]	[29.093]
% Farmers Voters x Political Concentration			-0.183
			[0.337]
Mean Dep. Var.	28.344	28.344	28.344
Observations	161	161	161
R^2	0.621	0.629	0.630
% Voters	Yes	Yes	Yes
Geography	Yes	Yes	Yes
Land Inequality	Yes	Yes	Yes
Transport	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes

Notes: This table reports the effects of elites' political power on the literacy rate in 1920. All columns report the results from OLS regressions where the dependent variable is the municipality percentage of literate people aged 5+ years over total population aged 5+ years in 1920. The main variable of interest is the percentage of farmers voters over the total number of voters in the municipality in 1905. While in column (2) we add the Herfindahl index of the share of the voters by occupation in 1905, a proxy for political power concentration, in column (3) we add the interaction term between the share of farmers voters and the political power concentration index. In all columns, we control for the overall share of the population registered to vote. Geographic controls include longitude, latitude, altitude, slope, land area, distance to the nearest coast, distance to the states' capital, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The land inequality control is the land Gini from the 1920 Census. The transportation control includes the distance to the nearest railroad in 1870, the distance to the nearest port in 1870, the distance to the nearest road in 1867, and the distance to the nearest river. The socioeconomics characteristics in 1872 include the share of the population density, literacy rate, share of children attending school, number of teachers per school-aged child, percentage of foreigners, percentage of slaves, the total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture, manufacturing over the total number of occupied workers. All regressions estimated for the 161 municipalities based on the 1905 administrative boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.5.2
The Human Capital Persistence

Despite the significant changes in the Brazilian economic structure, with increased urbanization and industrialization, and the revolution in the political system in the 1930s, the effects of the landed elites' political power on educational outcomes persisted over time. Figure 3.3 reports the persistence of the effect of the agrarian elites' political power (% of farmers voters over total registered voters in 1905) on the literacy rate between 1920 and 2010. Each point on the graph represents the estimated coefficient of equation 3-1 in our preferred specification, controlling for the proportion of people registered to vote, geography, land inequality, transportation, and baseline socioeconomic characteristics.

Figure 3.3: Human Capital Persistence: Literacy rate (%), 1920-2010



Notes: The figures show the effects of elites' political power on the literacy rate in 1920-2010. All coefficients present the results from OLS regressions where the dependent variable is the municipality percentage of literate people aged 5+ years over total population aged 5+ years in 1920-2010. The variable of interest is the percentage of farmers voters over the total number of voters in the municipality in 1905. We include the overall share of the population registered to vote, geographic, land inequality, transportation, and socioeconomic characteristics in 1872 as controls. All regressions estimated for the 161 municipalities based on the 1905 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality level.

If an increase in 1 percentage point in the share of farmers voters decreases the literacy rate in 0.35 percentage point in 1920, the magnitude of the impact increased between the 1940s and 1950s, and decreased from the 1960s, although

it remains negative and significant. An increase in 1 standard deviation in the share of farmers voters in 1905 led to a 6.11 percentage point decrease in the literacy rate in 1920, 6.28 in 1940, 6.98 in 1950, 4.54 in 1960, 3.14 in 1970, 1.92 in 1980, 1.05 in 1991, 0.70 in 2000, and 0.35 in 2010. The magnitude of the impact decreases as literacy becomes almost universal. Although the political power of landed elites has decreased over time[24] (Font, 1983), the effects of the concentration of political power on agricultural elites in the early 20th century on the literacy rate persisted until the 21st century.

One potential deficiency of the results presented below is that the political power of the agricultural elites and the spread of education across the municipalities from the state of São Paulo can both be correlated with unobservable characteristics biasing the estimated coefficients. In Table 3.5, we present results from an alternative specifications, based on equation 3-2, that controls for unobservable characteristics that are fixed over time. To do so, we build a panel dataset covering the original 87 municipalities across ten censuses between 1872 and 2010. The estimated coefficient gives us the impact of the agrarian elites' political power on education over time relative to the educational outcome in the baseline, 1872. We restrict the analysis to literacy rate since is the only educational outcome that can be compared across censuses.

In Table 3.5, the interest variable is the percentage of farmers voters over the total number of voters interacted by year dummies. In column (1) we only control for the municipality and year fixed-effects, the coefficients estimated suggest that an increase in one percentage point in the share of farmers voters decreases the literacy rate in 0.32 percentage point in 1920, 0.44 in 1940, 0.44 in 1950, 0.30 in 1960 and 0.18 in 1970 relative to the base year of 1872. After the 1970's the effects are not significant, indicating a convergence of literacy rates between municipalities. To check whether the results are robust to trends in initial characteristics, in columns (2)-(6) we present the results including gradually year fixed effects interacted with the share of voters in 1905, geographic characteristics, land inequality measures, transportation infrastructure, and baseline socioeconomic characteristics. In general, the inclusion of the new controls increases the magnitude of the estimated coefficients, however, practically not changing their statistical significance.

[24]For an analysis of the São Paulo political elite between 1889 and 1937, see Love and Barickman (1986).

Table 3.5: The Effects of Agricultural Elites' Political Power on Education: Panel-Data Specifications, 1872-2010

	Literacy rate (%), 1872-2010					
	(1)	(2)	(3)	(4)	(5)	(6)
% Farmers Voters x 1920	-0.323***	-0.322***	-0.402***	-0.393***	-0.403***	-0.411***
	[0.063]	[0.074]	[0.090]	[0.089]	[0.090]	[0.101]
% Farmers Voters x 1940	-0.436***	-0.425***	-0.462***	-0.430***	-0.433***	-0.429***
	[0.073]	[0.080]	[0.102]	[0.099]	[0.100]	[0.123]
% Farmers Voters x 1950	-0.436***	-0.425***	-0.443***	-0.413***	-0.417***	-0.431***
	[0.082]	[0.089]	[0.115]	[0.110]	[0.110]	[0.138]
% Farmers Voters x 1960	-0.298***	-0.286***	-0.312***	-0.296***	-0.315***	-0.339**
	[0.083]	[0.089]	[0.110]	[0.108]	[0.107]	[0.132]
% Farmers Voters x 1970	-0.175**	-0.156*	-0.200**	-0.190*	-0.210**	-0.227*
	[0.076]	[0.083]	[0.100]	[0.098]	[0.098]	[0.121]
% Farmers Voters x 1980	-0.091	-0.069	-0.081	-0.073	-0.098	-0.125
	[0.069]	[0.075]	[0.093]	[0.091]	[0.092]	[0.114]
% Farmers Voters x 1991	-0.033	-0.022	-0.019	-0.012	-0.035	-0.056
	[0.067]	[0.073]	[0.087]	[0.085]	[0.086]	[0.108]
% Farmers Voters x 2000	0.021	0.033	0.015	0.018	-0.006	-0.032
	[0.064]	[0.070]	[0.083]	[0.081]	[0.080]	[0.098]
% Farmers Voters x 2010	0.050	0.062	0.033	0.034	0.012	-0.018
	[0.064]	[0.070]	[0.081]	[0.080]	[0.080]	[0.097]
Observations	870	870	870	870	870	870
Municipalities	87	87	87	87	87	87
R^2	0.967	0.967	0.976	0.977	0.979	0.982
Municipality fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
% Voters	No	Yes	Yes	Yes	Yes	Yes
Geography	No	No	Yes	Yes	Yes	Yes
Land Inequality	No	No	No	Yes	Yes	Yes
Transport	No	No	No	No	Yes	Yes
Characteristics 1872	No	No	No	No	No	Yes

Notes: This table reports the effects of elites' political power on the literacy rate. All columns report the results from OLS regressions where the dependent variable is the municipality percentage of literate people aged 5+ years over total population aged 5+ years. The variable of interest is the percentage of farmers voters over the total number of voters in the municipality in 1905. All specifications include municipality fixed effects and year fixed effects. In column (2)-(6), we control for the overall share of the population registered to vote. Geographic controls include longitude, latitude, altitude, slope, land area, distance to the nearest coast, distance to the states' capital, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The land inequality control is the land Gini from the 1920 Census. The transportation control includes the distance to the nearest railroad in 1870, the distance to the nearest port in 1870, the distance to the nearest road in 1867, and the distance to the nearest river. The socioeconomics characteristics in 1872 include the share of the population density, share of children attending school, number of teachers per school-aged child, percentage of foreigners, percentage of slaves, the total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture, manufacturing over the total number of occupied workers. The controls are all interacted with year dummies. All regressions estimated for the 87 municipalities based on the 1872 census boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In the more saturated specification in column (6), an increase in one percentage point in the share of farmers voters decreases the literacy rate in 0.41 percentage point in 1920, 0.43 in 1940, 0.43 in 1950, 0.34 in 1960 and 0.23 in 1970 relative to the base year of 1872. After the 1970's the effects are not significant. Overall, these results corroborate the cross-section results presented in Table 3.2 and Figure 3.3: the agrarian elites' political power in 1905 led to a decrease in the literacy rates in the short and medium terms. In the long-run, with the universalization of literacy (in 2010 almost 94% of the population was literate, see Table 3.1), the effects are no longer significant or become very small.

The effects of the elites' political structure in the municipalities of the state of São Paulo in the early twentieth century on educational outcomes over time are not limited to the literacy rate, there are also effects on the percentage of people who completed the elementary and middle schools. In Table 3.6 we report the impact of the agrarian elites' political power in 1905 on the percentage of people with completed elementary and middle school education between 1940 and 2010. In all specifications, we control for the share of voters in 1905, geographic characteristics, land inequality, transportation infrastructure, and baseline socioeconomics characteristics. We find negative and significant effects of the concentration of political power of landed elites on the share of people with completed elementary/middle school education. An increase in one percentage point in the share of farmers voters decreases the percentage of people with completed elementary/middle school education at 0.19 percentage point in 1940, 0.34 in 1950, 0.12 in 1970, 0.13 in 1980, 0.17 in 1991 and 0.17 in 2010. Contrary to the results found for literacy rates, the effects on primary education seem to remain over time with an increase in the estimated effects. As there was a universalization of literacy, the difference between municipalities, in the long-run, seems to have displaced to completed educational degrees. Therefore, in addition to affecting the literacy rates in the short and medium terms, the political power of the agrarian elites in 1905, also influenced the higher educational level acquired by the population.

Table 3.6: The Effects of Agricultural Elites' Political Power on Elementary Education, 1940-2010

	% Completed Elementary/Middle Education (aged 10+)		% Completed Elementary/Middle Education (aged 25+)			
	1940	1950	1970	1980	1991	2010
% Farmers Voters	-0.187***	-0.335***	-0.117***	-0.128***	-0.166***	-0.169***
	[0.023]	[0.041]	[0.022]	[0.024]	[0.031]	[0.042]
Mean Dep. Var.	7.174	20.034	8.547	12.351	22.386	48.973
Observations	161	161	161	161	161	161
R^2	0.654	0.693	0.572	0.622	0.614	0.620
% Voters	Yes	Yes	Yes	Yes	Yes	Yes
Geography	Yes	Yes	Yes	Yes	Yes	Yes
Land Inequality	Yes	Yes	Yes	Yes	Yes	Yes
Transport	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the effects of agricultural elites' political power on completed grade. All columns report the results from OLS regressions where the dependent variable is the percentage of people aged 10+ with the elementary/middle school diploma (1940-1950). In 1970-2010 the dependent variable is the percentage of people aged 25+ with the elementary/middle school diploma. The variable of interest is the percentage of farmers voters over the total number of voters in the municipality in 1905. In all columns, we control for the overall share of the population registered to vote. Geographic controls include longitude, latitude, altitude, slope, land area, distance to the nearest coast, distance to the states' capital, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The land inequality control is the land Gini from the 1920 Census. The transportation control includes the distance to the nearest railroad in 1870, the distance to the nearest port in 1870, the distance to the nearest road in 1867, and the distance to the nearest river. The socioeconomics characteristics in 1872 include the share of the population density, literacy rate, share of children attending school, number of teachers per school-aged child, percentage of foreigners, percentage of slaves, the total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture, manufacturing over the total number of occupied workers. All regressions estimated for the 161 municipalities based on the 1905 administrative boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.5.3
Persistence Mechanism: The Supply of Educational Inputs

In this subsection we explore mechanisms that can explain the persistent effect of agrarian elites over time. We focus on the effects on educational inputs: the proportion of children attending school, and investment in educational infrastructure, such as teachers and the schools. In Table 3.7 we report the impact of the agricultural elites' political power in 1905 on the percentage of children attending elementary/middle school over the total number of children between 1940 and 2010.

Table 3.7: The Effects of Agricultural Elites' Political Power on School Attendance, 1940-2010

	% Children Attending School						
	1940	1960	1970	1980	1991	2000	2010
% Farmers Voters	-0.381***	-0.213***	-0.131***	-0.086**	-0.059**	-0.025***	0.002
	[0.064]	[0.044]	[0.037]	[0.042]	[0.028]	[0.009]	[0.007]
Mean Dep. Var.	34.510	45.998	74.262	73.884	84.578	96.224	97.636
Observations	161	161	161	161	161	161	161
R^2	0.612	0.556	0.558	0.549	0.598	0.502	0.172
% Voters	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geography	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Land Inequality	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Transport	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the effects of agricultural elites' political power on school attendance. All columns report the results from OLS regressions where the dependent variable is the percentage of children aged 7-14 years (or 5-14 years) attending school over the total number of children aged 7-14 years (or 5-14 years), for 1940-2010. The variable of interest is the percentage of farmers voters over the total number of voters in the municipality in 1905. In all columns, we control for the overall share of the population registered to vote. Geographic controls include longitude, latitude, altitude, slope, land area, distance to the nearest coast, distance to the states' capital, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The land inequality control is the land Gini from the 1920 Census. The transportation control includes the distance to the nearest railroad in 1870, the distance to the nearest port in 1870, the distance to the nearest road in 1867, and the distance to the nearest river. The socioeconomics characteristics in 1872 include the share of the population density, literacy rate, share of children attending school, number of teachers per school-aged child, percentage of foreigners, percentage of slaves, the total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture, manufacturing over the total number of occupied workers. All regressions estimated for the 161 municipalities based on the 1905 administrative boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

For almost all available census years, we find negative and significant effects at least 5% for both of our variables of interest. Only for 2010, we do not find any significant results, perhaps due to the universalization of basic education. An increase in one standard deviation in the agrarian political power in 1905 led to a 6.65 percentage points decrease in the share of children attending elementary school in 1940 (which represents a decrease in 19% compared to the average percentage of children attending school). Although it remains significant, the magnitude of the effect is reduced over time. An increase in one standard deviation in the agrarian political power in 1905 led to a 0.44 percentage point decrease in the share of children attending elementary school in 2000.

The increase in the number of children attending school was followed by an expansion in the number of teachers per school-aged child. In Table 3.8 we

show the effects of the agricultural elites' political power in 1905 on the number of teachers of elementary/middle school over the total number of aged-school children in 1920 and between 1970 and 2010 (unfortunately, we don't have data for 1940-1960).

Table 3.8: The Effects of Agricultural Elites' Political Power on Number of Teachers, 1920-2010

	Teachers per School Aged Child					
	1920	1970	1980	1991	2000	2010
% Farmers Voters	-0.126***	-0.111	-0.095	-0.011	-0.023	-0.062
	[0.020]	[0.072]	[0.079]	[0.072]	[0.042]	[0.136]
Mean Dep. Var.	5.148	29.252	28.085	37.375	12.131	52.051
Observations	161	161	161	161	161	161
R^2	0.565	0.292	0.255	0.296	0.150	0.401
% Voters	Yes	Yes	Yes	Yes	Yes	Yes
Geography	Yes	Yes	Yes	Yes	Yes	Yes
Land Inequality	Yes	Yes	Yes	Yes	Yes	Yes
Transport	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the effects of agricultural elites' political power on the number of teachers. All columns report the results from OLS regressions where the dependent variable is the number of teachers (elementary and middle schools) over the total number of children aged 7-14 years (or 6-14 years) *1,000, for 1920-2010. The variable of interest is the percentage of farmers voters over the total number of voters in the municipality in 1905. In all columns, we control for the overall share of the population registered to vote. Geographic controls include longitude, latitude, altitude, slope, land area, distance to the nearest coast, distance to the states' capital, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The land inequality control is the land Gini from the 1920 Census. The transportation control includes the distance to the nearest railroad in 1870, the distance to the nearest port in 1870, the distance to the nearest road in 1867, and the distance to the nearest river. The socioeconomics characteristics in 1872 include the share of the population density, literacy rate, share of children attending school, number of teachers per school-aged child, percentage of foreigners, percentage of slaves, the total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture, manufacturing over the total number of occupied workers. All regressions estimated for the 161 municipalities based on the 1905 administrative boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Although since 1970 we find no more significant effects, we find negative and significant effects for 1920. An increase in 1 percentage point in the share of farmers voters in 1905 decreases the share of teachers in 0.13 percentage

point in 1920, or an increase in 1 standard deviation in the share of farmers voters led to a 2.20 percentage point decrease in the number of teachers, a reduction of more than 42% given the average 5.15 teachers per school-aged child in 1920. Since the 1970s, we find no more significant impact. We can conclude, therefore, that after the 1970s, there seems to be a convergence in the supply of educational inputs for elementary and middle schools between municipalities in the state of São Paulo. This convergence is consistent with the fact that the effect on the literacy rate is no longer significant in the medium term.

In addition to the impact on the number of teachers, we also find effects on the number of schools per school-aged child. Table 3.9 reports the effect of the agricultural elites' political power in 1905 on the number of schools over the total number of aged-school child in 1920, 1940 and 2010. Although we do not find any significant effects for 1940 and 2010, we find for 1920. An increase in 1 percentage point in the share of farmers voters in 1905 led to a decrease in 0.007 in the number of schools per aged-school child for 1920, or an increase in 1 standard deviation in the share of farmers voters led to a 0.12 decrease in the number of schools (a reduction of 41.5%). While the effects on the proportion of children attending school are greatly reduced in the medium term, the effects on the supply of teachers and schools are no longer significant. We note, therefore, that from the medium-term onward there is a convergence in the supply of inputs for elementary education.

Table 3.9: The Effects of Agricultural Elites' Political Power on Number of Schools, 1920-2010

	Schools per School Aged Child		
	1920	1940	2010
% Farmers Voters	-0.007***	-0.014	0.015
	[0.002]	[0.010]	[0.012]
Mean Dep. Var.	0.294	5.708	7.074
Observations	161	161	161
R^2	0.350	0.319	0.490
% Voters	Yes	Yes	Yes
Geography	Yes	Yes	Yes
Land Inequality	Yes	Yes	Yes
Transport	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes

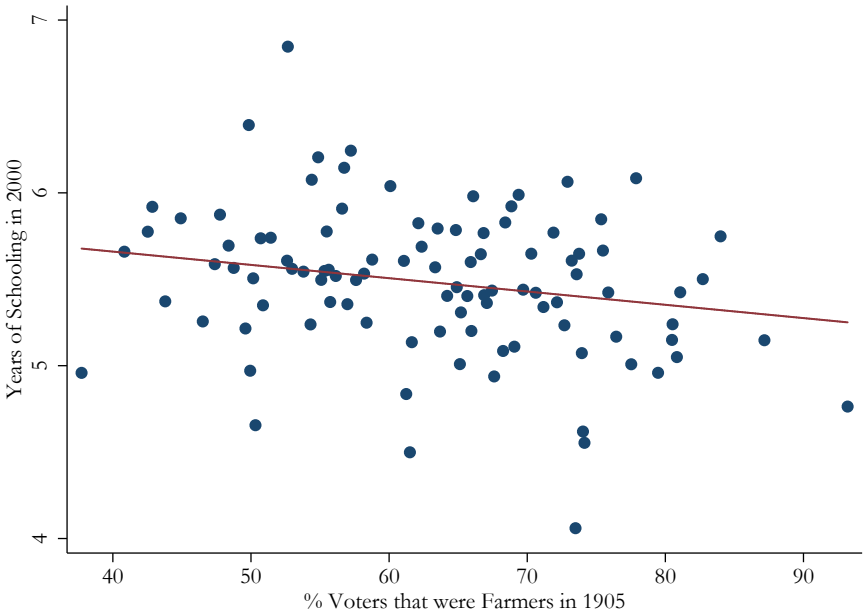
Notes: This table reports the effects of agricultural elites' political power on the number of schools. All columns report the results from OLS regressions where the dependent variable is the number of schools (elementary, middle and high schools) over the total number of children aged 7-14 years (or 6-14 years) *1,000, for 1920-2010. The variable of interest is the percentage of farmers voters over the total number of voters in the municipality in 1905. In all columns, we control for the overall share of the population registered to vote. Geographic controls include longitude, latitude, altitude, slope, land area, distance to the nearest coast, distance to the states' capital, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The land inequality control is the land Gini from the 1920 Census. The transportation control includes the distance to the nearest railroad in 1870, the distance to the nearest port in 1870, the distance to the nearest road in 1867, and the distance to the nearest river. The socioeconomics characteristics in 1872 include the share of the population density, literacy rate, share of children attending school, number of teachers per school-aged child, percentage of foreigners, percentage of slaves, the total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture, manufacturing over the total number of occupied workers. All regressions estimated for the 161 municipalities based on the 1905 administrative boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

As we analyzed in this section, the effects of the political power of agricultural elites on educational outcomes were negative. The political power of agrarian elites is associated with a decrease in the literacy rate in 1920. The effects persisted in both the short and medium-term. In the long run, given the universalization of literacy and the spread of educational inputs, such as attendance, teachers, and schools (Tables 3.7-3.9), the effect persisted no longer on the literacy rate, but on other educational measures like completed elementary/middle education (Table 3.6). In the next subsection, we look at the consequences of this persistence in the long-term, both for educational and economic outcomes.

3.5.4
The Long-Term Effects: Education and Income

Figures 3.4 and 3.5 illustrate, respectively, the results found of the consequences of the political power of agrarian elites in 1905 over the years on education in 2000 and log of per capita income in 2010.

Figure 3.4: Modern Education and Agricultural Elites' Political Power in 1905

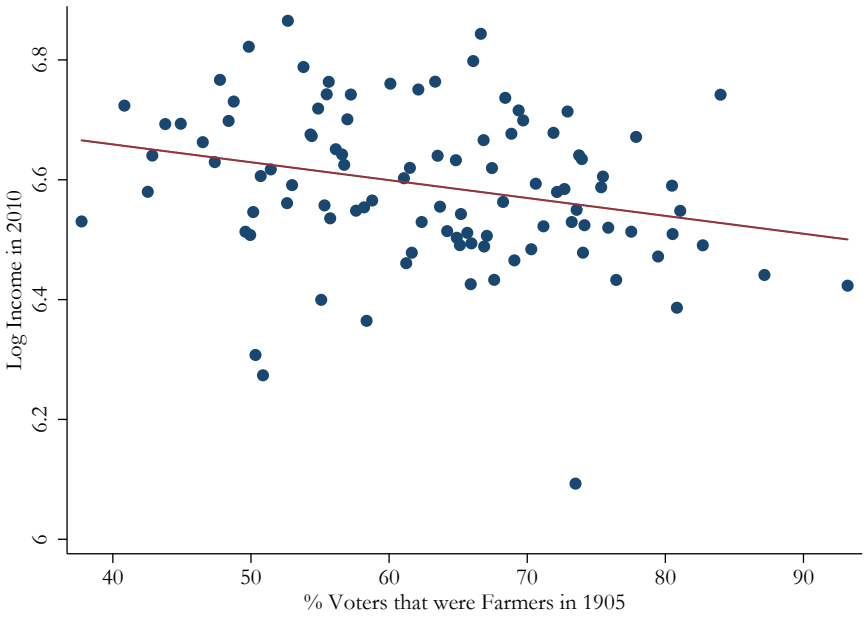


Notes: The figure displays the non-parametric relationship between years of schooling in 2000 and the percentage of farmers' voters over the total number of voters in the municipality in 1905, conditional on the % of voters, geography, land inequality, transportation and baseline economic controls. Observations are sorted into 100 bins of equal size and the dots indicate the mean value in each group. The red line presents the linear fit line.

In both cases, we find a negative and significant relationship, indicating

the persistence of educational effects, and its consequence on per capita income in modern times. These negative relationships that we will explore in this subsection.

Figure 3.5: Modern Income and Agricultural Elites' Political Power in 1905



Notes: The figure displays the non-parametric relationship between log income per capita in 2010 and the percentage of farmers' voters over the total number of voters in the municipality in 1905, conditional on the % of voters, geography, land inequality, transportation and baseline economic controls. Observations are sorted into 100 bins of equal size and the dots indicate the mean value in each group. The red line presents the linear fit line.

Table 3.10 reports the long-term effects of the agrarian elites' political power in 1905 on the educational outcomes in 2000 and 2010. First, we find a negative and significant coefficient in years of schooling, an increase in one standard deviation in the agrarian political power in 1905 led to a 0.14 decrease in the years of schooling in 2000, a reduction of 2.5% of the average years of schooling. Second, we find significant negative effects on primary, secondary, and higher education: the percentage of the population with the middle, high school, or college degrees. An increase in one standard deviation in the agrarian political power led to a reduction, respectively, in 2.95, 2.69 and 1.17 in the share of people with middle school, high school, and bachelor degrees, a reduction of 6.02%, 8.02% and 11.59% of the average share of completed educational degree. The long-term consequences of the political power structure in 1905 are greater for college education than for elementary/middle since the last one have been practically universalized, and

due to the spread of educational inputs for elementary/middle education as we showed in the last subsection.

Table 3.10: The Long-Term Effects: Political Elites and Education in 2000 and 2010

	Education in 2000 and 2010			
	Years of schooling	% With middle school degree	% With high school degree	% With bachelor degree
% Farmers Voters	-0.008**	-0.169***	-0.154***	-0.067***
	[0.003]	[0.042]	[0.038]	[0.016]
Mean Dep. Var.	5.484	48.973	33.492	10.088
Observations	161	161	161	161
R^2	0.665	0.620	0.600	0.603
% Voters	Yes	Yes	Yes	Yes
Geography	Yes	Yes	Yes	Yes
Land Inequality	Yes	Yes	Yes	Yes
Transport	Yes	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes	Yes

Notes: This table reports the effects of elites' political power on education in 2000 and 2010. All columns report the results from OLS regressions where the dependent variable is listed in the top of the columns. The variable of interest is the percentage of farmers voters over the total number of voters in the municipality in 1905. In all columns, we control for the overall share of the population registered to vote. Geographic controls include longitude, latitude, altitude, slope, land area, distance to the nearest coast, distance to the states' capital, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The land inequality control is the land Gini from the 1920 Census. The transportation control includes the distance to the nearest railroad in 1870, the distance to the nearest port in 1870, the distance to the nearest road in 1867, and the distance to the nearest river. The socioeconomics characteristics in 1872 include the share of the population density, literacy rate, share of children attending school, number of teachers per school-aged child, percentage of foreigners, percentage of slaves, the total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture, manufacturing over the total number of occupied workers. All regressions estimated for the 161 municipalities based on the 1905 administrative boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

While the effects of the political power of agrarian elites in 1905 on the literacy rate persisted until the 1970s (Table 3.5), in the long run, the effect was reversed in differences in years of schooling and the proportion of people with completed primary, secondary and higher school education. In Table 3.11, and taking into account the results found in the Tables 3.7-3.9, we show that this change in impact between educational outcomes is due to a convergence in the supply of educational inputs for primary education and persistence in the

difference for technical and university educational inputs. Table 3.11 reports the long-term effects of the agrarian elites' political power in 1905 on the number of teachers (or professors) over the total number of people aged 25 years old or more in 2010 by educational stage.

Table 3.11: The Long-Term Effects: Political Elites and Educational Inputs in 2010

	Educational Inputs in 2010: Teachers and Professors			
	Elementery/Middle school teachers	High school teachers	Technical school teachers	University professors
% Farmers Voters	-0.005	0.013	-0.005*	-0.020***
	[0.031]	[0.010]	[0.002]	[0.007]
Mean Dep. Var.	11.012	3.423	0.296	0.884
Observations	161	161	161	161
R^2	0.370	0.333	0.292	0.249
% Voters	Yes	Yes	Yes	Yes
Geography	Yes	Yes	Yes	Yes
Land Inequality	Yes	Yes	Yes	Yes
Transport	Yes	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes	Yes

Notes: This table reports the effects of elites' political power on the number of teachers and professors in 2010. All columns report the results from OLS regressions where the dependent variables are the share of teachers and professors over the total number of people aged 25 years old or more *1,000. The variable of interest is the percentage of farmers voters over the total number of voters in the municipality in 1905. In all columns, we control for the overall share of the population registered to vote. Geographic controls include longitude, latitude, altitude, slope, land area, distance to the nearest coast, distance to the states' capital, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The land inequality control is the land Gini from the 1920 Census. The transportation control includes the distance to the nearest railroad in 1870, the distance to the nearest port in 1870, the distance to the nearest road in 1867, and the distance to the nearest river. The socioeconomics characteristics in 1872 include the share of the population density, literacy rate, share of children attending school, number of teachers per school-aged child, percentage of foreigners, percentage of slaves, the total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture, manufacturing over the total number of occupied workers. All regressions estimated for the 161 municipalities based on the 1905 administrative boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

There is no impact of the political power of agrarian elites in 1905 on the number of elementary or high school teachers (columns 1 and 2). However, the effects on technical school teachers and university professors are negative and significant at least at 10% (columns 3 and 4). The results suggest that there is

a convergence in the supply of inputs for primary and secondary education, but the negative effects on the supply of inputs for technical and higher education persist, explaining the negative effect found on the years of schooling in the Table 3.10.

In Table 3.12 we show the impact of the agrarian elites' political power in 1905 on income per capita in 2010.

Table 3.12: The Long-Term Effects: Political Elites and Economic Development, 2010

	Log Income p.c., 2010					
	(1)	(2)	(3)	(4)	(5)	(6)
% Farmers Voters	-0.007***	-0.006***	-0.004***	-0.004***	-0.004***	-0.003***
	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
Mean Dep. Var.	6.591	6.591	6.591	6.591	6.591	6.591
Observations	161	161	161	161	161	161
R^2	0.232	0.253	0.602	0.615	0.630	0.669
% Voters	No	Yes	Yes	Yes	Yes	Yes
Geography	No	No	Yes	Yes	Yes	Yes
Land Inequality	No	No	No	Yes	Yes	Yes
Transport	No	No	No	No	Yes	Yes
Characteristics 1872	No	No	No	No	No	Yes

Notes: This table reports the effects of elites' political power on economic development in 2010. All columns report the results from OLS regressions where the dependent variable is listed in the top of the columns. The variable of interest is the percentage of farmers voters over the total number of voters in the municipality in 1905. In columns (2)-(6), we control for the overall share of the population registered to vote. Geographic controls include longitude, latitude, altitude, slope, land area, distance to the nearest coast, distance to the states' capital, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The land inequality control is the land Gini from the 1920 Census. The transportation control includes the distance to the nearest railroad in 1870, the distance to the nearest port in 1870, the distance to the nearest road in 1867, and the distance to the nearest river. The socioeconomics characteristics in 1872 include the share of the population density, literacy rate, share of children attending school, number of teachers per school-aged child, percentage of foreigners, percentage of slaves, the total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture, manufacturing over the total number of occupied workers. All regressions estimated for the 161 municipalities based on the 1905 administrative boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

As a consequence of lower investments in education, we find that municipalities where agrarian elites had more political power in 1905 have lower income per capita in 2010. In addition to being negative and significant, the results are stable with the introduction of new controls. The introduction of

controls for land inequality, transport infrastructure, and baseline characteristics hardly changes the magnitude of the estimated coefficient. In our most robust specification, column (6), a one percentage point increase in the proportion of farmers over total voters generates a 0.3% decrease in average per capita income in 2010 (or one standard deviation increase led to a 5.23% decrease in average per capita income in 2010). Overall, we find significant and negative effects of the political structure of the early 20th century on economic development more than a century later, as a result of the low educational investment made by the economic elites.

3.6 Robustness Check

In this section, we provide evidence on the robustness of the main results of the paper. In particular, we show that our main results are robust to controlling for the share of immigrants in 1920. Despite the socioeconomic controls for 1872 already include the share of immigrants, immigration could explain the evolution of mass schooling over time. In addition to being more educated, there is evidence that immigrants demanded higher investment in education and that this had implications for the long-term development (Colistete, 2016; Rocha et al., 2017; Witzel de Souza, 2018). If the political power structure of local elites is correlated with the decisions of immigrants about their municipality of residence, then part of the effects obtained in our results may be due to immigration, and not to the political power of the elites. Therefore, the inclusion of the percentage of immigrants in the total population in 1920, despite being endogenous, helps us to verify whether our results come from a correlation between the political power structure of the municipalities and the inflow of immigrants over time.

To address the concern that migratory flows can explain our main results, we include the share of immigrants in 1920 as an additional control variable. Table 3.13 reports the impact of agricultural elites' political power on literacy rate in 1920, years of schooling in 2000, and log per capita income in 2010. In all columns, we add controls for franchise extension, geography, land inequality, transport, and baseline socioeconomic characteristics. Also, in columns (2), (4), and (5) we include the share of foreigners in 1920. Although the inclusion of the new control reduces the magnitude of the estimated effect, the coefficients remain statistically significant. Therefore, even controlling for the percentage of immigrants in 1920, we continue to verify that the agrarian elites' political power in 1905 is associated with a decrease in the literacy rate in 1920, a reduction in years of schooling in 2000, and with lower per capita income in

2010. Overall, the estimates obtained in our main results do not appear to be derived from differences in migratory flows between municipalities.

Table 3.13: Robustness Checks: Controlling for Immigration in 1920

	Literacy Rate, 1920		Years of Schooling, 2000		Log Income p.c., 2010	
	(1)	(2)	(3)	(4)	(5)	(6)
% Farmers Voters	-0.350***	-0.284***	-0.008**	-0.007*	-0.003***	-0.002*
	[0.054]	[0.058]	[0.003]	[0.003]	[0.001]	[0.001]
% Foreigners in 1920		0.397***		0.007		0.007***
		[0.121]		[0.006]		[0.002]
Mean Dep. Var.	28.344	28.344	5.484	5.484	6.591	6.591
Observations	161	161	161	161	161	161
R^2	0.621	0.651	0.665	0.667	0.669	0.684
% Voters	Yes	Yes	Yes	Yes	Yes	Yes
Geography	Yes	Yes	Yes	Yes	Yes	Yes
Land Inequality	Yes	Yes	Yes	Yes	Yes	Yes
Transport	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics 1872	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the effects of elites' political power on educational and economic outcomes in 1920 and 2010. All columns report the results from OLS regressions where the dependent variable is listed in the top of the columns. The variable of interest is the percentage of farmers voters over the total number of voters in the municipality in 1905. In columns (2), (4), and (6) we add the share of foreigners in 1920 as an additional control. In all columns, we control for the overall share of the population registered to vote. Geographic controls include longitude, latitude, altitude, slope, land area, distance to the nearest coast, distance to the states' capital, and percentage of municipality area covered by different types of soil - cambisol, latosol, argisol, spondosol. The land inequality control is the land Gini from the 1920 Census. The transportation control includes the distance to the nearest railroad in 1870, the distance to the nearest port in 1870, the distance to the nearest road in 1867, and the distance to the nearest river. The socioeconomics characteristics in 1872 include the share of the population density, literacy rate, share of children attending school, number of teachers per school-aged child, percentage of foreigners, percentage of slaves, the total number of workers in public administration and legal professions relative to the total population, the share of workers in agriculture, manufacturing over the total number of occupied workers. All regressions estimated for the 161 municipalities based on the 1905 administrative boundaries. Standard errors are given in brackets and are clustered at the 1872 municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.7 Conclusion

Throughout the 20th century, the state of São Paulo became the richest region in Brazil. This development process took place through the transition from an agricultural to an industrialized economy, in which the production of coffee for export had a decisive role since it led to the accumulation of wealth, the growth of urbanization, the improvement of transport infrastructure, the displacement of the agricultural frontier and the creation of a monetized economy based on free labor (Furtado, 1959; Dean, 1969). Along with the economic

growth of the state, agricultural producing elites have become important political actors on the national scene, influencing national and regional public policies, including those on the expansion of public education (Love and Barickman, 1986; Colistete, 2016). Given the importance of landowners for the economic development of the state of São Paulo, we exploit differences in the political power of agricultural elites among municipalities in the state in the early 20th century to document the short, medium and long-term effects of political power concentration on the expansion of mass schooling.

We use data from ten Brazilian censuses between 1872 and 2010, as well as many educational and socioeconomic data from various historical sources, to analyze the impact of the agrarian elites' political power on educational outcomes in the short/medium terms and its persistence over time. The results show that an increase in one standard deviation in the share of farmers voters in 1905 led to a decrease in 6.11 percentage points in the literacy rate in 1920, which represents a reduction of almost 22% in relation to the average rate for the year. We have also shown that this difference in literacy rates is not the result of blocking technological innovations, limiting the diffusion of infrastructure projects, or labor coercion by landowners. Finally, we bring evidence that the negative impact on the literacy rate is due only to the agrarian elites' political power, and not by the concentration of political power within the economic elites.

Despite the change in the economic structure of the state of São Paulo and the decrease in the political importance of landowners, the effects on literacy persisted in the medium term until the 1970s. From the 1970s onwards, differences started to focus on other educational outcomes like the percentage of the population with some completed educational stage or years of schooling. We find that an increase in one standard deviation in the share of farmers voters in 1905 led to a decrease in 0.14 in the years of schooling in 2000, which represents a reduction of 2.5% about the average years of schooling for the year. The consequence of this persistent impact on educational outcomes is that the political power of agricultural elites in the early 20th century is associated with a lower level of economic development today: one standard deviation increase in the landowners' political power in 1905 led to a 5.23% decrease in average per capita income in 2010. The evidence suggests that municipalities, where the agrarian elite had more political power, invested less in education, resulting in fewer children attending school and fewer teachers and schools per school aged-child. Over time and with the universalization of elementary education, the persistence of impact has shifted from literacy to the number of years of schooling.

The overall results underline the importance of human capital investment's decisions by economic elites and their consequences on long-term economic development. The interest groups that control the state structure may favor different political choices that have long-term consequences. In the particular case of the state of São Paulo, landowners, to maintain the political status quo, did not invest in public education, resulting in lower levels of development in the long-run.

Bibliography

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *American economic review*, 91(5):1369–1401.

Acemoglu, D., Johnson, S., and Robinson, J. A. (2002). Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution. *The Quarterly journal of economics*, 117(4):1231–1294.

Acemoglu, D. and Robinson, J. A. (2000a). Political Losers as a Barrier to Economic Development. *The American Economic Review*, 90(2):126–130.

Acemoglu, D. and Robinson, J. A. (2000b). Why did the West extend the franchise? Democracy, inequality, and growth in historical perspective. *The Quarterly Journal of Economics*, 115(4):1167–1199.

Acemoglu, D., Rubin, P. Q., and Robinson, J. A. (2009). Economic and Political Inequality in Development: The Case of Cundinamarca, Colombia. *Institutions and Economic Performance*.

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2015). The Economics of Density: Evidence From the Berlin Wall. *Econometrica*, 83(6):2127–2189.

Alvarez-Cuadrado, F. and Poschke, M. (2011). Structural change out of agriculture: Labor push versus labor pull. *American Economic Journal: Macroeconomics*, 3(3):127–58.

Andersson, D., Berger, T., and Prawitz, E. (2020). Making a market: Infrastructure, integration, and the rise of innovation.

Andersson, J. and Berger, T. (2019). Elites and the expansion of education in nineteenth-century Sweden. *The Economic History Review*, 72(3):897–924.

Andrabi, T. and Kuehlwein, M. (2010). Railways and Price Convergence in British India. *The Journal of Economic History*, 70(2):351–377.

Ashraf, Q. H., Cinnirella, F., Galor, O., Gershman, B., and Hornung, E. (2018). Capital-Skill Complementarity and the Emergence of Labor Emancipation.

Atack, J. (2013). On the Use of Geographic Information Systems in Economic History: The American Transportation Revolution Revisited. *The Journal of Economic History*, 73(2):313–338.

Atack, J., Bateman, F., Haines, M., and Margo, R. A. (2010). Did Railroads Induce or Follow Economic Growth? Urbanization and Population Growth in the American Midwest, 1850–1860. *Social Science History*, 34(2):171–197.

Atack, J., Haines, M. R., and Margo, R. A. (2008). Railroads and the Rise of the Factory: Evidence for the United States, 1850-70. Technical report, National Bureau of Economic Research.

Baland, J.-M. and Robinson, J. A. (2008). Land and Power: Theory and Evidence from Chile. *American Economic Review*, 98(5):1737–65.

Bandiera, O., Mohnen, M., Rasul, I., and Viarengo, M. (2019). Nation-building through compulsory schooling during the age of mass migration. *The Economic Journal*, 129(617):62–109.

Becker, S. O. and Hornung, E. (2019). The Political Economy of the Prussian Three-class Franchise.

Becker, S. O., Hornung, E., and Woessmann, L. (2011). Education and Catch-up in the Industrial Revolution. *American Economic Journal: Macroeconomics*, 3(3):92–126.

Berger, T. (2019). Railroads and Rural Industrialization: evidence from a Historical Policy Experiment. *Explorations in Economic History*, 74:101277.

Berger, T. and Enflo, K. (2017). Locomotives of local growth: The short- and long-term impact of railroads in Sweden. *Journal of Urban Economics*, 98:124–138.

Birchal, S. (1999). *Entrepreneurship in Nineteenth-century Brazil: The Formation of a Business Environment*. Springer.

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

Bogart, D. (2018). Party Connections, Interest Groups and The Slow Diffusion of Infrastructure: Evidence From Britain'S First Transport Revolution. *The Economic Journal*, 128(609):541–575.

Bourguignon, F. and Verdier, T. (2000). Oligarchy, democracy, inequality and growth. *Journal of development Economics*, 62(2):285–313.

Brasil (1850). *Relatório do Ministério da Fazenda*. Typ. Nacional.

Brooks, L. and Lutz, B. (2019). Vestiges of Transit: Urban Persistence at a Microscale. *Review of Economics and Statistics*, 101(3):385–399.

Bustos, P., Caprettini, B., and Ponticelli, J. (2016). Agricultural Productivity and Structural Transformation: Evidence from Brazil. *American Economic Review*, 106(6):1320–65.

Cano, W. (1977). *Raízes da concentração industrial em São Paulo*, volume 53. Difel São Paulo.

Chaudhary, L., Musacchio, A., Nafziger, S., and Yan, S. (2012). Big BRICs, weak foundations: The beginning of public elementary education in Brazil, Russia, India, and China. *Explorations in Economic History*, 49(2):221–240.

Cinnirella, F. and Hornung, E. (2016). Landownership concentration and the expansion of education. *Journal of Development Economics*, 121:135–152.

CMBEU (1954). *Comissão Mista Brasil - Estados Unidos para o Desenvolvimento Econômico: Relatório Geral*. Rio de Janeiro.

Colistete, R. P. (2016). O atraso em meio à Riqueza: Uma História Econômica da Educação Primária em São Paulo, 1835 a 1920.

Colistete, R. P. (2019). Contando o Atraso Educacional: Despesas e Matrículas na Educação Primária de São Paulo (1880-1920). *Dados*, 62(2).

Corvalan, A., Querubín, P., and Vicente, S. (2020). The political class and redistributive policies. *Journal of the European Economic Association*, 18(1):1–48.

Cvrcek, T. and Zajicek, M. (2019). The rise of public schooling in nineteenth-century Imperial Austria: Who gained and who paid? *Cliometrica*, 13(3):367–403.

Davis, D. R. and Weinstein, D. E. (2002). Bones, Bombs, and Break Points: The Geography of Economic Activity. *American Economic Review*, 92(5):1269–1289.

Davis, D. R. and Weinstein, D. E. (2008). A search for multiple equilibria in urban industrial structure. *Journal of Regional Science*, 48(1):29–65.

de Carvalho Filho, I. and Colistete, R. P. (2010). Education Performance: Was It All Determined 100 Years Ago? Evidence from São Paulo, Brazil.

de Carvalho Filho, I. and Monasterio, L. (2012). Immigration and the origins of regional inequality: Government-sponsored European migration to southern Brazil before World War I. *Regional Science and Urban Economics*, 42(5):794–807.

Dean, W. (1969). *The Industrialization of São Paulo, 1800-1945*, volume 17. University of Texas Press.

Dell, M. (2010). The Persistent Effects of Peru's Mining *Mita*. *Econometrica*, 78(6):1863–1903.

Dijkstra, E. W. et al. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271.

Donaldson, D. (2018). Railroads of the Raj: Estimating the Impact of Transportation Infrastructure. *American Economic Review*, 108(4-5):899–934.

Donaldson, D. and Hornbeck, R. (2016). Railroads and American Economic Growth: A "Market Access" Approach. *The Quarterly Journal of Economics*, 131(2):799–858.

Droller, F. (2018). Migration, Population Composition and Long Run Economic Development: Evidence from Settlements in the Pampas. *The Economic Journal*, 128(614):2321–2352.

Easterlin, R. A. (1981). Why isn't the Whole World Developed? *The Journal of Economic History*, 41(1):1–17.

Engerman, S. and Sokoloff, K. (1997). Factor Endowments, Institutions and Differential Paths of Growth among the New World Economies," in Stephen Haber, ed., How Latin America Fell Behind, Stanford: Stanford University Press.

Engerman, S. L., Mariscal, E. V., and Sokoloff, K. L. (1999). The persistence of inequality in the Americas: schooling and suffrage, 1800–1945. *University of California at Los Angeles*.

Engerman, S. L. and Sokoloff, K. L. (2002). Factor endowments, inequality, and paths of development among new world economics. Technical report, National Bureau of Economic Research.

Fajgelbaum, P. and Redding, S. J. (2018). Trade, Structural Transformation and Development: Evidence from Argentina 1869-1914. *NBER Working Paper*.

Findlay, R. and O'rourke, K. H. (2009). *Power and Plenty: Trade, War, and the World Economy in the Second Millennium*, volume 30. Princeton University Press.

Fishlow, A. (1965). *American Railroads and the Transformation of the Antebellum Economy*. Harvard University Press, Cambridge.

Fogel, R. W. (1964). *Railroads and American economic growth*. Johns Hopkins Press, Baltimore.

Font, M. A. (1983). *Planters and the State: The Pursuit of Hegemony in São Paulo, Brazil (1889-1930)*. PhD thesis, University of Michigan.

Foster, A. D. and Rosenzweig, M. R. (2007). Economic development and the decline of agricultural employment. *Handbook of development economics*, 4:3051–3083.

Funari, P. P. P. (2017). Inequality, Institutions, and Long-Term Development: A Perspective from Brazilian Regions. In *Has Latin American Inequality Changed Direction?*, pages 113–142. Springer, Cham.

Furtado, C. (1959). *Formação econômica do Brasil*. Compainha das Letras.

Gallego, F. A. (2010). Historical origins of schooling: The role of democracy and political decentralization. *The Review of Economics and Statistics*, 92(2):228–243.

Galor, O. (2005). From Stagnation to Growth: Unified Growth Theory. *Handbook of economic growth*, 1:171–293.

Galor, O. and Moav, O. (2006). Das Human-Kapital: A Theory of the Demise of the Class Structure. *The Review of Economic Studies*, 73(1):85–117.

Galor, O., Moav, O., and Vollrath, D. (2009). Inequality in Landownership, the Emergence of Human-Capital Promoting Institutions, and the Great Divergence. *The Review of economic studies*, 76(1):143–179.

Galor, O. and Weil, D. N. (2000). Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond. *American economic review*, 90(4):806–828.

Glaeser, E. L., La Porta, R., Lopez-de Silanes, F., and Shleifer, A. (2004). Do Institutions Cause Growth? *Journal of economic Growth*, 9(3):271–303.

Go, S. and Lindert, P. (2010). The Uneven Rise of American Public Schools to 1850. *The Journal of Economic History*, 70(1):1–26.

Goldsmith-Pinkham, P., Sorkin, I., and Swift, H. (2018). Bartik Instruments: What, When, Why, and How. Technical report, National Bureau of Economic Research.

Goñi, M. (2018). Landed Elites and Education Provision in England: Evidence from School Boards, 1870-99.

Grandi, G. (2007). *Café e expansão ferroviária: a Companhia EF Rio Claro, 1880-1903*. Annablume.

Grandi, G. (2013). *Estado e capital ferroviário em São Paulo: a Companhia Paulista de Estradas de Ferro entre 1930 e 1961*. PhD thesis, Universidade de São Paulo.

Hanlon, W. W. (2017). Temporary Shocks and Persistent Effects in Urban Economies: Evidence from British Cities after the U.S. Civil War. *Review of Economics and Statistics*, 99(1):67–79.

Herrendorf, B., Rogerson, R., and Valentinyi, A. (2014). Growth and structural transformation. In *Handbook of economic growth*, volume 2, pages 855–941. Elsevier.

Hornbeck, R. (2012). The Enduring Impact of the American Dust Bowl: Short- and Long-Run Adjustments to Environmental Catastrophe. *American Economic Review*, 102(4):1477–1507.

Hornbeck, R. and Rotemberg, M. (2019). Railroads, Reallocation, and the Rise of American Manufacturing. Technical report, National Bureau of Economic Research.

Hornung, E. (2015). Railroads and Growth in Prussia. *Journal of the European Economic Association*, 13(4):699–736.

Huillery, E. (2009). History Matters: The Long-Term Impact of Colonial Public Investments in French West Africa. *American Economic Journal: Applied Economics*, pages 176–215.

IBGE (2003). *Estatísticas do século XX*. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro.

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

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

Keller, W. and Shiue, C. H. (2008). Institutions, technology, and trade. *NBER Working Paper*.

Lamounier, M. L. (2012). *Ferrovias e Mercado de Trabalho no Brasil do Século XIX*. Edusp, São Paulo.

Leff, N. H. (1972). Economic Retardation in Nineteenth-Century Brazil. *The Economic History Review*, 25(3):489–507.

Lindert, P. H. (2004). *Growing Public: Social Spending and Economic Growth since the Eighteenth Century*. Cambridge University Press.

Lindert, P. H. (2010). The unequal lag in Latin American schooling since 1900: follow the money. *Revista de Historia Economica-Journal of Iberian and Latin American Economic History*, 28(2):375–405.

Love, J. L. (1970). Political Participation in Brazil, 1881-1969. *Luso-Brazilian Review*, 7(2):3–24.

Love, J. L. and Barickman, B. J. (1986). Rulers and Owners: A Brazilian Case Study in Comparative Perspective. *The Hispanic American Historical Review*, 66(4):743–765.

Luna, F. V. and Klein, H. S. (2014). *The Economic and Social History of Brazil since 1889*. Cambridge University Press.

Matos, O. N. d. (1990). *Café e Ferrovias: A Evolução Ferroviária de São Paulo e o Desenvolvimento da Cultura Cafeeira*. Pontes, Campinas.

Mattoon Jr, R. H. (1977). Railroads, Coffee, and the Growth of Big Business in São Paulo, Brazil. *Hispanic American Historical Review*, pages 273–295.

Melander, E. (2018). Mobility and Mobilisation: Railways and the Spread of Social Movements.

Michaels, G. and Rauch, F. (2018). Resetting the Urban Network: 117–2012. *The Economic Journal*, 128(608):378–412.

Miguel, E. and Roland, G. (2011). The Long-Run Impact of Bombing Vietnam. *Journal of development Economics*, 96(1):1–15.

Musacchio, A., Fritscher, A. M., and Viarengo, M. (2014). Colonial Institutions, Trade Shocks, and the Diffusion of Elementary Education in Brazil, 1889–1930. *The Journal of Economic History*, 74(3):730–766.

Nafziger, S. (2011). Did Ivan's vote matter? The political economy of local democracy in Tsarist Russia. *European Review of Economic History*, 15(3):393–441.

Nunn, N. (2008). Slavery, Inequality, and Economic Development in the Americas: An Examination of the Engerman-Sokoloff Hypothesis. In Helpman, E., editor, *Institutions and Economic Performance*, pages 148–180. Harvard University Press, Cambridge.

Nunn, N. (2009). The Importance of History for Economic Development. *Annu. Rev. Econ.*, 1(1):65–92.

Nunn, N. (2014). Historical Development. In *Handbook of economic growth*, volume 2, pages 347–402. Elsevier.

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

Paula, D. A. d. (2000). *Fim de Linha. A extinção de ramais da Estrada de Ferro Leopoldina, 1955-1974*. PhD thesis, Universidade Federal Fluminense.

Pérez, S. (2017). Railroads and the Rural to Urban Transition: Evidence from 19th-Century Argentina.

Porta, R. L., Lopez-de Silanes, F., Shleifer, A., and Vishny, R. W. (1998). Law and Finance. *Journal of political economy*, 106(6):1113–1155.

Ramcharan, R. (2010). Inequality and redistribution: evidence from US counties and states, 1890–1930. *The Review of Economics and Statistics*, 92(4):729–744.

Redding, S. J., Sturm, D. M., and Wolf, N. (2011). History and Industry Location: Evidence from German Airports. *Review of Economics and Statistics*, 93(3):814–831.

Redding, S. J. and Turner, M. A. (2015). Transportation costs and the spatial organization of economic activity. In *Handbook of regional and urban economics*, volume 5, pages 1339–1398. Elsevier.

Rocha, R., Ferraz, C., and Soares, R. R. (2017). Human Capital Persistence and Development. *American Economic Journal: Applied Economics*, 9(4):105–36.

Saes, F. A. M. (1981). *As Ferrovias de São Paulo, 1870-1940*. Editora Hucitec.

Saxonhouse, G. R. and Wright, G. (2010). National Leadership and Competing Technological Paradigms: The Globalization of Cotton Spinning, 1878–1933. *The Journal of Economic History*, 70(3):535–566.

Silva, M. M. (1949). *Geografia dos Transportes no Brasil*, volume 7. Serviço Gráfico do Instituto Brasileiro de Geografía e Estatística.

Spolaore, E. and Wacziarg, R. (2013). How Deep Are the Roots of Economic Development? *Journal of economic literature*, 51(2):325–69.

Squicciarini, M. P. and Voigtländer, N. (2015). Human Capital and Industrialization: Evidence from the Age of Enlightenment. *The Quarterly Journal of Economics*, 130(4):1825–1883.

Stein, S. J. (1957). *The Brazilian cotton manufacture: Textile enterprise in an underdeveloped area, 1850–1950*. Harvard University Press.

Summerhill, W. (2010). Colonial Institutions, Slavery, Inequality, and Development: Evidence from Sao Paulo, Brazil.

Summerhill, W. R. (1998). Market Intervention in a Backward Economy: Railway Subsidy in Brazil, 1854-1913. *Economic History Review*, pages 542–568.

Summerhill, W. R. (2003). *Order Against Progress: Government, Foreign Investment, and Railroads in Brazil, 1854-1913*. Stanford University Press, Stanford.

Summerhill, W. R. (2005). Big Social Savings in a Small Laggard Economy: Railroad-Led Growth in Brazil. *The Journal of Economic History*, 65(1):72–102.

Suzigan, W. (1986). *Indústria brasileira: origem e desenvolvimento*. Brasiliense São Paulo.

Tang, J. P. (2014). Railroad Expansion and Industrialization: Evidence from Meiji Japan. *The Journal of Economic History*, 74(3):863–886.

Tapia, F. J. B. and Martinez-Galarraga, J. (2018). Inequality and Education in Pre-Industrial Economies: Evidence from Spain. *Explorations in Economic History*, 69:81–101.

Tyrefors, B., Lindgren, E., and Pettersson-Lidbom, P. (2019). The Political Economics of Growth, Labor Control and Coercion: Evidence from a Suffrage Reform. Technical report, Research Institute of Industrial Economics.

Valencia Caicedo, F. (2019). The Mission: Human Capital Transmission, Economic Persistence, and Culture in South America. *The Quarterly Journal of Economics*, 134(1):507–556.

Villela, A. V. and Suzigan, W. (1975). *Política do Governo e Crescimento da Economia Brasileira, 1889-1945*. Ipea/Inpes, Rio de Janeiro.

Vollrath, D. (2013). Inequality and School Funding in the Rural United States, 1890. *Explorations in Economic History*, 50(2):267–284.

Wantchekon, L., Klašnja, M., and Novta, N. (2015). Education and Human Capital Externalities: Evidence from Colonial Benin. *The Quarterly Journal of Economics*, 130(2):703–757.

Weber, E. (1976). *Peasants into Frenchmen: the modernization of rural France, 1870-1914*. Stanford University Press.

Witzel de Souza, B. G. (2018). Immigration and the path dependence of education: the case of German-speakers in São Paulo, Brazil (1840–1920). *The Economic History Review*, 71(2):506–539.

Yamasaki, J. (2017). Railroads, Technology Adoption, and Modern Economic Development: Evidence from Japan.

A
Data Appendix: Chapter 1

A.1
Merging the 1872-1950 Censuses

The unit of analysis is the municipality. To link the municipalities from the 1872 administrative division with those from the 1920/1940/1950 administrative division we use the official information of municipality's partitions and unions from IBGE[1]. We keep the unit of analysis constant over time, based on the 1872 administrative division, and aggregate the data from the 1920/1940/1950 census to the municipalities from the 1872 census.

To make the data compatible over time we proceed as follows:

- For the 1950/1940/1920 municipalities entirely contained within a given 1872 municipality, we just aggregated the data;
- For the 1950/1940/1920 municipalities contained within more than one 1872 municipality, we located the main urban center (downtown, or capital district) [2] in 1950/1940/1920 and determined to which 1872 municipality it belonged. Thus, we aggregated the data of the entire municipality based on the location of the main urban center, since it is the place of greatest population concentration.

Finally, for those municipalities incorporated into others over time, we aggregate your data to those of the incorporating municipalities to keep our sample based on the 1872 administrative division.

A.2
The Construction of the Brazilian Railway Network 1860-2017

- The first step was download the shapefiles of the modern railway network in operation on the website of the Ministry of Infrastructure: `http://www.infraestrutura.gov.br/component/content/`

[1]The evolution of the Brazilian administrative division can be found at: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15771-evolucao-da-divisao-territorial-do-brasil.html?=t=o-que-e

[2]The location of the main urban center ("sede municipal") from 1872 to 1991 can be found at: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15771-evolucao-da-divisao-territorial-do-brasil.html?=t=downloads

article.html?id=5124. I found data available for 2017, including railroads operating, deactivated, under construction, planned, and under studies. The modern network will be used to build the old one. The hypothesis is that the routes have changed little over time, given the high initial cost;

– In addition to the railway network shapefile, we will use a set of georeferenced data from Brazilian municipalities to georeference old maps. The structure of the evolution of the Brazilian state/municipal network was obtained from IBGE: https://www.ibge.gov.br/geociencias/downloads-geociencias.html,indivision_territorial_1872_1991;
– After having the modern data, we obtain maps of the old lines. The maps were obtained from IBGE or historical reports. With the old maps, we can reconstruct the evolution of the Brazilian railway network between 1860 and 2017;
– The most complete material found was for 1954, the height of the Brazilian railway network. With the report *I Centenário das Ferrovias Brasileiras* it was possible to digitize the railway network by railroad companies for all of Brazil. It is the year with the most accurate data. Therefore, 1954 will be the base year for our digitalization process;
– The digitalization process of the 1954 railway network will be carried out as follows: a) We georeferenced the maps of each railroad using the "Georeferencing" function of ArcGIS, associating points of the maps to the municipal headquarters; b) With the georeferenced maps, we build the railway network using the functions "Snapping" and "Edit Features", or copying the 2017 network when applicable;
– After digitizing the 1954 railroad network, we divide the work into two stages: Pre-1954 and Post-1954:

Pre-1954

– With the 1954 railway network, we used the historical maps of the *Plano Nacional de Aviação de 1949*, *Carta Geográfica do Brasil de 1922*, *Estatísticas das Estradas de Ferro do Brasil de 1920*, and information from the website http://www.estacoesferroviadas.com.br/ to characterize the expansion between 1920 and 1945;
– To digitize the railway network between 1860 and 1920, we used the following sources of information: *Plano Nacional de Aviação de 1949*, *Carta do Império do Brasil de 1875*, railroad network maps from 1913, 1907, and 1875, and information from the website http://www.estacoesferroviadas.com.br/;

Post-1954

– To digitize the railway network between 1954 and 2017, we used the fol-

lowing sources of information: *Ferrovias do Brasil 1970* from DNEF, *Anuário dos Transportes de 1985*, and information from two websites `http://www.estacoesferroviadas.com.br/` and `http://vfco.brazilia.jor.br/Mapas.Ferrovias.shtml`, in addition to the current railway network obtained on the website of the Ministry of Infrastructure.

A.3
The Construction of the Least-Cost Paths

To calculate the least-cost paths we use two geographic information: the average slope and the presence of rivers. The average slope data was downloaded into high-resolution cells from the Consortium for Spatial Information (CGIAR-SI), available at the following link: `http://srtm.csi.cgiar.org/srtmdata/`. The river data were downloaded from the National Water Agency (ANA), available at `http://dadosabertos.ana.gov.br/`. With the slope and river data, we discretize Brazil into a raster of grid cells (0.5 Km x 0.5 Km) and assume a cost function of displacement between cells that is directly proportional to the average slope within the cell and add a penalty if there is a river crossing it. The penalty corresponds to 1.7 degrees or 3%, the maximum slope allowed in many late 19th-century railroad projects in Brazil. Therefore, the cost function for each cell i has the following form:

$$C_i = Slope_i + 1.7 * River_i$$

Where $Slope_i$ is the average slope in the cell i, and $River_i$ is a dummy for the presence of a river in the same cell. With the data and the cost function, we calculate the least-cost paths using the Dijkstra's (1959) algorithm available in ArcGis by defining the following start points and destination targets:

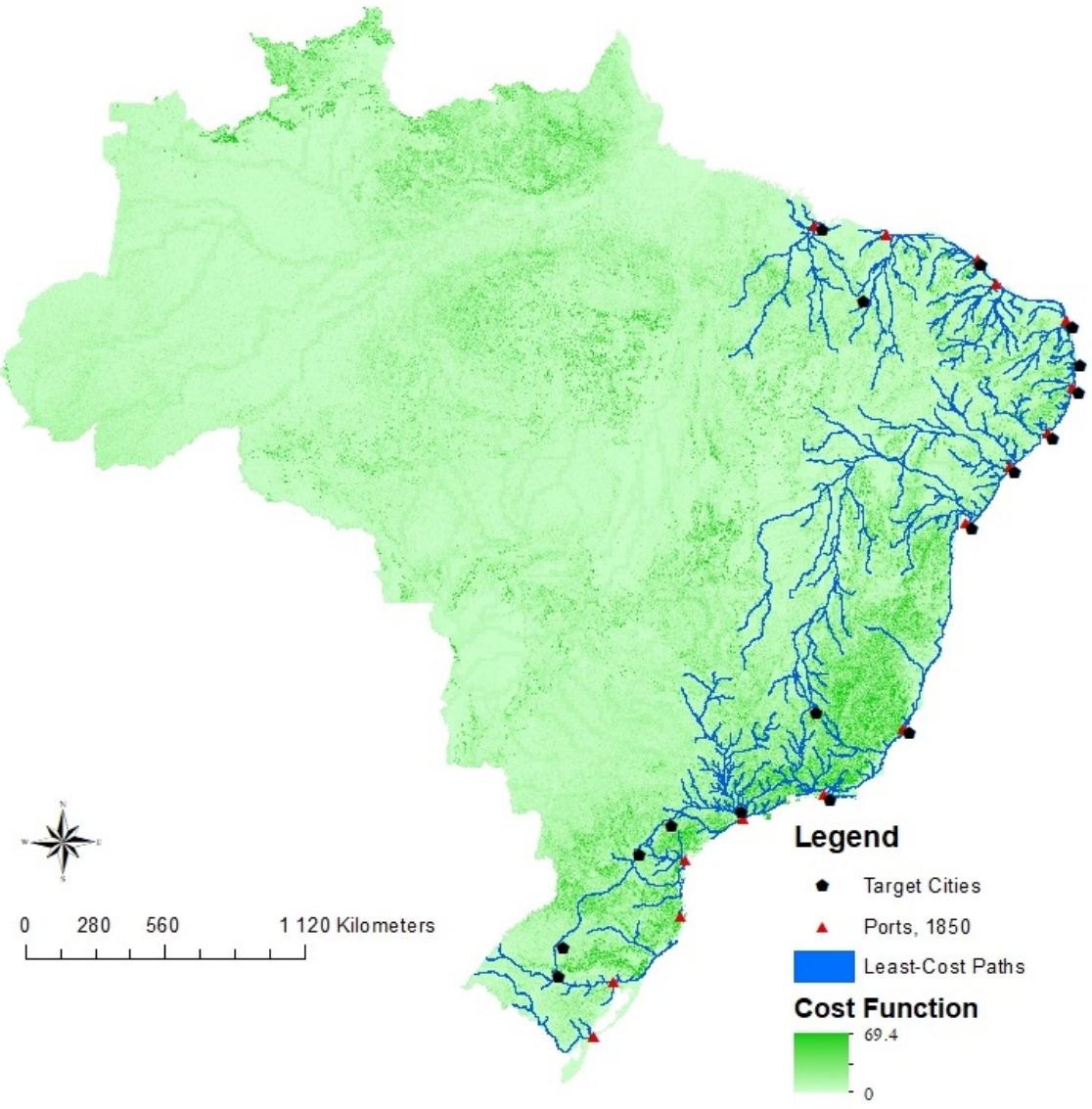
- *Cities - Port Connection:* We calculate the least-cost paths between each municipality that existed in 1872 and the existing ports in 1850 (municipalities and ports existing at the beginning of the expansion of railways in Brazil). The list of the most important Brazilian ports in 1850 is as follows: Rio de Janeiro, Salvador, Recife, São Luiz, Rio Grande, Maceió, Santos, São José do Norte, Paranaguá, Aracajú, Fortaleza, Desterro, Porto Alegre, Aracaty, Parnayha, Vitória, Parnahyba, and Natal;
- *State's Capitals Connection:* We calculate the least-cost paths between state's capitals and target cities as expressed in acts 10,432 from 1889, and 24,497 from 1934. In particular, we create bilateral cost-minimizing

routes between the following cities:

- To the South: Rio de Janeiro - São Paulo - Itararé - Imbituva - Cruz Alta - Santa Maria;
- To the North: Rio de Janeiro - Vitória | Rio de Janeiro - Belo Horizonte - Salvador - Aracajú - Maceió - Recife - João Pessoa - Natal - Fortaleza - Terezina - São Luís.

Finally, for each municipality, we compute the percentage of grid cells within its boundaries that lie along at these least-cost paths. Figure A.1 shows the least-cost paths constructed, the blue lines. Besides, we present ports and target cities on the map. The cost of moving from one cell to another is represented by the green area, the darker is the green, higher is the associated transposition cost.

Figure A.1: The Least-Cost Paths



Notes: The figure displays the least-cost paths, the blue lines, the target cities and ports, and the costs associated with the transposing of each cell, in green. Darker greens are related to higher costs.

A.4
Definition of Variables

A.4.1
Structural Transformation and Population

– *Labor Force Participation:* Share of workers in agriculture/manufacturing/service over total number of occupied workers. Original data from 1872/1920/1940/1950 Censuses. In general, we follow a similar occupation classification from Villela and Suzigan (1975)

to build the employment structure.
- *Population:* Population data from 1872/1920/1940/1950 Censuses.
- *Population Density:* Number of individuals by total area, in Km^2. Area data from 1920/1940/1950 Censuses.

A.4.2
Manufacturing, Price Dispersion, and Machinery Imports

- *Number of Factories:* Number of manufacturing plants. Original data from 1920 Industrial Census.
- *Factorie Dummy:* Takes the value 1 if the number of factories > 0, and 0 otherwise. Original data from 1920 Industrial Census.
- *Log Factories:* Log of total number of plants (+1). Original data from 1920 Industrial Census.
- *Factories' Size:* Total number of manufacturing workers over total number of factories. Original data from 1920 Industrial and Population Census.
- *Price Dispersion:* Absolute value of the log price difference between municipalities: $P_{ij} = |log(P_i) - log(P_j)|$ for all $i \neq j$. Original data from *Questionários sobre as Condições da Agricultura dos Municípios do Brasil*. Cross-section data collected between 1910 and 1913.
- *Spindles Import Dummy:* Takes the value 1 if the total number of spindles imported from British producers at the municipality is > 0, and 0 otherwise. Original data from Saxonhouse and Wright (2010).
- *Number of Spindles:* Total number of spindles imported from British machinery producers, in log (+1). Original data from Saxonhouse and Wright (2010).

A.4.3
Railroads

- *Railroad Distance:* Linear distance from the municipality to the nearest railroad line, in log. See main text for data sources and details on the construction of the railway network.
- *Railroad Dummy:* Takes 1 if linear distance from the municipality to the nearest railroad line $\leq$ 10 Km., and 0 otherwise. See main text for data sources and details on the construction of the railway network.
- *Railroad Stations:* Total number of railroads stations. Original data from http://www.estacoesferroviarias.com.br/.

A.4.4
Instrumental Variable

- $LeastCostPath_i * Extension_t$*:* Percentage of grid points within each municipality i that lies on the least-cost paths interacted with the Brazilian railroad network extension in year t. See main text and Appendix A.3 for data sources and details on the construction of the instrument.

A.4.5
Baseline Controls

- *Literate:* Number of literate individuals aged 6+ over total population aged 6+. Original data from 1872 Population Census.
- *% Foreigners:* Number of foreigners over total population. Original data from 1872 Population Census.
- *% Slaves:* Number of slaves over total population. Original data from 1872 Population Census.
- *Public Administration:* Total number of workers in public administration over total population*1000. Original data from 1872 Population Census.
- *Legal Professionals:* Total number of workers in legal professions over total population*1000. Original data from 1872 Population Census.

A.4.6
Geographic Controls

- *Latitude/Longitude:* Latitude and Longitude from municipalities in decimal degrees. Calculated with ArcGIS using shapefiles from IBGE.
- *Altitude:* Log of the average elevation measured in meters over sea level. Calculated with ArcGIS using high resolution spatial data from CGIAR-CSI:
  http://srtm.csi.cgiar.org/srtmdata/.
- *Slope:* Log of the average slope measured in degrees. Calculated with ArcGIS using high resolution spatial data from CGIAR-CSI: http://srtm.csi.cgiar.org/srtmdata/.
- *Area:* Log of the municipality area. Original data from 1920 Census and collapsed to the administrative division of 1872.
- *Soil Types:* Share of municipality area covered by latosol, argisol, cambisol, and neosol. Original data from Embrapa.
- *Distance to the Coast:* Log of the linear distance from the municipality to the nearest coast. Calculated with ArcGIS using shapefiles from IBGE and CPRM.
- *Distance to the Capital:* Log of the linear distance from the municipality to the nearest state's capital. Calculated with ArcGIS using shapefiles

from IBGE.

A.4.7
Transportation Controls

- *Distance to the Port:* Log of the linear distance from the municipality to the nearest port in 1850. Calculated with ArcGIS using shapefiles from IBGE and port list from *Relatório do Ministério da Fazenda, 1850*.
- *Distance to the Road:* Log of the linear distance from the municipality to the nearest road in 1856. Calculated with ArcGIS using shapefiles from IBGE and historical map from *Nova Carta Chorographica do Império do Brazil, 1856*.
- *Distance to the River:* Log of the linear distance from the municipality to the nearest river. Calculated with ArcGIS using shapefiles from IBGE and ANA.

B
Data Appendix: Chapter 2

B.1
Merging the 1950-2010 Censuses

The unit of analysis is the municipality. To link the municipalities from the 1950 administrative division with those from the 1960/1970/1980/1991/2000/2010 administrative division we use the official information of municipality's partitions and unions from IBGE[1]. We keep the unit of analysis constant over time, based on the 1950 administrative division, and aggregate the data from the 1960/1970/1980/1991/2000/2010 census to the municipalities from the 1950 census. Finally, we match each municipality that existed in 1950 to the original municipality it belonged to in the 1872 administrative division.

To make the data compatible over time we proceed as follows:

- For the 1960/1970/1980/1991/2000/2010 municipalities entirely contained within a given 1950 municipality, we just aggregated the data;
- For the 1960/1970/1980/1991/2000/2010 municipalities contained within more than one 1950 municipality, we located the main urban center (downtown, or capital district) [2] in 1960/1970/1980/1991/2000/2010 and determined to which 1950 municipality it belonged. Thus, we aggregated the data of the entire municipality based on the location of the main urban center, since it is the place of greatest population concentration.

Finally, to connect the 1950 administrative division to the 1872 census municipalities we proceed in the same way as described previously.

[1]The evolution of the Brazilian administrative division can be found at: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15771-evolucao-da-divisao-territorial-do-brasil.html?=t=o-que-e

[2]The location of the main urban center ("sede municipal") from 1872 to 1991 can be found at: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15771-evolucao-da-divisao-territorial-do-brasil.html?=t=downloads

B.2
The Construction of the Brazilian Railway Network 1860-2017

- The first step was download the shapefiles of the modern railway network in operation on the website of the Ministry of Infrastructure: `http://www.infraestrutura.gov.br/component/content/article.html?id=5124`. I found data available for 2017, including railroads operating, deactivated, under construction, planned, and under studies. The modern network will be used to build the old one. The hypothesis is that the routes have changed little over time, given the high initial cost;
- In addition to the railway network shapefile, we will use a set of georeferenced data from Brazilian municipalities to georeference old maps. The structure of the evolution of the Brazilian state/municipal network was obtained from IBGE: `https://www.ibge.gov.br/geociencias/downloads-geociencias.html,indivision_territorial_1872_1991`;
- After having the modern data, we obtain maps of the old lines. The maps were obtained from IBGE or historical reports. With the old maps, we can reconstruct the evolution of the Brazilian railway network between 1860 and 2017;
- The most complete material found was for 1954, the height of the Brazilian railway network. With the report *I Centenário das Ferrovias Brasileiras* it was possible to digitize the railway network by railroad companies for all of Brazil. It is the year with the most accurate data. Therefore, 1954 will be the base year for our digitalization process;
- The digitalization process of the 1954 railway network will be carried out as follows: a) We georeferenced the maps of each railroad using the "Georeferencing" function of ArcGIS, associating points of the maps to the municipal headquarters; b) With the georeferenced maps, we build the railway network using the functions "Snapping" and "Edit Features", or copying the 2017 network when applicable;
- After digitizing the 1954 railroad network, we divide the work into two stages: Pre-1954 and Post-1954:

Pre-1954

- With the 1954 railway network, we used the historical maps of the *Plano Nacional de Aviação de 1949*, *Carta Geográfica do Brasil de 1922*, *Estatísticas das Estradas de Ferro do Brasil de 1920*, and information from the website `http://www.estacoesferroviadas.com.br/` to characterize the expansion between 1920 and 1945;
- To digitize the railway network between 1860 and 1920, we used the following sources of information: *Plano Nacional de Aviação de 1949*,

Carta do Império do Brasil de 1875, railroad network maps from 1913, 1907, and 1875, and information from the website `http://www.estacoesferroviadas.com.br/`;

Post-1954

– To digitize the railway network between 1954 and 2017, we used the following sources of information: *Ferrovias do Brasil 1970* from DNEF, *Anuário dos Transportes de 1985*, and information from two websites `http://www.estacoesferroviadas.com.br/` and `http://vfco.brazilia.jor.br/Mapas.Ferrovias.shtml`, in addition to the current railway network obtained on the website of the Ministry of Infrastructure.

B.3 The Construction of the Least-Cost Paths

To calculate the least-cost paths we use two geographic information: the average slope and the presence of rivers. The average slope data was downloaded into high-resolution cells from the Consortium for Spatial Information (CGIAR-SI), available at the following link: `http://srtm.csi.cgiar.org/srtmdata/`. The river data were downloaded from the National Water Agency (ANA), available at `http://dadosabertos.ana.gov.br/`. With the slope and river data, we discretize Brazil into a raster of grid cells (0.5 Km x 0.5 Km) and assume a cost function of displacement between cells that is directly proportional to the average slope within the cell and add a penalty if there is a river crossing it. The penalty corresponds to 1.7 degrees or 3%, the maximum slope allowed in many late 19th-century railroad projects in Brazil. Therefore, the cost function for each cell i has the following form:

$$C_i = Slope_i + 1.7 * River_i$$

Where $Slope_i$ is the average slope in the cell i, and $River_i$ is a dummy for the presence of a river in the same cell. With the data and the cost function, we calculate the least-cost paths using the Dijkstra's (1959) algorithm available in ArcGis by defining the following start points and destination targets:

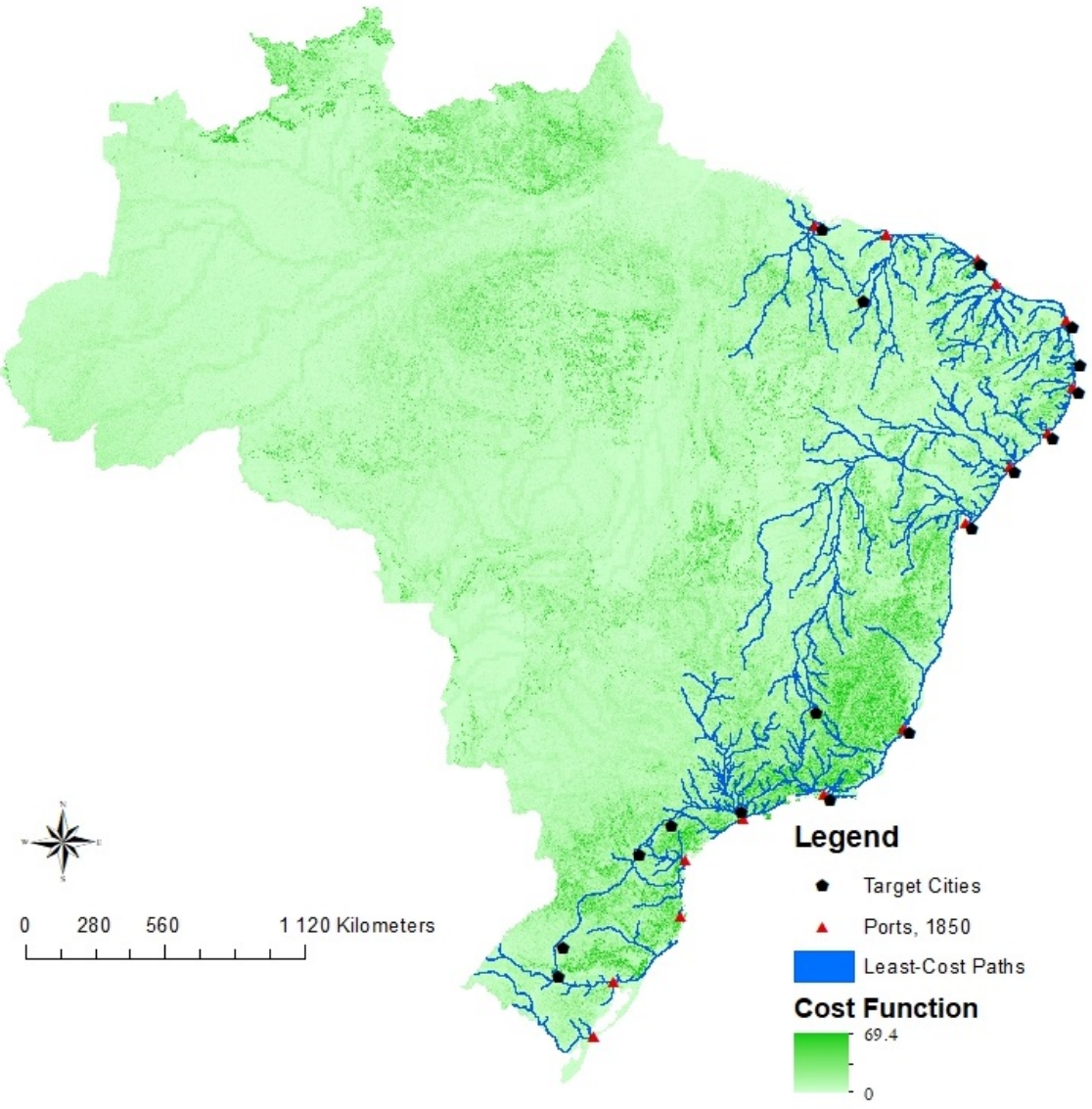
– *Cities - Port Connection:* We calculate the least-cost paths between each municipality that existed in 1872 and the existing ports in 1850 (municipalities and ports existing at the beginning of the expansion of railways in Brazil). The list of the most important Brazilian ports in 1850 is as follows: Rio de Janeiro, Salvador, Recife, São Luiz, Rio Grande, Maceió, Santos, São José do Norte, Paranaguá, Aracajú, Fortaleza,

Desterro, Porto Alegre, Aracaty, Parnayha, Vitória, Parnahyba, and Natal;

- *State's Capitals Connection:* We calculate the least-cost paths between state's capitals and target cities as expressed in acts 10,432 from 1889, and 24,497 from 1934. In particular, we create bilateral cost-minimizing routes between the following cities:
  - To the South: Rio de Janeiro - São Paulo - Itararé - Imbituva - Cruz Alta - Santa Maria;
  - To the North: Rio de Janeiro - Vitória | Rio de Janeiro - Belo Horizonte - Salvador - Aracajú - Maceió - Recife - João Pessoa - Natal - Fortaleza - Terezina - São Luís.

Finally, for each municipality, we compute the percentage of grid cells within its boundaries that lie along at these least-cost paths. Figure B.1 shows the least-cost paths constructed, the blue lines. Besides, we present ports and target cities on the map. The cost of moving from one cell to another is represented by the green area, the darker is the green, higher is the associated transposition cost.

Figure B.1: The Least-Cost Paths



Notes: The figure displays the least-cost paths, the blue lines, the target cities and ports, and the costs associated with the transposing of each cell, in green. Darker greens are related to higher costs.

B.4
Definition of Variables

B.4.1
Development and Structural Transformation

- *Income per capita:* Log of the income per capita in 2010. Income data from 2010 population census.
- *GNP per capita:* Log of the GNP per capita in 2010. GNP data from IBGE.

– *Labor Force Participation:* Share of workers in agriculture/manufacturing/service over total number of occupied workers. Original data from 1950-2010 Censuses. In general, we follow a similar occupation classification from Villela and Suzigan (1975) to build the employment structure.

B.4.2
Agglomeration, Urbanization and Manufacturing

– *Population:* Population data from 1950-2010 Censuses.
– *Population Density:* Number of individuals by total area, in Km^2. Area data from 1950-2010 Censuses.
– *Share Migrants:* Number of migrants over the total population. Original migrants data from 1970-1991 Censuses.
– *Share Urban Population:* Number of people living in urban areas over the total population. Original urban data from 1950-1991 Censuses.
– *Log Factories:* Log of total number of plants (+1) in 1950. Original data from 1950 Industrial Census.
– *Factories' Size:* Log of the total number of manufacturing workers over total number of factories in 1950. Original data from 1950 Industrial and Population Census.
– *Factories' Production:* Log of the average factorie's production in 1950. Original data from 1950 Industrial Census.

B.4.3
Railroads

– *Railroad Distance:* Linear distance from the municipality to the nearest railroad line, in log. See main text for data sources and details on the construction of the railway network.
– *Railroad Dummy:* Takes 1 if linear distance from the municipality to the nearest railroad line $\leq$ 10 Km., and 0 otherwise. See main text for data sources and details on the construction of the railway network.
– *Railroad Stations:* Total number of railroads stations. Original data from http://www.estacoesferroviarias.com.br/.

B.4.4
Instrumental Variable

– *Least Cost Path:* Log of the percentage of grid points within each municipality that lie on the least-cost paths. See main text and Appendix B.3 for data sources and details on the construction of the instrument.

B.4.5
Baseline Controls

- *Log Population:* Log of the number of people living in the municipality. Original data from 1872 Population Census.
- *Literate:* Number of literate individuals aged 6+ over total population aged 6+. Original data from 1872 Population Census.
- *% Foreigners:* Number of foreigners over total population. Original data from 1872 Population Census.
- *% Slaves:* Number of slaves over total population. Original data from 1872 Population Census.
- *Public Administration:* Total number of workers in public administration over total population*1000. Original data from 1872 Population Census.
- *Legal Professionals:* Total number of workers in legal professions over total population*1000. Original data from 1872 Population Census.
- *% Emp. Agriculture:* Share of workers in agriculture over total number of occupied workers. Original data from 1872 Population Census.
- *% Emp. Manufacturing:* Share of workers in manufacturing over total number of occupied workers.. Original data from 1872 Population Census.

B.4.6
Geographic Controls

- *Latitude/Longitude:* Latitude and Longitude from municipalities in decimal degrees. Calculated with ArcGIS using shapefiles from IBGE.
- *Altitude:* Log of the average elevation measured in meters over sea level. Calculated with ArcGIS using high resolution spatial data from CGIAR-CSI:
  http://srtm.csi.cgiar.org/srtmdata/.
- *Slope:* Log of the average slope measured in degrees. Calculated with ArcGIS using high resolution spatial data from CGIAR-CSI: http://srtm.csi.cgiar.org/srtmdata/.
- *Area:* Log of the municipality area. Original data from 1950 Census.
- *Soil Types:* Share of municipality area covered by latosol, argisol, cambisol, and neosol. Original data from Embrapa.
- *Distance to the Coast:* Log of the linear distance from the municipality to the nearest coast. Calculated with ArcGIS using shapefiles from IBGE and CPRM.
- *Distance to the Capital:* Log of the linear distance from the municipality to the nearest state's capital. Calculated with ArcGIS using shapefiles from IBGE.

B.4.7
Transportation Controls

- *Distance to the Port:* Log of the linear distance from the municipality to the nearest port in 1850. Calculated with ArcGIS using shapefiles from IBGE and port list from *Relatório do Ministério da Fazenda, 1850*.
- *Distance to the Road:* Log of the linear distance from the municipality to the nearest road in 1856. Calculated with ArcGIS using shapefiles from IBGE and historical map from *Nova Carta Chorographica do Império do Brazil, 1856*.
- *Distance to the River:* Log of the linear distance from the municipality to the nearest river. Calculated with ArcGIS using shapefiles from IBGE and ANA.
- *Roads Dummy:* Takes 1 if linear distance from the municipality to the nearest road in 1970-2010 $\leq$ 10 Km., and 0 otherwise. Calculated with ArcGIS using shapefiles from Ministry of Infrastructure/DNIT.

C
Data Appendix: Chapter 3

C.1
Merging the 1872-2010 Censuses to the 1905 Administrative Division

The unit of analysis is the municipality. To link the municipalities from the 1905 administrative division with those from the 1920-2010 administrative division we use the official information of municipality's partitions and unions from IBGE[1]. We keep the unit of analysis constant over time, based on the 1905 administrative division from the *Anuário Estatístico do Estado de São Paulo, 1905*, and aggregate the data from the 1920-2010 census to the municipalities from the 1905 statistical report. Finally, we match each municipality that existed in 1905 to the original municipality it belonged to in the 1872 administrative division.

To make the data compatible over time we proceed as follows:

- For the 1920-2010 municipalities entirely contained within a given 1905 municipality, we just aggregated the data;
- For the 1920-2010 municipalities contained within more than one 1905 municipality, we located the main urban center (downtown, or capital district) [2] in 1920-2010 and determined to which 1905 municipality it belonged. Thus, we aggregated the data of the entire municipality based on the location of the main urban center, since it is the place of greatest population concentration.

Finally, to connect the 1905 administrative division to the 1872 census municipalities we proceed in the same way as described previously.

C.2
Definition of Variables

[1]The evolution of the Brazilian administrative division can be found at: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15771-evolucao-da-divisao-territorial-do-brasil.html?=t=o-que-e

[2]The location of the main urban center ("sede municipal") from 1872 to 1991 can be found at: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15771-evolucao-da-divisao-territorial-do-brasil.html?=t=downloads

C.2.1
Income and Educational Outcomes

- *Income per capita:* Log of the income per capita in 2010. Income data from the 2010 population census.
- *Years of Schooling:* The average number of years of schooling for people aged 25 years old or more. Original data from the 2000 population census.
- *% Middle School Degree:* Percentage of people aged 25 years old or more with a middle school degree. Original data from the 2010 population census.
- *% High School Degree:* Percentage of people aged 25 years old or more with a high school degree. Original data from the 2010 population census.
- *% Bachelor Degree:* Percentage of people aged 25 years old or more with a bachelor's degree. Original data from the 2010 population census.
- *Share of Elementery/Middle School Teachers:* Total number of elementery/middle school teachers over total population aged 25 years old or more *1000. Original data from the 2010 population census.
- *Share of High School Teachers:* Total number of high school teachers over total population aged 25 years old or more *1000. Original data from the 2010 population census.
- *Share of Technical School Teachers:* Total number of technical school teachers over total population aged 25 years old or more *1000. Original data from the 2010 population census.
- *Share of Professors:* Total number of universities' professors over total population aged 25 years old or more *1000. Original data from the 2010 population census.
- *% Completed Elementary Education, 1940-1950:* Percentage of people aged 10 years old or more with completed elementary education. Original data from the 1940-1950 population censuses.
- *% Completed Elementary Education, 1970-2010:* Percentage of people aged 25 years old or more with completed elementary education. Original data from the 1970-2010 population censuses.
- *Number of Schools, 1920:* The total number of elementary, middle and high schools over the total number of children aged 6-14 years old *1000. Original data from the 1920 population census.
- *Number of Schools, 1940 and 2010:* The number of elementary, middle and high schools over the total number of children aged 7-14 years old *1000. Original data from the 1940 statistical report, and the 2010 educational census.
- *Number of Teachers, 1920:* The number of elementary and middle schools teachers over the total number of children aged 6-14 years old *1000. Original data from the 1920 population census.

- *Number of Teachers, 1970-2010:* The number elementary and middle schools teachers over the total number of children aged 7-14 years old *1000. Original data from the 1970-2010 population censuses.
- *% Children Attending School, 1960:* The number of children aged 5-14 years old attending school over total population aged 5-14 years old. Original data from the 1960 population census.
- *% Children Attending School, 1940/1970-2010:* The number of children aged 7-14 years old attending school over total population aged 7-14 years old. Original data from the 1940/1970-2010 population censuses.
- *Literacy Rate (%):* The percentage of literate people aged 5 years old or more over the total population aged 5 years old or more. Original data from the 1920-2010 population censuses.
- *% Farms with Machinery:* The percentage of farms with agricultural machines over the total number of farms. Original data from the 1920 agricultura census.
- *Log Coffee Production:* Log of the average coffee production by hectares. Original data from the 1920 agricultura census.
- *Wages in Construction and Agriculture:* Log of the average wage of rural workers. Original data from the 1920 agricultura census.
- *Log Distance Railway, 1920:* Linear distance from the municipality to the nearest railroad line, in log. Original data from historical reports.

C.2.2
Political Power Measures

- *% Farmers Voters:* The percentage of farmers voters over the total number of voters in 1905. Original data from the *Anuário Estatístico do Estado de São Paulo, 1904-1905.*
- *Political Concentration Index:* The Herfindahl index of the share of the voters by occupation in 1905. Original data from the *Anuário Estatístico do Estado de São Paulo, 1904-1905.*
- *% Voters:* The percentage of registred voters over total population in 1905. Original data from the *Anuário Estatístico do Estado de São Paulo, 1904-1905.*

C.2.3
Baseline Controls

- *Population Density:* Number of individuals by total area. Original data from 1872 population census.
- *Literacy Rate (%):* Number of literate individuals aged 6+ over total population aged 6+. Original data from 1872 population census.
- *% Children Attending School:* The number of children aged 6-15 years

old attending school over total population aged 6-15 years old. Original data from the 1872 population census.
- *Number of Teachers:* The number of teachers over the total number of children aged 6-15 years old *1000. Original data from the 1872 population census.
- *Foreigners:* Number of foreigners over total population. Original data from 1872 population census.
- *Slaves:* Number of slaves over total population. Original data from 1872 population census.
- *Public Administration:* Total number of workers in public administration over total population *1000. Original data from 1872 population census.
- *Legal Professionals:* Total number of workers in legal professions over total population *1000. Original data from 1872 population census.
- *% Emp. Agriculture:* Share of workers in agriculture over total number of occupied workers. Original data from 1872 population census.
- *% Emp. Manufacturing:* Share of workers in manufacturing over total number of occupied workers.. Original data from 1872 population census.

C.2.4 Geographic Controls

- *Latitude/Longitude:* Average latitude and longitude of the municipalities of the 1920 administrative division that constitute the original municipality in 1905, in decimal degrees. Calculated to the 1920 administrative division with ArcGIS using shapefiles from IBGE.
- *Altitude:* Average elevation measured in meters over sea level of the municipalities of the 1920 administrative division that constitute the original municipality in 1905, in log. Calculated to the 1920 administrative division with ArcGIS using high resolution spatial data from CGIAR-CSI: http://srtm.csi.cgiar.org/srtmdata/.
- *Slope:* Average slope measured in degrees of the municipalities of the 1920 administrative division that constitute the original municipality in 1905, in log. Calculated to the 1920 administrative division with ArcGIS using high resolution spatial data from CGIAR-CSI: http://srtm.csi.cgiar.org/srtmdata/.
- *Area:* Log of the average area of the municipalities of the 1920 administrative division that constitute the original municipality in 1905. Original data from the 1920 population census.
- *Soil Types:* Share of municipality area covered by latosol, argisol, cambisol, and neosol. Original data from Embrapa to 2000. The data for the 1905 administrative division corresponds to the average of the 2000 measures according to the partitions and unions of the municipalities over

time.

- *Distance to the Nearest Coast:* Log of the linear distance from the municipality to the nearest coast. Calculated to the 1920 administrative division with ArcGIS using shapefiles from IBGE and CPRM. The data for the 1905 administrative division corresponds to the average of the 1920 measures according to the partitions and unions of the municipalities over time.
- *Distance to the Capital:* Log of the linear distance from the municipality to the nearest state's capital. Calculated to the 1920 administrative division with ArcGIS using shapefiles from IBGE. The data for the 1905 administrative division corresponds to the average of the 1920 measures according to the partitions and unions of the municipalities over time.

C.2.5
Land Inequality Control

- *Land Gini:* Land Gini index calculated using the following formula: $1 + (1/n) - \frac{2\sum_{i=1}^{n}(n-i+1)a_i}{n\sum_{i=1}^{n}a_i}$, where n is the number of farms, a_i is rural property size, and i denotes the rank, where farms are ranked in ascending order of a_i. The land Gini is calculated using the Stata program *ineqdec0*. Original data from the 1920 population census and adapted for the 1905 administrative division.

C.2.6
Transportation Controls

- *Distance to the Nearest Port:* Log of the linear distance from the municipality to the nearest port in 1850. Calculated to the 1920 administrative division with ArcGIS using shapefiles from IBGE and port list from *Relatório do Ministério da Fazenda, 1850*. The data for the 1905 administrative division corresponds to the average of the 1920 measures according to the partitions and unions of the municipalities over time.
- *Distance to the Nearest Road:* Log of the linear distance from the municipality to the nearest road in 1856. Calculated to the 1920 administrative division with ArcGIS using shapefiles from IBGE and historical map from *Nova Carta Chorographica do Império do Brazil, 1856*. The data for the 1905 administrative division corresponds to the average of the 1920 measures according to the partitions and unions of the municipalities over time.
- *Distance to the Nearest River:* Log of the linear distance from the municipality to the nearest river. Calculated to the 1920 administrative

division with ArcGIS using shapefiles from IBGE and ANA. The data for the 1905 administrative division corresponds to the average of the 1920 measures according to the partitions and unions of the municipalities over time.

C.2.7
Additional Data

– *% Foreigners in 1920:* Percentage of foreigners over total population. Original data from the 1920 population census.