The Agglomeration Effect of Road Endpoints *

Bruno Barsanetti†

December 2019
Preliminary Draft

Abstract

How does the gradual construction of transportation infrastructure affect the distribution of economic activity across the many sites it serves? I present a simple model of urban emergence and growth along an expanding railroad. Towns import intermediate goods through the railroad to produce a good that is consumed by their agricultural hinterland. Due to a larger agglomeration shadow, towns that are railroad endpoints for longer become persistently larger. I explore over 100 years of railroad expansion around the city of São Paulo to document this key prediction of the model. Each additional year that a municipality was a railroad endpoint is associated with a municipality urban population that is 0.083 log points larger. An instrumental variable approach based on the expansion rate of Brazilian railroads indicates that the association is causal.

*I appreciate the comments and suggestions of Cynthia Kinnan. All errors are mine.
†Northwestern University. E-mail: barsanetti@u.northwestern.edu
1 Introduction

It frequently takes many years to complete the construction of transportation structures such as roads, railroads, or subway lines. Moreover, there is often a gradual opening of the transportation structure to use.

In this paper, I examine how the gradual opening of a road permanently affects the spatial distribution of economic activity across the different sites it serves. The intuition for why the gradual opening matters is that, in each period, economic growth in a site depends on the competition with other sites along the road. Consider a site on the scheduled path of a road. When the timing of road construction changes, the productivity of other sites at the time the road reaches that site will be different. This occurs because firms learn over time, either from their own experience or from the experience of neighboring firms. As a consequence, economic activity in the site can be either persistently curbed or enhanced.

In the first part of the paper, I present a stylized model of town emergence and growth along an expanding railroad. The model shows how the gradual opening of the railroad affects the distribution of towns and town sizes along the railroad. It predicts that towns that are railroad endpoints for longer become persistently larger.

I model an economy in which a railroad is built over time. In this economy, towns serve a rural hinterland that produces an agricultural good that is exported. The railroad not only reduces the costs of moving the agricultural good to the port, but it also gives urban firms access to important inputs for the production of urban services they provide to the rural hinterland. Over time, towns accumulate knowledge and become more productive in the provision of urban services.

Neighboring towns compete for hinterland markets. New towns emerge on newly constructed railroad lines. If a town is an endpoint for longer, it is more productive when the railroad reaches a new site farther down the road. Therefore, the town casts a larger agglomeration shadow over the site, preventing its urban development. As a consequence, the endpoint town will persistently serve a larger hinterland. Since hinterland size determines town size, a town that was an endpoint for longer will be larger even after the railroad line expands to other sites.

In the second part of the paper, I empirically document the key predictions of the model in the context of the radial railroad network that was built around the city of São Paulo. The expansion of railroads in São Paulo is an ideal setting to examine urban emergence and growth around railroad lines. The city of São Paulo is on a plateau 760 meters above sea level. It was first connected by rail in 1867, when a railroad from the nearby port city of Santos was built. Since then, many railroad lines were built toward the western portions of the state and of Brazil. These areas were sparsely populated by the time of railroad construction, and most towns along the railroad emerged after the railroad arrival. Moreover, there were many lines built during a time interval of more than 1

---

1 As shown empirically, for instance, by Faber (2014).
100 years, which implies there is sufficient variation to test the predictions of the model.

I estimate the correlation between years as endpoint and city size in 2010. In line with the predictions of the model, municipalities that were endpoints for longer are larger. Distance to the next town along the railroad is also greater. There is no correlation with distance to the previous town, as the model also predicts. Each additional year that a municipality was a railroad endpoint is associated with a municipality urban population that is 0.083 log points larger. This association is hardly affected by the inclusion of control variables or by different sample selections.

To address concerns that these associations do not reflect a causal relation, I use the expansion rate of the Brazilian railroad network in the year following the railroad arrival as an instrumental variable. The idea behind the instrument is that, during the period when the railroads were being built in São Paulo, railroad construction projects were lumpy investments that competed for capital (Summerhill 2003). Therefore, when other lines were in the midst of large construction expenditures, capital became scarcer for the railroad line that had just arrived at an endpoint site. In this way, construction would slow down after reaching that endpoint town, which would be a railroad endpoint for longer.

Using the instrumental variable approach, I estimate a relationship between time as endpoint and town size which is similar to the documented correlation between the two variables. Therefore, the results suggest that there is a causal effect of time as endpoint on town sizes.

Further support for the interpretation that there is a causal relationship comes from a historical example. I discuss the case of São José do Rio Preto, a town which was a railroad endpoint for longer due to plausibly exogenous reasons. It became persistently larger than other towns along its rail line, as predicted by the model.

I also discuss the studies of contemporary geographers that highlighted the importance of railroad endpoints for urban growth in the west of the São Paulo state (Deffontaines 1938, Monbeig 1952). My paper builds on top of this older literature and makes two contributions to it: I provide an economic explanation and I quantify the long-term effects of railroad endpoints on city sizes.

This paper contributes to a large literature on the effects of transportation infrastructure; see Redding and Turner (2015) for a recent review. In particular, railroads have been shown to impact economic activity in developing economies (Jedwab and Moradi 2016, Donaldson 2018). Although there is evidence that the effects of transportation infrastructure are heterogeneous (as, for instance, in Jedwab and Storeygard 2019), there is little work on how the heterogeneity relates to position along the transportation infrastructure. A seminal paper is Fujita and Mori (1996), who present a model of why cross-road positions can sustain larger cities. In this paper, I focus instead on a different position: road endpoints. Another difference of my paper in relation to much of the literature is that I examine transportation investments in a dynamic setting. In this way, my paper relates to Balboni (2019), who studies road building in Vietnam in a dynamic spatial equilibrium model.
2 Model

Consider a railroad that expands along a route. The different sites \( y \) in this economy are located on the half-line \( Y \equiv [0, \infty) \). Site \( y = 0 \) is a port town through which this economy trades with the rest of the world (ROW).\(^2\) Time is also continuous: \( t \in T \equiv [0, \infty) \). Starting from the initial period \( t = 0 \), a railroad is built along the half-line \( Y \). The railroad construction schedule is exogenous. At time \( t \), the railroad has been built up to site \( y_r(t) \). Assume \( y_r(.) \) is absolutely continuous, so it can be written as \( y_r(t) = \int_0^t g_r(\theta) d\theta \); \( g_r(t) \) is the growth rate of the railroad at time \( t \). For each site \( y \), we can define the time the railroad reaches it as \( t_r(y) \equiv \min \{ t \in T | y_r(t) \geq y \} \).

There are three goods in this economy:

1. an agricultural good \( a \) that is exported to the ROW at the port town \( y = 0 \). Denote its price on site \( y \) and at time \( t \) as \( p(y,t) \). The price it is transacted with the ROW is \( p(0,t) = p \) for all \( t \in T \). The agricultural good can be costlessly transported by rail. When it is transported overland through a transportation mode other than rail, there is an iceberg cost: \( e_{\text{road}} \) units have to be shipped for one unit to arrive to a destination at a distance \( d \).

2. an intermediate good \( m \) that is imported from the ROW at the port town \( y = 0 \). Normalize the price of the intermediate good at \( y = 0 \) as 1. It can be costlessly transported by rail, but the cost of transporting it overland outside of the railroad is prohibitive (\( \tau_m = \infty \)).

3. urban services \( s \) that cannot be traded with the ROW and are transported overland at an iceberg cost \( \tau_s \) regardless of the transportation mode. Let \( q(y,t) \) be the price of this good in site \( y \) at time \( t \).

There are two types of agents in this economy. First, there are immobile farmers, uniformly distributed across the half-line \( Y \). Normalize the mass of farmers in each site \( y \) to 1. Each farmer produces one unit of the agricultural good \( a \) per period. They consume both the agricultural good \( a \) and the urban services \( s \), that they value according to the utility function \( u(a, s) = a^\beta s^{1-\beta} \).

There are also urban firms. They use one unit of the intermediate good \( m \) to produce \( e^{A(y,t)} \) units of urban services \( s \). Note that \( A(y,t) \) measures a site-specific productivity at period \( t \), which is the same across all firms that are located at that site \( y \). In each period, there is free entry of urban firms, that decide on which site to locate. I say that, at time \( t \), there is a town in site \( y \) if there is a positive production of the urban services at that site: \( s(y,t) > 0 \). I assume that, when firms are indifferent between entering or not into a site, they enter; this assumption guarantees uniqueness.\(^3\)

---

\(^2\) The space is modeled as a country with internal geography as in Coşar and Fajgelbaum (2016).

\(^3\) This assumption can be rationalized by perturbing the condition for the entry of firms: assuming entry if profits are greater than \(-\varepsilon\). Hence, there is entry when firm profits are zero even in the limit as \( \varepsilon \to 1 \). This perturbation captures the idea that firms can make vanishingly small mistakes in their location decisions and that the equilibrium must be robust to that.
Initially, productivity is $A(0, 0) = A_0$ at the port town. The port town is the only town at the initial period, as the intermediate good is not available at other sites. For other sites $t > 0$, assume that initial productivity is $A(y, 0) = 0$.

I assume that, over time, site-specific productivity increases in towns but not in non-town sites. This assumption captures the idea that there is learning in cities. It implies intertemporal local spillovers: the presence of firms in a site in the past implies that, in the future, that site will be more productive. The assumption is that productivity growth follows:

$$
\dot{A}(y, t) = \begin{cases} 
0 & \text{if } s(y, t) = 0 \\
g(A(y, t)) & \text{if } s(y, t) > 0
\end{cases}
$$

where $g(.)$ is a Lipschitz continuous, non-negative, and non-increasing function. I assume that the productivity growth of a newly formed town is sufficiently fast when compared to the expansion rate of the railroad. This assumption guarantees that, at any point in time, there is a finite number of cities in the economy.

**Assumption 1.** $\tau_s \cdot \sup g^*(t) < g(0)$

**Equilibrium**

I first examine the prices and the distribution of urban firms across space at time $t$. At time $t$, the railroad has been built up to $y^r(t)$. Using the transportation costs for the agricultural good, the price of the agricultural good is given by:

$$
\log p(y, t) = \begin{cases} 
\log(p) & \text{if } y \leq y^r(t) \\
\log(p) - \tau_a[y - y^r(t)] & \text{if } y > y^r(t)
\end{cases}
$$

(1)

Given these prices, the expenditure from site $y$ farmers on urban services will be given by $(1 - \beta)p(y, t)$. Hence, the demand for urban services will depend on the agricultural good prices.

What about the supply of urban services? It will come from towns that exist at time $t$. Denote the set of all such towns as $\mathcal{Y}(t) \equiv \{y \in Y | s(y, t) > 0\}$. Since the production of urban services requires the use of the intermediate good, all towns will be on sites that have been reached by the railroad: $\mathcal{Y}(t) \subset [0, y^r(t)]$. Moreover, the free entry condition implies that urban services prices at city $y \in \mathcal{Y}(t)$ satisfy $\log q(y, t) = -A(y, t)$. At any other site $y$, the price of urban services will be:

$$
\log q(y, t) = \min_{\hat{y} \in \mathcal{Y}(t)} \left( -A(\hat{y}, t) + \tau_s |y - \hat{y}| \right)
$$

(2)

Over time, the site-specific productivity $A(y, t)$ of each town increases. As a consequence, urban services prices must fall everywhere. This fact has an important implication: any site that, at a time $t$, (a) had been reached by the railroad ($y \leq y^r(t)$) and (b) was not a town on that period
(y \notin \mathbb{Y}(t)) will never be a town in a future period. To understand this implication, note that a site y can become a town only if its productivity is high enough so urban services can be produced as cheap as it takes to buy them from somewhere else: \( \log q(y, t) \geq -A(y, t) \). When a site is not a town, the right-hand side of the inequality does not change over time, while the left-hand side goes down.

I can also show that, once a town emerges on a site y, it will always exist. This occurs for two reasons. First, relative to towns that will emerge later, the town at y will always be at least weakly more productive. As a consequence, local firms will always be able to sell at a lower price than firms from younger towns. Second, relative to towns that already existed when y emerged, productivity growth at y will always be larger, since \( g(.) \) is non-increasing. Hence, if local firms can sell at a lower price than firms from existing towns at time \( t \), they would also be able to do so in a future time \( t' > t \).

As a consequence of these two properties, there will be a single equilibrium in this economy. There is always a finite number of towns, that expand over time. The location of the towns can be solved recursively, starting from the first town (the port town). The proposition below formalizes this. A proof is in the Appendix.

Proposition 1. Under Assumption 1, there is an unique equilibrium. The set of towns \( \mathbb{Y}(t) \) is finite and non-decreasing over time. At the initial period, \( \mathbb{Y}(0) \) consists only of \( y_0 = 0 \). A recursive procedure defines \( \mathbb{Y}(t) \):

1. Let \( \{y_0, y_1, \ldots, y_N\} \) be the set of towns at a given time. Define \( t_N = t^r(y_N) \).

2. For all \( t \in [t_N, t_{N+1}) \), \( \mathbb{Y}(t) = \{y_0, y_1, \ldots, y_N\} \). A next town is founded only in period \( t_{N+1} \), that satisfies:

\[
t_{N+1} = \min \left\{ t > t_{N+1} \mid t_s[y^r(t) - y_N] = A(y_N, t) \right\}
\]

3. At time \( t_{N+1} \), \( \mathbb{Y}(t_{N+1}) = \{y_0, y_1, \ldots, y_N, y_{N+1}\} \), where \( y_{N+1} = y^r(t_{N+1}) \).

According to Proposition 1, the next town to emerge will be at the first site where, due to transportation costs, the price of urban services provided by the last existing town on the railroad is high enough so that it becomes profitable for a firm to establish itself at that site. Assumption 1 is a bound on the railroad expansion rate that implies the site of the next town will not be reached immediately.

Proposition 1 characterizes, at any time \( t \), the set of towns in equilibrium. To completely characterize the equilibrium, we need to find prices \( q(y, t) \) and town sizes \( s(y, t) \). Prices are given by equations (1) and (2). And, given the prices, it is possible to define which sites are served
Town hinterland areas

Note: The green area indicates the hinterland served by \( y_0 \); the red, by \( y_1 \); the navy blue, by \( y_2 \); and the brown, by \( y_N \).

by each town (town hinterland). If \( \mathcal{Y}(t) = \{y_0, y_1, ..., y_N\} \), then town \( y_n \) serves a hinterland \([d_n(t), d_{n+1}(t)]\). This situation is shown on the line below.

The hinterland border \( d_n(t) \) is the site where farmers are indifferent between consuming from \( y_{n-1} \) and \( y_n \). It is given by:

\[
d_n(t) = \frac{y_n + y_{n-1}}{2} - \frac{A(y_n, t) - A(y_{n-1}, t)}{2\tau_s}
\]

(3)

Given each town \( y_n \) hinterland \([d_n(t), d_{n+1}(t)]\), its total production of urban services is given by the demand from its rural hinterland. Consumption from farmers in site \( y \in [d_n(t), d_{n+1}(t)] \) will be \( \frac{(1-\beta)p(y,t)}{q(y,t)} \). Accounting for the iceberg transportation costs, using the expression for the prices of the agricultural good and the urban services, total production of a non-endpoint town \( y_n \) (i.e. a town that was not the last one to emerge along the road, \( y_n < y_N \)) will be:

\[
s(y_n, t) = p(1-\beta)e^{A(y_n, t)} \left[ d_{n+1}(t) - d_n(t) \right]
\]

(4)

For the endpoint town \( y_N \), total production of urban services is:

\[
s(y_N, t) = p(1-\beta)e^{A(y_n, t)} \left[ y^r(t) - d_N(t) + \frac{1}{\tau_a} \right]
\]

(5)

These equations, jointly with Proposition 1, characterize the equilibrium.

**Key Model Predictions**

Given the characterization of the equilibrium, what are the key predictions of the model? In particular, what are the persistent effects of being an endpoint on town size?

I compare towns that were railroad endpoint for different time intervals. That is, I compare towns according to the time for the railroad to reach the next town to emerge along the railroad. According to the recursive procedure from Proposition 1, this will be a function solely of the railroad construction schedule \( g^r(t) \) in the time periods succeeding the period when the town emerged.

Denote a town site by \( y_n \). From Proposition 1, the next town to emerge will be at the closest site \( y_{n+1} \) that satisfies the following equation:

\footnote{For completeness, \( d_0(t) = 0 \) and \( d_{N+1}(t) = \infty \).}
$$\tau_s(y_{n+1} - y_n) = A(y_N, t^r(y_{n+1}))$$

The equation above can be illustrated by the figure below:

**Distance to the next town**

![Graph](image)

Note: The green line indicates $A(y_n, t^r(y))$ under slow railroad growth, while the navy blue indicates it under fast railroad growth.

When the railroad is built slowly, the last endpoint town will be larger at the time the railroad reaches a site $y$ farther down the road. As a consequence, urban firms will have less incentives to produce there. As a consequence, the next town to emerge will be farther away.

**Prediction 1.** As a town is a railroad endpoint for longer, the distance to the next town farther down the railroad increases.

The model also has a prediction on town sizes. If a town is an endpoint for longer, it will be persistently larger. This is the case because the town will supply a larger hinterland. This occurs for two reasons. First, the next town on the railroad will be farther away, directly increasing hinterland size according to equation (3). Second, the next town will have had emerged later, so it will be (at least weakly) less productive at any future time. According to equation (3), the productivity difference also increases hinterland size. As a consequence, the town will be larger.

**Prediction 2.** As a town is a railroad endpoint for longer, its size permanently increases.
3 Empirical Investigation

In this section, I empirically examine the two key predictions of the model in a setting that resembles the model environment: the expansion of the São Paulo rail network towards the western parts of the state.

The city of São Paulo is on a plateau 760 meters above sea level. The harbor of Santos is geographically close, but transportation costs to the harbor were high until the construction of the São Paulo Railway in 1867. After the opening of this railroad, the city of São Paulo became a railroad hub as new lines were built connecting it with other parts of Brazil. Many of the new lines were built toward low population density regions in the west of the state.

Before the railroad era, overland transportation costs in Brazil were high due to the hilly and forested terrain and the absence of navigable rivers (Leff 1982, Summerhill 2003). The main alternative transportation mode was the mule. As a consequence, railroad building integrated the west of the state with markets near the coast, effectively incorporating it into a fast-growing economy. Coffee production -the main export industry in the state -expanded along the railroads. At the same time, a myriad of new towns emerge along the rail lines. These changes coincided with a period of unprecedented economic growth, industrialization, and mass migration from Europe (Monbeig 1950).

Note that the setting shares similarities with the economy I model in the previous section. Railroad lines were built from a harbor towards the interior of the country. The economy exported an agricultural good, coffee. Railroad construction reduced transportation costs. Finally, new towns emerged along the railroad. In these towns, merchants and other businesses imported goods from the city of São Paulo or from the rest of the world to sell them to local farmers.

Finally, the São Paulo setting is particularly suitable to examine the effects of railroad construction. Still today, the state geography is influenced by the railroads that were built in the century following 1867 (Menucci Giesbrecht 2001). This is the case even after many railroads were abandoned. Appendix Figure A.1 maps the sample railroads against nightlights in 2010. Note that most towns are located along the railroads; this is especially true in the western parts of the state.

Data

Information on the expansion of the São Paulo railroad network comes from two sources. The railroad lines and the location of each station are from shapefiles released by the Agência Nacional de Transportes Terrestres (ANTT), the Brazilian federal agency responsible for regulating railway and road transportation. The list of railroad stations includes both active stations and stations no longer in use.

I match each station in the ANTT dataset with station opening dates from Estações Ferroviárias
do Brasil (ESB), a comprehensive website with information on railroad stations.\textsuperscript{5} The available information includes the year of station opening, the construction date of the existing building, and its current state and use. Opening years are obtained from a variety of sources, including railroad schedules. Using this information, for each station I construct the year of arrival of the railroad line. Note this is different than the year of station opening, as some stations were opened on preexisting railroad lines. Finally, I use the station opening years to construct the year of railroad arrival for each municipality. In this way, for each municipality I can define the total number of years as endpoint as the difference between the railroad arrival year for the next municipality on the railroad line and the municipality own arrival year.\textsuperscript{6}

The sample consists of municipalities along the railroad lines that were built toward the west of the state of São Paulo; some lines extend into the neighboring states of Paraná and Minas Gerais.\textsuperscript{7} These lines were built and originally operated by 8 different companies. I exclude municipalities that are endpoints today. All railroad lines either started in the city of São Paulo or in some other railroad line that began in São Paulo. In this way, for each municipality I can calculate its distance, along the railroad, to the Estação da Luz, the central station in the city of São Paulo.

As dependent variables, I use urban population from the 2010 Brazilian Population Census and estimates of urban GDP for the same year from the Instituto Brasileiro de Geografia e Estatística. I define urban GDP as the combined GDP in the manufacturing, services, and government sectors; agriculture is excluded from it. I also use auxiliary data sources, which are described in Appendix E.

The map in Figure 1 shows the sample municipalities, the railroad lines, and the year each municipality was a railroad endpoint. Table 1 displays summary statistics. The average railroad arrival date was 1911. Of the 203 sample municipalities, 49 were railroad endpoints for at least 2 years. The city that was an endpoint for longer was São José do Rio Preto: it was an endpoint for 22 years. On average, municipalities were railroad endpoints for 2.2 years.

On average, municipalities in the sample had 80 thousand urban inhabitants in 2010.\textsuperscript{8} The average urbanization rate was 90%. Per capita GDP was $16,032 dollars, which was higher than the 2010 Brazilian per capita GDP of $14,320. The sample municipalities are on a plateau and their average altitude is 540 meters. The average distance through rail to the São Paulo main station is 404 km; maximum distance is 694 km.

\textsuperscript{5}See also Menucci Giesbrecht (2001), a book by the website curator. To the extent of my knowledge, Rocha \textit{et al.} (2017) is the only economics paper to use these data.

\textsuperscript{6}In some cases, there is a branch of the railroad lines and there are two or three next municipalities; I define years as endpoint as the difference to the first next municipality to be reached by a railroad.

\textsuperscript{7}Therefore, I excluded two railroads that connected São Paulo to other important capitals: the one connecting São Paulo to Rio de Janeiro, the federal capital, and the one connecting it to the southern capitals of Curitiba and Porto Alegre.

\textsuperscript{8}In Brazil, an area is urban if it is administratively classified as so by the municipality it is in; see IBGE (2017) for a discussion of the urban-rural classification in Brazil.
Empirical Facts

To document the association between the number of years a municipality was a railroad endpoint and city size, I estimate the following equation by ordinary least squares:

\[ y_i = \beta EndpointYears_i + X_i'\theta + u_i \]  \hspace{1cm} (6)

where \( i \) indicates a municipality, \( y_i \) is a dependent variable such as the logarithm of total urban population in 2010, \( EndpointYears_i \) is the number of years that the municipality was a railroad endpoint, and \( X_i \) is a vector of control variables. I weight each municipality by its total area, so the coefficient \( \beta \) represents the effects on a square kilometer.

Figure 2 displays the relationship between years as endpoint and the logarithm of urban population (in Panel a), of urban GDP (in Panel b), or of distance to the next municipality (in Panel c). Note there is an increasing relationship throughout the sample.

The estimated coefficients of equation 6 are shown in Table 2. In Panel A, the dependent variable is the logarithm of urban population in 2010. In Panel B, it is the logarithm of urban GDP in 2010. And in Panel C, it is the logarithm of distance to the next municipality along the railroad. In column (1), I report the coefficients on years as endpoint when there are no control variables. In column (2), I report the coefficients after the inclusion of the baseline controls: the year of railroad arrival, the logarithm distance to São Paulo, and a dummy equals to one if the municipality was already incorporated by the time of railroad arrival. The results indicate that an additional year as railroad endpoint is associated with a larger urban population in 2010 by 0.083 log points, a larger urban GDP by 0.098 log points, and a larger distance to the next municipality by 0.032 log points. All these coefficients are statistically different than zero at a 1% significance level. Hence, the correlations in the data are in line with the two predictions of the model.

To further evaluate how the empirical patterns are supported by the model, in column (3) I examine the relation of years as endpoint with variables for which the model predicts no effect. In Panel A, the dependent variable is the inverse hyperbolic sine of rural population.\(^9\) In Panel B, it is the logarithm of agriculture GDP. And in Panel C, it is the distance to the previous municipality along the railroad. In either case, the coefficients are much smaller than the coefficients in column (2) and not statistically different than zero.

In column (4), I estimate an “intensive margin” of years as endpoint by including in the sample only those municipalities that were railroad endpoints for at least 2 years. The coefficients are slightly larger than the baseline estimates in column (2).

In column (5), I use a sub-sample which better satisfies the model assumptions. In the model, the railroad expands over a rural region where there was no pre-existing town. The towns all

\(^9\)Three sample municipalities have no rural population. This is why I use the inverse hyperbolic sine instead of the logarithm.
emerge after the railroad arrival. This is not exactly the case in the setting I study, as 40% of the sample consists of municipalities that were already incorporated at the time of railroad arrival. These are towns that already existed by the time of the railroad arrival. However, the other 60% of the sample consists of municipalities that were not yet incorporated when the railroad arrived. They either correspond to towns that did not exist before the railroad and that developed around a station that served a rural area, or to towns that were very small by the time of railroad arrival.\textsuperscript{10} When I restrict the sample to these municipalities, the relation between years as endpoints and urban population is larger and statistically significant.

Summing up, there are positive associations between years as endpoint and urban population, urban GDP, or distance to the next municipality. They are consistent with the model predictions.

**Potential Biases**

The correlations just discussed might not necessarily reflect the causal effect of years as endpoint on town sizes. Four potential sources of bias are:

1. **Unobservable geographic fundamentals.** Larger cities are often in sites that have advantageous geographic fundamentals. These fundamentals were probably known to the railroad companies, which would have had an incentive to build a line until the site but less incentives to extend the line beyond it. As a consequence, municipalities with advantageous fundamentals would be railroad endpoints for longer. The ordinary least squares estimates would then pick up the relationship through the omitted fundamental, leading to an *upward bias*.

2. **Shocks to distance to the next town.** Towns that serve a larger rural hinterland will be larger. Indeed, this is what the model predicts. If, conditional on the location of a town, some exogenous factor implies that the location of the next town along the railroad will be farther away, then the town will serve a larger hinterland. The exogenous factor could be a geographic fundamental that makes it attractive to have the next town farther away. Since it is farther away, it could take longer to build the railroad segment until the next town. As a consequence, this exogenous factor would both increase time as endpoint and increase town size, as the served hinterland would be larger. As a consequence, it could introduce an *upward bias* to the estimates.

3. **Reverse causality.** Since railroads were costly investments, railroad companies often used their own profits to finance its construction. A town that, after railroad arrival, grows faster would increase the freight transportation of the railroad, attracting revenue to the company.

\textsuperscript{10}Santa Gertrudes is an example of a town founded after the railroad arrival; originally, it was only a railroad stop to serve a farm in the area. President Bernardes, on the other hand, is an example of a small village that grew after the railroad arrival.
The increased revenue could have been used to speed up construction, so the fast-growing town would be an endpoint for a shorter period. It would also be larger in the long run, as it had grown faster.\footnote{As predicted, for instance, by random urban growth theories; see Gabaix (1999).} As a consequence of it, there would be a downward bias of the estimates.

4. \textbf{Measurement error}. From the ESB data, I only observe the year of railroad arrival, not the exact date (including month and day). Hence, two stations might have been opened in different years, but the time difference between their opening dates might be smaller than for two stations that opened in the same year. For instance, it would be the case when a station opens in December and the next station in January, as opposed to when the former opens in February and the latter in October. This limitation introduces classic measurement error, leading to a downward bias.

Note that the overall direction of the bias is unclear. In the next subsection, I use an instrumental variable approach that suggests a small downward bias and that there is indeed a causal effect of years as endpoint on town size and distance to the next town. But before discussing the IV results, it is useful to assess some of these biases.

Note that (1) and (2) are omitted variables biases. With enough information on fundamentals of a municipality and of its next municipality, there would be no such bias. However, there are many geographic fundamentals that could matter, and some are hard to observe. To assess whether these unobservable fundamentals are a concern, in column (6) of Table 2 I include a variety of observable fundamentals as controls. They are mean altitude and squared mean altitude, mean terrain ruggedness index (TRI) and squared TRI, the logarithm of the potential yield for coffee (the main export crop of the São Paulo economy at the time of the railroad expansion), and the logarithm of the potential yield for maize (the main staple crop). I also include a quadratic polynomial of latitude and longitude. In comparison with column (2), the new coefficients in column (6) barely change. On the other hand, the explanatory power of the model, as measured by the R-squared, increases substantially.

In column (7), I also include the observable geographic fundamentals for the next municipality along the railroad.\footnote{In the few cases when there are more than one next municipality, I use the information from the first next municipality according to railroad arrival year.} Again, the results are robust to the inclusion of this variable.

Following Oster (2019), I can calculate a lower bound on the selection on unobservables, relative to the selection on observables, for the positive effect of years as endpoint on town size to be completely due to the omitted variable bias. This bound is displayed in columns (6) and (7) of Table 2, at the bottom of each panel. For the calculation of the bounds, I use the conservative assumption that the inclusion of the unobservable variables would lead to a model that completely explains the variation in the dependent variable. That is, the new equation would have an R-squared equal
to one. When only a municipality own fundamentals are included in the vector of observables, selection on unobservables would have to be at least as high as selection on observables. And when we also include the fundamentals of the next municipality, selection on observables would have to be at least 3 times as large! However, the same degree of relative selection is not necessary to explain the coefficients in Panel C, where the logarithm of distance to the next municipality is the dependent variable. In this case, selection on unobservables would need to be only 66% of selection on observables.

In this way, the results shown in columns (6) and (7) indicate that the first two biases are unlikely to be a concern, especially when the dependent variable is the logarithm of urban population or of urban GDP. Therefore, any resulting bias is likely to be downward instead of upward. Indeed, this is what the two-stage least squares coefficients in the next section suggest.

**Instrumental Variable Estimates**

To address concerns about bias, in this sub-section I use an instrumental variable to estimate the relationship between years as endpoint and town size using two-stage least squares estimation. The excluded instrument I use is the expansion rate of the Brazilian railroad network in the year after the railroad arrival.

To be precise, let \( t \) be the year of railroad arrival at municipality \( i \). The instrumental variable I use is \( Z_i = \log(\frac{T_{i+1}}{T_i}) \), where \( T_i \) is the total length of the Brazilian rail network at December of year \( t \). Since total rail network growth could depend on the total rail network length, I also include \( \log(T_i) \) as a control variable.

For this exercise, the sample consists of municipalities that were railroad endpoints for at least two years. The sample restriction insures that there is no mechanical relationship between years as endpoint and the growth of the national rail network, as no new segments of the railroad where a sample municipality was an endpoint were constructed in the year after the railroad arrival.

What is the economic intuition behind this exercise? During the era of railroad construction, Brazil was a poor country with underdeveloped capital markets. Railroads were large slumpy investments for which capital was scarce. Summerhill (2003) shows that, in the mid-1880s, the investments required to build even a small railroad were larger than the entire capital stock in the Brazilian cotton manufacturing sector. Furthermore, most railroad companies were Brazilian and depended on national capital. Therefore, when there were more railroad investments elsewhere, capital became scarcer for the expansion of the railroad line beyond an endpoint site. The site would thus be an endpoint for longer.

For \( Z_i \) to satisfy the exclusion restriction, it cannot directly affect town size. This is a reasonable assumption, since the instrument explores time variation that is unrelated with municipality characteristics. In this way, it is unlikely to be related with geographic fundamentals that determine...
city size. The hypothesis is untestable, but in Appendix Table A.1 I show that the instrumental variable does not correlate with any observable geographic fundamental. This result is suggestive that the exclusion restriction holds.

Column (1) of Panel A of Table 3 reports the first stage coefficient. There is a positive and statistically significant relationship between the national rail network growth and years as endpoint. When the rail network growth increases by one standard deviation (0.051), time as endpoint increases in 5.147 years.

The two-stage least square estimates are shown in columns (2) to (4) of Panel A of Table 3. The results indicate positive and statistically significant effects of years as endpoint on urban population, urban GDP, and distance to the next municipality. The point estimates are larger than the ordinary least squares estimates, suggesting that the overall bias of the coefficients in Table 2 was negative. Moreover, there is no effect on distance to the previous municipality. In this way, the two-stage least squares results support the key model predictions.

There are some concerns about the identification strategy behind the instrumental variables approach. First, railroad growth could be related with overall economic conditions which might directly impact the growth of endpoint municipalities. To account for this concern, in Panel B of Table 3 I include the national GDP growth at the year after railroad arrival as a control variable. The coefficients barely change. Second, scarcity in the available capital could also affect other investments. In this case, it would restrict the growth of the endpoint town. However, this violation of the exclusion restriction would lead to a downward bias of the two-stage least squares coefficients. Third, the small sample could be a problem if the instrument is different in some municipality with better geographic fundamentals. In Panel C, I include as controls the same fundamentals as in column (7) of Table 2; the results are robust to it.

Summing up, the instrumental variable approach lends credence to the hypotheses that, as predicted by the theoretical model, more years as endpoint increase town size and distance to the next town. The two-stage least squares coefficients are larger than the ordinary least squares ones, which indicates that the downward biases dominate the upward biases.

**Historical Example**

Contemporary geographers highlighted the importance of railroad endpoints for the emergence and growth of cities in São Paulo; seminal references are Deffontaines (1938) and Monbeig (1949, 1952). Such towns were often described as *bocas do sertão* -“gateways to the frontier” -, the last outposts of the modern and industrializing economy that expanded from the east of the state. In these towns, a variety of merchants and firms provided goods and traded with the farmers who were expanding the agricultural frontier. Monbeig (1949, 1952) argued that, even after the railroad line

---

13I use the GDP series constructed by Araújo et al. (2008).
was extended beyond the *boca do sertão* town and it no longer was an endpoint, the town kept its economic importance. This occurred due to the strength of the businesses established at the town; as Monbeig (1949, p. 64) noted: “The former “boca do sertão” kept, thanks to its well-known traders and bankers, their hospitals, and their schools, its influence over the former frontier.”

Often, it would become an important city in the regional urban hierarchy.

São José do Rio Preto was one of these cities. It is located on a railroad line built by the *Estrada de Ferro Araraquara* (EFA), a Brazilian private company that began its operations in 1896.

The EFA railroad was a promising enterprise (Silva and Tosi 2014). It would serve the expanding agricultural frontier in the northwest of the São Paulo state, where not only coffee was an important activity, but where cattle ranching was also a main industry. Moreover, in 1900 the federal government awarded to the company a concession that allowed it to build until the Mato Grosso capital of Cuiabá. The concession also granted tax benefits to the company. As a consequence, the company increased its indebtedness to finance railroad building.

However, the company soon faced a variety of setbacks. After 1906, the state government imposed a tax on new coffee plantations, which negatively affected the movement of the railroad (Silva and Tosi 2014). Furthermore, the concession to Cuiabá was unexpectedly reviewed and canceled. With its financial conditions aggravated by the mismanagement of an administration that took over the company in 1909, the railroad line reached São José do Rio Preto in 1912.

EFA was then acquired by the São Paulo Northern Railroad Company, a Delaware company that actually represented L. Behrens and Söhne, a German bank that held many EFA bonds. This new company, however, had no investment capacity to expand or even maintain the railroad. After a railroad workers’ strike in 1919, the state government canceled EFA concession and took over the railroad. Government investment was also halted until a long judicial litigation led to the settlement of the government takeover. As a consequence, São José do Rio Preto remained a railroad endpoint from 1912 to 1933.

The period as endpoint is considered a period of fast growth for the city. An excerpt from the municipality entry in the Brazilian Encyclopedia of Municipalities (IBGE 1958, v. 30, p.189) reads: “The arrival of the EFA railroad in 1912 was the beginning of a ‘golden age’ in the development of Rio Preto, which became a center of commerce for the region due to its position on the railroad endpoint.”

Figure 3 shows the gradual opening of the *Estrada de Ferro Araraquara*. The figure also shows

---

14 Author’s translation of: “a antiga ‘boca do sertão’ conseguia manter pela força do hábito, graças a seus negociantes e seus bancos, já conhecidos, os seus hospitais, e seus colégios, uma influência sobre a ex-zona pioneira.”

15 The first railroad segment was completed by 1898.

16 L. Behrens and Söhne had to been affected by World War I in Europe.

17 Author’s translation of: “a chegada dos trilhos da Estrada de Ferro Araraquara em 1912, marca o início da ”Era Áurea”, no desenvolvimento do então Rio Prêto transformando-o num empório comercial da região, em virtude de sua situação de ponta de linha férrea.”
each municipality urban population in 2010. In relation to the other municipalities, São José do Rio Preto was a railroad endpoint for much longer. The difference is unlikely to be due to large differences in geographic fundamentals, as the financial difficulties faced by the company were not due to characteristics of São José do Rio Preto. Today, São José do Rio Preto is the largest town along the railroad. The distance to the next municipality, Mirassol, is also larger than the average distance to the next municipality along the EFA. These facts point out for the urban history of São José do Rio Preto as illustrative of the model predictions.

4 Conclusion

In this paper, I argue that the timing of gradual construction of transportation infrastructure can have persistent effects on the spatial distribution of economic activity. First, I present a stylized model of urban emergence along an expanding railroad. The model has two key predictions: town size and distance to the next town along the railroad are increasing in the time a town spent as a railroad endpoint. Second, I confirm these predictions in the context of railroad expansion into the west of the São Paulo state. Geographers have already pointed out the importance of railroad endpoints for urban growth in this setting. I quantify the positive relationships and use an instrumental variable approach to provide suggestive evidence that the positive relationships reflect a causal effect.

The intuition for why a town that was a railroad endpoint for longer will be persistently larger is that, since it takes a longer time for the railroad to reach a site farther down the road, the firms in the endpoint town will be more productive, reducing the incentives for new firms to establish at that site. An important limitation of this paper is that I do not distinguish the reasons behind productivity growth. It could be simply a consequence of firms becoming more productive over time as a consequence of experience in a local market. For instance, they learn about their customer needs. However, firms could also become more productive as a consequence of the experience of other neighboring firms; that is, as a consequence of dynamic agglomeration economies, as in Rauch (1993). In this case, productivity growth in railroad endpoints should be actually faster than in other towns, as firms in endpoint towns serve a larger market. Future theoretical and empirical work in other settings could address this question.\footnote{Without historical data on productivity and without a second instrument for years as endpoint, I cannot distinguish the two hypotheses. Accordingly, the model abstracts from the specific mechanism by assuming that productivity growth rate within towns is deterministic.}

The question of how the gradual construction of transportation infrastructure impacts the spatial distribution of economic activity is relevant to policy makers, particularly in developing economies where funds for investment in infrastructure are scarce. The timing of road construction is a policy choice. I show it is a consequential one. There is a need for richer models to understand the optimal
time of road building when funds are scarce.

References


A Figures

Figure 1: Sample municipalities, railroad lines, and years as endpoint

Note: Railroad lines in blue; Estação da Luz, the main station in the city of São Paulo, is indicated by the blue circle. The black lines delineate sample municipalities, while the dotted gray line delineates the state of São Paulo. Years as endpoint are represented in shades of gray, with darker shades indicating more years.
Figure 2: Years as endpoint and town size

(a) Urban population

(b) Urban GDP

(c) Distance to next municipality

Note: Each circle is a municipality. If years as endpoint are larger or equal than 2, the circle is solid and blue; otherwise, it is hollow and gray. The dashed line indicates the predicted values of the linear fit, with constant but no controls. The solid line indicates the linear fit, conditional on years as endpoint larger or equal than 2. For the linear regressions, municipalities are weighted by their area.
Figure 3: Railroad arrival dates and town sizes along the *E.F. Araraquara*

Note: The navy line indicates the year of railroad arrival. The circles indicate the total urban population in 2010. The green ticks on the horizontal axis indicate the distance of each municipality to São Paulo.
## B Tables

Table 1: Summary statistics

<table>
<thead>
<tr>
<th>Panel A: all municipalities</th>
<th>Obs.</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years as endpoint</td>
<td>203</td>
<td>2.24</td>
<td>4.18</td>
</tr>
<tr>
<td>Years as endpoint ≥ 2</td>
<td>203</td>
<td>0.30</td>
<td>0.46</td>
</tr>
<tr>
<td>Population ('000s)</td>
<td>203</td>
<td>84.28</td>
<td>146.36</td>
</tr>
<tr>
<td>Population density</td>
<td>203</td>
<td>153.69</td>
<td>389.28</td>
</tr>
<tr>
<td>Urban population ('000s)</td>
<td>203</td>
<td>80.86</td>
<td>143.89</td>
</tr>
<tr>
<td>Urbanization rate</td>
<td>203</td>
<td>0.90</td>
<td>0.09</td>
</tr>
<tr>
<td>Per capita GDP (2010 PPP US$)</td>
<td>203</td>
<td>16032.40</td>
<td>9245.92</td>
</tr>
<tr>
<td>Distance to São Paulo (km)</td>
<td>203</td>
<td>404.29</td>
<td>160.30</td>
</tr>
<tr>
<td>Distance to next municipality (km)</td>
<td>203</td>
<td>18.78</td>
<td>9.64</td>
</tr>
<tr>
<td>Year of railroad arrival</td>
<td>203</td>
<td>1911.54</td>
<td>24.09</td>
</tr>
<tr>
<td>Year of incorporation</td>
<td>203</td>
<td>1912.20</td>
<td>50.71</td>
</tr>
<tr>
<td>Incorporated before railroad arrival</td>
<td>203</td>
<td>0.40</td>
<td>0.49</td>
</tr>
<tr>
<td>Altitude (m)</td>
<td>203</td>
<td>540.36</td>
<td>128.16</td>
</tr>
<tr>
<td>Terrain ruggedness index (m)</td>
<td>203</td>
<td>7.22</td>
<td>2.53</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: conditional on ≥ 2 years as endpoints</th>
<th>Obs.</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years as endpoint</td>
<td>49</td>
<td>6.79</td>
<td>5.26</td>
</tr>
<tr>
<td>Population ('000s)</td>
<td>49</td>
<td>138.53</td>
<td>214.35</td>
</tr>
<tr>
<td>Population density</td>
<td>49</td>
<td>218.63</td>
<td>331.35</td>
</tr>
<tr>
<td>Urban population ('000s)</td>
<td>49</td>
<td>134.24</td>
<td>211.15</td>
</tr>
<tr>
<td>Urbanization rate</td>
<td>49</td>
<td>0.94</td>
<td>0.05</td>
</tr>
<tr>
<td>Per capita GDP (2010 PPP US$)</td>
<td>49</td>
<td>18200.46</td>
<td>8068.09</td>
</tr>
<tr>
<td>Distance to São Paulo (km)</td>
<td>49</td>
<td>381.95</td>
<td>148.83</td>
</tr>
<tr>
<td>Distance to next municipality (km)</td>
<td>49</td>
<td>22.67</td>
<td>11.30</td>
</tr>
<tr>
<td>Year of railroad arrival</td>
<td>49</td>
<td>1910.19</td>
<td>25.03</td>
</tr>
<tr>
<td>Year of incorporation</td>
<td>49</td>
<td>1892.79</td>
<td>61.02</td>
</tr>
<tr>
<td>Incorporated before railroad arrival</td>
<td>49</td>
<td>0.64</td>
<td>0.48</td>
</tr>
<tr>
<td>Altitude (m)</td>
<td>49</td>
<td>565.63</td>
<td>112.38</td>
</tr>
<tr>
<td>Terrain ruggedness index (m)</td>
<td>49</td>
<td>7.31</td>
<td>2.26</td>
</tr>
</tbody>
</table>

Note: Observations are weighted by municipality area.
Table 2: Years as endpoint and town size

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: Log urban population (columns 1-2, 4-9), IHS rural population (column 3)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years as endpoint</td>
<td>0.107**</td>
<td>0.083**</td>
<td>0.018</td>
<td>0.103**</td>
<td>0.117**</td>
<td>0.084**</td>
<td>0.106**</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.025)</td>
<td>(0.021)</td>
<td>(0.026)</td>
<td>(0.041)</td>
<td>(0.024)</td>
<td>(0.025)</td>
</tr>
<tr>
<td>Railroad arrival year</td>
<td>-0.004</td>
<td>-0.010</td>
<td>0.019</td>
<td>-0.005</td>
<td>-0.032*</td>
<td>-0.032*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.008)</td>
<td>(0.012)</td>
<td>(0.009)</td>
<td>(0.013)</td>
<td>(0.013)</td>
<td></td>
</tr>
<tr>
<td>Log distance to São Paulo</td>
<td>-0.515+</td>
<td>0.519</td>
<td>-1.741**</td>
<td>-0.238</td>
<td>-1.148</td>
<td>-1.075</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.277)</td>
<td>(0.463)</td>
<td>(0.548)</td>
<td>(0.348)</td>
<td>(0.960)</td>
<td>(0.923)</td>
<td></td>
</tr>
<tr>
<td>Incorporated municipality</td>
<td>0.659*</td>
<td>0.588**</td>
<td>0.321</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.256)</td>
<td>(0.179)</td>
<td>(0.340)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.119</td>
<td>0.276</td>
<td>0.103</td>
<td>0.375</td>
<td>0.096</td>
<td>0.399</td>
<td>0.462</td>
</tr>
<tr>
<td>Observations</td>
<td>203</td>
<td>203</td>
<td>203</td>
<td>49</td>
<td>143</td>
<td>203</td>
<td>203</td>
</tr>
</tbody>
</table>

Proportional selection for $\beta = 0$:

<table>
<thead>
<tr>
<th></th>
<th>1.014</th>
<th>3.096</th>
</tr>
</thead>
</table>

**Panel B: Log urban GDP (columns 1-2, 4-9), log agriculture GDP (column 3)**

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years as endpoint</td>
<td>0.125**</td>
<td>0.098**</td>
<td>0.011</td>
<td>0.108**</td>
<td>0.132**</td>
<td>0.099**</td>
<td>0.123**</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.028)</td>
<td>(0.014)</td>
<td>(0.028)</td>
<td>(0.047)</td>
<td>(0.028)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.112</td>
<td>0.304</td>
<td>0.287</td>
<td>0.395</td>
<td>0.107</td>
<td>0.420</td>
<td>0.472</td>
</tr>
<tr>
<td>Observations</td>
<td>203</td>
<td>203</td>
<td>203</td>
<td>49</td>
<td>143</td>
<td>203</td>
<td>203</td>
</tr>
</tbody>
</table>

Proportional selection for $\beta = 0$:

<table>
<thead>
<tr>
<th></th>
<th>1.161</th>
<th>3.178</th>
</tr>
</thead>
</table>

**Panel C: Log distance to next mun. (columns 1-2, 4-9), log distance to previous mun. (column 3)**

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years as endpoint</td>
<td>0.045**</td>
<td>0.032**</td>
<td>-0.013</td>
<td>0.044*</td>
<td>0.026*</td>
<td>0.031**</td>
<td>0.035**</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.009)</td>
<td>(0.010)</td>
<td>(0.022)</td>
<td>(0.013)</td>
<td>(0.009)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.123</td>
<td>0.245</td>
<td>0.108</td>
<td>0.222</td>
<td>0.054</td>
<td>0.292</td>
<td>0.330</td>
</tr>
<tr>
<td>Observations</td>
<td>203</td>
<td>203</td>
<td>203</td>
<td>49</td>
<td>143</td>
<td>203</td>
<td>203</td>
</tr>
</tbody>
</table>

Proportional selection for $\beta = 0$:

<table>
<thead>
<tr>
<th></th>
<th>0.423</th>
<th>0.657</th>
</tr>
</thead>
</table>

Intensive margin? Y
Unincorporated mun. only? Y
Baseline controls? Y Y Y Y Y
Fundamentals as controls? Y Y
Fundamentals of next mun. as controls? Y

Note: Coefficients from OLS. Observations are weighted by municipality area. Robust standard errors in parentheses. Statistical significance denoted by: + 10%, * 5%, ** 1%
Table 3: Two-stage least squares estimates

<table>
<thead>
<tr>
<th></th>
<th>First stage</th>
<th>Log urban GDP</th>
<th>Log urban next mun.</th>
<th>Log distance previous mun.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>Panel A: baseline 2SLS results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log growth of national railroad network, next year</td>
<td>100.289** (25.267)</td>
<td>0.158* (0.065)</td>
<td>0.187* (0.079)</td>
<td>0.054* (0.027)</td>
</tr>
<tr>
<td>Years as endpoint</td>
<td>0.158*</td>
<td>0.187*</td>
<td>0.054*</td>
<td>-0.002</td>
</tr>
<tr>
<td>Observations</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>Robust F-statistic of the excluded instrument</td>
<td>15.754</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Panel B: includes national GDP growth as controls</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log growth of national railroad network, next year</td>
<td>101.673** (25.019)</td>
<td>0.153* (0.064)</td>
<td>0.179* (0.078)</td>
<td>0.058* (0.029)</td>
</tr>
<tr>
<td>Years as endpoint</td>
<td>0.153*</td>
<td>0.179*</td>
<td>0.058*</td>
<td>0.004</td>
</tr>
<tr>
<td>Observations</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>Robust F-statistic of the excluded instrument</td>
<td>16.515</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Panel C: includes obs. fundamentals as controls</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log growth of national railroad network, next year</td>
<td>106.269** (23.170)</td>
<td>0.187** (0.039)</td>
<td>0.227** (0.049)</td>
<td>0.063* (0.027)</td>
</tr>
<tr>
<td>Years as endpoint</td>
<td>0.187**</td>
<td>0.227**</td>
<td>0.063*</td>
<td>0.003</td>
</tr>
<tr>
<td>Observations</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>Robust F-statistic of the excluded instrument</td>
<td>21.036</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Coefficients from 2SLS, except column (1), which is estimated by OLS. Control variables in all regressions are the year of railroad arrival, log distance to São Paulo, an indicator of whether the municipality was already incorporated by the time of railroad arrival, and the logarithm of the total railroad extension in Brazil. In Panel B, the GDP growth in the next year after the railroad arrival is also included as a control. Observations are weighted by municipality area. Robust standard errors in parentheses. Statistical significance denoted by: + 10%, * 5%, ** 1%
C  Proof of Proposition 1

The proof has three steps.

STEP 1: I show that, if \( y \notin \mathbb{Y}(t')(y) \), then \( y \notin \mathbb{Y}(t) \) for any \( t > t'(y) \). That is, I show that all existing cities emerge at the time the railroad reaches their site.

To see this is the case, note that for any \( y \notin \mathbb{Y}(t) \cap [0, y'(t)] \), it must be the case that \( \log q(y, t) < -A(y, t) \). In particular, if \( y \notin \mathbb{Y}(t'(y)) \), then \( \log q(y, t'(y)) < 0 \). In this way, there is some \( t > t'(y) \) such that:

\[
\log q(y, t') < \log q(y, t'(y)) \leq -g(0)[t' - t'(y)] < -A(y, t')
\]

where the first inequality comes from the fact that prices are decreasing, the second inequality that \( t' - t'(y) \) is sufficiently small, and the third inequality by the fact that \( g(0)[t' - t'(y)] \) is an upper bound on productivity growth. Hence, for any \( t \in [t'(y), t'] \), \( y \notin \mathbb{Y}(t) \). As a consequence, \( A(y, t') = 0 \). We can then iteratively repeat this same argument to show that, for any \( t \in [t', t' + k(t' - t'(y))] \), \( y \notin \mathbb{Y}(t) \), where \( k \) is an arbitrary integer. Hence, \( y \notin \mathbb{Y}(t) \) for any \( t > t'(y) \).

STEP 2: define \( \tilde{\mathbb{Y}}(t) = \bigcup_{t' \leq t} \mathbb{Y}(t) \). This is the set of all sites that are or ever were a town. I show that, for any \( t, \tilde{\mathbb{Y}}(t) \) is a finite set. As a consequence, \( \mathbb{Y}(t) \) must also be finite, as it is a subset of \( \tilde{\mathbb{Y}}(t) \). First, note that as \( \tilde{\mathbb{Y}}(t) \subset [0, y'(t)] \), then \( \tilde{\mathbb{Y}}(0) = \mathbb{Y}(t) = \{y_0 = 0\} \). Let \( A(0, 0) \) be the initial productivity at the port town. The log price of urban services from \( y_0 \) at site \( y'(t) \) will be no higher then:

\[
\log p(y'(t), t) \leq -A(0, 0) + \tau_s y'(t).
\]

Therefore, define:

\[
t_1 = \frac{A(0, 0)}{\tau_s \sup g'(t)}
\]

as the least time it takes for the railroad to reach a site where it is profitable for a firm to compete with firms at the port town, assuming no productivity growth in the later. Hence, for \( t \leq t_1 \), it must be that \( \tilde{\mathbb{Y}} = \{0\} \).

Now consider a time \( t \) such that \( y'(t) \in \mathbb{Y}(t) \). Define \( \Delta t \) as the positive number that satisfies:

\[
A(y'(t), t + \Delta t) = \left( \tau_s \sup g'(t) \right) \Delta t
\]

Since \( A(y'(t), t + \Delta t) \) is concave in \( \Delta t \) and \( \frac{\partial A(y'(t), t)}{\partial \Delta t} = g(0) > \tau_s \sup g'(t) \) (by Assumption 1), then \( \Delta t \) is well-defined. Note that for any \( t' \in (t, t + \Delta t) \), \( \log q(y'(t'), t') < 0 \), so no town will emerge in that interval. As a consequence, we can bound the cardinality of \( \tilde{\mathbb{Y}}(t) \) by:

\[
\# \tilde{\mathbb{Y}}(t) \leq \begin{cases} 
1, & \text{if } t < t_1 \\
2, & \text{if } t \in [t_1 + (k - 1)\Delta t, t_1 + k\Delta t]
\end{cases}
\]

STEP 3: now I define a sequence of finite-dimensional dynamical systems that have each a unique solution and that, jointly with an updating rule for the set of existing towns, characterizes the unique equilibrium.

Dynamical system 1: it is a one dimensional dynamical system with \( \mathbf{A}(t) = (A_0(t)) \). It is characterized by the following differential equation and the initial value:
\[
\begin{align*}
\dot{A}_0(t) &= g(A_0(t)) \\
A_0(0) &= A(0, 0)
\end{align*}
\]

The equation above defines an initial value problem (call it \(P_0\)) which is Lipschitz continuous in \(A\) and continuous in \(t\). Hence, by the Picard-Lindelöf Theorem, \(P_0\) has a unique and continuous solution \(A_0(t)\).

Now define \(t_1 \equiv \min\{t > 0 | \tau_y(t) \geq A_0(t)\}\). If no such \(t_1\) exists, then there will never exist any town other than \(y_0\). Otherwise, then \(\mathcal{Y}(t_1) = \{y_0, y_1\}\), where \(y_1 = y^\tau(t_1)\). Moreover, since \(A_0(t)\) is continuous, so is \(\tau_y(t) - A_0(t)\). Hence, \(\tau_y(t_1) \geq A_0(t_1)\).

**Dynamical system \(N\):** at \(t_N\), let \(\mathcal{Y}(t_N) = \{y_0, y_1, ..., y_m\}\), where \(y_m = y^\tau(t_N)\).\(^{19}\) We define \(A(t) = (A_0(t), ..., A_m(t))\) and the initial value problem \(P_N\) (with initial time \(t_N\)) as:

\[
\begin{align*}
\dot{A}_i(t) &= g(A_i(t)) \quad i = 0, 1, ..., m \\
A_i(t_N) &= A(y_i, t_N) \quad i = 0, 1, ..., m
\end{align*}
\]

which, by the Picard-Lindelöf Theorem, has a unique and continuous solution \(A(t)\).

Now I show that, along \(A(t)\), all towns in \(\mathcal{Y}(t_N)\) will also be a town at a future time (Claim 1). This occurs if, for any towns \(y_i\) and \(y_j\), the following is true:

\[-A(y_i, t) \leq -A(y_j, t) + \tau_y |y_i - y_j|\]

The equation above holds for \(t_N\). Suppose \(A(y_j, t_N) \leq A(y_i, t_N)\), then it must be the case at any future time \(t > t_N\); hence, the inequality above will hold as well. Now suppose that \(A(y_j, t_N) > A(y_i, t_N)\), then \(g(A(y_j, t_N)) \leq g(A(y_i, t_N))\), so the left-hand side of the equation above decreases faster than the right-hand side; hence, the inequality is also satisfied at any \(t > t_N\). This proves the claim.

Now define:

\[t_{N+1} = \min\{t > t_{N+1} | \tau_y(y^\tau(t) - y_m) \geq A_m(t)\}\]

which will actually hold in equality, since \(A_m(.)\) is continuous.

This completes the proof of uniqueness of the equilibrium. As a consequence of Claim 1, \(\mathcal{Y}(t) = \tilde{\mathcal{Y}}(t)\). Moreover, the recursive procedures to identify the sequence \(\{t_1, t_2, \ldots\}\) is the same recursive procedure explained in Proposition 1.

\(^{19}\)Note that, since \(\mathcal{Y}(t) \subset \tilde{\mathcal{Y}}(t)\), then \(m \leq N\). I will later show that \(\mathcal{Y}(t) = \tilde{\mathcal{Y}}(t)\).
Figure A.1: Sample railroads and nightlights

Note: Railroad lines in blue; *Estação da Luz*, the main station in the city of São Paulo, is indicated by the blue circle. The dotted gray line delineates the state of São Paulo. Nightlights in orange refer to 2010.
Table A.1: Balance of the instrument with respect to observable fundamentals

<table>
<thead>
<tr>
<th></th>
<th>Mean altitude (1)</th>
<th>Mean TRI (2)</th>
<th>Pot. yield: coffee (3)</th>
<th>Pot. yield: maize (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log growth of national railroad network, next year</td>
<td>-306.148</td>
<td>-1.351</td>
<td>-1.467</td>
<td>0.348</td>
</tr>
<tr>
<td></td>
<td>(304.834)</td>
<td>(6.500)</td>
<td>(1.239)</td>
<td>(3.527)</td>
</tr>
<tr>
<td>Log of national railroad extension</td>
<td>8.090</td>
<td>0.150</td>
<td>-0.048</td>
<td>0.171</td>
</tr>
<tr>
<td></td>
<td>(53.886)</td>
<td>(0.987)</td>
<td>(0.134)</td>
<td>(0.521)</td>
</tr>
<tr>
<td>Railroad arrival year</td>
<td>-2.745*</td>
<td>0.064*</td>
<td>-0.000</td>
<td>0.032**</td>
</tr>
<tr>
<td></td>
<td>(1.049)</td>
<td>(0.024)</td>
<td>(0.003)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>Log distance to São Paulo</td>
<td>-42.795</td>
<td>-4.838**</td>
<td>0.005</td>
<td>-0.916</td>
</tr>
<tr>
<td></td>
<td>(51.912)</td>
<td>(1.610)</td>
<td>(0.168)</td>
<td>(0.615)</td>
</tr>
<tr>
<td>Incorporated municipality</td>
<td>19.673</td>
<td>-0.984</td>
<td>0.081</td>
<td>0.270</td>
</tr>
<tr>
<td></td>
<td>(34.171)</td>
<td>(0.656)</td>
<td>(0.076)</td>
<td>(0.348)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.410</td>
<td>0.374</td>
<td>0.064</td>
<td>0.252</td>
</tr>
<tr>
<td>Observations</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
</tbody>
</table>

Note: Coefficients from OLS. Observations are weighted by municipality area. Robust standard errors in parentheses. Statistical significance denoted by: + 10%, * 5%, ** 1%
E Data Appendix

Railroad data. The coordinates of each railroad station are from the shapefile by the Agência Nacional e Transportes Terrestres. The shapefile contains information on the municipality of each station, the current operator, the station code, and the station name. I match each station by name with the Estações Ferroviárias do Brasil; see http://www.estacoesferroviarias.com.br. For each station, I assign the year it started operating. I then define the year of railroad arrival as the minimum between the year of operation start of that station or of any other station that follows from it. All railroads in the sample were constructed from São Paulo towards the west, so this procedure indeed identifies the year of railroad arrival at each site. For each station, we can define the station that precedes it; for each station, one can go back and reach the São Paulo main station, Estação da Luz, after some iterations. I define the distance to the previous station as the Euclidean distance between the station and the previous station. The station distance to the São Paulo is the sum of all such distances until reaching the Estação da Luz.

For each municipality, railroad arrival year and distance to São Paulo refer to the first station in the municipality. Time as endpoint is the difference between railroad arrival year in that municipality and the lowest railroad arrival year in a next municipality. Distance to the next municipality is the difference between distance to São Paulo in that municipality and the lowest distance to São Paulo in a next municipality.

There are eight railroad companies in the sample: São Paulo Railway, Companhia Paulista de Estradas de Ferro, Estrada de Ferro Sorocabana (includes the Ituana), Estrada de Ferro São Paulo-Minas, Companhia Mogiana de Estradas de Ferro, Estrada de Ferro Araraquara, Estrada de Ferro Noroeste do Brasil, Estrada de Ferro São Paulo-Paraná. They include lines in the neighboring states of Paraná and Minas Gerais that were a continuation of the São Paulo rail network. I include only the railroads that are present in both the ANTT shapefile and Estações Ferroviárias do Brasil. As a consequence, I do not have the Estrada de Ferro Douradense nor the abandoned segment of the Noroeste after Araçatuba; this segment was replaced by an alternative route that I have in my sample. I also cannot trace back the evolution of the Mogiana in the state of Minas Gerais. Finally, I do not include railroad segments built by FEPASA after the 1970s; the purpose of these segments was to connect some of these lines after they were all taken over by the government.

Sample selection. I include all municipalities along the sample railroads, with the exception of: (a) São Paulo, which is the origin of the network; (b) Bauru, the only municipality where there is a crossing of railroads (the Paulista, Sorocabana and Noroeste all go through the city); (c) the railroad endpoints today.

Population and GDP. Urban and rural populations are from the 2010 Brazilian census. Municipality GDP estimates are from IBGE. Municipality coordinates indicate the centroid of the municipality territory. Municipality incorporation year was manually collected at IBGE Cidades; see https://cidades.ibge.gov.br. Agricultural suitability. Coffee and maize potential yields are from the Global Agro-ecological Zones (GAEZ) dataset from FAO. In all cases, I use the potential yield under rainfed, intermediate-input agriculture. I calculate averages over the municipality territory.

Digital elevation model. The altitude and ruggedness variable is created using as source the Japan Aerospace Exploration Agency (JAXA) ALOS Global Digital Surface Model. The terrain ruggedness index (TRI) is the square root of the average squared differences in altitude between each pixel and its eight surrounding pixels. The mean of the variables is taken over the municipality area.

Historical series. The historical series of the Brazilian total rail network extension are from the Anuário Estatístico do Brasil, a series of publications from IBGE. The estimates of Brazilian GDP are from Araújo et al. (2008).