WARWICK
ECONOMICS

Railways and cities in India

James Fenske, Namrata Kala & Jinlin Wei

(This paper also appears as CAGE Discussion paper 559)

April 2021 No: 1349

(Accepted for publication in the Journal of Development Economics)

Warwick Economics Research Papers

ISSN 2059-4283 (online)

ISSN 0083-7350 (print)

RAILWAYS AND CITIES IN INDIA

JAMES FENSKE[†], NAMRATA KALA[⋆], AND JINLIN WEI[‡]

ABSTRACT. Using a new dataset on city populations in colonial India, we show that the railroad network increased city size in the period 1881 to 1931. Our baseline estimation approach includes fixed effects for city and year, and we construct instrumental variables for railroad proximity based on distance from a least cost path spanning cities that existed prior to the start of railroad construction. Cities that increased market access due to the railroad grew, particularly those cities that were initially small and isolated.

1. INTRODUCTION

How did the spread of the railroad shape the size of cities in colonial India? Governments in developing countries today make large investments in transportation infrastructure; in India, for example, the government's flagship road-building program aims to connect more than 175,000 settlements to all-weather roads.[1] Evidence of the impact of these investments, however, is often limited to developed countries and can only consider outcomes over a short time horizon, creating scope for historical evidence to improve our understanding of their effects (Berger and Enflo, 2017; Donaldson and Hornbeck, 2016). The growth of cities is a particular challenge in developing countries like those we consider; the overwhelming bulk of urbanization over the next three decades will occur in Asia and Africa, where congestion, contagion, and other difficulties of density are particularly acute (Bryan et al., 2020). Urbanization in developing countries also displays features distinct from those in developed countries, potentially challenging conventional models of spatial equilibrium (Henderson and Kriticos, 2018; Henderson et al., 2018; Henderson and Turner, 2020). In this paper, we seek to understand one aspect of the origins of urbanization in South Asia.

We introduce a new dataset on cities of at least 1,000 persons in colonial India. Our data are taken from the 1931 census of India, and cover modern-day Bangladesh, Burma,

[†]UNIVERSITY OF WARWICK
[⋆]SLOAN SCHOOL OF MANAGEMENT, MASSACHUSETTS INSTITUTE OF TECHNOLOGY
[‡]UNIVERSITY OF WARWICK
E-mail addresses: J.Fenske@warwick.ac.uk, kala@mit.edu, Jinlin.Wei@warwick.ac.uk.
Date: April 18, 2021.
We are grateful to Dave Donaldson and seminar audiences at George Mason University and the University of Warwick for comments.
[1]https://pmgsy.nic.in/

India and Pakistan. There are 2,456 distinct cities for which population is reported, and the data cover the years 1881, 1891, 1901, 1911, 1921, and 1931. We have geocoded these data ourselves, and one contribution of this paper is the introduction of this dataset. Our baseline specification is a fixed effects model, estimated using ordinary least squares (OLS). We include fixed effects for city and year, and ask whether proximity to the railroad predicts log population size. Our OLS results suggest a negative elasticity of city size with respect to distance from a railroad that is between -0.017 and -0.019, corresponding to a standardized magnitude of roughly 5% of a standard deviation. So, while railway access spurred city growth in colonial India, the impact of railways on urban geography is less than that estimated in other developing and developed countries.

Because of possible biases in this fixed effects estimation, we employ a number of instrumental variables (IV) strategies. Our principal instrument is based on the use of a least cost path similar to the one constructed by Bogart et al. (2020) for nineteenth-century England. This path connects pairs of cities that existed prior to the railway that are selected based on their market potential. The paths between them are chosen to minimize construction costs that are parameterized using data on terrain slope at the grid cell level. We use the fact that proximity to this least cost path predicts the speed at which cities gained railway proximity to construct our instrument, and find elasticities that are much larger than our OLS estimates, ranging from -0.113 to -0.191. This difference in magnitudes may be plausibly attributed to the negative selection of certain railway lines, heterogeneous responses to railway connection, and measurement error in railway proximity.

To understand the mechanisms that connect railways to cities in colonial India, we turn first to a major concept that links transportation costs with equilibrium population in several models of economic geography: market access (e.g. Donaldson and Hornbeck (2016); Redding and Sturm (2008)). This is a measure of the access that firms and consumers in a given location have to the firms and consumers in all other locations, scaled down by the costs of reaching these other locations. That is, market access measures the degree to which one city is exposed to supply and demand forces from all other cities. We estimate elasticities of city size with respect to market access that range from 0.385 to 0.628 via OLS and 1.028 and 1.370 via IV. In heterogeneity analyses, we show that railways increased city size most where their impact on market access was greatest: initially smaller and more isolated cities. Similarly, their impact was attenuated for cities with alternative transport links such as ports and rivers, in regions suitable for cash crop (i.e. cotton) cultivation, and where military motives directed railroad placement.

We show that our results are robust to computing market access based on distant markets, to alternative functional forms for physical proximity, and to alternative parameterizations of market access. We show that they are not driven by outliers in terms of railway proximity, city size, or statistical influence. They do not depend on the inclusion of modern-day Burma in the sample. They survive comparing two cities in the same district in the same year. Alternative constructions of our least cost path instrument give results similar to our baseline.

1.1. **Contribution.** We engage first with a literature on the economic effects of transportation infrastructure. Studies that have evaluated the modern effects of transportation infrastructure have linked roads, highways, and railways to several outcomes. These include education (Adukia et al., 2020; Aggarwal, 2018) innovation (Agrawal et al., 2016), the structure of employment (Asher and Novosad, 2020; Pérez, 2018), city growth and shape (Baum-Snow, 2007; Baum-Snow et al., 2012), trade and migration (Morten and Oliveira, 2016), urbanization and specialization (Forero et al., 2020) and economic growth (Banerjee et al., 2020; Faber, 2014).

Existing work on transportation infrastructure and the growth of cities in developing countries largely uses recent data or data from colonial Africa, which had little pre-rail urbanization. We consider a developing-country historical context in which pre-rail urbanization was extensive when compared with pre-colonial Africa.

Within the literature on the impacts of transportation infrastructure, studies focused on economic history date at least to Fogel (1964). In more recent work, railways and other transportation infrastructure have been linked to industrialization (Atack et al., 2011), land values (Donaldson and Hornbeck, 2016), city growth (Atack et al., 2010; Jedwab et al., 2017; Jedwab and Moradi, 2016), and long-run development (Bertazzini, 2018; Okoye et al., 2019).

This historical literature gives us reason to expect that the expansion of the railway network under colonial rule may have had an effect on the size of cities. However, the already relatively mature state of urbanization in India compared to sub-Saharan Africa at the time the railway was introduced, the poor performance of colonial industry, the already-known low level and slow growth of Indian urbanization before independence, and the disparities in methods used across studies give us reason to expect that results that have been found in other contexts need not necessarily apply to colonial India. Our results are particularly resonant with those of Okoye et al. (2019), who consider the modern-day effects of transportation in Nigeria, finding effects in the North but not the South. The North and South have had many historical differences, one of which is the relatively high levels of pre-colonial urbanization among the Yoruba (Bascom, 1955).

We further engage with a literature on the long-run causes of Indian development, with a particular focus on the impact of colonial rule. There exists a tension between two sets of results in this literature. On the one hand, indicators of economic development such as income per capita, real wages, and industrialization suggest that India's development stagnated or even declined for much of the colonial period (Allen et al., 2011; Broadberry et al., 2015; Broadberry and Gupta, 2006; Clingingsmith and Williamson, 2008; Gupta, 2019). This has led many to question whether the actions taken by India's British rulers promoted economic development. On the other hand, there are several findings that suggest that many colonial activities had measurable economic benefits during the colonial period, and that many of effects of colonial activities – both beneficial and harmful – have persisted to the present (Banerjee and Iyer, 2005; Castelló-Climent et al., 2018; Chaudhary and Garg, 2015; Iyer, 2010).

Given our particular focus on railways, there is a debate on whether colonial railroads were to India's economic benefit.[2] Critics of the railways have focused on low productivity, high freight rates, and guaranteed returns to investors; they have argued that investment that should have gone into irrigation was misdirected into the railways (Hurd, 2012; Hurd and Kerr, 2012; Sweeney, 2011). Bogart and Chaudhary (2015, p. 157) have cited the slow growth of urbanization in India as one reason to doubt the transformative impact of the railroad. Further, the failure of India's railways to generate backward linkages may help explain the country's disappointing industrial performance under colonial rule (Parthasarathi, 2011). Recent empirical work, has, on the other hand, shown that India's railways drove price convergence (Andrabi and Kuehlwein, 2010), reduced vulnerability to famine (Burgess and Donaldson, 2010, 2017) and increased agricultural incomes (Donaldson, 2018). Hurd (1983) calculates a social savings for the Indian railway much larger than what Fogel (1964) computed for the United States, but smaller than has been identified in other developing countries (Bogart and Chaudhary, 2015). The effects, of course, were heterogeneous depending on what alternative modes of transportation existed (Roy, 2012). Our results add further evidence of the effect of India's colonial railroads on that country's economic transformation.

In section 2, we provide background on India's cities and railroads and outline the potential conceptual links between them. In section 3, we describe our data. In section 4, we outline our empirical strategy. In section 5, we present our results. Section 6 concludes.

[2]See Bogart and Chaudhary (2015) and Kuehlwein (2021) for reviews of the literature on the economic impact of the colonial railway in India.

2. Background and Conceptual Framework

2.1. **Urbanization in colonial India.** Owing in part to their advantages in soil, rainfall, and natural transportation, the floodplains of the Ganges and Indus were historically more urban than peninsular India (Roy, 2011, p. 21). Gujurat too was historically more urban than other regions (Roy, 2011, p. 56), as were the wet, rice-growing areas (Tomlinson, 2013, p. 29). While information on urbanization in India prior to the census of 1872 is limited, Visaria and Visaria (1983, p. 519) cite estimates from Gadgil (1959) that the net growth of urbanization from 1800 to 1872 was negative, with growth in the presidency cities of Calcutta, Bombay and Madras being offset by the decline of older capital towns such as Lucknow. While the region had many urban centres, thousands of which appear in our data, urbanization was low when compared, for example, to Europe (Tomlinson, 2013, p. 3). The fraction of the population living in towns or cities of at least 5,000 was 8.7% in 1872 (Visaria and Visaria, 1983, p. 519); de Vries (1984, p. 76), by contrast, estimates that 10.8% of the population of Western Europe lived in towns of at least 5,000 in 1600. This measure of Indian urbanization increased slowly and without acceleration to 11.1% in 1931 (Visaria and Visaria, 1983, p. 519).[3] Of this urban population in 1931, some 27.4% lived in cities of 100,000 or more (Bose and Bhatia, 1980, p. 50).

In the census reports, colonial officials proposed a wide range of contradictory factors that drove differences in urbanization and its growth across regions of India, including race, rainfall, plague, famines, and accidents of history (Bose and Bhatia, 1980, p. 76). The increases in urbanization that existed over the period 1881 to 1931 were driven largely by rural-urban migration, and not by differences in fertility and mortality (Visaria and Visaria, 1983, p. 521). Many of these migrants were recruited by labor contractors (Gupta, 2015, p. 74). Some of these workers migrated out of caste-based "attached" labor relationships (Roy, 2011, p. 131). Some migrated seasonally (Roy, 2011, p. 136). As a result, the population of India's urban centers was disproportionately male (Visaria and Visaria, 1983, p. 521); female migration was constrained by the need to mind children and land (Roy, 2015, p. 189). Workers might retire to their native villages, creating multi-generational links with urban mills (Wolcott, 2015, p. 200). Urbanization increased in the 1920s, in part due to postwar industrial protection (Visaria and Visaria, 1983, p. 520).

What role did cities play in the Indian economy? In the colonial period, several small-scale industries had a distinctively urban character (Roy, 2011, p. 173-179). Large-scale industry was almost entirely in urban areas (Roy, 2011, p. 183), particularly in Bombay, Madras, Calcutta, Agra and Kanpur (Roy, 2012, p. 195). Cotton mills in Bombay served export markets, while upcountry mills supplied domestic demand (Rothermund, 2002,

[3]These estimates are also quoted in Tomlinson (2013, p. 4).

FIGURE 1. Persistence of Urban Populations

p. 68). Colonial cities were also large centers of consumer demand (Tomlinson, 2013, p. 115).

Patterns of urbanization in colonial India show persistence similar to what has been found in other contexts (Bleakley and Lin, 2012; Davis and Weinstein, 2002); many of South Asia's larger cities were already established during the colonial period. Consider the 581 prominent cities of Bangladesh, Burma, India, and Pakistan reported in the World Cities Database.[4] 450 of these are within 10 kilometers of a city reported in the 1931 census. Taking the sum of the colonial cities within 10 kilometers as a rough measure of the modern-day city's population in 1931, we estimate an elasticity of contemporary city size of 0.757 with respect to colonial city population, and show the corresponding scatterplot in Figure 1. While the process by which past cities have merged and the presence of modern cities that were outside the borders of colonial India makes this procedure inexact, and while many new cities such as Chandigarh and Islamabad have emerged since the colonial period, it is clear that the relative sizes of the cities that existed in the late colonial period have remained remarkably stable over the past century.

2.2. **Railroads in colonial India.** In 1853, Governor-General Dalhousie proposed constructing 5,000 miles of railway in India (Rothermund, 2002, p. 32). By 1930, more than 40,000 miles of track had been built (Donaldson, 2018). Several concerns prompted the

[4] `https://simplemaps.com/data/world-cities`

construction of the railway. Rothermund (2002, p. 32) cites political unification and access to raw cotton. Bogart and Chaudhary (2015, p. 141) claim that commercial viability was paramount until the 1870s, after which military and famine concerns became more important.

How did the railway affect the Indian economy? Because engines and coal were imported, Rothermund (2002, p. 33) argues that the railroad did not provide linkage effects that might spur growth in other sectors of the economy. McAlpin (1974) argues that precautionary food storage dampened farmers' substitution towards cash crops. Other writers have claimed that the railroads did matter. It is through these impacts that the railway might be expected to affect city growth and size. Roy (2012, p. 189-190) argues that falling transportation costs benefitted industries, such as cotton textiles, in which India had an advantage; further, money earned in rail-facilitated cotton cultivation was later invested in Bombay mills. Empirical work has found that the extension of the railway system reduced price gaps over space (Andrabi and Kuehlwein, 2010; Hurd, 1975), increased trade and real incomes (Donaldson, 2018), and reduced vulnerability to famine (Burgess and Donaldson, 2017).

2.3. **Conceptual Framework.** A number of theoretical and structural contributions have noted that a critical link between population and transportation costs in spatial equilibrium is market access (e.g. Allen and Donaldson (2020); Baum-Snow et al. (2016); Donaldson and Hornbeck (2016)). In particular, Redding and Sturm (2008) note two important dimensions of market access: while "firm market access" captures the proximity of firms to demand in all markets, consumer market access captures the access consumers have to the goods produced in all markets. One increases the wages firms can pay, while the other reduces the cost of living. This importance of market access in the literature will motivate our focus on market access measures in our empirical analysis.

Beyond this core mechanism of greater market access, a number of papers have identified other related channels that could link transforation infrastructure to urbanization and the growth of cities.[5] These include factor mobility and the ability of rural labor to access external labor markets (Asher and Novosad, 2020; Banerjee et al., 2020; Bogart et al., 2020; Morten and Oliveira, 2016), consumption cities in resource-exporting countries (Gollin et al., 2016), complementarity with market-oriented minority communities (Jedwab et al., 2017; Johnson and Koyama, 2017), relaxation of the land constraint on the growth of large cities (Dittmar, 2011a; Nagy, 2020), structural change (Fajgelbaum and Redding, 2018), towns that serve as trading stations for agricultural products (Jedwab and Moradi, 2016), and better conditions for manufacturing production (Atack et al.,

[5]See Hanlon and Heblich (2020) for a review.

2011; Hornbeck and Rotemberg, 2019). If transportation infrastructure leads to the spatial concentration of production, output can fall in peripheral areas connected to the network (Faber, 2014).

In the specific context of colonial India, other effects of the railways identified in other studies, such as price convergence (Andrabi and Kuehlwein, 2010), reduced famine mortality (Burgess and Donaldson, 2010), greater agricultural incomes (Burgess and Donaldson, 2017), and human capital (Chaudhary and Fenske, 2020) may also have acted as supporting mechanisms through which railways facilitated urbanization. While we will not be able to test for all of these supporting or ancillary mechanisms in our empirical analysis, we will use the variables available to us in order to test for heterogeneous responses to railway access – for example, by initial city size or by access to alternative transportation modes – that will allow us to evaluate the degree to which some of these reinforce or attenuate our main effect of interest.

3. DATA

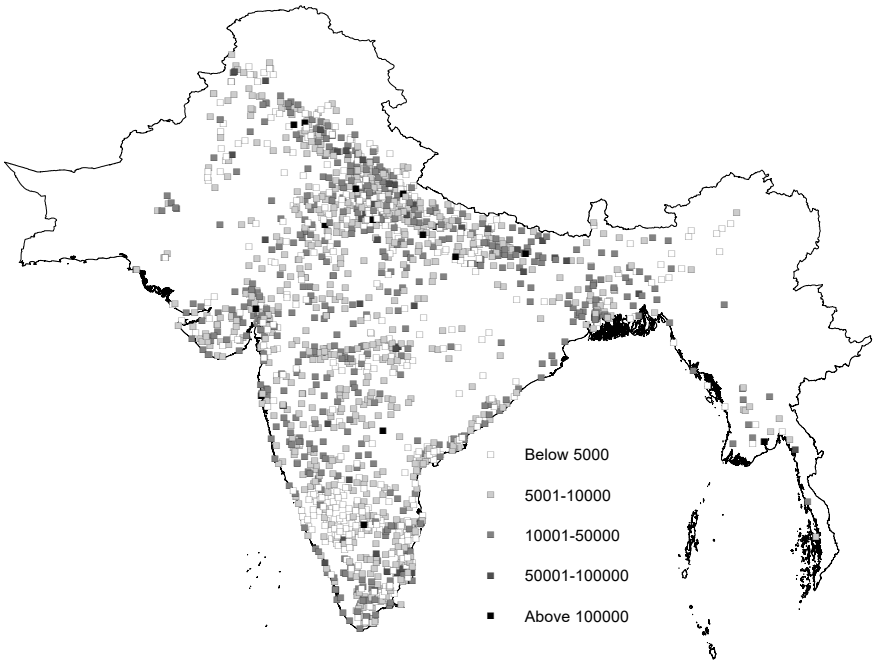
3.1. **Indian Cities.** We have digitized data on city populations from the 1931 Census of India. These cover modern-day Bangladesh, Burma, India, and Pakistan. In particular, for each provincial volume of the census, these are reported in Table 4 of the section containing the Imperial Tables. There are 2,456 distinct cities in the data, and populations are reported for the years 1881, 1891, 1901, 1911, 1921, and 1931. The Census itself states that these data cover cities with populations of at least 1,000 persons, and indeed only 80 of 14,736 possible entries report populations less than 1,000. 2,043 of 14,736 possible entries are missing in the original data and likely reflect years in which these settlements had populations of less than 1,000. For consistency, then, we code as missing all observations of populations less than 1,000. We have located latitude and longitude coordinates for all but three cities in these data, using GeoHack and Google Earth as our principal sources.[6]

Because these data are all taken from the 1931 Census, the original data assigns these to the districts that existed in 1931. We do not, then, need to address the creation, dissolution, or modification of districts and their boundaries over time. However, for consistency with how the Census reports data on total district populations, we have collapsed some districts into aggregate units.[7] Cities, similarly, are aggregated into a single unit

[6]The three cities we have not been able to locate are Raswas (Bhopal District), Qadirabad (Aurangabad District) and Kodaikal (Raichur District).

[7]These aggregated units are Agency Division (Madras), Bangalore City And District, Benaskantha Agency, Baroda, Cochin State, Eastern Kathiawar Agency, Godavari, Gwalior, Kolar Gold Fields and District, Mysore City and District, Other Seventeen Salute States (Western India States Agency), Rest Of Bombay Presidency, Rest of Central India Agency, Southern Maratha States (Bombay Presidency), Travancore, and Western Kathiawar Agency.

FIGURE 2. City populations in 1881

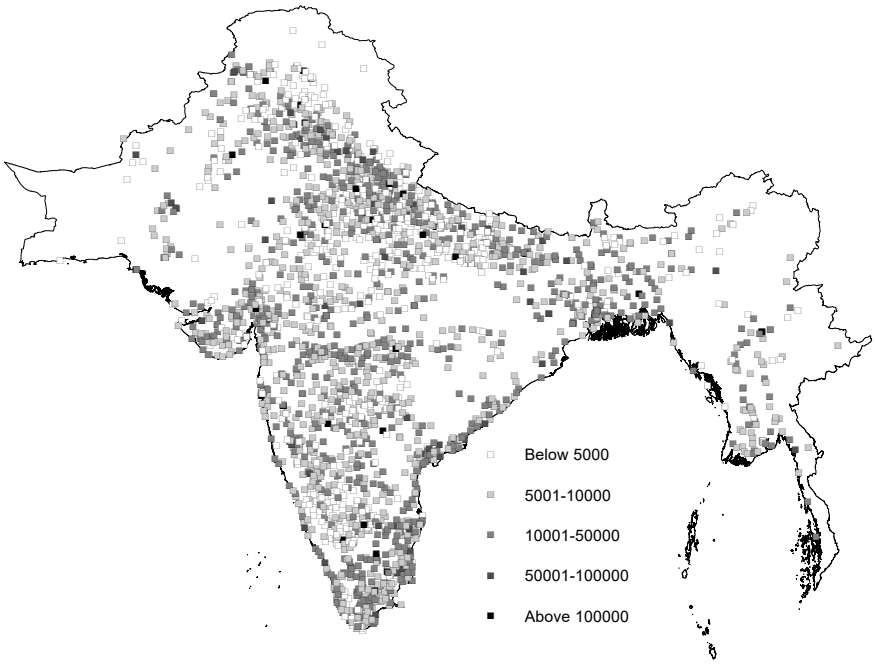


if the populations of their constituent parts are not reported separately. For example, Dehra Dun urban, suburban, and cantonment are treated as the single city Dehra Dun, because separate populations are not reported prior to 1921. Where the populations of constituent units are consistently reported separately in the original data (for example, Barrackpore, North Barrackpore, and Barrackpore Cantonment), we treat these as separate observations.

In Table 1 we report summary statistics for the cities in our data. The number of cities for which populations are reported rises from 1,786 in 1881 to 2,429 in 1931. The summary statistics reflect that, on average, city populations grew moderately from 1881 to 1931. Of the cities for which populations are reported in 1881, the population mean was 13,113 in 1881. This rises to 15,951 for the 2,429 cities reported in in 1931. The largest city in 1881 was Bombay, with a population of 773,196. By 1931, Calcutta was the largest city, with a population of 1,196,734. The standard deviation of city sizes also grew over time, from 32,882 to 46,175. We present maps of city populations for 1881 and 1931 in Figures 2 and 3.

Creation of these data is one of the contributions of this paper, and it is our hope that these data will be of use to other researchers. Existing work in both economics and economic history has used similar data on cities for other parts of the world. It has been used, for example, as a proxy measure for development (Acemoglu et al., 2005;

FIGURE 3. City populations in 1931



Bosker et al., 2013; Wrigley, 1985). City populations have been used to assess the importance, among other variables, of the printing press (Dittmar, 2011b), the Protestant reformation (Cantoni, 2015), medieval universities (Cantoni and Yuchtman, 2014), and the French Revolution (Acemoglu and Cantoni, 2011).

Our new dataset has several interesting features that we further describe in Appendix A. While cities grew on average over the period 1881-1931, many cities shrank. Indeed, from 1901 to 1911, the average growth rate reported was negative. Log city sizes and city growth both have bell-curve distributions but display too much positive skewness and excess kurtosis to be described accurately by a normal distribution. Neither Zipf's law (a linear relationship between log city size and log rank that emerges if city sizes follow a power law distribution) nor Gibrat's law (growth uncorrelated with initial size) describe our data. The largest and smallest cities are both too small to fit Zipf's law. Larger cities grow more slowly before 1921. We also report geographic correlates of city growth: northern cities grow more slowly, as do those in areas suitable for cotton cultivation and those distant from rivers. Those in areas suitable for wheat cultivation grow more rapidly.

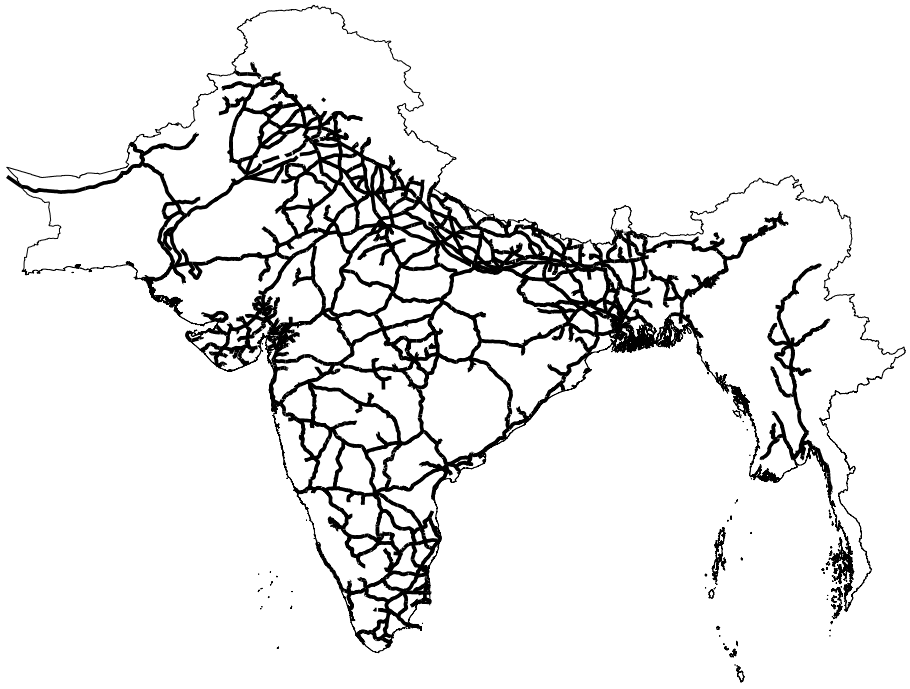
3.2. **Railroads.** In order to assess the impact of the expansion of the colonial railway system on the growth of Indian cities, we have followed a procedure similar to that in Donaldson (2018) in order to construct a polyline shapefile of the Indian railway system

FIGURE 4. Railroads by 1881

in which the opening date is known for each segment. We begin with the 1934 edition of *History of Indian Railways Constructed and In Progress*. For each of the roughly 2,000 railway lines listed, we record the opening dates and identify start points and end points, again mostly using GeoHack and Google Earth. We then take a polyline file of the modern Indian railway from `www.gadm.org`. We fracture this polyline using the start and end points of the colonial railway segments. We assign each railway line from *History of Indian Railways Constructed and In Progress* the polyline segments between its start and end points. If a polyline segment belongs to several railway lines, we assign it to the railway line that opens the earliest. There are some railway lines that are in the *History of Indian Railways Constructed and In Progress* that are not in the modern map of railroads, such as that between Nidamangalam and Manargudi. We add these to the polyline file using straight lines. Some of these lines that are not in the modern map of railroads are very short (e.g. "Bhagalpur Kachery To Bhagalpur Station, E.I. Ry."). We ignore these short lines.

We plot the railway maps we obtain for 1881 and 1931 using Figures 4 and 5. While there was already a substantial railroad network in place by 1881, it became much more dense by 1931. Comparing these maps with Figures 2 and 3, the relationship between expansion of the railroad and city growth over the 1881 to 1931 interval is not obvious. The railway system did expand into regions in which rapid city growth is visible, such as

FIGURE 5. Railroads by 1931



Punjab and Assam, but the railway system was also built up substantially in areas that saw much slower urban growth, such as Rajasthan and Uttar Pradesh.

3.3. **Additional variables.** We create data on a number of geographic controls. At the city level, the correlates we consider are latitude, longitude, log distance to a major river, and log distance to coast. We compute these distances using ArcMAP, using polylines of rivers and the coastline taken from `www.naturalearthdata.com`.

The other geographic correlates we consider are originally available as raster data, and so we compute them at the district level rather than individually for each city. To match raster points to districts, we begin by converting the map of districts from the 1931 census to a shapefile. Because this map has a low resolution, we are concerned that this will lead to measurement error for geographic controls, particularly for small or irregularly shaped districts. We address this by identifying all modern-day third-level administrative divisions (e.g. tehsils) that intersect these historic districts, and averaging over the raster points within this set of units. For example, historic Agra district is merged to the Agra, Bah, Faehabad, Khairagarh, and Kiraoli tehsils of modern Agra district, as well as the Etmadpur and Firozabad tehsils of modern Firozabad district.

In particular, we include ruggedness, malaria, altitude, precipitation, temperature, and suitability for dryland rice, wetland rice, wheat and cotton as additional correlates.

Ruggedness is from Nunn and Puga (2012) and captures the roughness of the terrain.[8] Our measure of malaria is that originally created by Kiszewski et al. (2004).[9] We have used altitude data that are originally taken from the CGIAR's SRTM30 dataset.[10] We rely on the FAO-GAEZ data portal for means of precipitation, temperature, and suitabilities for specific crops.[11]

There are three additional variables that we will consider in our tests for possible heterogeneous responses to railway access, but that we do not treat as controls in our baseline specification: presence of a medieval port, proximity to events during the Indian Rebellion of 1857, and exposure to famines. For medieval ports, we take the list of ports from Jha (2013) and code a dummy for whether a city in our data is within 10 kilometers of a city on this list. For events during the Rebellion of 1857, we begin with the list of events in Jaques (2007), as geocoded by Dincecco et al. (2020). We code a city as exposed to the Rebellion if an event occurred within 20 kilometers – roughly the range an army can cover in one day. We code famine events using the lists and maps of major nineteenth century famines from Srivastava (1968). These provide information at the district-by-year level on the existence of a famine and have previously been used by Burgess and Donaldson (2010, 2017). We code a city as exposed to a famine if there was a famine in its district within the previous decade, i.e. the time period between observations of city populations.

4. EMPIRICAL STRATEGY

4.1. **Fixed Effects.** Our main empirical specification is a fixed effects model. For city i in year t, we use OLS to estimate:

$$\ln P_{i,t} = \alpha + \beta \ln RailwayDistance_{i,t} + \delta_i + \eta_t + x'_{i,0}\eta_t + \epsilon_{it}. \tag{1}$$

In equation (1), the variable $P_{i,t}$ is the population of city i in census year t, where $t \in \{1881, 1891, ..., 1931\}$. $RailwayDistance_{i,t}$ is the distance of the city to the railway in kilometers. Because the city fixed effects will remove any time-invariant geographical controls, we follow the same procedure as in several studies where time-varying historical control variables are difficult to obtain (e.g. Juhász (2018); Waldinger (2015)) and interact our controls $x_{i,0}$ with the year fixed effects. The baseline controls we include in $x_{i,0}$ are latitude, longitude, log distance to river, log distance to coast, ruggedness,

[8] `http://diegopuga.org/data/rugged/tri.zip`.
[9] We are grateful to Marcella Alsan for providing us with these data.
[10] `http://www.diva-gis.org/gdata`.
[11] `http://www.fao.org/nr/gaez/en/`.

malaria, altitude, precipitation, temperature, and suitability for dryland rice, wetland rice, wheat, and cotton. We cluster standard errors by city.

The identifying variation in this specification comes from comparing the change over time in a city's size as it gains proximity to the railway network, over and above common trends in population growth given by the year fixed effects. Time-invariant variables that predict how a city gained proximity to the railway over time will not confound these estimates unless they predict differential *trends* in city growth rather than differing levels of city size. Time-varying variables that correlate with railway proximity may, by contrast, bias our estimates. These could include, for example, colonial investments such as canals insofar as these are not consequences of the railway network. Because of this possible bias, we employ a number of instrumental variables specifications.

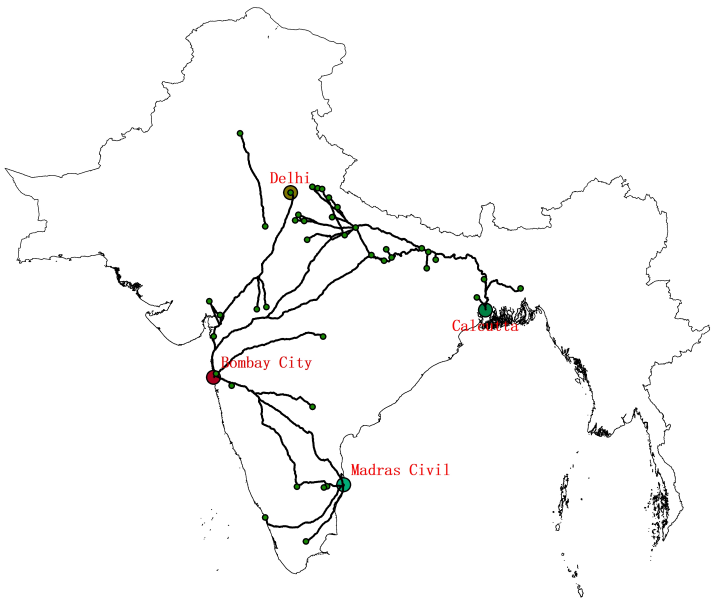
4.2. **Instrumental Variables.** In order to mitigate possible omitted variables bias, we employ a number of alternative instruments for $\ln RailwayDistance_{i,t}$ and estimate equation (1) using instrumental variables. Our main instrument is based on work by Bogart et al. (2020) for the United Kingdom, and takes as its base the distance between each city in our data and a least cost path that connects cities that existed in India before the beginning of railway construction and that, based on market potential and the costs of construction, were likely to be connected to the railway network.

The cities data in the census go back only to 1881, while the first railway line reported in *History of Indian Railways Constructed and In Progress* (Victoria Terminus To Thana), opens in 1853. To find a set of cities that predate the Indian railway, we turn to Chandler and Fox (1974). They do not report data in tabular format, but instead provide a list of cities and estimates of their populations at various dates that differ across cities. We have identified 97 cities in British India (including Burma) that Chandler and Fox (1974) list as having a population of at least 10,000 in 1850, in "c. 1850," or in the closest years before and after 1850 that are reported.

We construct our least cost path in three steps, following Bogart et al. (2020). First, we begin with the subset of 76 Indian cities whose populations are recorded in Chandler and Fox (1974) in the mid-nineteenth century, before 1853 (the start of railway construction), and that also appear in the census. We then compute the market potential of any pair of cities as $G_{ij} = \frac{Population_i \times Population_j}{Distance_{ij}}$, where $Population_i$ and $Population_j$ are the populations of each city, and $Distance_{ij}$ is the distance between them in kilometers.

The second step is to create least cost paths connecting this set of market pairs. Rather than using straight lines to connect cities, we follow Bogart et al. (2020) and create paths between cities that minimize the cost of construction. We begin with raster data on

FIGURE 6. Least Cost Path



slope at the grid cell level.[12] We parameterize the cost of any hypothetical line crossing a cell by letting the cost of construction increase by a factor of three for every 1 percentage point increase in the slope of a cell. That is, if the cost of crossing a flat grid cell is 1, the cost of crossing a cell with a slope of x% is $1 + 3 \times x$. For example, the cost of crossing a cell with a slope of 2% is 7. This again follows Bogart et al. (2020), and is based on the relationship they estimate between construction costs and elevation change for 36 non-London railways during the nineteenth century in England. These costs are unitless, and the choice of unit will not affect the optimal route placement since the least cost path will minimize costs expressed in any unit.

The third step is to select from this set of least cost paths a subset that is to be included in the data. We sort each pair of towns by market potential G_{ij}, and select routes until the total length of the least cost path network is as large as the actual railway network in 1881. The resulting hypothetical railway network is shown in Figure 6.

This hypothetical network resembles the early stages of the network that was actually constructed, and so proximity to this least cost path predicts how quickly the cities in our data – including the vast majority that are not recorded in Chandler and Fox (1974) –

[12]We work with grid cells that are 180m $\times$ 180m at the equator. Our underlying data source is Shuttle Radar Topography Mission with a resolution of 90 metres (SRTM 90). This is the same underlying source used in Bogart et al. (2020). We aggregate the raster data to a resolution of 180 metres because South Asia's vast size makes computations with 90 metre cells computationally demanding.

gained access to the railway network. Cities closer to this least cost path became closer to the railway network in earlier years.

Once this least cost path is constructed, we use it to construct an instrumental variable. We compute the distance in kilometres of each city in the data to this least cost path. We then use the interaction the log of (one plus) distance to this least cost path with year (i.e. t) to instrument for $\ln RailwayDistance_{i,t}$. We make two notes here. First, because the cities are treated as a set of points with zero area, and the railways are treated as a set of lines with zero thickness, no city has zero distance from the railway. However, the least cost path is built to connect a subset of these city points. As a result, some have zero distance from the least cost path. This motivates the use of the log of one plus distance rather than simply log distance in the instrument.

Second, because distance from the least cost path is time-invariant, it is collinear with our city fixed effects. Hence, interacting this with year to construct our instrument bases identification on how proximity to the hypothetical plan predicts differential time trends in railroad proximity. In 1881, cities distant from the least cost path were distant from the railroad. Over time, this relationship flattened as cities more distant from the least cost path grew closer to the railroad. It is the flattening of this relationship that we exploit for exogenous variation in our instrumental variables analysis.

In addition to our principal instrumental variable, we construct three alternative instruments for robustness, based on alternative least cost paths. Two of our three alternative least cost paths are based on alternative scenarios from Bogart et al. (2020). These scenarios allow terrain slope to have differing effects on the costs of construction. In the first alternative, which we call A1, the cost of building across a grid cell is equal to one plus its slope: 1 if it is flat, 2 if the slope is 1%, 3 if it is 2%, and so forth. We cap the cost of crossing any one cell at 51. This corresponds to what Bogart et al. (2020) call "Scenario 1," and allows for a less convex relationship between terrain slope and construction cost than in their baseline. The second alternative, which we call A2, corresponds to what Bogart et al. (2020) call "Scenario 3." This scenario assumes that any cell with a gradient greater than 6% requires a tunnel, and so caps costs at 19.

Our third alternative least cost path, which we call A3, is similar to the baseline scenario in Bogart et al. (2020), but based on data from data on Indian construction costs. We take data from Bogart and Chaudhary (2013) on the real value of capital of 21 Indian railway companies from 1851 to 1912. We assume that the value of capital of a railway line in year t is equal to the construction costs of all the branches that have been finished by year t. Combining these data with the lengths and accumulated slopes of railway lines in the years 1861, 1871, 1881, 1891, 1901 and 1911, we can then estimate

the relationship between cost and elevation change for each line i using the following regression, based on Bogart et al. (2020):

$$Cost_i = \beta_0 + \beta_1 Length_i + \beta_2 Slope_i + \epsilon_i$$

Here, $Length_i$ is measured as the number of raster cells crossed, while $Slope_i$ is the total slope of the line, in percentage points. Our estimates suggest that $\hat{\beta}_2 \approx 0.6\hat{\beta}_1$. We then replace the cost parameterization of $1 + 3 \times x$ with $1 + 0.6 \times x$ when computing the cost of any given line.

For each of these three alternative least cost paths, we again take the interaction the log of (one plus) distance to the path with year as an alternative instrument for $\ln RailwayDistance_{i,t}$.

4.3. **Market access.** The existing literature on economic geography stresses market access as the critical link between equilibrium population and transportation costs. As an alternative to our main empirical specification, which considers physical proximity to a railway, we can estimate:

$$\ln P_{i,t} = \alpha + \beta \ln MarketAccess_{i,t} + \delta_i + \eta_t + x'_{i,0}\eta_t + \epsilon_{it}. \tag{2}$$

All terms here are defined as in (1), except that we have replaced $RailwayDistance_{i,t}$ with $MarketAccess_{i,t}$. Whereas physical proximity to a railway measures whether a city has access to a railway, market access measures the sizes of the markets that each city is connected to, deflated by the costs of reaching them. We follow Donaldson and Hornbeck (2016) and define market access as:

$$MarketAccess_{i,t} = \sum_{j \neq i} \tau_{i,j,t}^{-\theta} P_{j,t} \tag{3}$$

Here, the market access for city i in year t is depends on the costs of reaching each other city j in year t, $\tau_{i,d,t}$, and the population of each other city j in year t, $P_{j,t}$. This is a close approximation of the market access measures that emerge as sufficient statistics for transportation infrastructure in structural models of economic geography (e.g. Donaldson (2018); Eaton and Kortum (2002); Redding and Sturm (2008)).

To compute market access, we need three quantities: $\tau_{i,j,t}$, θ, and $P_{j,t}$. We compute $\tau_{i,j,t}$, the cost of travel between any city i and any other city j, by following Donaldson (2018). We compute least cost paths connecting any two cities i and j in the data. Transportation modes allowed in these routes include wagons, coastal shipping, rivers, and railways. Connections to oceanic transportation routes are only accessible via ports.

Normalizing the cost of shipment by railways to 1, the relative costs of travel by wagons, coastal shipping, and rivers are 2.375, 6.188, and 2.250, respectively. These are based on estimates in Donaldson (2018). For θ, we will take 1 as our baseline, and report alternative values of 3.6 7.8, and 8.28. The baseline value follows the original parameterization in Harris (1954), and the alternatives come from Donaldson and Hornbeck (2016), Eaton and Kortum (2002), and Donaldson (2018). 8.28 is the preferred value from Eaton and Kortum (2002), while the mean result in Donaldson (2018) is 7.80. We will show below that lower values of θ have more predictive power in our data. City sizes $P_{j,t}$ are reported in our data, and for this calculation we assume that the population of a city is 0 if it is not recorded in the census in any specific year. Note that $\tau_{i,j,t}$ will only change over time due to the expansion of the railway network.

As with equation (1), we will estimate (2). We will use both OLS and IV, and we will employ the same instruments for $\ln MarketAccess_{i,t}$ that we used for that we used for $\ln RailwayDistance_{i,t}$.

5. RESULTS

In this section, we present our estimates of equations (1) and (2). We begin by presenting results connecting distance from the railway to city size, before then presenting results in which we use market access to measure a city's connection to the transportation network. We explore the heterogeneity of our results, and report our principal robustness checks.

5.1. **Distance from railroad.**

In Table 2, we present OLS and IV estimates of equation (1). The first column reports OLS estimates without controls, while the second column interacts baseline geographic characteristics with our year fixed effects. Columns (3) and (4) present analogous specifications for our instrumental variables estimates. The corresponding first stage estimates are in columns (5) and (6). Note that we divide the instrument by 1,000 in order to ease the presentation of coefficients. Note that there are fewer observations in our IV estimations because we purposefully exclude the nodes of the least cost paths – this focuses identification on cities that were connected to the railway, incidentally based on their proximity to a path connecting two other cities.

Our OLS estimates suggest an elasticity of city size with respect to railway proximity that is negative, but that is not large. These range from -0.017 to -0.019. Put differently, a one standard deviation reduction in distance from the railroad increases city size between 4.47% and 5.06% of a standard deviation. Similarly, the share of city growth that is explained by railway proximity is small – the R^2 before controls are added is less than 1%. Our instrumental variables estimates are larger in magnitude. Here, the implied

elasticities range from -0.113 to -0.191, and the effect sizes expressed in standard deviations range from -29.8% to -50.3%.

We can put the magnitudes of our results in the context of estimates from other studies, though there is no single specification that is preferred in the literature and so existing studies each estimate different, though related, parameters. The most comparable work to ours examines how railway access affects urban populations. Berger and Enflo (2017) find that a dummy for connection to a railway in nineteenth-century Sweden raised city size by 23.4 log points in their OLS estimations and 31.8 log points in their IV estimations. Similarly, a railway within 10 kilometers raised urban population by 0.74 standard deviations in Ghana in 1931 (Jedwab and Moradi, 2016) and 0.37 standard deviations in Kenya in 1962 (Jedwab et al., 2017). Okoye et al. (2019) estimate that an individual living within 20km of a colonial railroad in modern Nigeria is 18.5 percentage points more likely to live in an urban area. Atack et al. (2010) find, by contrast, that urbanization increased by 3.7 percentage points relative to a baseline mean of 6.7% in US counties that gained rail access during the 1850s. Bogart et al. (2020) find in their OLS estimates that the change in log population between 1851 and 1891 was 16.6 log points greater for localities of England with a railway station in 1851. Their IV estimate of the same effect is 34.9.

More broadly, our results can be compared to other estimates of the impacts of transportation infrastructure, such as the the 16.4 log point increase in agricultural income in Indian districts connected to a railroad (Donaldson, 2018) or the 0.488 (OLS) to 3.95 (IV) log point increase in population growth in peripheral Chinese counties connected to the National Trunk Highway System (Faber, 2014).

While none of these studies directly reports an elasticity of city size with respect to distance from a railway, our OLS elasticity estimates imply that a colonial Indian city would need to become very distant from a railroad to experience the same reduction in size predicted by disconnection in the studies above.[13] Similarly, in Table 5, below, we will report robustness to alternative measures of distance from a railway, including discrete distance bands. Our smallest distance band (0-2km) will be roughly comparable to the dummies for connection or proximity used by studies such as Berger and Enflo (2017), Jedwab and Moradi (2016), or Okoye et al. (2019). In that table, it remains clear that our estimates for colonial India are smaller than in other contexts.

There are a number of possible reasons for these smaller magnitudes. One is that, in contrast to early twentieth century Africa and the United States before westward expansion of white migrants, the population density of India was already relatively high

[13]For example, for population to fall 23.4 log points as in Berger and Enflo (2017), a city would need to be $(100 \times 0.234/0.017)$ 1376 log points further from a railway according to our estimates in column (2).

and many urban centers existed that predated the railway. The capacity of the railway to reset the urban network will, then, have been less in India. Further, the slow growth of urbanization in India, outlined in Section 2, means there is less urban growth to be explained in our data.

There are also a number of possible reasons why our IV results are larger than our OLS results. One explanation would be a bias towards zero due to omitted variables that predict railway proximity but that retard city growth. Variables that predict absence of a railway and favor city growth would have the same effect. Chaudhary and Fenske (2020) discuss several motives for railway placement in colonial India that could create this type of bias, including "protective" lines that connect famine-prone areas to the transportation network, lines from Delhi towards Afghanistan built for military purposes, lines connecting ports to cotton-growing regions that were likely to remain agricultural, and lines connecting small hill stations that British officials used as summer retreats.

Another potential explanation is the difference between the local average treatment effect estimated by IV and the average treatment effect for the whole population of cities. We will show below in Table 4 that the impact of a railway is attenuated by a number of city characteristics, including an above-median population in 1881, suitability for cotton production, proximity to rivers, proximity to the Mutiny of 1857, and proximity to the railway in 1881. The instrumental variables approach focuses identification on compliers – cities that gained access to railways earlier because of their proximity to the least cost path. If these cities are less likely to have characteristics that attenuate the effects of railways, this would inflate the IV estimates relative to the OLS estimates.

Another possible explanation would be attenuation bias due to measurement error in railway proximity. Narrowly, treating railways as polylines and cities as massless points will lead to mis-measurement of the distance of cities from railways, and this will be exacerbated by changes over time in the locations of cities and of specific railway lines. Conceptually, it is possible that physical proximity does not fully capture the dimensions of the railway network that are most important and so mismeasures these. We will show below in Table 3 that the inflation of coefficients when moving from OLS to IV estimates using market access measures is smaller than in Table 2, which is consistent with this interpretation.

Another possibility would be weak instruments. We do not believe this is a likely explanation: the Kleibergen-Papp F statistics in our regression are greater than 70, well above the conventional cutoff of 10. Yet another possible explanation would be violations of the exclusion restriction. Given our baseline inclusion of both city and year

fixed effects, and since we show below that we obtain similar magnitudes with alternative instruments, we believe this is unlikely to explain the difference between OLS and IV estimates.

5.2. **Market access.** In Table 3, we report OLS and IV estimates of equation (2), where we now use market access to measure how a city is exposed to the railway network. In columns (1) and (2) we report OLS estimates with and without controls, respectively. In columns (3) and (4) we present our analogous IV results. Columns (5) and (6) show first stage estimates. Finally, columns (7) and (8) report our OLS estimates using an alternative measure of market access that follows Donaldson and Hornbeck (2016). Using equation (3) to compute market access, we now exclude any markets j that are within 100km of a given city. We call this new measure "access to distant markets." This isolates a component of market access that is unlikely to be affected by unobserved changes that are close to the city i for which market access is measured. We treat this as an alternative to instrumental variables in generating exogenous variation in market access.

Our OLS estimates suggest an elasticity of city size with respect to market access of between 0.385 and 0.628. Expressed as a standardized effect size, this suggests that a one standard deviation increase in market access would increase city size by 22.2% to 36.2% of a standard deviation. The IV results are larger in magnitude, corresponding to an elasticity between 1.028 and 1.370, and a standardized effect size between 59.3% and 79.0% of a standard deviation. Using access to distant markets as an alternative measure gives estimates larger than the OLS estimates in columns (1) and (2), but smaller than the IV estimates in columns (3) and (4). Here, the elasticities range from 0.579 to 0.886, and the standardized effect sizes range from 25.3% to 38.7%.

The magnitudes of our OLS estimates of the elasticity of city size with respect to market access fall within the range of other estimates of the impact of market access on economic outcomes in the past. Donaldson and Hornbeck (2016), for example, estimate an elasticity of the value of agricultural land with respect to market access of 0.511 – an elasticity that more than doubles in their IV estimation. Hornbeck and Rotemberg (2019) find an elasticity of 0.129 of county productivity with respect to market access in the nineteenth century United States. We are not aware of existing estimates of the elasticity of city size with respect to market access to which we can directly compare our results, though Hornbeck and Rotemberg (2019) find an elasticity of county populations with respect to market access of 0.259.

These results give us additional evidence on the difference between the OLS and IV coefficients in Table 2. The IV results remain larger than the OLS results when using market access, but the degree of inflation is less. This is still consistent with negative selection of cities into railway access, but suggests the problem of attenuation bias due

to measurement error is larger when using the log of distance from the railway rather than market access as a measure of exposure. The larger elasticities and standardized effects obtained using access to distant markets is further evidence that OLS estimates may be biased downwards due to negative selection. The first stage F statistic is much larger in the market access regressions, suggesting that weak instruments do not explain the divergence of the OLS and IV estimates, and that distance from the least cost path is a better predictor of time trends in market access than of time trends in proximity to the railway network. We will show below in Table 4 that many of the same variables that predict differential response to railroad proximity also predict differential response to market access, suggesting again that compliers may differ from the average city and that this may explain the divergence between OLS and IV estimates.

5.3. **Heterogeneity.** To explore the channels by which railway access increased city size in India, we use Table 4 to test whether seven variables predict heterogeneous responses: greater city size in 1881, suitability for cotton cultivation, presence of a medieval port, having a river within 2 kilometers, experiencing an event related to the Indian Rebellion of 1857 within 20 kilometers, being above-median distance from the railway system in 1881, and exposure to the famines of the nineteenth and early twentieth centuries.

We report OLS estimates of both equation (1) and equation (2), augmented to include the interaction between the relevant measure of railway access (log distance from a line or market access) with the possible source of heterogeneity. In all cases except one, the source of heterogeneity is time-invariant and so it is absorbed by the city fixed effects. The exception is famine exposure. Because this is time-varying, we also include it as a control but do not report the coefficient.

Columns (1) and (2) of Table 4 allow the impact of railways to vary for cities that are above median size in 1881. This will capture the degree to which railways reinforced existing agglomeration or allowed smaller cities to grow. Note that we can only perform this test on the sub-sample of cities that have populations reported in 1881, which reduces sample size in these columns. The interaction is positive when we use the log of railway distance to measure proximity and negative when we use market access, suggesting that the effects of railways are attenuated in the set of cities that are already large in 1881. This implies that railways led to city growth not by reinforcing existing agglomeration, but by letting smaller cities grow.

In columns (3) and (4) of Table 4, we perform a related test and divide our sample by median distance from a railway in 1881. Cities above median distance from a railway will have typically had the lowest levels of market access at the start of our data series. Across specifications, it is clear that the effect sizes are largest for these most initially isolated cities. Indeed, using distance from a railway as a measure of access, it appears

as though greater proximity only increased city size for the initially most isolated cities. This further reinforces our interpretation that railways increased market access for initially small and isolated cities, rather than reinforcing the advantages of initially large and more connected locations.

In columns (5) and (6) of Table 4, we examine the possible differential response of former medieval ports. These ports may capture both historical prosperity or the predetermined presence of alternative transportation links. If railways substitute for other forms of transportation, their effects could be mitigated in these cities. This would be similar to what Okoye et al. (2019) find in Nigeria. By contrast, if connecting sea-borne trade hubs to a railway reinforces network externalities, their impacts could be greater. Across specifications, the coefficient signs suggest attenuation, but they are not significantly different from zero in three of four cases. This provides little evidence, then, of network externalities as the main driving force behind our results.

Columns (7) and (8) of Table 4 consider a related dimension of heterogeneity – proximity to a river. Our use of a 2 kilometer threshold here follows earlier versions of Bogart et al. (2020). Our logic here resembles that in the previous test: like a port, a river may substitute for a railway, attenuating its impact, or it may reinforce network externalities. In the specification that employs physical proximity, the presence of a river significantly attenuates railway access. This does not appear to be the case when we use the market access measure of railway connection. This again provides little evidence of a major role for network externalities or reinforcement of existing agglomeration in accounting for our main results.

Columns (9) and (10) of Table 4 consider possible heterogeneity by suitability for cotton cultivation. In particular, we create a dummy equal to one for cities located in districts with above-median cotton suitability. Especially during the civil war in the United States (1861-65), British officials in India believed railways could ensure a reliable supply of cotton for use by the textile industry in Britain (Thorner, 1951, 1955). While these districts may have become more specialized in cash crop agriculture due to the railways, limiting urbanization, secondary towns that served the farming sector may still have grown in these regions. In three of four relevant specifications, cotton suitability appears to attenuate the impact of railroads. This suggests that the agglomeration effects due to services that serve the agricultural sector, such as those Jedwab and Moradi (2016) find in Ghana, are less important in the Indian case.

In columns (11) and (12) of Table 4, we test whether cities that were connected to the railway for military reasons responded differently. Particularly after the Sikh wars of the 1840s, the British were concerned that railways would be needed to move troops to politically unstable regions (Hurd, 1983; Parliamentary Papers, 1854). We use spatial

variation in the Indian Rebellion of 1857, which occurred only shortly after the start of railway construction and for which there is rich data on the locations of major events, to measure military motives for railway construction. Across specifications, the coefficients suggest an attenuating effect of Rebellion exposure; cities that were connected to the railroad for reasons other than economic potential responded less in terms of city growth.

Finally, in columns (13) and (14) of Table 4, we consider cities that were vulnerable to famine and that were connected to a railway. Particularly after the 1870s, the British constructed railway lines that could aid in famine relief for famine-prone areas (Hurd, 1983; Parliamentary Papers, 1854). Across specifications, we find coefficient signs suggesting that railways had smaller effects on city sizes in these areas, but these heterogeneous responses are not significant at conventional levels using the market access measure.

In sum, then, our results are consistent with the railway increasing the size of Indian cities through a market access channel. The heterogeneous results we find suggest that railways increased city growth by facilitating the growth of smaller and initially isolated cities, rather than reinforcing existing agglomeration effects. We do not find evidence that secondary towns serving the cotton sector nor reinforcement of network externalizes in port and river trade help explain the result. The impacts were attenuated where railways were built for military reasons, though we find no similar evidence for famines.

5.4. **Robustness.**

5.4.1. *Principal Robustness Checks.* Here, we discuss the robustness of our results. We begin by showing the robustness of our results on the proximity of railways to alternative functional forms. In Figure 7, we show that the relationship between log city size and log distance from a railroad is approximately linear. We begin by residualizing the data on log population and log distance from a railroad relative to the fixed effects for both city and year. We then show a binned scatterplot of these partial residuals against each other. While the best quadratic fit of these data is not perfectly linear, the curvature is slight. This validates our baseline log-log specification in Equation (1).

We further explore functional forms in Table 5. First, in columns (1) and (2), we estimate equation (1) by OLS, but we replace $\ln RailwayDistance_{i,t}$ with dummies for falling within three distance bands: 0-2 kilometers, 2-10 kilometers, and 10-20 kilometers. This resembles the empirical approach used by, for example, Jedwab and Moradi (2016) and Jedwab et al. (2017). Here, we find that cities within 2 kilometers of a railroad are 7.1 to 7.5 log points larger in size.

In Figure 8, we take an even more general approach. We again estimate equation (1) by OLS, but now we replace $\ln RailwayDistance_{i,t}$ with a full set of dummies for falling

FIGURE 7. Log city size and log distance from railroad: partial residuals and quadratic fit

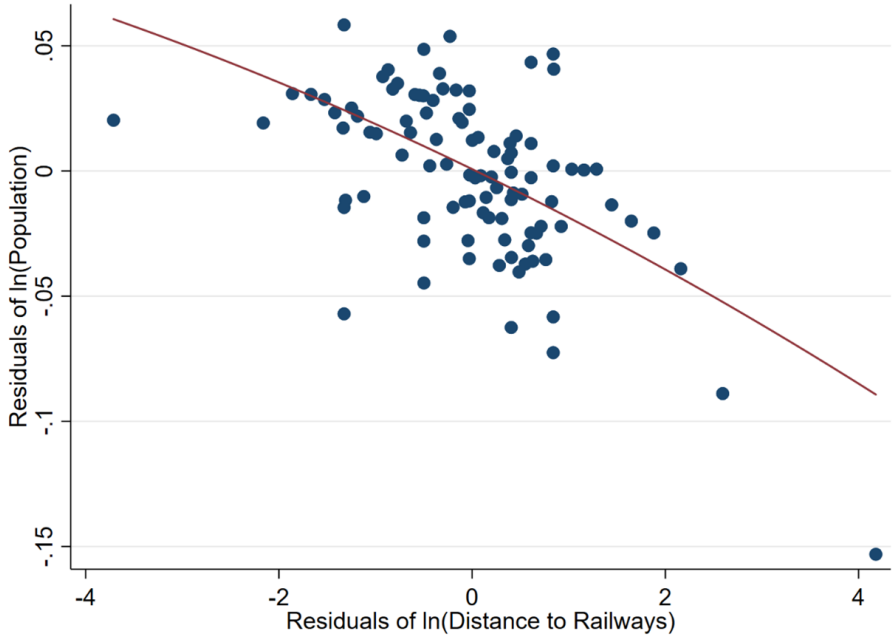
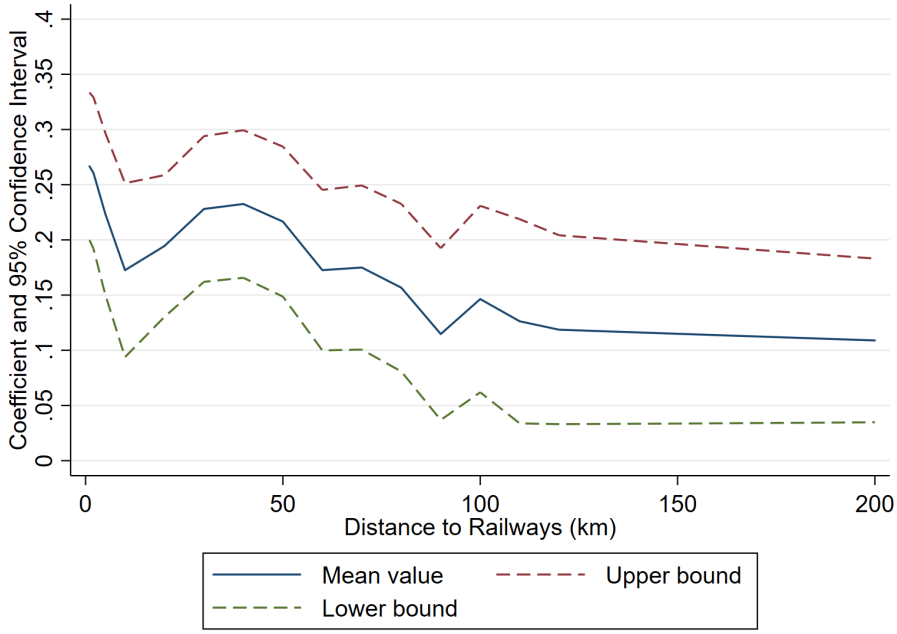


FIGURE 8. Log city size by distance from railroad



within distance bands of the railroad. We use bands that are 10 kilometers wide up to a distance of 120 kilometers, and then use bands of 120-150 km and 150-200 km due to the sparsity of cities at these greater distances. We plot the coefficient estimates and 95% confidence intervals from this regression in the Figure. In this estimation, cities immediately adjacent to a railway are a bit more than 25 log points larger in population. This declines as distance from the railway increases, flattening out at distances greater than 100 kilometers. Coefficients are larger in this exercise than in the first two columns of Table 5, as the set of baseline cities against which these coefficients are to be compared is now much more distant from a railway – at least 200, rather than at least 20 kilometers.

We also use Table 5 to consider a more subtle issue of functional form: the possible influence of outliers due to cities that are very close to a railway line. At very low distances, the logarithmic transformation can rapidly approach negative infinity. We show that this does not drive our main results, replacing observed values of railway distance below a cutoff with the cutoff itself. We consider four cutoffs: 1m, 1km, 2km, and 5km. This is a procedure similar to winsorizing. In columns (3) through (6) of Table 5, we show that the results from this exercise give coefficients very similar to those from Tables (1) and (2) of Table 2 – possible outliers very close to railway lines do not drive our results.

In Table 6, we consider the robustness of our market access results. We begin by changing the value of θ in equation 3. Often referred to as the "trade elasticity," this parameter governs the speed at which access to a market declines as transportation costs increase. Greater values of θ imply a more rapid decline in market access for a given increase in transportation costs. Following Donaldson and Hornbeck (2016), we consider three alternatives to our baseline value of 1 – 3.6, 7.8, and 8.28, which lie within the typical range of gravity estimates reported in the meta-survey by Head and Mayer (2014). In columns (3) through (8) of Table 6, it is clear that these do not change the general conclusion that greater market access due to the expansion of the railway network increases city size.

Quantitatively, the elasticity estimates, standardized magnitudes, and the "within" R^2 measure of goodness of fit net of city and year fixed effects all fall as θ increases. The impact of market access on city size, then, is smaller when compared with columns (1) and (2) of Table 3. We take this as evidence that the trade elasticity that best describes the relationship between market access and city growth in colonial India is smaller than that found in modern trade studies. The fundamental parameters underlying θ differ between demand-side and supply-side derivations of structural gravity models, and so lower values of θ are consistent with a number of interpretations (Head and Mayer,

2014). These include lower substitutability of goods in consumption, greater marginal costs of serving individual markets, and greater heterogeneity in productivity across consumers or producers.

In the second panel of Table 6, we show that an alternative parameterization of our market access measure gives results that are qualitatively similar to those in Table 3. Recall that, in our baseline computation, we normalized the cost of shipping on railways to 1, and set the relative costs of travel by wagons, coastal shipping, and rivers to 2.375, 6.188, and 2.250, following Donaldson (2018). In panel 2, we adopt the alternative relative costs of 4.5, 2.25, and 3.0, respectively. This alternative parameterization also follows Donaldson (2018), and gives results that, while again qualitatively similar to our baseline, are also quantitatively smaller, whether interpreted as elasticities or in standardized magnitudes.

In the bottom panel of Table 6, we again turn to our alternative measure of market access that exploits changes in access to distant markets. Here, we show the robustness of this measure to alternative values of the trade elasticity, θ. We consider the same alternatives as before: 3.6 7.8, and 8.28. Columns (1) and (2) reproduce the baseline results from Table 6. For greater values of θ, the qualitative conclusion of a positive impact of market access on city size remains. As with our baseline measure of market access, the estimated elasticities and standardized coefficients fall as θ increases.

5.4.2. *Additional Robustness Checks.* We report a number of additional robustness exercises in the appendix.

In Table A1, we show the robustness of our results to alternative sample restrictions, specifications, and estimators. In columns (1) through (4) of the first panel, we report results using both railway proximity and market access, but discarding any cities that remained distant from a railroad – more than 100 km – throughout the entire sample period. The results are largely unchanged. Our results are not, then, driven by these possible outliers. In columns (5) through (8) of the first panel, we discard modern-day Burma from the results. This too does little to affect the results, showing that our results hold for the core regions of what is conventionally considered to be colonial India.

In columns (1) through (4) of the second panel of Table A1, we rule out the possibility that cities that are reported despite having populations below 1,000 are driving our results. We truncate all populations below 1,000 and recode them as 1,000. Results are again very similar to our baseline. In columns (5) through (8) of the second panel, we make this same truncation, but use a tobit estimator to account for the fact population is bounded below by 1,000. This too does little to our main results. Here, we gain observations by treating cities whose populations are not yet reported as if they are 1,000.

In columns (1) through (4) of the third panel of Table A1, we discard possible outliers – those in the top and bottom 1% of the sample by statistical influence on $\hat{\beta}$. Finally, in columns (5) through (8) of the bottom panel, we include fixed effects for district $\times$ year. This focuses identification on cities observed in the same district in the same year with differing degrees of railway access. Here too, the results are similar to our baseline results, suggesting that time-varying unobservables at the district level do not explain our results.

In Tables A2 and A3, we show that alternative instrumental variables give results similar to our baseline estimations. The construction of these alternative instruments has been described in more detail above in Section 4.2. Table A2 shows both first and second stage results using the log of distance from a railway to measure a city's railroad access. Table A3 does the same using market access. In both tables, columns (1) and (2) show results in which we continue to use our baseline least cost path to construct our instrument, but now we interact distance from the path, rather than its logarithm, with year.

Columns (3) and (4) use least cost path "A1", in which the cost of construction rises more rapidly with terrain slope than in our baseline. Columns (5) and (6) use instead the least cost path "A2," in which construction costs are capped at high slopes due to the use of tunnels. Columns (7) and (8) use least cost path "A3," based on Indian construction costs.

Across both tables, a general pattern emerges. Each of these least cost paths is a strong predictor of the speed with which cities gained access to the railway network, measured either with physical proximity or with market access. Similarly, our second stage results are similar to our baseline results in Tables 2 and 3, suggesting our results are not driven by the selection of one possible IV strategy relative to another.

Finally, in Table A4, we expand on the district $\times$ year fixed effect specifications reported in Table A1. In particular, we show that our market access results continue to hold in this specification using alternative values of both θ – the trade elasticity – and of the relative costs of transportation. In the top panel, we use the same alternative values of θ as in Table 6: 3.6 7.8, and 8.28. In the bottom panel, we replace our baseline relative costs of travel (2.375, 6.188, and 2.250 for wagons, coastal shipping, and rivers, relative to rail) with the same alternatives that we reported in Table 6: 4.5, 3.0 and 2.25. Both exercises follow Donaldson (2018) and Donaldson and Hornbeck (2016). The conclusion that greater market access due to changes in the railway network increases city size, even controlling for district $\times$ year fixed effects, is evident across all specifications in this table.

6. Conclusion

Indian urbanization grew slowly in the colonial period despite the construction of an extensive railway network. We have confirmed that colonial India's railways spurred the growth of urban population, but that this impact was modest. We have introduced a new decadal dataset on city sizes and locations in colonial India, spanning from 1881 to 1931. We have evaluated the effects of railroad proximity on city size. Both our OLS and IV results suggest that cities closer to railroads increased in population, and that greater market access does a good job of explaining this increase. The effect sizes we find are smaller than those estimated in other contexts, which is consistent with our result that impacts were greatest for initially small, isolated cities, while the levels of urbanization and population density in India prior to the railroad exceeded those in several other contexts that have been examined in past work.

Our exercise is limited by the lack of city population preceding the railroad that is of the same resolution and comprehensiveness as what is present in the colonial census. It is our hope that future researchers will extend our data to cover even earlier years as new sources of data are discovered, and will use the data we provide in order to better understand the development of the South Asian economy.

References

Acemoglu, D. and Cantoni, D. (2011). The Consequences of Radical Reform: The French Revolution. *The American Economic Review*, 101:3286–3307.

Acemoglu, D., Johnson, S., and Robinson, J. (2005). The rise of Europe: Atlantic trade, institutional change, and economic growth. *The American Economic Review*, 95(3):546–579.

Adukia, A., Asher, S., and Novosad, P. (2020). Educational investment responses to economic opportunity: evidence from Indian road construction. *American Economic Journal: Applied Economics*, 12(1):348–76.

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

Agrawal, A., Galasso, A., and Oettl, A. (2016). Roads and innovation. *Review of Economics and Statistics*, 99(3):417–434.

Allen, R. C., Bassino, J.-P., Ma, D., Moll-Murata, C., and Van Zanden, J. L. (2011). Wages, prices, and living standards in China, 1738–1925: in comparison with Europe, Japan, and India. *The Economic History Review*, 64(s1):8–38.

Allen, T. and Donaldson, D. (2020). Persistence and Path Dependence in the Spatial Economy. *National Bureau of Economic Research Working Paper No. w28059*.

Anderson, G. and Ge, Y. (2005). The size distribution of Chinese cities. *Regional Science and Urban Economics*, 35(6):756–776.

Andrabi, T. and Kuehlwein, M. (2010). Railways and price convergence in British India. *The Journal of Economic History*, 70(02):351–377.

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

Atack, Haines, M., and Margo, R. (2011). Railroads and the Rise of the Factory: Evidence for the United States, 1850-1870. *Economic Evolution and Revolutions in Historical Time, edited by Paul Rhode, Joshua Rosenbloom, and David Weiman (eds.)*, pages 162–79.

Atack, J., Bateman, F., Haines, M., and Margo, R. A. (2010). Did railroads induce or follow economic growth? *Social Science History*, 34(2):171–197.

Banerjee, A., Duflo, E., and Qian, N. (2020). On the road: Access to transportation infrastructure and economic growth in China. *Journal of Development Economics*, 145:102442.

Banerjee, A. and Iyer, L. (2005). History, institutions, and economic performance: the legacy of colonial land tenure systems in India. *The American Economic Review*, 95(4):1190–1213.

Bascom, W. (1955). Urbanization among the Yoruba. *American Journal of Sociology*, 60(5):446–454.

Baum-Snow, N. (2007). Did highways cause suburbanization? *The Quarterly Journal of Economics*, 122(2):775–805.

Baum-Snow, N., Brandt, L., Henderson, J. V., Turner, M. A., and Zhang, Q. (2012). Roads, railroads and decentralization of Chinese cities. *Review of Economics and Statistics*, 99(3):435–448.

Baum-Snow, N., Henderson, J. V., Turner, M. A., Zhang, Q., and Brandt, L. (2016). Highways, market access and urban growth in China. *International Growth Centre Working Paper C-89114-CHN-1*.

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.

Bertazzini, M. C. (2018). The Long-term Impact of Italian Colonial Roads in the Horn of Africa, 1935-2015. *LSE Economic History Working Paper 72/2018. Forthcoming in The Journal of Economic Geography*.

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

Bogart, D. and Chaudhary, L. (2013). Engines of growth: the productivity advance of Indian railways, 1874–1912. *The Journal of Economic History*, 73(2):339–370.

Bogart, D. and Chaudhary, L. (2015). Railways in Colonial India: An Economic Achievement? In Chaudhary, L., Gupta, B., Roy, T., and Swamy, A. V., editors, *A New Economic History of Colonial India*, chapter 9, pages 140–160. New York.

Bogart, D., You, X., Alvarez, E., Satchell, M., and Shaw-Taylor, L. (2020). Railways, population divergence, and structural change in 19th century England and Wales. *Working Paper*.

Bose, A. and Bhatia, J. (1980). *India's urbanization, 1901-2001*. Tata McGraw-Hill.

Bosker, M., Buringh, E., and van Zanden, J. L. (2013). From Baghdad to London: Unraveling Urban Development in Europe, the Middle East, and North Africa, 800–1800. *Review of Economics and Statistics*, 95(4):1418–1437.

Broadberry, S., Custodis, J., and Gupta, B. (2015). India and the Great Divergence: An Anglo-Indian comparison of GDP per capita, 1600–1871. *Explorations in Economic History*, 55:58–75.

Broadberry, S. and Gupta, B. (2006). The early modern great divergence: wages, prices and economic development in Europe and Asia, 1500–1800. *The Economic History Review*, 59(1):2–31.

Bryan, G., Glaeser, E., and Tsivanidis, N. (2020). Cities in the developing world. *Annual Review of Economics*, 12:273–97.

Burgess, R. and Donaldson, D. (2010). Can openness mitigate the effects of weather shocks? Evidence from India's famine era. *American Economic Review*, 100(2):449–53.

Burgess, R. and Donaldson, D. (2017). Railroads and the demise of famine in colonial India. *Working Paper*.

Cantoni, D. (2015). The economic effects of the Protestant Reformation: testing the Weber hypothesis in the German lands. *Journal of the European Economic Association*, 13(4):561–598.

Cantoni, D. and Yuchtman, N. (2014). Medieval universities, legal institutions, and the commercial revolution. *The Quarterly Journal of Economics*, 129(2):823–887.

Castelló-Climent, A., Chaudhary, L., and Mukhopadhyay, A. (2018). Higher education and prosperity: From Catholic missionaries to luminosity in India. *The Economic Journal*, 128(616):3039–3075.

Chandler, T. and Fox, G. (1974). *3000 years of urban growth*. Academic Press.

Chaudhary, L. and Fenske, J. (2020). Did Railways Affect Literacy? Evidence from India. *CAGE Working Paper 529/2020*.

Chaudhary, L. and Garg, M. (2015). Does history matter? Colonial education investments in India. *The Economic History Review*, 68(3):937–961.

Clingingsmith, D. and Williamson, J. G. (2008). Deindustrialization in 18th and 19th century India: Mughal decline, climate shocks and British industrial ascent. *Explorations in Economic History*, 45(3):209–234.

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

de Vries, J. (1984). *European Urbanization, 1500-1800*. Routledge.

Dincecco, M., Fenske, J., and Menon, A. (2020). The Columbian Exchange and conflict in Asia. *Centre for Competitive Advantage in the Global Economy Working Paper 527 2020*.

Dittmar, J. (2011a). Cities, markets, and growth: the emergence of Zipf's law. *Working Paper: Institute for Advanced Study*.

Dittmar, J. E. (2011b). Information technology and economic change: the impact of the printing press. *The Quarterly Journal of Economics*, 126(3):1133–1172.

Donaldson, D. (2018). Railroads of the Raj: Estimating the impact of transportation infrastructure. *The 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.

Doornik, J. A. and Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, 70(s1):927–939.

Eaton, J. and Kortum, S. (2002). Technology, geography, and trade. *Econometrica*, 70(5):1741–1779.

Eeckhout, J. (2004). Gibrat's law for (all) cities. *The American Economic Review*, 94(5):1429–1451.

Faber, B. (2014). Trade integration, market size, and industrialization: evidence from China's National Trunk Highway System. *Review of Economic Studies*, 81(3):1046–1070.

Fajgelbaum, P. and Redding, S. J. (2018). Trade, Structural Transformation and Development: Evidence from Argentina 1869-1914. *National Bureau of Economic Research Working Paper No. 20217*.

Fogel, R. W. (1964). *Railroads and American Economic Growth: Essays in Econometric History*. Johns Hopkins Press.

Forero, A., Gallego, F. A., González, F., and Tapia, M. (2020). Railroads, specialization, and population growth: evidence from the first globalization. *Forthcoming in the Journal of Population Economics*.

Gabaix, X. (2009). Power laws in economics and finance. *Annual Review of Economics*, 1(1):255–294.

Gabaix, X. (2016). Power laws in economics: An introduction. *The Journal of Economic Perspectives*, 30(1):185–205.

Gabaix, X. and Ibragimov, R. (2011). Rank- 1/2: a simple way to improve the OLS estimation of tail exponents. *Journal of Business & Economic Statistics*, 29(1):24–39.

Gabaix, X. and Ioannides, Y. M. (2004). The evolution of city size distributions. *Handbook of Regional and Urban Economics*, 4:2341–2378.

Gadgil, D. R. (1959). *The industrial evolution of India in recent times*. Oxford University Press.

Gollin, D., Jedwab, R., and Vollrath, D. (2016). Urbanization with and without industrialization. *Journal of Economic Growth*, 21(1):35–70.

Gupta, B. (2015). The rise of modern industry in colonial India. In Chaudhary, L., Gupta, B., Roy, T., and Swamy, A. V., editors, *A New Economic History of Colonial India*, chapter 5, pages 67–83. New York.

Gupta, B. (2019). Falling Behind and Catching up: India's Transition from a Colonial Economy. *The Economic History Review*, 72(3):803–827.

Hanlon, W. W. and Heblich, S. (2020). History and Urban Economics. *Forthcoming in Regional Science and Urban Economics*.

Harris, C. D. (1954). The Market as a Factor in the Localization of Industry in the United States. *Annals of the association of American geographers*, 44(4):315–348.

Head, K. and Mayer, T. (2014). Gravity equations: Workhorse, toolkit, cookbook. In Gopinath, G., Helpman, E., and Rogoff, K., editors, *Handbook of International Economics*, volume 4, chapter 3, pages 131–195. Elsevier, New York.

Henderson, J. V. and Kriticos, S. (2018). The development of the African system of cities. *Annual Review of Economics*, 10:287–314.

Henderson, J. V., Squires, T., Storeygard, A., and Weil, D. (2018). The global distribution of economic activity: nature, history, and the role of trade. *The Quarterly Journal of Economics*, 133(1):357–406.

Henderson, J. V. and Turner, M. A. (2020). Urbanization in the developing world: too early or too slow? *Journal of Economic Perspectives*, 34(3):150–73.

Hornbeck, R. and Rotemberg, M. (2019). Railroads, reallocation, and the rise of American manufacturing. *National Bureau of Economic Research Working Paper No. 26594*.

Hurd, J. (1975). Railways and the Expansion of Markets in India, 1861–1921. *Explorations in Economic History*, 12(3):263–288.

Hurd, J. (2012). A Huge Railway System But No Sustained Economic Development: The Company Perspective, 1884-1939: Some Hypotheses. *Kerr, Ian (ed) 27 Down: New Departures in Indian Railway Studies*, pages 314–352.

Hurd, J. and Kerr, I. J. (2012). *India's railway history: a research handbook*. Brill.

Hurd, J. M. (1983). Railways. In Kumar, D. and Desai, M., editors, *The Cambridge Economic History of India: Vol. II, c.1757–c.1970*, pages 737–761. Cambridge University Press, Cambridge.

Iyer, L. (2010). Direct versus indirect colonial rule in India: Long-term consequences. *The Review of Economics and Statistics*, 92(4):693–713.

Jaques, T. (2007). *Dictionary of battles and sieges: A guide to 8,500 battles from antiquity through the twenty-first century*. Greenwood Publishing Group.

Jedwab, R., Kerby, E., and Moradi, A. (2017). History, path dependence and development: Evidence from colonial railroads, 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.

Jha, S. (2013). Trade, institutions, and ethnic tolerance: Evidence from South Asia. *American Political Science Review*, 107(4):806–832.

Johnson, N. D. and Koyama, M. (2017). Jewish communities and city growth in preindustrial Europe. *Journal of Development Economics*, 127:339–354.

Juhász, R. (2018). Temporary protection and technology adoption: Evidence from the Napoleonic blockade. *American Economic Review*, 108(11):3339–76.

Kiszewski, A., Mellinger, A., Spielman, A., Malaney, P., Sachs, S. E., and Sachs, J. (2004). A global index representing the stability of malaria transmission. *The American journal of tropical medicine and hygiene*, 70(5):486–498.

Kuehlwein, M. (2021). Railroads and Trade in 19th-Century India. In *Oxford Research Encyclopedia of Asian History*.

McAlpin, M. B. (1974). Railroads, Prices, and Peasant Rationality: India 1860–1900. *The Journal of Economic History*, 34(03):662–684.

Morten, M. and Oliveira, J. (2016). The Effects of Roads on Trade and Migration: Evidence from a Planned Capital City. *NBER Working Paper No. 22158*.

Nagy, D. K. (2020). Hinterlands, city formation and growth: Evidence from the US westward expansion. *Economics Working Papers 1717, Department of Economics and Business, Universitat Pompeu Fabra*.

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

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.

Parliamentary Papers (1854). *131. Railways (India)*. Copy of memorandum by Major Kennedy, on the question of a general system of railways for India, referred to in the minute by the Governor-General in Council of 20 April 1853, relative to railway undertakings in that country. *Vol.XLVIII, 32 pp*.

Parthasarathi, P. (2011). *Why Europe grew rich and Asia did not: Global economic divergence, 1600–1850*. Cambridge University Press.

Pérez, S. (2018). Railroads and the rural to urban transition: Evidence from 19th-century Argentina. *Working Paper, University of California, Davis*.

Redding, S. J. and Sturm, D. M. (2008). The costs of remoteness: Evidence from German division and reunification. *The American Economic Review*, 98(5):1766–97.

Rosen, K. T. and Resnick, M. (1980). The size distribution of cities: an examination of the Pareto law and primacy. *Journal of Urban Economics*, 8(2):165–186.

Rothermund, D. (2002). *An economic history of India*. Routledge.

Roy, T. (2011). Economic History of India, 1857-1947. *OUP Catalogue*.

Roy, T. (2012). *India in the world economy: from antiquity to the present*. Cambridge University Press.

Roy, T. (2015). The growth of a labour market in the twentieth century. In Chaudhary, L., Gupta, B., Roy, T., and Swamy, A. V., editors, *A New Economic History of Colonial India*, chapter 11, pages 179–194. Routledge, New York.

Soo, K. T. (2005). Zipf's law for cities: a cross-country investigation. *Regional Science and Urban Economics*, 35(3):239–263.

Srivastava, H. S. (1968). *The History of Indian Famines and Development of Famine Policy, 1858-1918*. Sri Ram Mehra.

Studer, R. (2008). India and the Great Divergence: Assessing the efficiency of grain markets in eighteenth-and nineteenth-century India. *Journal of Economic History*, 68(02):393–437.

Sweeney, S. (2011). *Financing India's Imperial Railways, 1875–1914*. Routledge.

Thorner, D. (1951). Great Britain and the development of India's railways. *The Journal of Economic History*, 11(4):389–402.

Thorner, D. (1955). The pattern of railway development in India. *The Journal of Asian Studies*, 14(2):201–216.

Tomlinson, B. R. (2013). *The economy of modern India: from 1860 to the twenty-first century*. Cambridge University Press.

Visaria, L. and Visaria, P. (1983). Population (1757-1947). In Kumar, D. and Desai, M., editors, *The Cambridge Economic History of India*, volume 2, pages 463–532. Cambridge University Press, Cambridge.

Waldinger, M. (2015). The economic effects of long-term climate change: Evidence from the little ice age. *Working Paper: London School of Economics*.

Wolcott, S. (2015). Industrial labour in late colonial India. In Chaudhary, L., Gupta, B., Roy, T., and Swamy, A. V., editors, *A New Economic History of Colonial India*, chapter 12, pages 195–217. Routledge, New York.

Wrigley, E. A. (1985). Urban growth and agricultural change: England and the continent in the early modern period. *The Journal of Interdisciplinary History*, 15(4):683–728.

Table 1. Summary statistics

	(1)	(2)	(3)	(4)	(5)	(6)
	Year	N	Mean	s.d.	Min	Max
Population	1881	1,786	13,113	32,882	1,020	773,196
	1891	1,948	13,841	34,779	1,021	821,764
	1901	2,017	14,354	36,943	1,003	949,144
	1911	2,122	14,027	40,064	1,000	1,043,307
	1921	2,301	14,069	42,071	1,011	1,175,914
	1931	2,429	15,951	46,175	1,030	1,196,734
Distance to railroad (km)	1881	1,786	69.10	89.43	0	578.5
	1891	1,948	37.23	53.73	0	546.6
	1901	2,017	23.23	31.83	0	546.6
	1911	2,122	18.24	29.87	0	465.7
	1921	2,301	16.30	28.17	0	465.7
	1931	2,429	13.59	24.66	0	363.2
Distance to railroad ≤ 2 km	1881	1,786	0.143	0.351	0	1
	1891	1,948	0.235	0.424	0	1
	1901	2,017	0.308	0.462	0	1
	1911	2,122	0.382	0.486	0	1
	1921	2,301	0.416	0.493	0	1
	1931	2,429	0.457	0.498	0	1
2 km≤ Distance to railroads ≤ 10 km	1881	1,786	0.104	0.305	0	1
	1891	1,948	0.137	0.344	0	1
	1901	2,017	0.153	0.360	0	1
	1911	2,122	0.166	0.372	0	1
	1921	2,301	0.167	0.373	0	1
	1931	2,429	0.179	0.384	0	1
10 km ≤ Distance to railroads ≤ 20 km	1881	1,786	0.0857	0.280	0	1
	1891	1,948	0.119	0.324	0	1
	1901	2,017	0.139	0.346	0	1
	1911	2,122	0.143	0.350	0	1
	1921	2,301	0.142	0.349	0	1
	1931	2,429	0.135	0.342	0	1
Market access (θ=1)	1881	1,786	3.499	28.21	0.384	822.3
	1891	1,948	3.168	10.20	0.454	337.0
	1901	2,017	3.373	9.820	0.620	415.4
	1911	2,122	3.531	10.45	0.649	456.9
	1921	2,301	3.757	10.34	0.704	472.2
	1931	2,429	4.574	11.32	0.895	525.0

Table 2. Railroad distance and city size

	(1)	(2)	(3)	(4)	(5)	(6)
	ln(Population)				*ln(Railroad Distance)*	
	OLS		*IV*		*First Stage*	
ln(Railroad Distance)	-0.019***	-0.017***	-0.191***	-0.113***		
	(0.003)	(0.003)	(0.030)	(0.026)		
ln(1+LCP distance) X Year / 1000					-6.904***	-7.463***
					(0.717)	(0.861)
Observations	12,484	12,484	12,228	12,228	12,228	12,228
Within R2	0.00983	0.126				
KPF			92.67	74.97		
City and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Controls	None	Main X Year	None	Main X Year	None	Main X Year
LHS SD	0.849	0.849	0.849	0.849		
RHS SD	2.236	2.236	2.236	2.236		
Standardized β	-0.0506	-0.0447	-0.503	-0.298		

Note: Standard errors clustered by city are reported in parentheses. Main controls are latitude, longitude, log distance to river, log distance to coast, ruggedness, malaria, altitude, precipitation, temperature, and suitability for dryland rice, wetland rice, wheat and cotton. *** indicates significance at the 1% level, ** at 5% and * at 10%.

Table 3. Market access and city size

	(1)	(2)	(3)	(4)
	ln(Population)			
	OLS		*IV*	
ln(Market Access (θ=1))	0.628***	0.385***	1.370***	1.028***
	(0.095)	(0.082)	(0.166)	(0.214)
Observations	12,484	12,484	12,228	12,228
Within R2	0.0629	0.137		
KPF			304.2	231.1
City and Year Fixed Effects	Yes	Yes	Yes	Yes
Controls	None	Main X Year	None	Main X Year
LHS SD	0.849	0.849	0.849	0.849
RHS SD	0.489	0.489	0.489	0.489
Standardized β	0.362	0.222	0.790	0.593
	(5)	(6)	(7)	(8)
	ln(Market Access (ϑ=1))		*ln(Population)*	
	First Stage		*OLS*	
ln(1+LCP distance) X Year	0.963***	0.822***		
	(0.055)	(0.054)		
ln(Access to Distant Markets (θ=1))			0.886***	0.579***
			(0.052)	(0.059)
Observations	12,228	12,228	12,484	12,484
Within R2			0.0581	0.136
KPF				
City and Year Fixed Effects	Yes	Yes	Yes	Yes
Controls	None	Main X Year	None	Main X Year
LHS SD			0.849	0.849
RHS SD			0.371	0.371
Standardized β			0.387	0.253

Note: Standard errors clustered by city are reported in parentheses. Main controls are latitude, longitude, log distance to river, log distance to coast, ruggedness, malaria, altitude, precipitation, temperature, and suitability for dryland rice, wetland rice, wheat and cotton. *** indicates significance at the 1% level, ** at 5% and * at 10%.

Table 4. Heterogeneity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	ln(Population)							
ln(Railroad Distance)	-0.034***	-0.031***	0.007*	0.005	-0.019***	-0.018***	-0.021***	-0.019***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
Interaction	0.025***	0.025***	-0.041***	-0.035***	0.000	0.028***	0.022***	0.022***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.010)	(0.010)	(0.008)	(0.008)
Observations	10,640	10,640	12,484	12,484	12,484	12,484	12,484	12,484
	ln(Population)							
ln(Market Access (θ=1))	0.666***	0.413***	0.535***	0.300***	0.630***	0.386***	0.633***	0.387***
	(0.109)	(0.092)	(0.104)	(0.081)	(0.096)	(0.082)	(0.097)	(0.082)
Interaction	-0.134***	-0.145***	0.125***	0.124***	-0.038	-0.205	-0.082	-0.049
	(0.034)	(0.031)	(0.040)	(0.037)	(0.109)	(0.128)	(0.067)	(0.080)
Observations	10,640	10,640	12,484	12,484	12,484	12,484	12,484	12,484
Interaction Variable	Above-median initial size		Above-median railway distance in 1881		Medieval Port		River within 2 km	
City and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	None	Main X Year	None	Main X Year	None	Main X Year	None	Main X Year

	(9)	(10)	(11)	(12)	(13)	(14)
	ln(Population)					
ln(Railroad Distance)	-0.033***	-0.020***	-0.021***	-0.018***	-0.021***	-0.018***
	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
Interaction	0.033***	0.007	0.055***	0.017*	0.011***	0.006**
	(0.005)	(0.005)	(0.009)	(0.009)	(0.003)	(0.003)
Observations	12,484	12,484	12,484	12,484	12,484	12,484
	ln(Population)					
ln(Market Access (θ=1))	0.736***	0.525***	0.695***	0.448***	0.625***	0.384***
	(0.070)	(0.071)	(0.060)	(0.059)	(0.096)	(0.082)
Interaction	-0.300***	-0.207***	-0.490***	-0.271***	0.019	0.011
	(0.034)	(0.043)	(0.072)	(0.071)	(0.016)	(0.017)
Observations	12,484	12,484	12,484	12,484	12,484	12,484
Interaction Variable	Cotton suitability		Mutiny within 20 km		Famine Exposure	
City and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Controls	None	Main X Year	None	Main X Year	None	Main X Year

Note: Standard errors clustered by city are reported in parentheses. Main controls are latitude, longitude, log distance to river, log distance to coast, ruggedness, malaria, altitude, precipitation, temperature, and suitability for dryland rice, wetland rice, wheat and cotton. *** indicates significance at the 1% level, ** at 5% and * at 10%. In columns (13) and (14) famine exposure is included as an un-interacted control.

Table 5. Robustness to functional Form

	(1)	(2)	(3)	(4)	(5)	(6)
	ln(Population)					
1(distance≤2km)	0.071***	0.075***				
	(0.014)	(0.013)				
1(2km<distance≤10km)	0.011	0.007				
	(0.020)	(0.018)				
1(10km<distance≤20km)	0.006	0.004				
	(0.016)	(0.015)				
ln(Railroad Distance)			-0.020***	-0.024***	-0.027***	-0.034***
			(0.003)	(0.003)	(0.004)	(0.005)
Observations	12,484	12,484	12,484	12,484	12,484	12,484
Within R2	0.00544	0.125	0.127	0.128	0.128	0.127
Minimum distance			1 m	1 km	2 km	5 km
City and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Controls	None	Main X Year	Main X Year	Main X Year	Main X Year	Main X Year
Standardized β 2km	0.0388	0.0408				
Standardized β 2-10km	0.00449	0.00303				
Standardized β 10-20km	0.00233	0.00145				
Standardized β			-0.0488	-0.0500	-0.0494	-0.0479

Note: Standard errors clustered by city are reported in parentheses. Main controls are latitude, longitude, log distance to river, log distance to coast, ruggedness, malaria, altitude, precipitation, temperature, and suitability for dryland rice, wetland rice, wheat and cotton. *** indicates significance at the 1% level, ** at 5% and * at 10%. Distances below the minimum distance are recoded to equal the minimum distance.

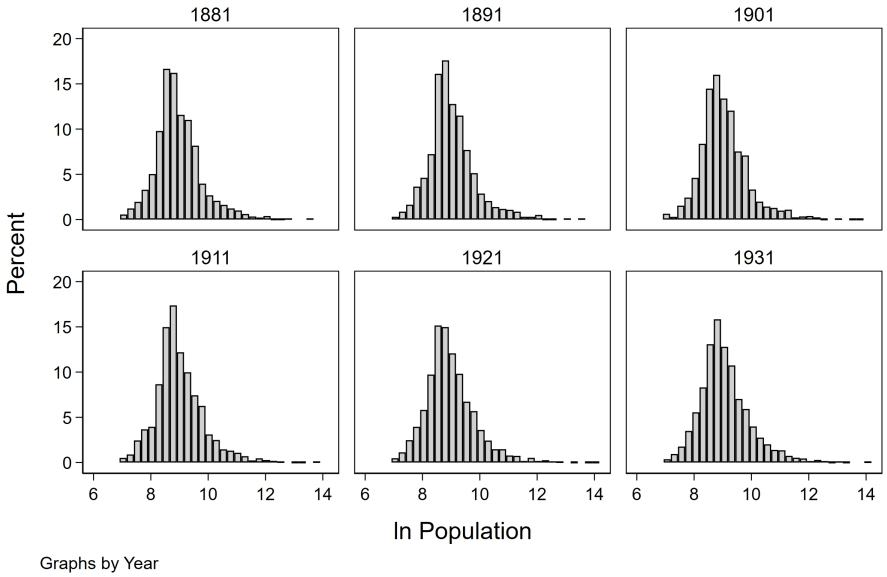
Table 6. Robustness: Market access

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	ln(Population)							
ln(Market Access)	0.628***	0.385***	0.048***	0.035***	0.014***	0.011***	0.013***	0.010***
	(0.095)	(0.082)	(0.010)	(0.008)	(0.003)	(0.003)	(0.003)	(0.002)
Observations	12,484	12,484	12,484	12,484	12,484	12,484	12,484	12,484
Within R2	0.0629	0.137	0.0224	0.130	0.0117	0.126	0.0112	0.125
Standardized β	0.362	0.222	0.200	0.145	0.132	0.102	0.129	0.0997
	ln(Population)							
ln(Market Access (Alternative Parameters))	0.298***	0.201***	0.029***	0.024***	0.010***	0.009***	0.009***	0.008***
	(0.081)	(0.060)	(0.005)	(0.004)	(0.002)	(0.002)	(0.002)	(0.001)
Observations	12,484	12,484	12,484	12,484	12,484	12,484	12,484	12,484
Within R2	0.0323	0.132	0.0154	0.129	0.0108	0.127	0.0105	0.127
Standardized β	0.187	0.126	0.120	0.101	0.0952	0.0858	0.0938	0.0849
	ln(Population)							
ln(Access to Distant Markets)	0.886***	0.579***	0.135***	0.108***	0.038***	0.033***	0.035***	0.030***
	(0.052)	(0.059)	(0.011)	(0.011)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	12,484	12,484	12,484	12,484	12,484	12,484	12,484	12,484
Within R2	0.0581	0.136	0.0311	0.136	0.0173	0.131	0.0168	0.131
Standardized β	0.387	0.253	0.184	0.148	0.102	0.0892	0.0988	0.0865
θ	1	1	3.6	3.6	7.8	7.8	8.28	8.28
City and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	None	Main X Year	None	Main X Year	None	Main X Year	None	Main X Year

Note: Standard errors clustered by city are reported in parentheses. Main controls are latitude, longitude, log distance to river, log distance to coast, ruggedness, malaria, altitude, precipitation, temperature, and suitability for dryland rice, wetland rice, wheat and cotton. *** indicates significance at the 1% level, ** at 5% and * at 10%.

Appendix: Not for publication

FIGURE A1. Distribution of log city size



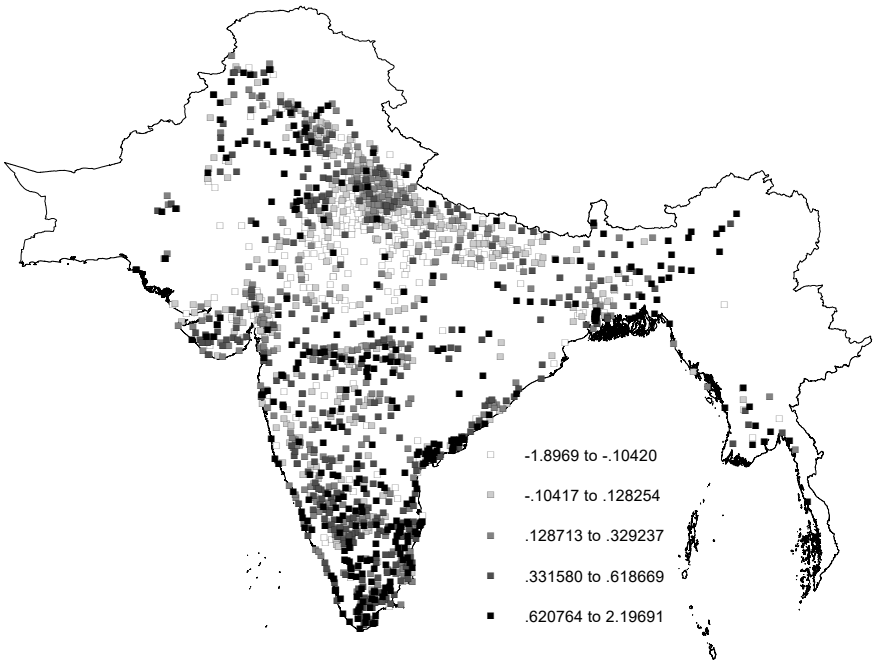
APPENDIX A. ADDITIONAL DATA DESCRIPTION

Because we introduce a new data source on Indian cities that we hope will be used by later researchers, we describe a number of features of the data here that have been of interest to other researchers working in economic geography. We focus on the distributions of both city size and city growth, on Zipf's law, on Gibrat's law, and on the geographic correlates of city growth.

Similar to the results Eeckhout (2004) reports for the United States, we find that log city size has a bell-curve shape, but that it is statistically distinguishable from a normal distribution. In Figure A1, we present histograms of log population size for each year in the data. Though these are single-peaked and relatively stable over time, it is clear from visual examination that they are to some extent right-skewed. In Table A5, we report results of a Doornik and Hansen (2008) test for normality for each year in our data. The p-values we find are all very small ($p < 0.0001$). This is driven by both the presence of positive skewness and of excess kurtosis in the data, also reported in Table A5.

Focusing now on city growth, we report summary statistics on the change in log population between years t and $t+10$ in Table A6, both for the whole sample and by decade. Although growth is generally positive, it is clear that many cities shrink. Indeed, the average city that is reported in 1901 shrank in size by 1911. City growth is more rapid in the 1920s, as the mean of $\Delta \ln P_i$ is, at 0.13, more rapid than for any other decade in the data. In Figure A2, we plot the quintiles of the change in log population over the interval 1881 to 1931 on a map. Again, while there is growth on average, many cities shrink

FIGURE A2. Quintiles of $\ln(Population_{1931}/Population_{1881})$

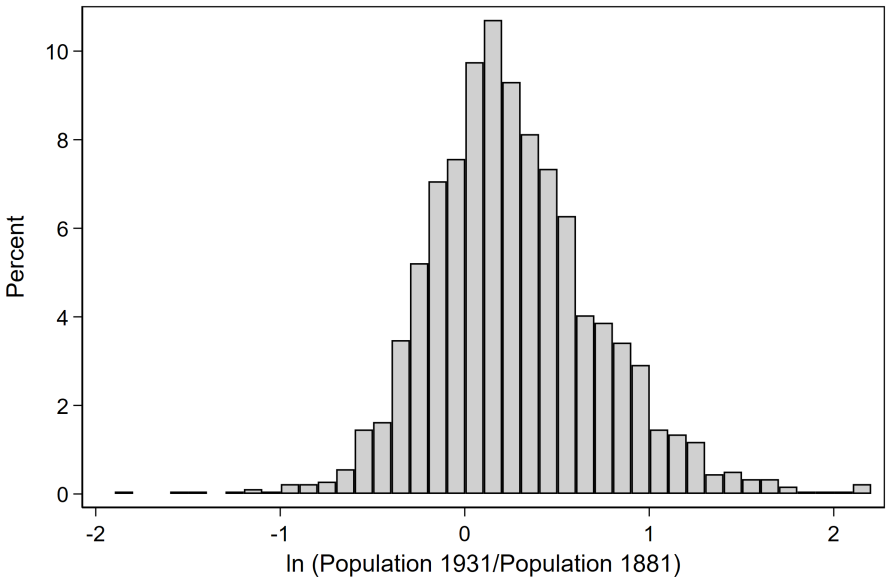


in size over this interval, particularly in the regions of modern-day Uttar Pradesh, Rajasthan, and Madhya Pradesh. In Figure A3, we plot the distribution of city growth over the interval 1881 to 1931. Much like city size, the distribution of city growth has a bell-curve shape, but displays positive skewness and excess kurtosis, and so is statistically distinguishable from a normal distribution.

A.1. **Zipf's law and Gibrat's law.** Gabaix (2016) lists the linear relationship between log city rank and log city size as one of the "nontrivial and true" power laws in economics that are both well established in the data and well understood in theory. Similar relationships have been found for firm sizes, stock market movements, income and wealth distributions, and CEO compensation (Gabaix, 2009, 2016). For city sizes, Zipf's law has been shown to fit the data reasonably well in the United States and United Kingdom, and poorly in many pre-modern or non-capitalist contexts (Dittmar, 2011a). Even where it does not fit exactly or can be rejected statistically (e.g. Rosen and Resnick (1980); Soo (2005)), Zipf's law still often fits the data closely (Gabaix and Ioannides, 2004).

Denoting R_i as a city's population rank in a given year and P_i as its population, we use Figures A4 and A5 to show the relationship between $\ln R_i$ and $\ln P_i$. In addition to plotting the raw data, we show the line of best fit obtained from using Ordinary Least Squares (OLS) to estimate:

FIGURE A3. Distribution of $\ln(Population_{1931}/Population_{1881})$



$$\ln R_i = \beta_0 + \beta_1 \ln P_i + \epsilon_i \tag{4}$$

Both figures are similar, and two patterns stand out. First, while cities with log population sizes between roughly 8 and 11 (or absolute populations between roughly 3,000 and 60,000) fall on the regression line, larger cities fall below it. This is similar to the pattern Dittmar (2011a) finds for European cities prior to 1500. Big cities are "too small" to fit a power law distribution. He accounts for this by appealing to the restrictions placed on the growth of large cities by the state of trade, agricultural productivity, and knowledge-based activities. These explanations could similarly account for late nineteenth century India, where markets were fragmented (Studer, 2008), agricultural production per head was stagnant, and the non-agricultural sector was relatively small (Broadberry et al., 2015). The second pattern that emerges is that the smaller cities are also "too small" relative to their ranks, and and also fall below the regression line.

In table A7, we report the results of estimating (4) for each year in the data. Though the figures above show that the largest and smallest cities in the data fail to lie on the regression line, the bulk of the data do lie close to the line, giving R^2 statistics above 0.9 in all cases. The coefficient estimates on β_1 are close to -1, though we can reject the hypothesis that $\hat{\beta}_1 = -1$ in all cases.

Though it is common to test Zipf's law formally by adding $\ln(P_i^2)$ as an additional control in (4), Gabaix (2009) notes that robust standard errors will be biased downwards.

FIGURE A4. Zipf's law in 1881

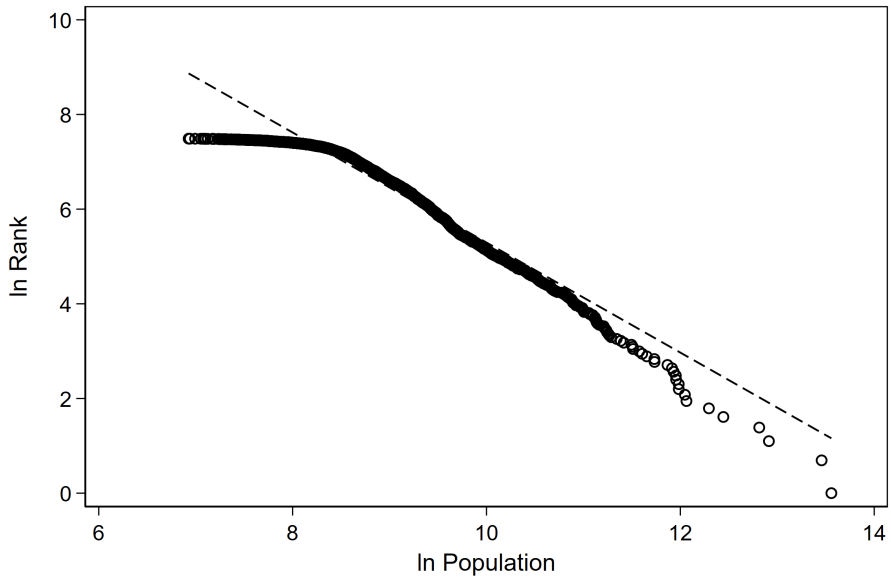
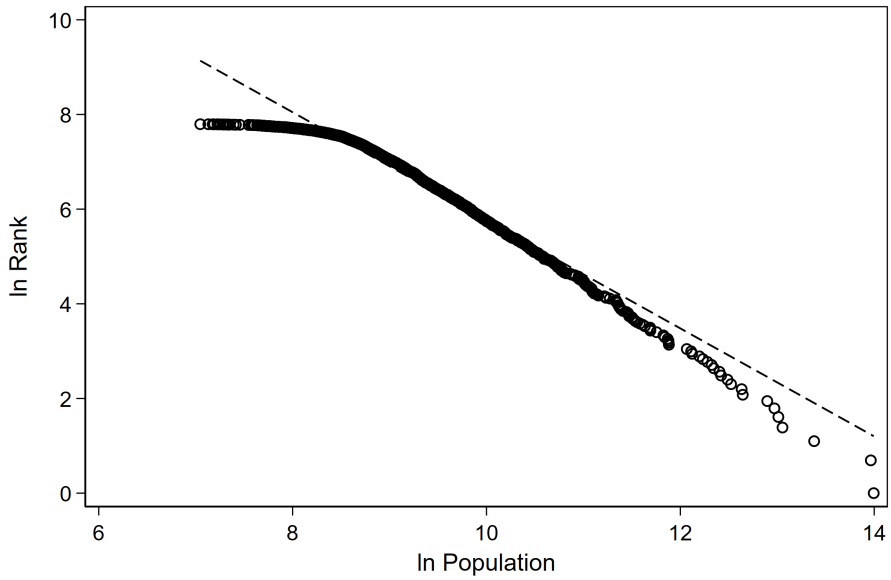


FIGURE A5. Zipf's law in 1931



Following Gabaix and Ibragimov (2011), and Dittmar (2011a), we instead implement the following test for Zipf's law:

$$\ln\left(R_i - \frac{1}{2}\right) = \beta_0 + \beta_1 \ln P_i + \beta_2 \left(\ln P_i - \frac{\text{Cov}((\ln P_i)^2, \ln P_i)}{2\text{Var}(\ln P_i)}\right)^2 + \epsilon_i \tag{5}$$

The Gabaix and Ibragimov (2011) test for a power law rejects at the 95% level if, and only if, $|\frac{\hat{\beta_2}}{\hat{\beta_1}^2}| > \frac{1.95}{\sqrt{2n}}$, where n is the sample size. We present results of this test in table A8. For every year in our data, we are able to reject the hypothesis that city sizes follow a power law distribution. This confirms formally the visual evidence above that Zipf's law does not describe the size distribution of Indian cities.

As Eeckhout (2004) notes, it is common to test Zipf's law only on a truncated sample of cities, often the 100 or so largest within a country. We also implement the Gabaix and Ibragimov (2011) test for the top 100 cities in each year, and again present results of this test in table A8. In this truncated sample, we fail to reject that the distribution of city sizes follows a power law.[14]

One of the mechanisms that can give rise to a power-law distribution of city sizes is proportional random growth (Gabaix, 2009, 2016): that is, if city growth follows Gibrat's law such that growth is independent of initial size. These theories struggle, however, with the transitory effects of large shocks, such as the bombing of Japanese cities during the Second World War (Davis and Weinstein, 2002). The correlation between city growth and initial city size has been investigated in several contexts, including China (Anderson and Ge, 2005) and the United States (Eeckhout, 2004).

In Figure A6, we show the nonparametric relationship between city growth and initial city size. In particular, taking $\ln\left(\frac{P_i^{1931}}{P_i^{1881}}\right)$ as a measure of population growth over the interval 1881 to 1931, and treating $\ln(P_i^{1881})$ as initial population, we present a kernel-weighted local polynomial smoothing, with 95% confidence intervals, of the relationship between growth and initial population. There is a downwards-slowing relationship between city size and growth for cities with a size of roughly 10,000 inhabitants ($\ln(10,000) \approx 9.21$) or less that flattens out for larger cities.

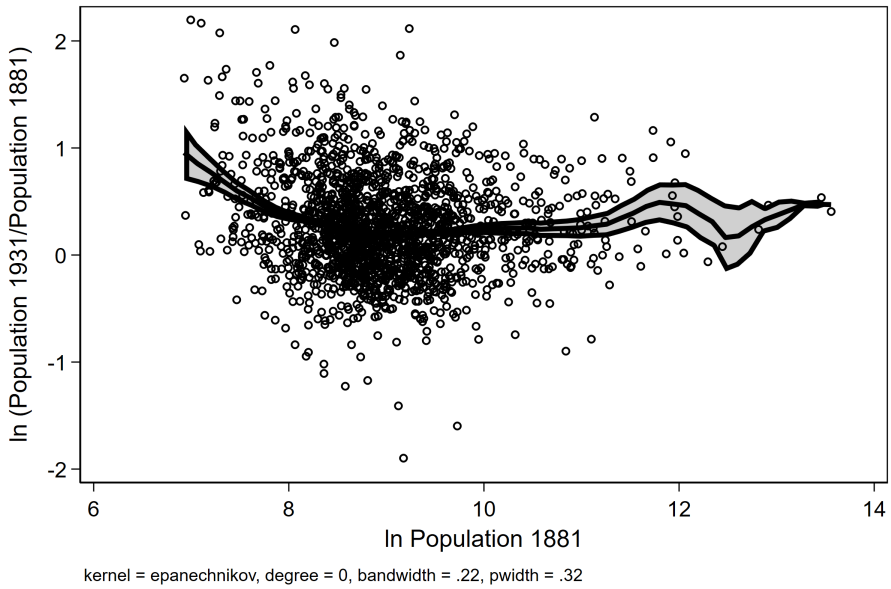
In Table A9, we present regression results that show we can reject the hypothesis of proportionate growth for most of our time period. For $t \in \{1881, 1891, ..., 1921\}$ and $t' \in \{t+10, ..., 1931\}$, we report regressions of population growth on initial population that take the form:

$$\ln\left(\frac{P_i^{t'}}{P_i^t}\right) = \beta_0 + \beta_1 \ln P_i + \epsilon_i \tag{6}$$

We estimate (6) using OLS and report heteroskedasticity-robust standard errors. We again restrict the sample to cities with populations of at least 1,000 in year t. For most intervals in our data, we find a significant negative relationship between initial city size and growth. Excepting over the intervals 1901-31, 1911-31, and 1921-31, smaller cities

[14]There are 101 cities in 1921, since the 100th and 101st largest cities, Ranchi and Bezwada, both have populations of 44,159.

FIGURE A6. $\ln(Population_{1931}/Population_{1881})$



Notes: This figure reports a kernel-weighted local polynomial smoothing of the relationship between $\ln(Population_{1931}/Population_{1881})$ and $\ln(Population_{1881}$, with 95% confidence intervals.

grow more quickly in our data. This is inconsistent with a common claim made in the literature that only large cities grew, while smaller towns stagnated or shrank (e.g. Bose and Bhatia (1980, p. 109)).

A.2. **Correlates of city size and growth.** In Table A10, we present regression results that describe the geographic correlates of city size and growth in our sample. We regress measures of city size (log population in 1881 and 1931) and city growth (the change in log population over the intervals 1881-1901 and 1881-1931) on the geographic correlates. We standardize all geographic variables by their means and standard deviations in the sample of 1881 cities, so the magnitudes of the reported coefficients can be interpreted as the predicted change in the outcome due to a one standard deviation increase in the variable of interest.

Latitude correlates with city size both in 1881 and 1931: Northern cities are larger on average. While greater altitude and temperature predict greater city size in 1881, they fail to do so by 1931. Greater distance from the coast predicts smaller cities in both periods. Greater suitability for wheat predicts larger cities in both periods, while greater suitability for cotton predicts smaller cities. Dryland rice suitability predicts larger cities only in 1931.

While Northern cities are larger in both periods, they grow more slowly. Malaria ecology positively predicts city growth in the 1881 to 1901 intervals, though not over 1881-1931. Dryland rice suitability predicts faster growth over the 1881-1931 interval, but not from 1881 to 1901. Wheat suitability positively predicts growth, while cotton suitability negatively predicts growth over both intervals. Cities further from rivers grow more rapidly over both intervals. Over both intervals, initial population predicts slower growth.

In sum: we find evidence of growth on average, but that many cities in our data shrink, particularly before 1921. Contrary to some claims in the literature and to the hypothesis of random proportional growth, it is the smallest towns in our data that grow the most rapidly. Both the largest and smallest cities in the data are too small to fit well with Zipf's law, though we cannot reject a power law distribution for the top 100 cities.

APPENDIX B. APPENDIX TABLES

Table A1. Additional robustness

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	ln(Population)				*ln(Population)*			
ln(Railroad Distance)	-0.019***	-0.017***			-0.019***	-0.017***		
	(0.003)	(0.003)			(0.003)	(0.003)		
ln(Market Access (θ=1))			0.631***	0.387***			0.633***	0.393***
			(0.097)	(0.083)			(0.100)	(0.085)
Robustness check	Drop cities more than 100km from a railroad in 1931				Drop Burma			
Observations	12,394	12,394	12,394	12,394	12,116	12,116	12,116	12,116
Standardized β	-0.0507	-0.0444	0.364	0.223	-0.0493	-0.0438	0.365	0.227
	ln(Population)				*ln(Population)*			
ln(Railroad Distance)	-0.025***	-0.016***			-0.025***	-0.012**		
	(0.005)	(0.004)			(0.006)	(0.006)		
ln(Market Access (θ=1))			0.639***	0.393***			1.085***	0.539***
			(0.099)	(0.085)			(0.145)	(0.108)
Robustness check	Truncate population at 1000				Tobit estimation			
Observations	14,591	14,591	12,484	12,484	14,592	14,592	14,592	14,592
Standardized β	-0.0511	-0.0328	0.364	0.224				
	ln(Population)				*ln(Population)*			
ln(Railroad Distance)	-0.021***	-0.018***			-0.009***	-0.009***		
	(0.003)	(0.002)			(0.003)	(0.003)		
ln(Market Access (θ=1))			0.922***	0.599***			0.115**	0.117**
			(0.039)	(0.045)			(0.056)	(0.056)
Robustness check	Drop influential observations				Include District X Year FE			
Observations	12,187	12,188	12,189	12,191	11,956	11,956	11,956	11,956
Standardized β	-0.0511	-0.0450	0.470	0.305	-0.0224	-0.0230	0.0664	0.0672
City and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	None	Main X Year	None	Main X Year	None	Main X Year	None	Main X Year

Note: Standard errors clustered by city are reported in parentheses. Main controls are latitude, longitude, log distance to river, log distance to coast, ruggedness, malaria, altitude, precipitation, temperature, and suitability for dryland rice, wetland rice, wheat and cotton. *** indicates significance at the 1% level, ** at 5% and * at 10%.

Table A2. Alternative Instruments for distance

	(1)	(2)	(3)	(4)
	ln(Population)			
Second Stage				
ln(Railroad Distance)	-0.239***	-0.161***	-0.193***	-0.106***
	(0.045)	(0.046)	(0.033)	(0.028)
Standardized B	-0.630	-0.424	-0.508	-0.279
	ln(Railroad Distance)			
First Stage				
Instrument	-0.048***	-0.048***	-6.361***	-6.689***
	(0.007)	(0.010)	(0.766)	(0.929)
Observations	12,228	12,228	68.97	51.79
KPF	41.75	23.83	62.41	47.69
Instrument	LCP distance X Year / 1000		ln (1 + LCP A1 distance) X Year / 1000	
City and Year Fixed Effects	Yes	Yes	Yes	Yes
Controls	None	Main X Year	None	Main X Year
	(5)	(6)	(7)	(8)
	ln(Population)			
Second Stage				
ln(Railroad Distance)	-0.200***	-0.129***	-0.152***	-0.081***
	(0.030)	(0.026)	(0.028)	(0.024)
Standardized B	-0.527	-0.340	-0.401	-0.214
	ln(Railroad Distance)			
First Stage				
Instrument	-6.831***	-7.131***	-7.253***	-7.753***
	(0.766)	(0.942)	(0.779)	(0.968)
Observations	12,228	12,228	12,228	12,228
KPF	79.51	57.24	86.72	64.14
Instrument	ln (1 + LCP A2 distance) X Year / 1000		ln (1 + LCP A3 distance) X Year / 1000	
City and Year Fixed Effects	Yes	Yes	Yes	Yes
Controls	None	Main X Year	None	Main X Year

Note: Standard errors clustered by city are reported in parentheses. Main controls are latitude, longitude, log distance to river, log distance to coast, ruggedness, malaria, altitude, precipitation, temperature, and suitability for dryland rice, wetland rice, wheat and cotton. *** indicates significance at the 1% level, ** at 5% and * at 10%

Table A3. Alternative Instruments for market access

	(1)	(2)	(3)	(4)
	ln(Population)			
Second Stage				
ln(Market Access (θ=1))	1.176***	0.889***	1.344***	0.953***
	(0.167)	(0.214)	(0.177)	(0.232)
Standardized B	0.678	0.512	0.775	0.549
	ln(Railroad Distance)			
First Stage				
Instrument / 1000	0.010***	0.009***	0.913***	0.745***
	(0.001)	(0.001)	(0.060)	(0.061)
Observations	12,228	12,228	12,228	12,228
KPF	339	151.5	235.2	148.5
Instrument	LCP distance X Year		ln (1 + LCP A1 distance) X Year	
City and Year Fixed Effects	Yes	Yes	Yes	Yes
Controls	None	Main X Year	None	Main X Year
	(5)	(6)	(7)	(8)
	ln(Population)			
Second Stage				
ln(Market Access (θ=1))	1.452***	1.196***	1.222***	0.832***
	(0.164)	(0.212)	(0.183)	(0.233)
Standardized B	0.838	0.689	0.705	0.480
	ln(Railroad Distance)			
First Stage				
Instrument / 1000	0.941***	0.770***	0.902***	0.759***
	(0.057)	(0.059)	(0.057)	(0.056)
Observations	12,228	12,228	12,228	12,228
KPF	274.7	171.4	254.8	186.2
Instrument	ln (1 + LCP A2 distance) X Year		ln (1 + LCP A3 distance) X Year	
City and Year Fixed Effects	Yes	Yes	Yes	Yes
Controls	None	Main X Year	None	Main X Year

Note: Standard errors clustered by city are reported in parentheses. Main controls are latitude, longitude, log distance to river, log distance to coast, ruggedness, malaria, altitude, precipitation, temperature, and suitability for dryland rice, wetland rice, wheat and cotton. *** indicates significance at the 1%

Table A4 Robustness: Alternative Measures of Market access

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	ln(Population)							
Baseline Parameters								
ln(Market Access)	0.115**	0.117**	0.014***	0.014***	0.006***	0.006***	0.006***	0.006***
	(0.056)	(0.056)	(0.005)	(0.005)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	11,956	11,956	11,956	11,956	11,956	11,956	11,956	11,956
Within R2	0.00152	0.00424	0.00208	0.00474	0.00245	0.00510	0.00246	0.00511
Standardized β	0.0664	0.0672	0.0596	0.0593	0.0589	0.0586	0.0589	0.0586
	ln(Population)							
Alternative Parameters								
ln(Market Access)	0.076**	0.075*	0.013***	0.013***	0.006***	0.006***	0.006***	0.006***
	(0.038)	(0.038)	(0.004)	(0.004)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	11,956	11,956	11,956	11,956	11,956	11,956	11,956	11,956
Within R2	0.00180	0.00442	0.00342	0.00603	0.00454	0.00714	0.00457	0.00717
Standardized β	0.0475	0.0468	0.0553	0.0548	0.0599	0.0594	0.0600	0.0595
θ	1	1	3.6	3.6	7.8	7.8	8.28	8.28
City and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	None	Main X Year	None	Main X Year	None	Main X Year	None	Main X Year

Note: Standard errors clustered by city are reported in parentheses. Main controls are latitude, longitude, log distance to river, log distance to coast, ruggedness, malaria, altitude, precipitation, temperature, and suitability for dryland rice, wetland rice, wheat and cotton. *** indicates significance at the 1% level, ** at 5% and * at 10%.

Table A5 Doornik and Hansen (2008) Tests for Normality

	(1)	(2)	(3)
Year	*p-value*	*Skewness*	*Kurtosis*
1881	0	1.033138	5.661367
1891	0	1.10976	5.849988
1901	0	1.059113	5.712863
1911	0	1.011092	5.622181
1921	0	1.027677	5.46955
1931	0	1.03191	5.367925

Table A6 Summary Statistics: City Growth

	(1)	(2)	(3)	(4)	(5)
D ln Population	0.060	0.22	-1.50	2.44	10,108
D ln Population: 1891	0.093	0.20	-1.13	1.55	1,768
D ln Population: 1901	0.050	0.22	-1.33	1.44	1,932
D ln Population: 1911	-0.018	0.23	-1.31	1.03	1,995
D ln Population: 1921	0.033	0.23	-1.06	2.33	2,115
D ln Population: 1931	0.13	0.19	-1.50	2.44	2,298

Table A7 Log rank and log city size

	(1)	(2)	(3)	(4)	(5)	(6)
	ln(Rank)					
ln Population	-1.163***	-1.176***	-1.162***	-1.134***	-1.112***	-1.091***
	(0.014)	(0.013)	(0.014)	(0.013)	(0.012)	(0.011)
Observations	1,788	1,949	2,018	2,124	2,303	2,431
R-squared	0.939	0.944	0.942	0.938	0.944	0.946
Year	1881	1891	1901	1911	1921	1931
p	0	0	0	0	0	0

Note: All regressions are OLS and include a constant. Robust standard errors in parentheses. p-values are for a test of whether the coefficient on ln(Population) = 1. *** indicates significance at the 1% level, ** at 5% and * at 10%.

Table A8 Results of Gabaix and Ibragimov (2011) Test

(1)	(2)	(3)	(4)	(5)	(6)
Year	β_1	β_2	n	$\lvert\beta_1/\beta_2^2\rvert$	$1.95/\sqrt{(2n)}$
			All Cities		
1881	-1.171	-0.157	1788	0.115	0.033
1891	-1.184	-0.147	1949	0.105	0.031
1901	-1.17	-0.151	2018	0.111	0.031
1911	-1.141	-0.151	2124	0.116	0.03
1921	-1.119	-0.145	2303	0.116	0.029
1931	-1.097	-0.141	2431	0.117	0.028
			Top 100 Cities		
1881	-1.55	-0.029	100	0.012	0.138
1891	-1.56	-0.06	100	0.025	0.138
1901	-1.549	-0.05	100	0.021	0.138
1911	-1.467	-0.02	100	0.009	0.138
1921	-1.446	-0.045	101	0.022	0.137
1931	-1.461	-0.076	100	0.035	0.138

Note: These results report the results of estimating the Gabaix and Ibragimov (2011) equation using OLS.

Table A9 City Size and Growth

(1) Start Year	(2) End Year	(3) β_1	(4) s.e.		(5) n
1881	1891	-0.03427	0.00599	*	1768
1881	1901	-0.06074	0.00859	*	1773
1881	1911	-0.07591	0.01179	*	1764
1881	1921	-0.08069	0.01354	*	1771
1881	1931	-0.06252	0.01462	*	1784
1891	1901	-0.03403	0.0061	*	1932
1891	1911	-0.0502	0.00933	*	1928
1891	1921	-0.04758	0.01065	*	1932
1891	1931	-0.03402	0.01192	*	1943
1901	1911	-0.01911	0.00642	*	1995
1901	1921	-0.01755	0.00832	*	2000
1901	1931	-0.00201	0.00978		2010
1911	1921	-0.01883	0.0068	*	2115
1911	1931	-0.00459	0.00779		2120
1921	1931	0.00523	0.00482		2298

Note: These results report the results of regressing the log of the ratios of the population in the end year to the population in the start year on a constant and the log of population in the start year using OLS with heteroskedasticity-robust standard errors. *Significant at 5%.

Table A10 Correlates of city size and growth

	(1)	(2)	(3)	(4)
	ln Population 1881	ln Population 1931	ln Population 1901/Population 1881	ln Population 1931/Population 1881
Latitude	0.253***	0.090*	-0.131***	-0.113***
	(0.053)	(0.049)	(0.019)	(0.029)
Longitude	-0.049	0.022	-0.002	-0.018
	(0.040)	(0.031)	(0.013)	(0.021)
Ruggedness	0.006	-0.038	0.009	-0.011
	(0.058)	(0.048)	(0.021)	(0.033)
Malaria	-0.003	-0.032	0.051***	0.011
	(0.046)	(0.028)	(0.018)	(0.032)
Altitude	0.231**	-0.002	-0.037	-0.106*
	(0.103)	(0.090)	(0.033)	(0.058)
Dry Rice Suitability	-0.025	0.202***	0.042**	0.189***
	(0.058)	(0.050)	(0.020)	(0.035)
Precipitation	0.056	-0.004	0.004	0.017
	(0.053)	(0.041)	(0.017)	(0.026)
Temperature	0.611***	0.193*	-0.016	-0.110
	(0.118)	(0.101)	(0.040)	(0.068)
Wetland Rice Suitability	0.088	-0.040	-0.051***	-0.059**
	(0.054)	(0.048)	(0.019)	(0.029)
Wheat Suitability	0.202***	0.208***	0.089***	0.119***
	(0.065)	(0.056)	(0.023)	(0.039)
Cotton Suitability	-0.062*	-0.096***	-0.046***	-0.107***
	(0.037)	(0.029)	(0.011)	(0.020)
Ln Coast Distance	-0.134***	-0.128***	0.005	0.015
	(0.038)	(0.034)	(0.011)	(0.018)
Ln River Distance	-0.019	-0.022	0.022***	0.032***
	(0.022)	(0.020)	(0.007)	(0.012)
ln Population 1881			-0.054***	-0.052***
			(0.008)	(0.014)
Observations	1,786	2,429	1,772	1,783
R-squared	0.060	0.057	0.168	0.156

Note: All regressions are OLS and include a constant. Robust standard errors in parentheses. p-values are for a test of whether the coefficient on ln(Population) = 1. *** indicates significance at the 1% level, ** at 5% and * at 10%.