# Railways, divergence, and structural change in 19th century England and Wales[1]

Dan Bogart[2] , Xuesheng You[3] , Eduard Alvarez[4] , Max Satchell[5], and Leigh Shaw-Taylor[6]

November 6, 2020

**Abstract**

Railways transformed inland transport during the nineteenth century. In this paper, we study how railways led to population change and divergence in England and Wales as it underwent dramatic urbanization. We make use of detailed data on railway stations, population, and occupational change in more than 9000 spatial units. A least cost path based on major towns and the length of the 1851 rail network is also created to address endogeneity. Our instrumental variable estimates show that having a railway station in a locality by 1851 led to significantly higher population growth from 1851 to 1891 and shifted the male occupational structure away from agriculture. Moreover, we estimate that having stations increased population growth more if localities had greater population density in 1801 and there were population losses for localities 3 to 15 km from stations. Overall, we find that railways reinforced the population hierarchy of the early nineteenth century and contributed to further spatial divergence. The implications for national income and labor productivity are large.

*Keywords*: Urbanization, railways, transport, reorganization, divergence

*JEL Codes*: N7, O1, R4

[1] Data for this paper was created thanks to grants from the Leverhulme Trust (RPG-2013-093), Transport and Urbanization c.1670-1911, from the NSF (SES-1260699), Modelling the Transport Revolution and the Industrial Revolution in England, from the ESRC (RES 000-23-0131), Male Occupational Change and Economic Growth in England 1750 to 1851, and the ESRC (RES-000-23-1579) the Occupational Structure of Nineteenth Century Britain. We thank Walker Hanlon, Gary Richardson, Petra Moser, Kara Dimitruk, Arthi Vellore, William Collins, Jeremy Atack, Stephan Heblich, James Fenske, and Elisabet Viladecans Marsal for comments on earlier drafts and seminar participants at Warwick, Bristol, UC Merced, UC Irvine, UC San Diego, NYU, Florida State, Trinity College Dublin, Queens Belfast, Los Andes, Vanderbilt, and EHA Meetings. We also thank Gill Newton, of the Cambridge Group for History of Population and Social Structure, who developed the Python code. We further thank Cranfield University for share soils data.

[2] Professor, Department of Economics, UC Irvine, dbogart@uci.edu

[3] Research Associate, Faculty of History, University of Cambridge, xy242@cam.ac.uk

[4] Senior Lecturer, Economics and Business, Universitat Oberta de Catalunya, ealvarezp@uoc.edu

[5] Research Associate, Dept. of Geography, University of Cambridge, aems2@cam.ac.uk

[6] Senior Lecturer, Faculty of History, University of Cambridge, lmws2@cam.ac.uk

## 1. Introduction

Britain's urbanization was exceptional during the nineteenth century. Between 1800 and 1900 the percentage of its population living in cities of 5000 or more increased from 19 to 67. In the whole of Europe, the urbanization rate increased far less from 11 to 30 between 1800 and 1900 and even in the United States urbanization rates increased less from 5 to 36 (Bairoch and Goertz 1986). Britain's urbanization process was remarkable in another respect. Between 1850 and 1900 its urban areas grew dramatically, but its rural areas had little growth. This was not true elsewhere in the world. Population growth was more balanced between urban and rural in much of continental Europe during the late nineteenth century (see Cameron 2003, p. 193).

In this paper we study how railways contributed to population change and spatial divergence in a key part of the British economy, England and Wales. We use theories on how transport improvements affect the spatial distribution of economies and emphasize agglomeration.[7] Our starting point is that commercial and industrial firms would have had an incentive to locate near railway stations because the low cost, high speed network could be easily accessed there. That implies that population should increase near railway stations because of greater employment opportunities. However, this process of re-location may not be uniform across the initial population distribution. If agglomeration is strong, then as transport costs decrease from high to moderate levels, the most densely populated areas grow at the expense of the least densely populated areas. This would suggest that when more densely populated areas got a railway station, the positive effect on their population growth will be larger. Other economic changes follow. Land rents rise with greater population, and thus land-intensive sectors become less profitable. This implies the occupational structure should move away from agriculture near stations and into either manufacturing or services.

We use a uniquely detailed and highly granular dataset of 9489 spatial 'units' constructed from parishes, townships, and hamlets reported in the British Census.[8] We observe populations in every decennial census year from 1801 to 1891 and male occupational shares in agriculture,

[7] See Fujita, Krugman, and Venables (2001), Lafourcade and Thisse (2011) for an introduction.
[8] Unfortunately, our population data do not include Scotland or Ireland, and thus we cannot study the whole UK.

secondary, and tertiary in 1851 and 1881. We also incorporate highly accurate GIS data on railway lines and stations in each census year, geographic characteristics, like coastline and coal, and pre-rail infrastructure networks like turnpike roads, ports, and inland waterways.

The empirical analysis estimates how being near a railway station in the mid-nineteenth century affected local population growth and changes in occupational structure over the following decades in England and Wales. Our baseline specification studies population change from 1851 to 1891 and uses an indicator for having a station within a unit's boundary by 1851 as the main railway variable. Endogeneity is a major challenge, especially as English and Welsh railways were built and owned by private companies pursuing profits. As a solution, we construct a least cost path (LCP) based on the length of the 1851 network and locations of large 1801 towns, which serve as nodes in the LCP. The main instrumental variable or IV is an indicator for having the LCP in a unit. It is a strong predictor for having a station by 1851 since it captures favorable routes for railways and stations were very dense along lines. If the sample excludes units near the nodes, the LCP indicator can also be reasonably excluded as an independent variable in the second stage analysis for population growth from 1851 to 1891.

The preferred estimates imply that having an 1851 station caused unit population to grow by an additional 0.87% per year from 1851 to 1891. This effect is large considering the average unit in our data lost 0.06% in population per year and the standard deviation in annual growth was 1.18%. We also estimate that having a station by 1851 led to a 0.121 *decrease* in the share of agricultural occupations between 1851 and 1881, equivalent to a 33% decline, and led to a 0.063 *increase* in the share of secondary occupations, or a 36% increase.

Our main extension examines how railways reinforced the population hierarchy of the early nineteenth century and contributed to further divergence. We find that having stations increased population growth significantly more when units had greater log population density in 1801. Moreover, station growth effects were close to zero for units in the bottom six deciles of the 1801 population distribution. Thus, railways contributed to further divergence between units that were small and large in 1801.

Another extension estimates population displacement effects based on distance to station. Being less than 3 km from 1851 stations led to higher population growth compared to units more than 20 km away, while being 3 to 15 km led to significantly *lower* population growth. Moreover, being in the 3 to 15 km zone led to increased shares of agricultural occupations and decreased shares of tertiary. Therefore, divergence also happened at a local scale near stations.

Our estimates also speak to the economy-wide effects of railways. We predict the 1891 population of all spatial units if none had stations by 1851. The 'counter-factual' total 1891 population is found to be 22% smaller. Moreover, the population share of the top 5% of units falls from 0.687 to 0.575, which accounts for most of the change in the top 5% share between 1851 and 1891. Concerning male occupations, we estimate that agriculture would have increased by 23%. These effects have implications for labor productivity which we estimate to decline by at least 4% due to reduced economies of density and less structural change.

Our results contribute to a large literature on railways and the English and Welsh economy. Several studies point to the importance of railways in affecting local populations.[9] Among the quantitative studies there is agreement that getting railway stations was associated with increased population density.[10] However, the causal effects of getting stations have not been established. We address endogeneity by constructing a novel LCP. Through heterogenous effects, we also estimate how railways fostered spatial divergence in Britain, which despite its remarkable features is relatively under-studied from a quantitative or theoretical approach.[11]

Our results also speak to the literature analyzing the aggregate impact of railways through the added consumer surplus from lower freight rates, fares, and higher passenger speeds.[12] Some estimates imply that the gains from shipment of freight were 10% of English and Welsh GDP in 1890 (Foreman-Peck 1991) and the gains to passengers in money and time saved were around 5% in 1890 (Leunig 2006). Our estimates for the labor productivity effects of population concentration and structural change suggests these are under-estimates.

[9] See Gourvish (1986) and Kellet (2012) for example.
[10] See Gregory and Martí Henneberg (2010), Casson (2013), Casson et. al. (2013), Alvarez et. al. (2013)
[11] Hanlon (2020) and Heblich, Redding, and Sturm (2018) for some exceptions.
[12] See Hawke (1970), Foreman-Peck (1991), and Leunig (2006) for examples.

We also make three contributions to a large literature studying railways, population, and economic change in different countries.[13] First, with few exceptions most studies use counties, districts, or cities as their spatial unit. We use small-scale spatial data, approximately at the village or town-level. Our study also introduces a richer set of geographic variables, like coal endowments, and a richer set of pre-railway infrastructures like roads and ports. Second, in constructing LCPs as instruments, most studies use straight lines to connect network nodes, however they are less accurate for small-scale spatial data. We use information on historical costs to create a non-linear LCP that incorporates sloping landscape.[14] Third, several studies analyze effects on firms and factories, but few examine effects on occupational change, one of the key transformations in development.[15] We estimate railways effects on changes in male agricultural, secondary, and tertiary employment.

How do our findings specifically relate to these other comparative studies? One of the few to analyze finely grained spatial data is Buchel and Kyburz (2020), who study municipalities in Switzerland. Employing a similar identification framework, these authors show that having a station increased annual population growth by 0.6%. Buchel and Kyburz also find evidence for population displacement 2 to 8 km from stations. By comparison, we find that in England and Wales having stations increased annual population growth by 0.87%, with displacement effects reaching 15 km from stations. We think the greater strength of agglomeration forces in the English and Welsh economy is likely to be one reason for the differences.

Berger (2019) is the only study we know of that estimates how railways affected occupational change using a similar framework. Studying Sweden, Berger shows that having a trunk railway line in a parish increased its manufacturing occupational share by 0.066. We find a nearly identical estimate for male secondary employment, which suggests that in two different environments, railways contributed to greater employment in manufacturing.

[13] See Tang (2014), Hornung (2015), Berger and Enflo (2017), Atack, Bateman, Haines, and Margo (2010), Donaldson and Hornbeck (2016), Hodgson (2018), Jedwab, Kerby, and Moradi (2015), and Donaldson (2018).
[14] Berger (2019) also uses slope and geographic impediments to create the LCP.
[15] Hornung (2015) studies number and size of firms, Attack, Haines, and Margo (2008) study factories, Tang (2014) studies firm capitalization.

Finally, our results contribute to a broader literature studying the effects of transport infrastructure, regional development, and structural change.[16] Most focus on local and regional outcomes in recent decades. Historical contexts complement this literature by demonstrating whether infrastructures create population gains as well as losses decades after they are built. The English historical context is particularly useful because it is closest to many current settings where infrastructure is built in developed economies with strong agglomeration forces.

The paper is organized as follows. Section 2 provides background. Sections 3, 4 and 5 introduce data and methods. Section 6 describes baseline results, while 7 and 8 examine heterogeneity and displacement. Sections 9 and 10 report counterfactuals and conclusions.

**2. Background on urbanization and railways**

England and Wales became highly urbanized in the nineteenth century. Decadal trends are reported in table 1 using the definition of urban as a census place with a population of 2500 or more (Law 1967). Up to 1841 urban growth exceeded rural growth, but both were positive. After 1841 urban growth remained high, while rural growth stagnated or declined. The greatest divergence between urban and rural occurred in the 1840s, 1860s, and 1870s. Many high growth areas were near the northern industrial centers of Manchester, Liverpool, and Leeds. The other high growth areas were near London, industrial Birmingham, and Cardiff in the South Wales coalfield. However, outside of these 'hotspots' there were few rapidly growing areas in Wales, the south, and east of England. Many villages and small towns had close to zero population growth after 1851.

Differences in net migration were the primary reasons for varied population growth. Shaw-Taylor and Wrigley (2014) show that industrial counties grew by 38% more than the average county between 1801 and 1851, but the rate of natural increase (birth rate minus death rate) in industrial counties was only 6% higher, and therefore the net migration rate must have been higher there. The primacy of migration was even stronger near London where

[16] A survey is provided by Duranton and Puga (2014). Also see Baum-Snow (2007), Duranton and Turner (2012), Michaels, Rauch, and Redding (2012), Faber (2014), Jedwab et. al. (2015), Garcia-Lopez et. al. (2015), Storeygard (2016), Ghani (2016), Holl (2016), Baum-Snow et. al. (2017), and Gibbons et. al. (2019), Pogonyi (et. al. 2019).

population growth was above average and the rate of natural increase below. Many individuals migrated within regions, but some moved great distances between regions. Better employment opportunities, especially outside of agriculture, appears to have been the key reason. [17]

Table 1: Decadal trends for urban and rural population in England and Wales, 1801-1891

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| year | Urban Pop. | % of Urban in total pop. | Urban growth rate over previous decade | Rural Pop. | % of Rural in total pop. | Rural growth rate over previous decade |
| 1801 | 3,009,260 | 33.8 | | 5,883,276 | 66.2 | |
| 1811 | 3,722,025 | 36.6 | 23.7 | 6,442,231 | 63.4 | 9.5 |
| 1821 | 4,804,534 | 40.0 | 29.1 | 7,195,702 | 60.0 | 11.7 |
| 1831 | 6,153,230 | 44.3 | 28.1 | 7,743,567 | 55.7 | 7.6 |
| 1841 | 7,693,126 | 48.3 | 25.0 | 8,221,022 | 51.7 | 6.2 |
| 1851 | 9,687,927 | 54.0 | 25.9 | 8,239,682 | 46.0 | 0.2 |
| 1861 | 11,784,056 | 58.7 | 21.6 | 8,282,168 | 41.3 | 0.5 |
| 1871 | 14,802,100 | 65.2 | 25.6 | 7,910,166 | 34.8 | -4.5 |
| 1881 | 18,180,117 | 70.0 | 22.8 | 7,794,322 | 30.0 | -1.5 |
| 1891 | 21,601,012 | 74.5 | 18.8 | 7,401,513 | 25.5 | -5.0 |
| 1901 | 25,371,849 | 78.0 | 17.5 | 7,155,994 | 22.0 | -3.3 |

Sources and Notes: Law (1967, p. 130). Urban includes census places with at least 2500 people.

There was an evolution in occupational structure related to increasing urbanization. The share of males in agriculture decreased from 0.27 in 1851 to 0.19 in 1871. The secondary share increased from 0.45 in 1851 to 0.46 in 1871. Tertiary increased the most from 0.23 to 0.28 (Shaw-Taylor and Wrigley 2014). The increase in secondary and tertiary was part of a long-term process starting with the early industrial revolution (Wallis, Colson, and Chilosi 2018).

*2.1 Development of railways*

England and Wales was the leader in developing railway networks and locomotives. But it is important to recognize that it had a well-developed transport network before.[18] By 1830 there were many good roads suitable for coaches and large wagons and a large inland waterway network for barges. There was a thriving coastal shipping trade based on sailing vessels. Coastal

[17] See Redford (1964), Boyer and Hatton (1997), Pooley and Turnbull (2005), Long (2005), Schurer and Day (2019), Day (2019).

[18] For a general summary see Bogart (2014).

ships could unload at hundreds of ports, 50 of which were major harbors with a total of 391 acres of wet dock space (Pope and Swann 1960). Along the coastline there were nearly 38 lighthouses with a visibility range of at least 15 miles.[19]

The pre-railway network was created and financed through local and private initiative. Government's role was to approve or reject proposals, assist land purchases, and regulate user-fees. Railways were developed using this system. Local business groups would introduce a bill in parliament that specified where the proposed railway would go and called for the creation of a company to finance construction and operate after. If approved, the company then collected subscription money from investors, and the construction process started.

The first steam powered freight and passenger rail service opened in 1830 between Liverpool and Manchester. It was followed by several others in the mid-1830s. At this early stage, railway companies were mainly interested in connecting the largest urban centers, because they had the most pre-existing passenger and freight services. By 1841, 9 of the 10 largest cities in England and Wales had railway connections, along with towns along their route.

The 'railway mania' of the mid-1840s saw the biggest expansion of the network. Between 1845 and 1847, 330 Railway Acts were passed to establish new railway companies or extend company networks. At the height of the mania the capital devoted to railways was more than twice as much as the British state spent on the military (Odlyzko, 2010). The mania was partly driven by the early railway company's strategy to maintain their position serving the large cities and by politicians wanting railway stations in their constituencies.[20]

The significance of the mania can be seen in the growth of railway lines. Between 1845 and 1851 railway km increased by 6,626, compared to an increase of 1,896 in the previous 6 years (You and Shaw-Taylor 2020). By 1851 regional networks had formed around the medium and large towns in addition to connections via the trunk lines (see figure 1). Yet there were still some regions that were under-served, most notably Wales and the southwest.

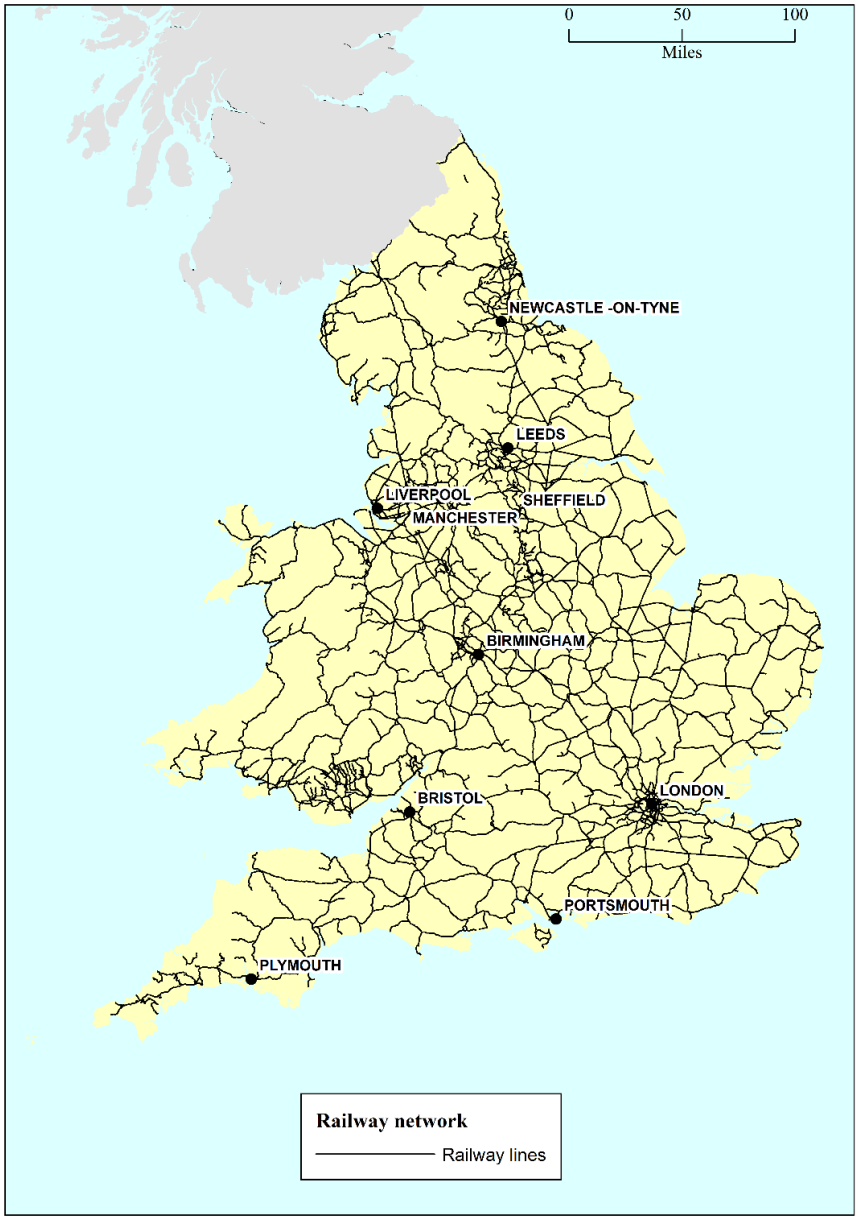
[19] See Buxton et. al. (2020).
[20] For the literature on the railway mania see Casson (2009), Odlyzko (2010), Campbell and Turner (2012, 2015)

Figure 1. The railway network in 1851


Source: Shaw-Taylor and You (2020).

The rail network further expanded and was nearly 25,000 km in 1881, or twice its size in 1851. Railway lines were now in every region of England and Wales (see figure 2). Within these regions there were some towns and rural areas that were better served than others, but none was very far from a railway. The network continued to be owned and operated by companies, but a process of consolidation left 7 companies with a network greater than 1000 km by 1910 (Crafts, Leunig, and Mulatu 2008).

Figure 2. The railway network in 1881




Source: Shaw-Taylor and You (2020).

Railways came to dominate most internal transport because they were superior in either speed or cost. Stagecoaches were displaced almost immediately when stations opened. Passenger miles increased at annual rates of 20% and 10% in the 1840s and 1850s. The growth rate of passenger miles fell to less than 5% by the 1860s reflecting a rate of increase closer to

GDP growth (Hawke 1970, p. 50).[21] In freight, canals offered some competition as barges charged similar freight rates, but they were slower and eventually lost traffic (Maw 2013). Railways managed to compete with coastal shipping, despite low shipping rates. One revealing statistic is that railways accounted for only 10% of the coal imported into London in 1851. The rest came by coast. But in 1870 railways accounted for 55% of the coal imported to London (Hawke 1970). Improvements to steamships led to a lower railway share by 1901, but they remained in an important shipper of coal in London and most towns (Armstrong 2009).

In our analysis one crucial issue relates to the routing of lines and placement of stations. The main consideration for lines built in the 1830s and 40s was to connect large cities by the most direct and flat route in order to save construction costs (Simmons 1986, pp. 169-171). Land acquisition costs were another consideration and when railways approached large towns, they often avoided built areas (Kellet 2012). When placing stations along the line, railway companies balanced the economic potential in the surrounding area against land acquisition costs (Casson 2009, Odlyzko 2010). In urban areas, stations were placed near centers if possible. In rural areas it was expected that individuals would travel to stations and were often placed at road junctions or near coaching inns to collect traffic. One example is Harrow & Wealdstone station, which the London and Birmingham railway opened in 1837. The nearest town, Harrow, had a small population in 1831. It was probably selected for a station because London coaches stopped there as early as the eighteenth century.

## 3. Data

Our population data come from British censuses available every decade starting in 1801. Individuals are counted at the smallest place where they lived, usually the parish or township. The census population counts have been digitized for each 'census year' from 1801 to 1911. Wrigley and Satchell (2011) performed the initial digitization and linking across time, which was carried forward by the Integrated Census Microdata project or I-CeM (Schurer and Higgs 2014). Male occupational shares for agriculture, mining, secondary, tertiary, and an unspecified sector

[21] Another revealing statistic is that there were 0.65 railway journey per head of population in 1841; 20 in 1881 and 32 in 1911. See Mitchell, *British Historical Statistics*, pp. 545-7.

have also been digitized at the smallest census place from 1851 to 1911 by I-CeM. Currently, 1851 and 1881 provide the best and most consistent occupational data for our analysis.

To address boundary changes, we have created 9764 consistent spatial 'units' between 1801 and 1891 linked with census population and male occupation data. See appendix A.1 for a detailed description. A few have missing variables and so our sample size is 9489. Note that spatial units in our data are contained within 55 counties, which were an important administrative unit of local government. The exception are units associated with metropolitan London, which we treat as four 'counties,' including south, west, east, and central London.

We also associate each unit with a center to calculate distance variables in GIS. The center corresponds to a town marketplace, if the unit had a town within its boundary at some point between 1600 and 1850.[22] If there was no town, the centroid is used, which arguably makes sense for a rural unit without a marketplace. Regardless, little error is introduced by using the town market or centroid since our units are only 15 square km on average.

Our railway data includes GIS shapefiles for railway lines and stations in every census year starting in 1831. The rail networks and stations are created using accurate historical maps.[23] From this we create two measures for access to railway stations: (1) an indicator if there was an open station within the boundaries of the unit in a particular year and (2) the distance from the center of each unit to its nearest station in a particular year.[24]

In addition to railways we create a rich set of variables on 'first and second nature geography'. First nature includes an indicator for being on exposed coalfields, an indicator for being on the coast, elevation and slope in the unit, rainfall, temperature, wheat suitability, latitude, longitude, and the share of land in 10 different soil types. Coastal is identified using an

[22] Satchell, Potter, Shaw-Taylor, Bogart (2017) provide a dataset on 1746 towns and their centers, which are based on marketplace, or other markers, like parish, if there was no marketplace. We should stress that some towns in this data were very small and did not have most urban characteristics, therefore they are best described as 'candidate towns.' 746 of our units have at least one candidate town in them. If there is a single candidate town, we choose its center. If there are multiple, the town center with the largest 1801 population is used.
[23] They are derived from derived from a railway atlas by Cobb (2005). See Martí-Henneberg, Satchell, You, Shaw-Taylor, and Wrigley (2017) created the GIS of England, Wales and Scotland railway stations 1807-1994.
[24] Note it was rare for stations to close in the nineteenth century (Simmons 1986, p. 325). But it did happen, which means a few units get more distant from stations.

intersection of the seacoast with unit boundaries. Elevation and slope are calculated in GIS (see appendix A.3). The wheat suitability index and annual rainfall and temperature (both averaged from 1961 to 1990) come from FAO.[25] The soils data comes from Cranfield University.[26] Satchell and Shaw-Taylor (2013) identify those areas with exposed coal bearing strata (i.e. not overlain by younger rocks). Exposed coalfields were more easily exploited compared to concealed coal (see appendix A.4). Variables for second-nature geography include distance to one of the ten largest cities in 1801[27], log population density in 1801, distance to turnpike roads in 1800, distance to inland waterways in 1800, and distance to ports in 1780. The last three are calculated using detailed pre-rail infrastructure data. [28]

Summary statistics for most variables are shown in table 2. Variables are divided into groups to make it clear when they enter regressions later. There are several features to note. First, despite the total English and Welsh population increasing between 1851 and 1891, the average difference in log 1891 and 1851 population across our units was negative. This is consistent with the share of the population in the top 5% of units increasing from 0.56 in 1851 to 0.69 in 1891. Second, the average difference in the 1881 and 1851 share of male secondary occupations was slightly negative, despite the national trend to slightly higher secondary shares. Like population, secondary occupations became more concentrated in top units.

Table 2: Summary statistics

| Variable | Obs. | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| Population and occupation variables | | | | | |
| Diff. Ln. 1831 and 1801 population | 9489 | 0.268 | 0.247 | -1.800 | 3.126 |

[25] See the Global Agro-Ecological Zones data at http://www.fao.org/nr/gaez/about-data-portal/agricultural-suitability-and-potential-yields/en/. We selected low input and rain fed for wheat suitability.
[26] Soils data (c) Cranfield University (NSRI) 2017 used with permission. The 10 soil categories are based on Avery (1980) and Clayden and Hollis (1985). They include (1) Raw gley, (2) Lithomorphic, (3) Pelosols, (4) Brown, (5) Podzolic, (6) Surface-water gley, (7), Ground-water gley, (8) Man made, (9) peat soils, and (10) other. See http://www.landis.org.uk/downloads/classification.cfm#Clayden_and_Hollis. Brown soil is the most common and serves as the comparison group in the regression analysis.
[27] The ten largest cities are London, Manchester, Birmingham, Liverpool, Leeds, Bristol, Newcastle, Plymouth, Portsmouth, and Sheffield (near Nottingham)
[28] Rosevear, Satchell, Bogart, Shaw Taylor, Aidt, and Leon (2017) created a GIS of turnpike roads, Satchell, Shaw-Taylor, and Wrigley (2017) created a GIS of inland waterways, and Alvarez, Dunn, Bogart, Satchell, Shaw-Taylor (2017) created a GIS of ports.

| | | | | | |
|---|---|---|---|---|---|
| Diff. Ln. 1891 and 1851 population | 9489 | -0.023 | 0.468 | -3.388 | 4.599 |
| Ln pop density 1851 | 9489 | 4.242 | 1.367 | 0.808 | 11.625 |
| Diff. 1881 and 1851 male agriculture share | 9,488 | -0.067 | 0.153 | -0.820 | 0.928 |
| Diff. 1881 and 1851 male secondary share | 9,489 | -0.007 | 0.072 | -0.707 | 0.639 |
| Diff. 1881 and 1851 male tertiary share | 9,489 | 0.045 | 0.092 | -0.700 | 0.806 |
| Rail variables | | | | | |
| At least one Station in unit by 1851 | 9489 | 0.107 | 0.309 | 0 | 1 |
| At least one Station in unit by 1891 | 9489 | 0.276 | 0.447 | 0 | 1 |
| Has LCP in unit | 9489 | 0.229 | 0.421 | 0 | 1 |
| Has stage coaching inn by 1802 | 9489 | 0.079 | 0.269 | 0 | 1 |
| Has LCP & stage coaching inn by 1802 | 9489 | 0.031 | 0.174 | 0 | 1 |
| First-nature controls | | | | | |
| Indicator exposed coal | 9489 | 0.080 | 0.271 | 0 | 1 |
| Indicator coastal unit | 9489 | 0.147 | 0.355 | 0 | 1 |
| Elevation | 9489 | 89.72 | 74.02 | -1.243 | 524.3 |
| Average elevation slope within unit | 9489 | 4.767 | 3.615 | 0.484 | 37.42 |
| SD elevation slope within unit | 9489 | 3.432 | 2.717 | 0 | 23.17 |
| Rainfall in millimeters | 9484 | 755.7 | 191.7 | 555 | 1424 |
| Temperature index | 9484 | 8.958 | 0.658 | 5.5 | 10 |
| Wheat suitability (low input level rain-fed) | 9484 | 2188.1 | 273.25 | 272 | 2503 |
| Latitude | 9484 | 259871 | 115236 | 13522 | 652900 |
| Longitude | 9484 | 443389 | 112073 | 136232 | 654954 |
| Land area in sq. km. | 9484 | 15.63 | 22.18 | 0.003 | 499.8 |
| Perc. of land with Raw gley soil | 9489 | 0.084 | 1.327 | 0 | 76.49 |
| Perc. of land with Lithomorphic soil | 9489 | 8.615 | 19.83 | 0 | 100 |
| Perc. of land with Pelosols soil | 9489 | 8.203 | 20.63 | 0 | 100 |
| Perc. of land with Podzolic soil | 9489 | 4.624 | 14.32 | 0 | 99.56 |
| Perc. of land with Surface-water gley soil | 9489 | 24.63 | 29.46 | 0 | 100 |
| Perc. of land with Ground-water gley soil | 9489 | 10.187 | 20.11 | 0 | 100 |
| Perc. of land with Man made soil | 9489 | 0.363 | 3.262 | 0 | 94.99 |
| Perc. of land with Peat soil | 9489 | 1.187 | 5.279 | 0 | 91.44 |
| Perc. of other soil | 9489 | 0.535 | 1.966 | 0 | 65.15 |
| Second nature controls | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Ln 1801 population per sq. km | 9489 | 3.877 | 1.310 | 0.483 | 11.43 |
| Distance to inland waterway in 1800 in km | 9489 | 8.121 | 7.063 | 0.006 | 48.67 |
| Distance to turnpike road in 1800 in km | 9489 | 2.431 | 3.185 | 0.00 | 27.95 |
| Distance to port in 1780 in km | 9489 | 33.39 | 22.33 | 0.078 | 99.71 |
| Distance to top 10 city in 1801 in km | 9487 | 68.29 | 38.69 | 0 | 184.14 |

Sources: see text.

The summary data also inform how station access increased and distance to station fell with time. In 1851 10.7% of units had at least one station open, rising to 27.6% in 1891. In 1851 the average unit center was 6.9 km from a station. In 1881 the average was only 3.8 km.

As a preview of our main result, a two-sided t-test shows that the difference in log 1891 and log 1851 population is 0.433 higher for units with a rail station open by 1851 versus all other units (p-value 0.00). However, a similar t-test shows that the difference in log 1831 and log 1801 population is 0.142 higher for units with a rail station open by 1851 versus all other units (p-value 0.00). Endogeneity is clearly a concern in this setting as the 1851 railway station indicator is correlated with population growth in the pre-railway era. Our methodology will address this using an instrumental variables strategy.

## 4. Methodology

We employ the commonly used 'changes-on-levels' specification in urban economics. As explained by Duranton and Puga (2014), it analyzes infrastructure levels and their effects on future population changes assuming a gradual adjustment process. Our baseline specification is a cross-section growth equation like (1)

$$\Delta Lnpop_{ij,1891-1851} = \beta * Station1851_i + \gamma \cdot x_i + c_j + \varepsilon_{ij} \qquad (1)$$

where the dependent variable $\Delta Lnpop_{ij,1891-1851}$ is the difference in log 1891 and log 1851 population for unit i in county j. The main variable, $station1851_i$, is an indicator that equals 1 if unit i has at least one open station within its boundary by 1851 and zero otherwise. In other words, the control group is all units with no open station in 1851. The idea is that rail transport services were so much cheaper or faster that some industrial and commercial firms had to be very near stations to be competitive. On the workers side, some had to live very near stations

because of jobs, and because non-rail commuting costs were very high. Due to positive net-migration, having a railway station in a unit is predicted to cause its population to grow more than in units without railway access all else equal.

The control vector $x_i$ always includes first nature characteristics and the natural log of unit population density in 1851, 1841, and 1831 to capture effects of base year levels and prior trends. In preferred specifications, second nature characteristics and 59 county fixed effects $c_j$ are added as controls. The standard errors are always clustered on counties.

The instrument for $station1851_i$ is an indicator for having a least cost path (LCP) pass through the unit. The LCP is created using historical construction cost information combined with elevation data as explained in the next section. We also introduce a second instrument, an indicator for having coaching inns by 1802 interacted with the LCP indicator. It is designed to capture where stations were most likely along the route of the LCP.

We focus on the effect of stations in the year 1851 for several reasons. First, we want to be comparable to previous studies, which estimate effects of 'first-wave' rail construction on population change over the next 20 to 50 years.[29] Second, the railway mania took place in the mid-1840s, and it led to the opening of the main trunk lines shortly after. Therefore, having a station in 1851 provides a measure of access to a new network connecting most large cities in the early nineteenth century. There is a potential concern that the effects of railway building after 1851 also affected growth. We address this by checking whether dating stations later in 1856, 1861, 1866, or 1871 affects the estimates.

Our analysis also uses a 'changes-on-changes' specification to study the longer term.

$$\Delta Lnpop_{ij,1891-1821} = \beta * \Delta\, station_{i,1891-1821} + \pi x_i + c_j + \varepsilon_{it} \qquad (2)$$

In equation (2) the dependent variable is the difference in log 1891 and log 1821 population. On the right-hand side $\Delta\, station_{i,1891-1821}$ is the difference in the indicator for stations in 1891

[29] For example, Hornung (2015) estimates the effect of having railway stations open by 1848 on city population growth from 1849 to 1871 in Prussia. Berger and Enflo (2017) use a panel version of (1) to estimate the effects of having a railway line in 1870 on parish population growth from 1850 to 1900 in Sweden.

and 1821. Since the station indicator is zero in 1821 for all units, $\Delta\, station_{i,1891-1821}$ is simply an indicator equal to 1 if a unit had an open station in 1891. The variable $x_i$ includes first nature controls and second nature controls. County fixed effects are also added as in our preferred changes-on-levels specification. The LCP is again used as an instrument for 1891 station access.

One of our key extensions estimates heterogenous effects of station access drawing on the idea of agglomeration. When railways stations arrived in low population density units, they brought increased competition from more productive units. The greater competition resulted in employment losses and offset some of the positive net-migration effects from getting stations. The prediction is that station effects on population growth will be smaller, and perhaps zero, for low density units. We estimate how much smaller using an interaction between $Station1851_i$ and several variables capturing 1801 population density.

Another extension estimates population displacement effects, which also build on agglomeration. Economies of scale create advantages to being very close to large centers, while being far away provides protection from competition. Locations in-between get neither and cannot sustain as many firms and hence population (Fujita, Krugman, and Venables 2001). The hypothesis is that beyond some station distance threshold, population growth is lower in units closer to stations than far from stations. The size and range of the 'displacement zone' is estimated with a modified version of equation (1) using station distance-bins.

Finally, we also study occupational change using the difference in 1881 and 1851 male agricultural, secondary, or tertiary employment shares as the dependent variables in equation (1). The idea is that land-intensive economic sectors should become less profitable relative to labor-intensive sectors as population grows. Therefore, the occupational share in agriculture is expected to decline in units with 1851 stations. If so, the occupational share must necessarily rise in secondary, tertiary, or both. It will depend on which sector benefits more from density.

## 5. The Least cost path and its properties

The key idea behind the LCP is that it selects "inconsequential units", which attracted railway lines only because of their favorable location along a route connecting large towns.[30] The first applications in the railways literature used straight lines to connect endpoints (e.g. Atack, Bateman, Haines and Margo 2010), but subsequent studies use slope and geographic impediments to create the LCP (e.g. Berger 2019). We build on this literature and create an LCP that fits our context.

The first step is to select town-pairs which are candidates to be connected by early railways. We start with all English and Welsh towns having a population greater than 5000 in 1801.[31] Their larger size meant they were likely to get at least one railway line connecting them with another town above 5000. But not all candidates will be connected. A profit-seeking company would see little value in building a railway to connect distant towns of moderate size. A simple gravity equation is used to approximate the relative value of connecting town pairs i and j. $G_{ij} = (pop_i * pop_j)/dist_{ij}$, where $dist_{ij}$ is the straight line distance and $pop_i$ is population. We ordered $G_{ij}$ from largest to smallest and connect all pairs with a value greater than a threshold defined momentarily.

The second step is to identify routes connecting candidate town-pairs. Railway companies were assumed to minimize the construction costs considering distance and elevation slope. We use construction cost data for railways in the 1830s and early 1840s and measure the distance of these lines and total elevation changes between the two main towns at the ends. The construction cost is then regressed on the distance and the elevation change (the details are in appendix A.2). Based on this analysis, we find a baseline construction cost per km when the slope is zero and for every 1% increase in slope the construction cost rises by three times the baseline (cost per km=1+3*slope%). This formula and GIS tools identify the least cost path connecting each candidate town pair.

The third step is to identify which least cost paths are included in the final LCP network. We start with the candidate town pair having the largest gravitational value $G_{ij}$ and include its

[30] Redding and Turner (2014) call this the 'inconsequential places' approach. See Chandra and Thompson (2000), Michaels (2008), Faber (2014), and Lipscombe et. al. (2013) for early applications.
[31] The town population data come from Law (1967) and Robson (2006).

path. Second, we add the path associated with the candidate town pair having the second largest $G_{ij}$. If the two routes are close to one another duplicate sections are combined. We continue in the same manner adding paths until the total LCP network size equals the size of the 1851 rail network. All town points selected in the final LCP network are labelled 'LCP nodes'. Our sample will be restricted based on these nodes as explained below.

The LCP network and actual 1851 railway network are shown in figure 3. The overlap is close in many cases and there is a 0.323 correlation between an indicator for having railway lines pass through a unit in 1851 and an indicator for having the LCP pass.[32] Most importantly, there is 0.251 correlation between the indicator for having the LCP and the indicator for having stations by 1851. The reason is that stations were so numerous along railway lines in England and Wales. On average there was one station for every 5.9 km of railway line in 1851.

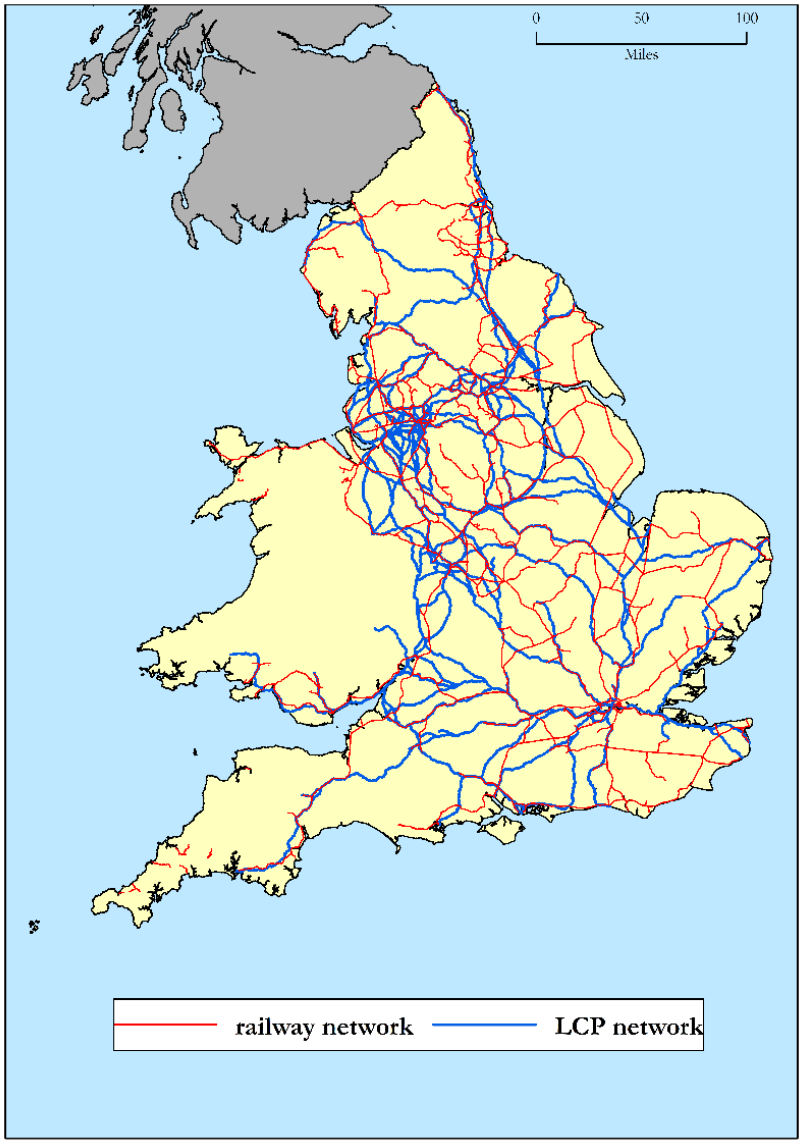
We have an additional variable to instrument for stations. Building on the idea that railway companies often placed them near nodes of the pre-existing network, we use an indicator for having a coaching inn in 1802. This comes from *Cary's New Itinerary*, which was a book for travelers identifying routes and inns to rest. There were 1228 inns throughout England and Wales by 1802 and these have been digitized and linked to GIS.[33] We interact inns in 1802 with the indicator for having the LCP in a unit (see table 2 for summary statistics). There is a 0.228 correlation between this interaction variable and the 1851 station indicator.[34]

Our exclusion restriction is more plausible if LCP instruments are not statistically related to population growth before railways. To show this, table 3 reports estimates from specifications where the dependent variable is the difference in log 1831 and log 1801 population and controls include the log 1801 population density and first nature variables. Second-nature and county fixed effects are included in some specifications. The standard errors are clustered on counties and the sample includes units whose center is more than 7 km from

[32] Note that railway lines built after 1851 are close to the LCP too, but the overlap is weaker. For example, there is a 0.279 correlation between having a railway line in 1861 and having the LCP.
[33] We thank Alan Rosevear for digitizing coaching inns from Cary.
[34] We also tried an instrument for the length of LCP divided by land area. But once we condition on having the LCP, this variable did not predict having an 1851 station.

an LCP node for reasons explained momentarily. The estimates in table 3 show that LCP variables are not significantly associated with the difference in log 1831 and 1801 population regardless of which controls are added. In appendix table A.5.1 we also show that a variable for the LCP interacted with coaching inns by 1802 is insignificant as well.

Figure 3: The LCP network and 1851 rail network compared




Sources: see text.

Table 3: Effects of least cost path (LCP) on population growth in pre-railway era

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Estimator | OLS | OLS | OLS | OLS | OLS | OLS |
| Dependent variable: | | | Δ1831,1801 ln pop | | | |
| LCP in unit | -0.00140 | 0.0103 | 0.00542 | | | |
| | (0.00854) | (0.00789) | (0.00802) | | | |
| Log distance to LCP | | | | 0.00560 | -0.000461 | 0.00361 |
| | | | | (0.00377) | (0.00397) | (0.00413) |
| County FE? | N | Y | Y | N | Y | Y |
| Second Nature controls? | N | N | Y | N | N | Y |
| Observations | 8,337 | 8,337 | 8,337 | 8,337 | 8,337 | 8,337 |
| R-squared | 0.064 | 0.110 | 0.116 | 0.065 | 0.110 | 0.116 |

Notes: Standard errors in parentheses are clustered on counties. *** p<0.01, ** p<0.05, * p<0.1. All models include first nature variables and ln pop 1801 density as controls. For definitions of first and second nature variables see table 1. For definitions of county fixed effects see the text. All units less than 7 km from an LCP node are dropped.

The sample is restricted to units more than 7 km from LCP nodes because at shorter distances the instruments are positively associated with pre-railway population growth following specifications in table 3.[35] Dropping these units makes our exclusion restriction more credible. The sample restriction means losing 12% of observations and not studying the effects of stations close to major 1801 towns, which are the nodes. But importantly we still have observations in all 1801 population density deciles.[36]

## 6. Estimates for baseline specifications

Ordinary least squares (OLS) and instrumental variable (IV) estimates for the effects of having stations by 1851 on the difference in log 1891 and log 1851 population are shown in the top panel of table 4. The bottom panel shows the first stage coefficient for having an LCP. The Kleibergen-Paap F-stat is above 48, and so weak instruments are not a concern. The IV estimate for 1851 stations in (2), which includes all the controls, is 0.349. When second nature controls

[35] To illustrate, consider specifications similar to column (3) in table 3 where we add dummy variables for being 0 to 1 km from the LCP node, 1 to 2 km from the LCP node and so on up to 14 to 15 km from the LCP node and interactions between these 15 variables and the indicator for having the LCP in the unit. Briefly, they reveal that for some distances less than 7 km, having the LCP is significantly associated with higher population growth. No such effects are found for having an LCP at distances more than 7 km from the node. For details see figure A.5.1. in the appendix.

[36] In our restricted sample, 5.1%, 9.7 and 10.3% are in the 10th, 9th, and 8th deciles of 1801 population density.

are omitted the IV estimate for 1851 stations is larger at 0.473 (see column 4). We do not prefer this second IV estimate because two important second nature factors, distance to major 1801 cities and distance to an 1800 inland waterway, are correlated with the LCP and they affect population growth too. When county fixed effects are also omitted the IV estimates become even larger at 0.956 (column 6). This estimate is also not preferred because there is unobserved heterogeneity across counties.

Table 4: Estimates for effect of 1851 station on population growth from 1851 to 1891

| | Dependent var.: Δ1891,1851 Ln Pop | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Estimator type | OLS | IV | OLS | IV | OLS | IV |
| Station in unit by 1851 | 0.166*** | 0.349* | 0.178*** | 0.473** | 0.231*** | 0.956*** |
| | (0.0213) | (0.206) | (0.0211) | (0.197) | (0.0292) | (0.175) |
| County FEs? | Y | Y | Y | Y | N | N |
| Second Nature controls? | Y | Y | N | N | N | N |
| Kleibergen-Paap F stat | | 48.939 | | 57.289 | | 95.428 |
| Observations | 8,337 | 8,341 | 8,341 | 8,341 | 8,341 | 8,341 |
| R-squared | 0.304 | | 0.287 | | 0.193 | |

| | First stage for Station in unit by 1851 | | |
|---|---|---|---|
| | (7) | (8) | (9) |
| | OLS | OLS | OLS |
| LCP in unit | 0.0737*** | 0.0800*** | 0.101*** |
| | (0.0132) | (0.0130) | (0.0146) |
| County FE? | Y | Y | N |
| Second Nature controls? | Y | N | N |
| Observations | 8,337 | 8,341 | 8,341 |
| R-squared | 0.216 | 0.211 | 0.188 |

Notes: Standard errors in parentheses are clustered on counties. *** p<0.01, ** p<0.05, * p<0.1. All specifications include first nature variables and 1851, 1841, and 1831 ln pop density as controls. For definitions of second nature variables see table 1. County FEs are described in the text. All units less than 7 km from an LCP node are dropped.

Concerning magnitudes, our preferred station coefficient in column (2), 0.349, is equivalent to 0.75 standard deviations of the dependent variable or in annual growth terms an increase of 0.875%.[37] The effects of stations are also large compared to other variables. These are not reported to save space, but coefficients for having coal and being a coastal unit in (2) were both significant, but less than the station effect at 0.171 and 0.168, respectively. The coefficient on distance to the nearest top ten 1801 city in km is -0.0028. This is 1/124 of the IV

[37] Here we calculate the total growth as 100*(exp(0.349)-1)%. Over 80 years this is equal to 0.875%

station effect, meaning getting a railway station in 1851 was equivalent to moving a unit 124 km closer to a major city, or like moving a unit from the midlands of England to near London.

It is notable that the IV estimates in table 4 are consistently larger than OLS. There are at least two explanations. First, getting stations in 1851 was associated with units having 'worse' 1851 to 1891 growth prospects after accounting for other factors. Second, there may have been an especially large impact of getting stations when a unit had an LCP. Unfortunately, it is difficult to tell which explanation is more likely.

The estimates for stations reported in table 4 are robust to different samples and specifications. If we exclude units within 1, 3, or 5 km of LCP nodes, the IV estimate gets much larger, so our sample restriction implies an under-estimate for the effects of stations (see appendix table A.5.2). If we expand our restriction, dropping units within 8 km of LCP nodes, the IV coefficient for (2) becomes 0.392 (S.E.0.205), similar to our preferred estimate 0.349. We also run specifications with a second instrument, the LCP interacted with the indicator for having coaching inns by 1802. The IV coefficient for 1851 stations is slightly smaller at 0.300 (See appendix table A.5.3). In table 5 we show estimates using indicators for having at least one station open in 1856, 1861, 1866, or 1871. The IV estimates are again similar to 0.349, indicating the choice of using 1851 to date station access is not crucial. [38] A different robustness check uses propensity score matching with the 1851 station indicator as the treatment and the difference in log 1891 and 1851 population as the outcome. The results are almost identical to our IV estimates.[39]

Our 'change on change' specification gives an estimate for the change in population between 1821 and 1891 caused by the change in station access over the same 70-year period. The sample includes all units more than 7 km from an LCP node and the controls include first

[38] Moreover, if we estimate our preferred specification in table 4, but drop units that would get their first station after 1851, the IV estimate for 1851 station is 0.337 (S.E. 0.195), again very similar.

[39] The simplest specification matches on a single variable: the log of 1801 population density. The matched sample is balanced and yields a statistically significant treatment effect of 0.323 (S.E. 0.029), which is very similar to our 0.349 IV estimate in table 4 column 2. Unfortunately, we were unable to achieve balanced matching on many co-variates. But, if we match on all second nature variables or selected first nature variables, the treatment effects are similar. These results are available upon request.

nature variables, second nature variables and county fixed effects. Having the LCP is the instrument. To summarize, the IV coefficient for station access by 1891 is 0.605 (SE 0.274, see appendix table A.5.4). It implies that having a railway station increased population by 0.867% per year between 1821 and 1891.[40] Notice this is nearly identical to the estimated increase in annual population growth implied by the IV coefficient for 1851 stations in (2) table 4. Thus, 'changes-on-levels' and 'changes-on-changes' specifications imply the same conclusion.

Table 5: Estimates for effect of station on population growth from 1851 to 1891 using different opening dates

| | Dependent var.: Δ1891,1851 Ln Pop | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | OLS | IV | OLS | IV | OLS | IV | OLS | IV |
| Station by 1856 | 0.183*** | 0.376* | | | | | | |
| | (0.0205) | (0.222) | | | | | | |
| Station by 1861 | | | 0.198*** | 0.366* | | | | |
| | | | (0.0177) | (0.218) | | | | |
| Station by 1866 | | | | | 0.203*** | 0.356* | | |
| | | | | | (0.0145) | (0.209) | | |
| Station by 1871 | | | | | | | 0.214*** | 0.353* |
| | | | | | | | (0.0161) | (0.202) |
| Kleibergen-Paap F | | 37.643 | | 34.880 | | 36.280 | | 34.534 |
| Observations | 8,337 | 8,337 | 8,337 | 8,337 | 8,337 | 8,337 | 8,337 | 8,337 |
| R-squared | 0.309 | | 0.317 | | 0.321 | | 0.328 | |

Notes: Standard errors in parentheses are clustered on counties. *** p<0.01, ** p<0.05, * p<0.1. All specifications include 1851, 1841, and 1831 ln pop density, first nature characteristics, second nature characteristics, and county fixed effects as controls. All units less than 7 km from an LCP node are dropped.

Next, we analyze the effect of 1851 stations on changes in male occupational structure. The specifications are like table 4, except the difference in the share of 1881 and 1851 male agricultural, secondary, or tertiary occupations are the dependent variables. We also add controls for 1851 male shares in secondary, tertiary, mining, or unspecified occupations to condition on occupational structure when railways were beginning to open. An alternative specification studies the difference in log shares as the dependent variable, but this is not our preferred as it leads to dropping several hundred units with zero shares in some sectors. Nevertheless, the results are similar and reported in appendix table A.5.5.

[40] Here we calculate the total growth as 100*(exp(0.605)-1)%. Over 80 years this is equal to 0.867%

The coefficients in columns (1) and (2) of table 6 show that getting stations in 1851 led to a significant decline in male agricultural shares. The IV coefficient -0.124 is equivalent to -0.80 standard deviations in the dependent variable. With respect to the difference in log shares it represents a 33% decline. Columns (3) and (4) show that getting stations led to a significant increase in male secondary shares. The IV coefficient 0.066 is equivalent to 0.89 standard deviations, or a 36% increase. Columns (5) and (6) show smaller or less precise effects for tertiary shares. These estimates imply that having railway stations led to occupational change, reducing employment in land-intensive economic sectors and increased employment in labor-intensive sectors, mainly manufacturing. The estimated effects are large and will have implications for interpreting the aggregate impact of railways as explained in section 9.

Table 6: Estimates for effect of 1851 station on difference in male occupational shares 1881 and 1851

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Estimator | OLS | IV | OLS | IV | OLS | IV |
| Dependent variable: | Δ male agriculture occupational share | | Δ male secondary occupational share | | Δ male tertiary occupational share | |
| Station in unit by 1851 | -0.0422*** | -0.124** | 0.0114*** | 0.0667** | 0.0253*** | 0.0384 |
| | (0.00465) | (0.0612) | (0.00286) | (0.0339) | (0.00339) | (0.0447) |
| Kleibergen-Paap F stat | | 48.139 | | 48.139 | | 48.139 |
| Observations | 8,337 | 8,337 | 8,337 | 8,337 | 8,337 | 8,337 |
| R-squared | 0.393 | | 0.212 | | 0.341 | |

Notes: Standard errors in parentheses are clustered on counties. *** p<0.01, ** p<0.05, * p<0.1. All specifications include county fixed effects, first nature variables, second nature variables, 1851, 1841, and 1831 ln pop density as controls, and 1851 male shares in agricultural, secondary, tertiary, mining, or unspecified occupations. For definitions of first and second nature variables see table 1. The instrument for station in unit by 1851 is an indicator if unit has LCP in its boundaries. All units less than 7 km from an LCP node are dropped.

## 7. Heterogenous effects based on 1801 population

The effects of railway stations should be greater if unit population density was already high. In this section, we estimate the size of heterogenous effects depending on 1801 population density. Column (1) in table 7 shows estimates for the baseline OLS after adding an interaction between 1851 station and log 1801 population density. Column (2) shows IV estimates using the LCP dummy interacted with the log of 1801 density as the second instrument. The interaction terms are positive and significant in both, but the IV is preferred due to endogeneity

of stations. To interpret the IV coefficients, we predict that at the median density having an 1851 station led to 0.21 increase in log 1891 and 1851 population. At the 25th and 75th percentiles the increases were significantly different at 0.14 and 0.33.

Table 7: Heterogeneous effects of getting a station by 1851 on population growth from 1851 to 1891

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Estimator | OLS | IV | OLS | IV | IV |
| Dependent variable: | Dependent var.: Δ1891,1851 Ln Pop | | | | |
| Station by 1851 | -0.035 | -0.668 | 0.214*** | 0.555*** | 0.609*** |
| | (0.101) | (0.643) | (0.0248) | (0.195) | (0.227) |
| Ln 1801 pop density | -0.089** | -0.094** | -0.097** | -0.081** | -0.085** |
| | (0.037) | (0.038) | (0.037) | (0.037) | (0.037) |
| Station by 1851* | 0.051** | 0.250* | | | |
| Ln 1801 pop density | (0.024) | (0.132) | | | |
| Below 60th pct. pop den. 1801 | | | -0.029** | 0.010 | 0.005 |
| | | | (0.011) | (0.018) | (0.018) |
| Station by 1851* Below | | | -0.108** | -0.497*** | -0.494*** |
| 60th pct. pop den. 1801 | | | (0.0414) | (0.172) | (0.221) |
| Drop units with more than 1 station | N | N | N | N | Y |
| Kleibergen-Paap F stat | | 17.433 | | 19.433 | 12.588 |
| Observations | 8,377 | 8,377 | 8,337 | 8,337 | 8,172 |
| R-squared | 0.305 | | 0.307 | | |

Notes: Standard errors in parentheses are clustered on counties. *** p<0.01, ** p<0.05, * p<0.1. All specifications include county fixed effects, first nature variables, second nature variables, 1851, 1841, and 1831 ln pop density as controls. For definitions of first and second nature variables see table 1. All units less than 7 km from an LCP node are dropped. In (2) the instruments are the indicator for LCP and the indicator for LCP interacted with log 1801 density. In (4) and (5) we add the instrument indicator for LCP interacted with dummy for unit below 60th percentile 1801 population.

Agglomeration theory often stresses a density threshold (see Lafourcade and Thisse 2011), so an alternative specification uses an indicator variable for units below the 60th percentile of 1801 population density. Columns (3) and (4) in table 7 report estimates for those specifications using the LCP interacted with the indicator for the 60th percentile as the second instrument. The IV estimates imply the effect of stations was to increase the difference in log 1891 and 1851 population by 0.059 (0.555-0.497) for units below the 60th percentile, but it is

not statistically different from zero. In other words, units in the bottom 6 deciles did not experience significant gains from a railway station.

The heterogenous effects are robust to considering different samples and specifications. Column (5) in table 7 restricts the sample to units with zero stations or only 1 station in 1851. The coefficients are similar, indicating getting multiple stations in high density units cannot account for the differences in outcomes. Moreover, using alternative dates for getting stations, like 1856 or 1861, and different percentiles like the 50$^{th}$ or 70$^{th}$, gives similar results (see appendix table A.5.6).

The findings in this section are consistent with agglomeration forces, which implies that peripheral areas will lose to core areas when transport costs begin to decline from high levels. We should also stress that accounting for heterogenous effects changes the quantitative interpretation. For units in the top 40% of 1801 density, getting stations led to an additional increase in annual population growth of 1.4%. Units in the bottom 60% got only a 0.1% increase in annual growth. These differences led to further divergence in the nineteenth century.

## 8. Local population displacement effects

While being very close to stations increased population growth, it is possible that after some distance threshold growth was lower in units closer to stations compared to units farther from stations. Equation (3) is used to estimate the threshold distance at which population growth between 1851 and 1891 was lower.

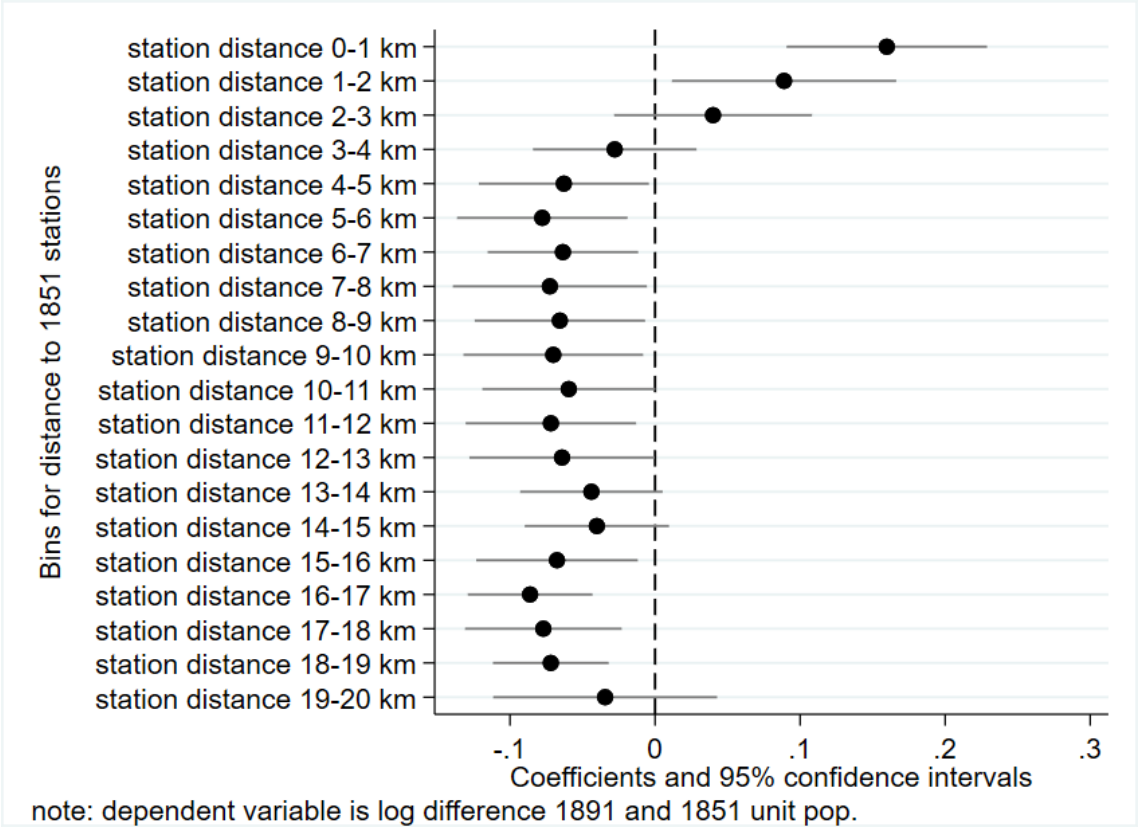
$$\Delta Lnpop_{ij,1891-1851} = \sum_{k=0}^{19} \beta_k Station1851distance[k, k+1] + \gamma \cdot x_i + c_j + \varepsilon_{ij} \qquad (3).$$

$Station1851distance[k, k+1]$ is an indicator variable for being between k and k+1 km from a station in 1851. These distance bins start with 0 to 1 km, 1 to 2 km and go on up to 19 to 20 km. The omitted comparison group includes units more than 20 km from an 1851 station. The same control variables are included as in our preferred specification in table 4. Units less than 7 km from an LCP node are also dropped as before.

The coefficients and their 95% confidence intervals for each distance bin are plotted in figure 4. Between 4 and 19 km the difference in log 1891 and 1851 population was 0.05 to 0.10

lower compared to units more than 20 km from stations. Thus, these estimates imply there was a 'displacement zone' starting not far from an 1851 station.

Figure 4. Effects of distance to 1851 stations on log difference in 1891 and 1851 population

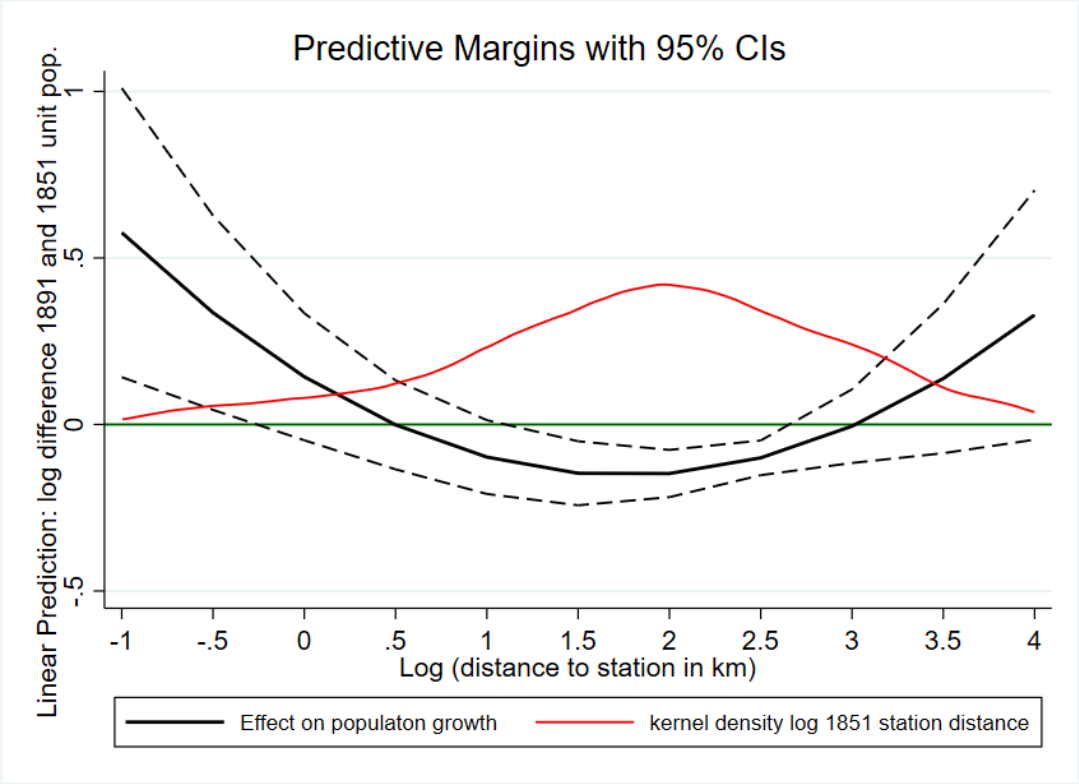



Notes: the coefficients are from specifications that include county fixed effects, first nature variables, second nature variables, 1851, 1841, and 1831 ln pop density as controls. For definitions of first and second nature variables see table 2. Standard errors are clustered on counties. All units less than 7 km from an LCP node are dropped.

One limitation of equation (3) is that station distance was selected, possibly biasing the estimates. Moreover, the large number of station distance bins makes it impractical to use instruments. To address this issue, we opt for a more parsimonious specification using log distance to 1851 stations and its square as the endogenous variables and log distance to LCP and its square as the instruments. The estimating equation is otherwise identical to that reported in column (2) table 4, with all controls. The IV estimates are summarized in Figure 5. Population growth from 1851 to 1891 is estimated to be positive and large less than 0.5 log

distance from stations or around 1.6 km. Population growth becomes negative and statistically different from zero between 1.25 and 2.7 log distance or 3 to 15 km. Comparing these effects with the kernel density (in red) shows that based on distance only a minority of units experienced positive growth effects from railways, around 15%. Many more, around 60% fell into the displacement zone and experienced population losses.

Figure 5. IV estimates for effect of log 1851 station distance on 1891 to 1851 population growth

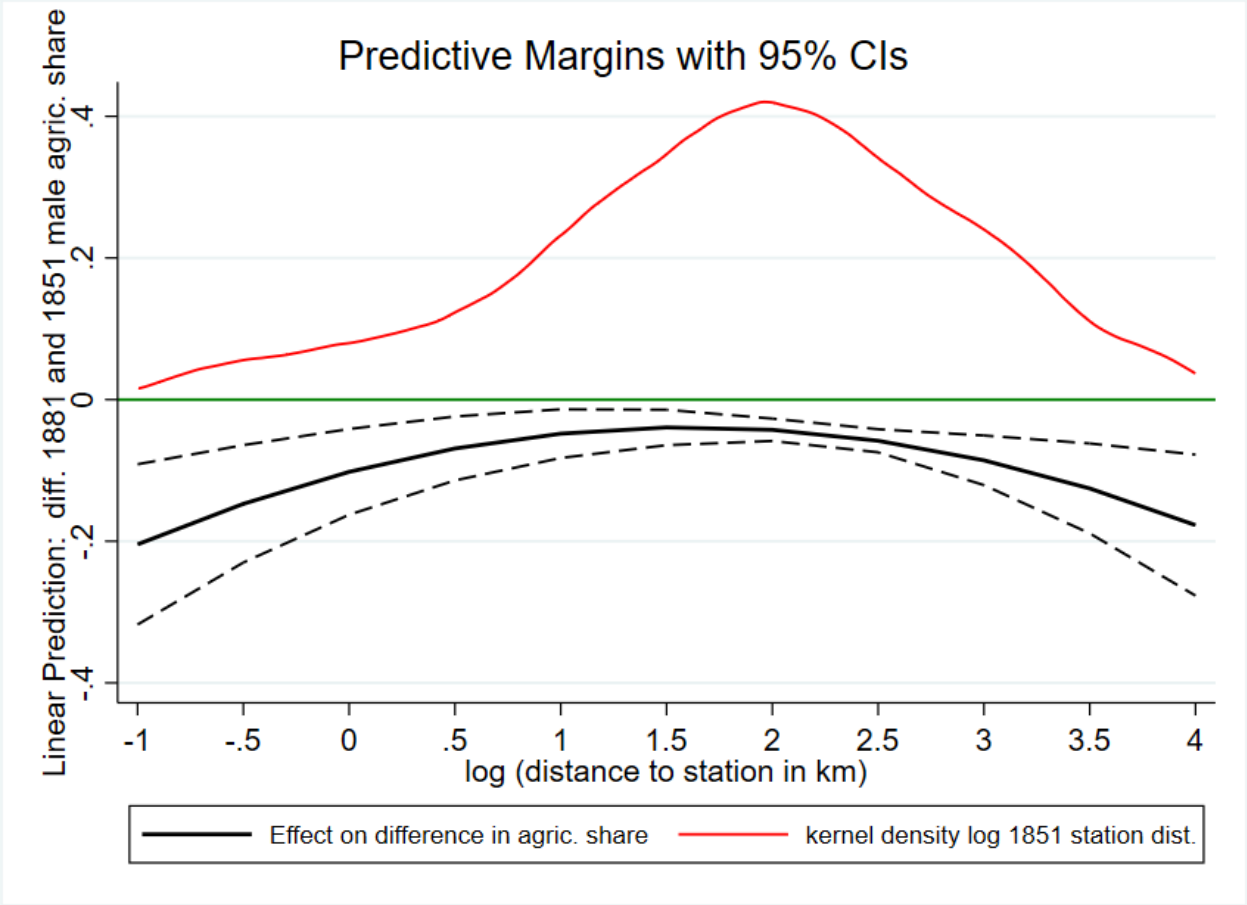



Sources: Author's calculation, see text.

A plot for the effect of log 1851 station distance on the change in the share of agricultural occupations between 1881 and 1851 is shown in figure 6. Like figure 5 the log distance to stations and its square is instrumented with log distance to the LCP and its square. The IV estimates imply agricultural shares declined significantly less for all units between a log distance of 1.5 and 2.5, or 5 to 12 km. Similar estimates for changes in the tertiary share reveal they increased significantly less between 5 and 12 km distance (see appendix figure A.5.2).

Together, these results suggest that in the displacement zone, the occupational structure became more land intensive and less oriented to services.

Figure 6. IV estimates for effect of log 1851 station distance on the difference in 1881 and 1851 agricultural shares




Sources: Author's calculation, see text.

## 9. Economy-wide population change and productivity

In this final section we show the population size and distribution would have been different if railways had not been invented. Then we quantify aggregate effects on labor income to further address how railways changed the whole of the British economy.

In the first step, total population levels in 1891 are predicted using the specification shown in column 4 of table 7 with the indicator for 1851 stations and its interaction with units

in the bottom 60th percentile of 1801 population density. We switch to using all 9489 units rather than the restricted sample more than 7 km from nodes. Our predictions are good. The correlation between our unit-level population prediction and its actual value in 1891 is 0.85.

In the second step, counter-factual population levels in 1891 are estimated if no unit had a station in 1851. The estimates imply that total 1891 population in England and Wales would have been about 22% lower if no units had stations by 1851. Also the share of 1891 population in the top 5% of units would have been 0.575 rather than 0.687. This estimate suggests railways can account for nearly all the change in population concentration. Recall that between 1851 and 1891 the actual population share in the top 5% rose from 0.564 to 0.687

Population changes caused by the railway depended on the 1801 population distribution. In the counterfactual most units in the bottom 90% would have had moderately *higher* population without railways. The median would have increased by 1.2%. In the top decile the effect of railways varied more, but they were larger. The median in the top 1801 decile would have 13.6% *lower* population in 1891 without railways.

In the third step, we estimate changes in occupational structure if there were no stations. Males in agricultural occupations in 1881 are predicted using the specification shown in column 2 of table 6. In cases where the predicted share is negative, the number of agricultural males is set to zero in a unit. Our estimates suggest there would have been 23.3% more males in agricultural occupations in 1881 if no units had railway stations in 1851. Also, there would have been 6.7% fewer males with secondary occupations and 9.8% fewer males with tertiary occupations. In other words, much less structural transformation.

What are the broader implications for national income? With 22% lower population, GDP would be significantly smaller. One could estimate about 11% lower assuming a labor share in income of 0.5 as is common in national accounting (Crafts and Mills 2004). The lower concentration of population in large units had implications for productivity too. We use Crafts and Leunig's (n.d.) estimate that the elasticity of labor productivity with respect to own population density was 0.025. We then calculate each unit change in productivity from the population change caused by no 1851 stations and then calculate the weighted average using

1891 population weights. This calculation implies that by shifting the population to lower density units, labor productivity in the English and Welsh economy would have fell by 0.58%. This effect is significant but not too large compared to the national income loss from total population change.

However, the productivity implications of occupational change are much larger. If we use Boyer and Hatton's (1997) estimate that rural unskilled wages were 27.2% lower in real terms than urban wages, then a 23.3% increase in agricultural male workers would represent a 23.3*(-0.272)=6.33% loss in male wage income. As males were at least half the labor force the total loss in wage income would have been above 3.5%. Broadly our estimate point to a large impact of railways on the English and Welsh economy mainly through population loss and less structural transformation.

## 10. Conclusion

In this paper, we study how railways led to population change and divergence in an economy whose urbanization rate increased dramatically from 1800 to 1900. We make use of detailed data on railway lines, stations, and population change in 9489 spatial units. Endogeneity is a major challenge in our context given that private companies built the network. To address this issue, we create a least cost path based on major nodal cities and the length of the 1851 rail network. Our instrumental variable estimates show that having a railway station in a unit by 1851 led to significantly higher population growth from 1851 to 1891 and shifted the male occupational structure away from agriculture to secondary. Moreover, in extensions, we estimate having stations increased population growth more if units had greater density in 1801. Also, there were population losses for units 4 to 15 km from stations, indicating a displacement effect. Overall, we find that railways reinforced the urban hierarchy of the early nineteenth century and contributed to further spatial divergence.

Our findings also have implications for the growth of the entire British economy. Through studying the effects on population concentration and occupational change, we find that railways' effects on national income are larger than is suggested by methods relying only on estimating consumer surplus from lower freight rates and higher speeds. Those methods

usually do not incorporate the effects of greater population concentration and occupational change, which we find to be quantitatively significant. Along with the population displacement effects, which were also quantitively significant, we conclude that railways had a very large impact on the English and Welsh economy in the second half of the nineteenth century.

**References**

Alvarez, Eduard, Xavi Franch, and Jordi Martí-Henneberg. "Evolution of the territorial coverage of the railway network and its influence on population growth: The case of England and Wales, 1871–1931." Historical Methods: A Journal of Quantitative and Interdisciplinary History 46.3 (2013): 175-191.

Alvarez, E, Dunn, O., Bogart, D., Satchell, M., Shaw-Taylor, L. , 'Ports of England and Wales, 1680-1911', 2017.
http://www.geog.cam.ac.uk/research/projects/occupations/datasets/documentation.html

Armstrong, John. The Vital Spark: The British Coastal Trade, 1700-1930. International Maritime Economic History Association, 2009.

Atack, Jeremy, Fred Bateman, Michael Haines, and Robert A. Margo. "Did railroads induce or follow economic growth?." Social Science History 34, no. 2 (2010): 171-197.

Atack, Jeremy, Michael R. Haines, and Robert A. Margo. Railroads and the Rise of the Factory: Evidence for the United States, 1850-70. No. w14410. National Bureau of Economic Research, 2008.

Atack, Jeremy, and Robert A. Margo. "The Impact of Access to Rail Transportation on Agricultural Improvement: The American Midwest as a Test Case, 1850-1860." Journal of Transport and Land Use 4.2 (2011).

Avery, Brian William. Soil classification for England and Wiles: higher categories. No. 631.44 A87. 1980.

Bairoch, Paul, and Gary Goertz. "Factors of urbanisation in the nineteenth century developed countries: a descriptive and econometric analysis." Urban Studies 23.4 (1986): 285-305.

Baldwin, Richard E., and Philippe Martin. "Agglomeration and regional growth." Handbook of regional and urban economics. Vol. 4. Elsevier, 2004. 2671-2711.

Baum-Snow, Nathaniel. "Did highways cause suburbanization?." The Quarterly Journal of Economics 122.2 (2007): 775-805.

Baum-Snow, N., Brandt, L., Henderson, J. V., Turner, M. A., & Zhang, Q. (2017). Roads, railroads, and decentralization of Chinese cities. Review of Economics and Statistics, 99(3), 435-448.

Beach, Brian, and W. Walker Hanlon. "Coal smoke and mortality in an early industrial economy." The Economic Journal 128.615 (2018): 2652-2675.

Berger, Thor. "Railroads and rural industrialization: Evidence from a historical policy experiment." Explorations in Economic History 74 (2019): 101277.

Berger, Thor, and Kerstin Enflo. "Locomotives of local growth: The short-and long-term impact of railroads in Sweden." Journal of Urban Economics (2015).

Bogart, Dan. “The Transport Revolution in Industrializing Britain,” in Floud, Roderick, Jane Humphries, and Paul Johnson, eds. The Cambridge Economic History of Modern Britain: Volume 1, Industrialisation, 1700–1870. Cambridge University Press, 2014.

Boyer, George R., and Timothy J. Hatton. "Migration and labour market integration in late nineteenth-century England and Wales." Economic History Review (1997): 697-734.

Büchel, Konstantin, and Stephan Kyburz. "Fast track to growth? Railway access, population growth and local displacement in 19th century Switzerland." Journal of economic geography 20.1 (2020): 155-195.

Bureau of Railway News and Statistics. Railway Statistics of the United States of America. Chicago: R. R. Donnelley and Sons, 1913 and 1916.

Buxton-Dunn, Oliver and Alvrez-Palau, Eduard and Bogart, Dan and Shaw-Taylor, Leigh (2020). *Historical light aids to navigation 1514-1911.* [Data Collection]. Colchester, Essex: UK Data Service. 10.5255/UKDA-SN-854172

Cameron, R. Concise Economic History of the World (New York: O.U.P., 1993) p. 193.

Campbell, Gareth, and John D. Turner. "Dispelling the Myth of the Naive Investor during the British Railway Mania, 1845–1846." Business History Review 86.01 (2012): 3-41.

Campbell, Gareth, and John D. Turner. "Managerial failure in mid-Victorian Britain?: Corporate expansion during a promotion boom." Business History 57.8 (2015): 1248-1276.

Cary, John. *Cary's New Itinerary: Or an Accurate Delineation of the Great Roads, Both Direct and Cross Throughout England and Wales; with Many of the Principal Roads in Scottland. From an Actual Admeasurement by---; Made by Command of His Majesty's Postmaster General, for Official Purposes. Under the Direction and Inspection of Thomas Hasker (etc.)*. Gosnell, 1802.

Casson, Mark. The world's first railway system: enterprise, competition, and regulation on the railway network in Victorian Britain. Oxford University Press, 2009.

Casson, Mark. "The determinants of local population growth: A study of Oxfordshire in the nineteenth century." Explorations in Economic History 50.1 (2013): 28-45.

Casson, Mark, A.E.M. Satchell, Leigh Shaw-Taylor, and E.A. Wrigley, “Railways and local population growth: Northampton and Rutland, 1801-1891” in Casson, Mark, and Nigar Hashimzade, eds. Large databases in economic history: research methods and case studies. Routledge, 2013.

Chandra, Amitabh, and Eric Thompson. "Does public infrastructure affect economic activity?: Evidence from the rural interstate highway system." Regional Science and Urban Economics 30.4 (2000): 457-490.

Church, Roy, Alan Hall, and John Kanefsky. History of the British Coal Industry: Volume 3: Victorian Pre-Eminence. Vol. 3. Oxford University Press, USA, 1986.

Clayden, Benjamin, and John Marcus Hollis. Criteria for differentiating soil series. No. Tech Monograph 17. 1985.

Cobb, M. H. "The Railways of Great Britain: A Historical Atlas at the Scale of 1 Inch to 1 Mile. 2 vols." *Shepperton: Allen* (2006).

Cormen, Thomas H., Charles E Leiserson, Ronald L Rivest and Clifford Stein: Introduction to Algorithms, Cambridge, MA, MIT Press (3rd ed., 2009) pp.695-6.

Crafts, Nicholas, and Timothy Leunig. " Transport improvements, agglomeration economies and city productivity: did commuter trains raise nineteenth century British wages?. working paper, n.d.

Crafts, Nicholas, and Terence C. Mills. "Was 19th century British growth steam-powered?: the climacteric revisited." Explorations in Economic History 41.2 (2004): 156-171.

Crafts, Nicholas, Timothy Leunig, and Abay Mulatu. "Were British railway companies well managed in the early twentieth century? 1." The Economic History Review 61.4 (2008): 842-866.

Day, Joseph. "The Process of Internal Migration in England and Wales, 1851-1911: Updating Ravenstein and the Step-Migration Hypothesis." *Comparative Population Studies* 44 (2019).

Desmet, Klaus, and Esteban Rossi-Hansberg. "Spatial development." The American Economic Review 104.4 (2014): 1211-1243.

Donaldson, Dave. "Railroads of the Raj: Estimating the impact of transportation infrastructure." American Economic Review 108.4-5 (2018): 899-934.

Donaldson, Dave, and Richard Hornbeck. "Railroads and American economic growth: A “market access” approach." The Quarterly Journal of Economics 131.2 (2016): 799-858.

Duranton, Gilles, and Matthew A. Turner. "Urban growth and transportation." The Review of Economic Studies 79.4 (2012): 1407-1440.

Duranton, G., & Puga, D. (2014). The growth of cities. In Handbook of economic growth (Vol. 2, pp. 781-853). Elsevier.

Faber, Benjamin. "Trade integration, market size, and industrialization: evidence from China's National Trunk Highway System." Review of Economic Studies 81.3 (2014): 1046-1070.

Fernihough, Alan, and Kevin Hjortshøj O'Rourke. Coal and the European industrial revolution. No. w19802. National Bureau of Economic Research, 2014.

Foreman-Peck, James. Railways and late Victorian economic growth. Cambridge University Press, 1991.

Fujita, Masahisa, Paul R. Krugman, and Anthony Venables. The spatial economy: Cities, regions, and international trade. MIT press, 2001.

Garcia-López, Miquel-Àngel, Adelheid Holl, and Elisabet Viladecans-Marsal. "Suburbanization and highways in Spain when the Romans and the Bourbons still shape its cities." Journal of Urban Economics 85 (2015): 52-67.

Ghani, Ejaz, Arti Grover Goswami, and William R. Kerr. "Highway to success: The impact of the Golden Quadrilateral project for the location and performance of Indian manufacturing." The Economic Journal 126.591 (2016): 317-357.

Gibbons, Stephen, Teemu Lyytikainen, Henry G. Overman, Rosa Sanchis-Guarner. "New road infrastructure: the effects on firms." Journal of Urban Economics 110 (2019): 35-50.

Gourvish, Terence Richard. Railways and the British economy, 1830-1914. Macmillan International Higher Education, 1980.

Gregory, Ian N., and Jordi Martí Henneberg. "The railways, urbanization, and local demography in England and Wales, 1825–1911." Social Science History 34.2 (2010): 199-228.

Hanlon, W. Walker. "Coal smoke, city growth, and the costs of the industrial revolution." The Economic Journal 130.626 (2020): 462-488.

Hsiang, Solomon M. "Temperatures and cyclones strongly associated with economic production in the Caribbean and Central America." Proceedings of the National Academy of sciences 107.35 (2010): 15367-15372.

Hawke, Gary Richard. Railways and economic growth in England and Wales, 1840-1870. Clarendon Press, 1970.

Heblich, Stephan, Stephen J. Redding, and Daniel M. Sturm. The Making of the Modern Metropolis: Evidence from London. No. w25047. National Bureau of Economic Research, 2018.

Hodgson, Charles. "The effect of transport infrastructure on the location of economic activity: Railroads and post offices in the American West." Journal of Urban Economics 104 (2018): 59-76.

Holl, Adelheid. "Highways and productivity in manufacturing firms." Journal of Urban Economics 93 (2016): 131-151.

Hornung, Erik. "Railroads and growth in Prussia." Journal of the European Economic Association 13.4 (2015): 699-736.

Jedwab, Remi, Edward Kerby, and Alexander Moradi. "History, path dependence and development: Evidence from colonial railroads, settlers and cities in Kenya." The Economic Journal (2015).

Jarvis A., H.I. Reuter, A. Nelson, E. Guevara (2008). Hole-filled seamless SRTM data V4, International Centre for Tropical Agriculture (CIAT), available from http://srtm.csi.cgiar.org.

Kellett, John R. The impact of railways on Victorian cities. Routledge, 2012.

Kitson, P. M., Shaw-Taylor, L., Wrigley, E. A., Davies, R. S., Newton, G., & Satchell, A. M. (2012). The creation of a 'census' of adult male employment for England and Wales for 1817. *Work in Progress. For details of the wider project see http://www. geog. cam. ac. uk/research/projects/occupations*.

Lafourcade, Miren, and Jacques-François Thisse. "New economic geography: the role of transport costs." A handbook of transport economics. Edward Elgar Publishing, 2011.

Law, Christopher M. "The growth of urban population in England and Wales, 1801-1911." Transactions of the Institute of British Geographers (1967): 125-143.

Leunig, Timothy. "Time is money: a re-assessment of the passenger social savings from Victorian British railways." The Journal of Economic History 66.3 (2006): 635-673.

Lipscomb, Molly, Mushfiq A. Mobarak, and Tania Barham. "Development effects of electrification: Evidence from the topographic placement of hydropower plants in Brazil." American Economic Journal: Applied Economics 5.2 (2013): 200-231.

Long, Jason. "Rural-urban migration and socioeconomic mobility in Victorian Britain." The Journal of Economic History 65.1 (2005): 1-35.

Maw, Peter. Transport and the Industrial City: Manchester and the Canal Age, 1750 1850. Manchester University Press, 2013.

Martí-Henneberg, J., Satchell, M., You, X., Shaw-Taylor, L., Wrigley E.A., 'England, Wales and Scotland railway stations 1807-1994 shapefile' (2017). http://www.geog.cam.ac.uk/research/projects/occupations/datasets/documentation.html

Michaels, Guy. "The effect of trade on the demand for skill: Evidence from the interstate highway system." The Review of Economics and Statistics 90.4 (2008): 683-701.

Michaels, Guy, Ferdinand Rauch, and Stephen J. Redding. "Urbanization and structural transformation." The Quarterly Journal of Economics 127.2 (2012): 535-586.

Mitchell, Brian R. Economic development of the British coal industry 1800-1914. Cambridge University Press, 1984.

Mitchell, Brian R. British Historical Statistics, Cambridge University Press, 1988.

Odlyzko, Andrew. "Collective hallucinations and inefficient markets: The British Railway Mania of the 1840s." University of Minnesota (2010).

Pascual Domènech, P. (1999). Los caminos de la era industrial: la construcción y financiación de la red ferroviaria catalana, 1843-1898 (Vol. 1). Edicions Universitat Barcelona.

Pogonyi, Csaba G. and Graham, Daniel J. and M. Carbo, Jose, Metros, Agglomeration and Firm Productivity. Evidence from London (March 11, 2019). Available at SSRN: https://ssrn.com/abstract=3350505 or http://dx.doi.org/10.2139/ssrn.3350505

Pooley, Colin, and Jean Turnbull. Migration and mobility in Britain since the eighteenth century. Routledge, 2005.

Pope, Alexander, and D. SWANN. "The pace and progress of port investment in England 1660–1830." Bulletin of Economic Research 12.1 (1960): 32-44

Poveda, G. (2003). El antiguo ferrocarril de Caldas. Dyna, 70 (139), pp. 1-10.

Purcar, Cristina. "Designing the space of transportation: railway planning theory in nineteenth and early twentieth century treatises." Planning Perspectives 22.3 (2007): 325-352.

Redding, Stephen J., and Matthew A. Turner. "Transportation costs and the spatial organization of economic activity." Handbook of regional and urban economics. Vol. 5. Elsevier, 2015. 1339-1398.

Redford, Arthur. *Labour migration in England, 1800-1850*. Manchester University Press, 1964.

Riley, S. J., S. D. Gloria, and R. Elliot (1999). A terrain Ruggedness Index that quantifies Topographic Heterogeneity, Intermountain Journal of Sciences, 5(2-4), 23-27.

Robson, Brian T. Urban growth: an approach. Vol. 9. Routledge, 2006.

Rosevear, A., Satchell, M., Bogart, D., Shaw Taylor, L., Aidt, T. and Leon, G., 'Turnpike roads of England and Wales, 1667-1892', 2017.
http://www.geog.cam.ac.uk/research/projects/occupations/datasets/documentation.html

Satchell, M. and Shaw-Taylor, L., 'Exposed coalfields of England and Wales' 2013.
http://www.geog.cam.ac.uk/research/projects/occupations/datasets/documentation.html

Satchell, M., Shaw-Taylor, L., Wrigley E.A., '1830 England and Wales navigable waterways shapefile' (2017).
http://www.geog.cam.ac.uk/research/projects/occupations/datasets/documentation.html

Satchell, M., Potter, E., Shaw-Taylor, L., Bogart, D., 'Candidate Towns of England and Wales, c.1563-1911', 2017.A description of the dataset can be found in M. Satchell, 'Candidate Towns of England and Wales, c.1563-1911 GIS shapefile'
http://www.geog.cam.ac.uk/research/projects/occupations/datasets/documentation.html

Shaw-Taylor, Leigh, and Xuesheng You. "The development of the railway network in Britain 1825-19111." 2020.

Schurer, K., Higgs, E. (2014). Integrated Census Microdata (I-CeM), 1851-1911. [data collection]. UK Data Service. SN: 7481, http://doi.org/10.5255/UKDA-SN-7481-1.

Schürer, K., and Joe Day. "Migration to London and the development of the north–south divide, 1851–1911." Social History 44.1 (2019): 26-56.

Shaw-Taylor, L. and Wrigley, E. A. "Occupational Structure and Population Change," in Floud, Roderick, Jane Humphries, and Paul Johnson, eds. The Cambridge Economic History of Modern Britain: Volume 1, Industrialisation, 1700–1870. Cambridge University Press, 2014.

Shaw-Taylor, Leigh, E.A. Wrigley, Peter Kitson, Ros Davies, Gill Newton and Max Satchell. The occupational structure of England and Wales c.1817-1881. Working Paper, 2010.

Simmons, Jack. The railway in town and country, 1830-1914. (1986).

Storeygard, Adam. "Farther on down the road: transport costs, trade and urban growth in sub-Saharan Africa." The Review of Economic Studies 83.3 (2016): 1263-1295.

Sugden, Keith, Sebastian Keibek and Leigh Shaw-Taylor. “Adam Smith revisited: coal and the location of the woollen manufacture in England before mechanization, c. 1500-1820”, CWPESH no. 33, 2018.

Tang, John P. "Railroad expansion and industrialization: evidence from Meiji Japan." The Journal of Economic History 74.03 (2014): 863-886.

Wallis, Patrick, Justin Colson, and David Chilosi. "Structural change and economic growth in the British economy before the Industrial Revolution, 1500–1800." *The journal of economic history* 78.3 (2018): 862-903.

Wellington, A.M. The Economic Theory of the Location of Railways: An Analysis of the Conditions Controlling the Laying Out of Railways to Effect the Most Judicious Expenditure of Capital. Ed. J. Wiley & sons, 1877.

Wrigley, Edward Anthony. Energy and the English industrial revolution. Cambridge University Press, 2010.

Wrigley, Edward Anthony, and Satchell, A.E.M.. The early English censuses. Oxford University Press, 2011.

## Appendix A.1: Linking population and occupational data across space

The English administrative units display highly inconsistent features. Several different hierarchal systems can coexist at the same time; different region can use different nomenclature; different systems can exist at different time slices; and boundaries of individual units within each system can change over time. Even though boundaries were never redrawn from scratch, different administrative system over time and boundary changes of individual units within any given systems over time mean that it would be difficult to carry out any analysis, either econometrically or cartographically, without having the data in a set of consistent geographical units.

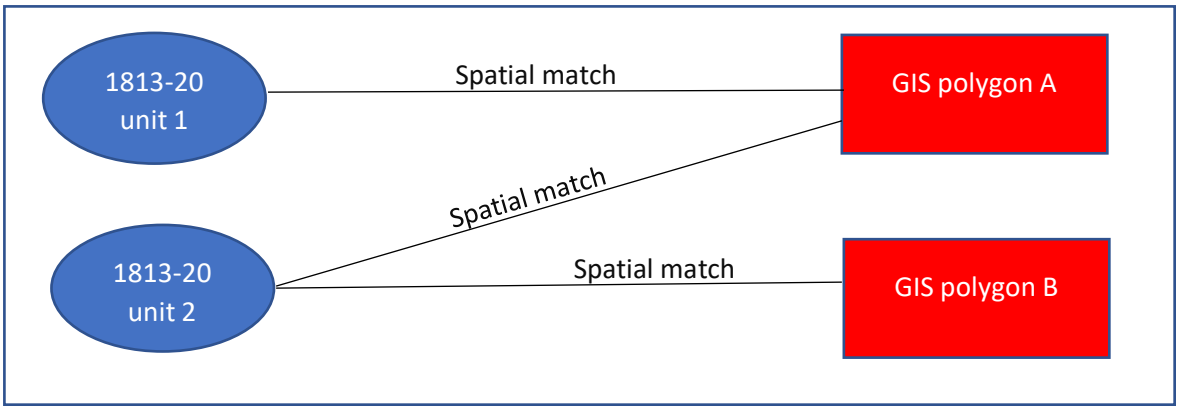
This problem becomes even more apparent drawing on evidence from several datasets at different slices: the baptism data between 1813 and 1820, the 1851 census data, the 1881 census data, and the population data between 1801 and 1891. Each of these datasets have data at different geographical unit. The name and the number of geographical units in each dataset are presented in the table below.

| | **Name of the geographical unit** | **Number of the geographical unit** |
|---|---|---|
| 1813-20 Baptism data | Ancient parish | 11,364 |
| 1851 census data | Civil parish | 16,397 |
| 1881 census data | Civil parish | 15,299 |
| 1801-91 population data | Continuous unit | 12,750 |

The method of creating a set of consistent geographical units based on the units in each dataset involves two steps. Firstly, we made spatial match between parish level Geographical

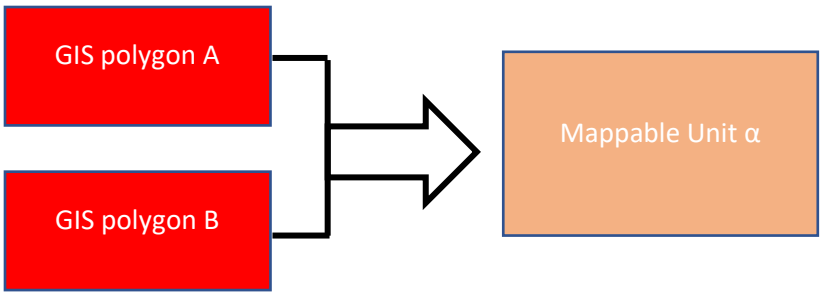
Information System (GIS) polygons and geographical unit from each dataset. The spatial match essentially made connections between the parish level GIS polygons and administrative units from each dataset through nominal linkage. The parish level GIS has c. 23,000 polygons. A separate note on the parish level GIS polygons can be found elsewhere. Part of spatial match process can be carried out automatically, but there are cases where spatial matches can not be made automatically and require manual linkage. Ms Gill Newton and Dr Max Satchell, both of the Cambridge Group for the History of Population and Social Structure (Cambridge Group), University of Cambridge, managed the process of spatial matching based on an approach suggested by Dr Peter Kitson, previously of the Cambridge Group. A number of students from the University of Cambridge also provided research assistance during the process. A brief account of the spatial match process can be found in Kitson, P., et al, 'The creation of a 'census' of adult male employment for England and Wales for 1817', http://www.econsoc.hist.cam.ac.uk/docs/CWPESH%20number%204%2017th%20December%202013,%20March%202012.pdf It should be noted that the nominal link between GIS polygons and administrative units from each dataset generated by the spatial match process cannot be used directly for mapping purpose. This is due to the fact that a particular GIS polygon can be linked to more than one administrative units from each given dataset. But the spatial match process is essential for the second step we need to create a set of consistent geographical units over time.

The second step is called Transitive Closure. Imagine the following situation using just 1813-20 baptism dataset as an example:

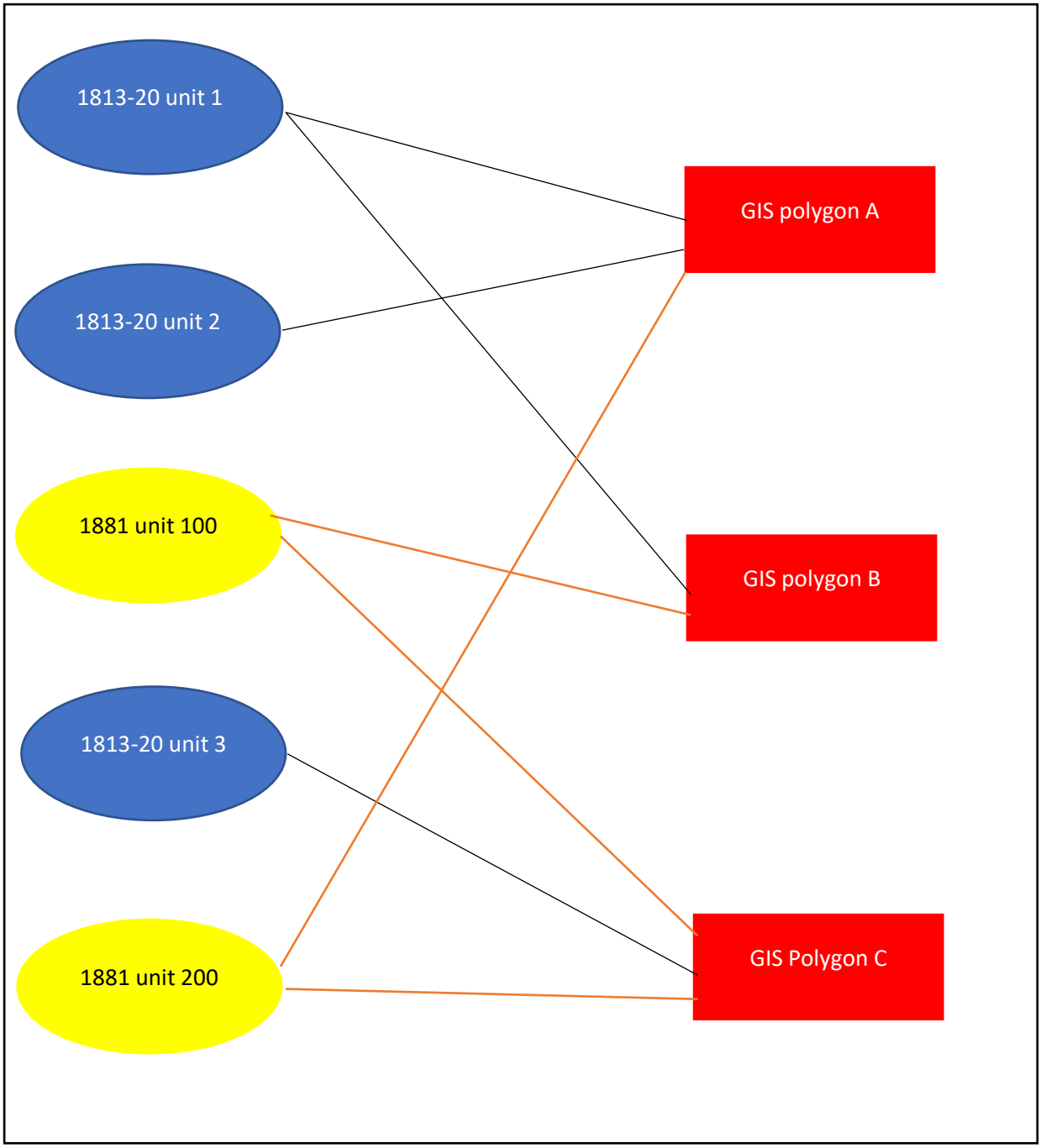



In this case, unit 1 from 1813-20 baptism dataset has a spatial match with the GIS polygon A, and polygon A only. And It does not have direct match with the GIS polygon B. But unit 2 from1813-20 baptism dataset has spatial matches with both GIS polygons A and B. Namely, part of the land enclosed by polygon A belonged to unit 1 with the other part belonging to unit 2. The problem is we do not know where exactly the divide within polygon A is:


So GIS polygon A is left undivided, and both polygon A and polygon B were grouped together to form a 'mappable unit', say mappable unit α, to present units 1 and 2:



The process presented above is the main function of Transitive Closure. When more datasets are added to the study, the situation becomes more complicated. But the basic idea remains the same. For example, imagine the following hypothetical situation:



If we are only dealing with 1813-20 baptism dataset, we can group polygons A and B together to form one mappable unit to represent units 1 and 2; and polygon C becomes a mappable unit on its own to represent unit 3. But once we add more datasets with different geographical units, in this case 1881 census data, we need to generate mappable units that are consistent across different datasets, i.e. over time as well. In this hypothetical case, Transitive Closure will group polygons A, B, and C together to form a single mappable unit. When dealing with 1813-20 baptism dataset, this mappable unit will draw data from units 1, 2 and 3. When dealing with 1881 census dataset, this mappable unit will draw data from units 100 and 200. In this way, the Transitive Closure process makes sure we are presenting and comparing observations from the same geographical units over time.

Transitive closure is a concept widely used in graph theory; for a formal definition and how to compute it, see for instance: Thomas H Cormen, Charles E Leiserson, Ronald L Rivest and Clifford Stein: Introduction to Algorithms, Cambridge, MA, MIT Press (3rd ed., 2009) pp.695-6. Ms Gill Newton, of the Cambridge Group, developed the Python code for Transitive Closure as part of the research project 'The occupational structure of Britain, 1379-1911' based at the Cambridge Group. Dr Xuesheng You, also of the Cambridge Group, implemented this code for this particular paper.

## Appendix A.2: The least cost path instrument

In this appendix, we describe how we identify the LCP connecting our nodes. The main criteria used to plan linear projects is usually the minimization of earth-moving works. Assuming that the track structure (composed by rails, sleepers and ballast) is equal for the entire length, it is in the track foundation where more differences can be observed. Thus, terrains with higher slopes require larger earth-moving and, in consequence, construction costs become higher (Pascual 1999, Poveda 2003, Purcar 2007). The power of traction of the locomotives and the potential adherence between wheels and rails could be the main reason. Besides, it is also important to highlight that having slopes over 2% might imply the necessity of building tunnels, cut-and-cover tunnels or even viaducts. The perpendicular slope was also crucial. During the construction of the track section, excavation and filling have to be balanced in order to minimize provisions, waste and transportation of land. Nowadays, bulldozers and trailers are used, but historically workers did it manually. It implied a direct linkage between construction cost, wages and availability of skilled laborers. In fact, it is commonly accepted in the literature that former railways were highly restricted by several factors. The quality of the soil, the necessity of construction tunnels and bridges or the interference with preexistences (building and land dispossession) were several. Longitudinal and perpendicular slope were the more significant ones and we focus on these below.

Slopes are determined using elevation data. Several DEM rasters have been analyzed in preliminary tests, but we finally chose the Shuttle Radar Topography Mission (SRTM) obtained in 90 meter measurements (3 arc-second). Although being a current raster data set, created in

2000 from a radar system on-board the Space Shuttle, the results offered in historical perspective should not differ much from the reality. The LCP tool calculates the route between an origin and a destination, minimizing the elevation difference (or cost in our case) in accumulative terms. The method developed was based on the ESRI Least-Cost-Path algorithm, although additional tasks were implemented to optimize the results and to offer different scenarios. The input data was the SRTM elevation raster, converted into slope. This conversion was necessary in order to input different construction costs.

The next step is to specify the relationship between construction costs and slope. One approach is to use the historical engineering literature. Wellington (1877) discusses elevation slope (i.e. gradients), distance, and operational costs of railways, but this is not ideal as we are interested in construction costs. We could not find an engineering text that specified the relationship between construction costs and slopes. As an alternative we use historical construction cost data. The following details our data and procedure.

A select committee on railways in 1844 published a table on the construction costs of 54 railways. See the Fifth report from the Select Committee on Railways; together with the minutes of evidence, appendix and index (BPP 1844 XI). The specific section with the data is appendix number 2, report to the lords of the committee of the privy council for trade on the statistics of British and Foreign railways, pp. 4-5. There were 45 with a clear origin and destination, to which we can measure total elevation change along the route (details are available). For these 45 railways we calculate the distance of the railway line in meters and the total elevation change (all meters of ascent and descent). We then ran the following regression of construction costs on distance in 100 meters and the elevation change in meters. This

regression produces unsatisfactory results, with total elevation change having a negative sign. We think the main reason is that the sample includes railways with London as an origin and destination. Land values in London were much higher than elsewhere and thus construction costs were higher there. Therefore, we omit railways with a London connection. We also think it is important to account for railways in mining areas as they were typically built to serve freight traffic rather than a mix with passenger.

Our extended model uses construction costs for 36 non-London railways. We regress construction costs on a distance in 100 meters, elevation change, and dummy for mining railways. The results imply that for every 100 meters of distance construction costs rise by £128.9 (st. err 45.27) and holding distance constant construction costs rise by £382.6 (st. err. 274.5) for every 1 meter increase in total elevation change. Construction costs for mining railways are £340,418 less (st. err. 179,815). For our LCP model we assume a non-mining railway, re-scale the figures into construction costs per 100 meters, and normalize so that costs per 100 meters are 1 at zero elevation change. The formula becomes: NormalizedCostper100meters=1+2.96*(ElevationChangeMeters/Distance100meters). The elevation change divided by distance can be considered as the slope in percent, in which case our formula becomes Cost=1+2.96*%slope. We think this is a reasonable approximation of the relationship between construction costs, distance, and elevation slope.

The LCP algorithm is implemented using ESRI python, using as initial variables the elevation slope raster, the reclassification table of construction costs, and the node origin-destination nodes. We implemented the least-cost-path function to obtain the LCP corridors.

These corridors were converted to lines, exported, merged and post-processed. Maps of our preferred LCP are shown in the text.

## Appendix A.3: Elevation, slope, and ruggedness variables

The aim of this appendix is to explain the creation of the elevation variables, including the original sources and method we followed to estimate them. There are several initiatives working on the provision of high-resolution elevation raster data across the world. The geographical coverage, the precision of the data and the treatment of urban surroundings concentrate the main differences between databases.

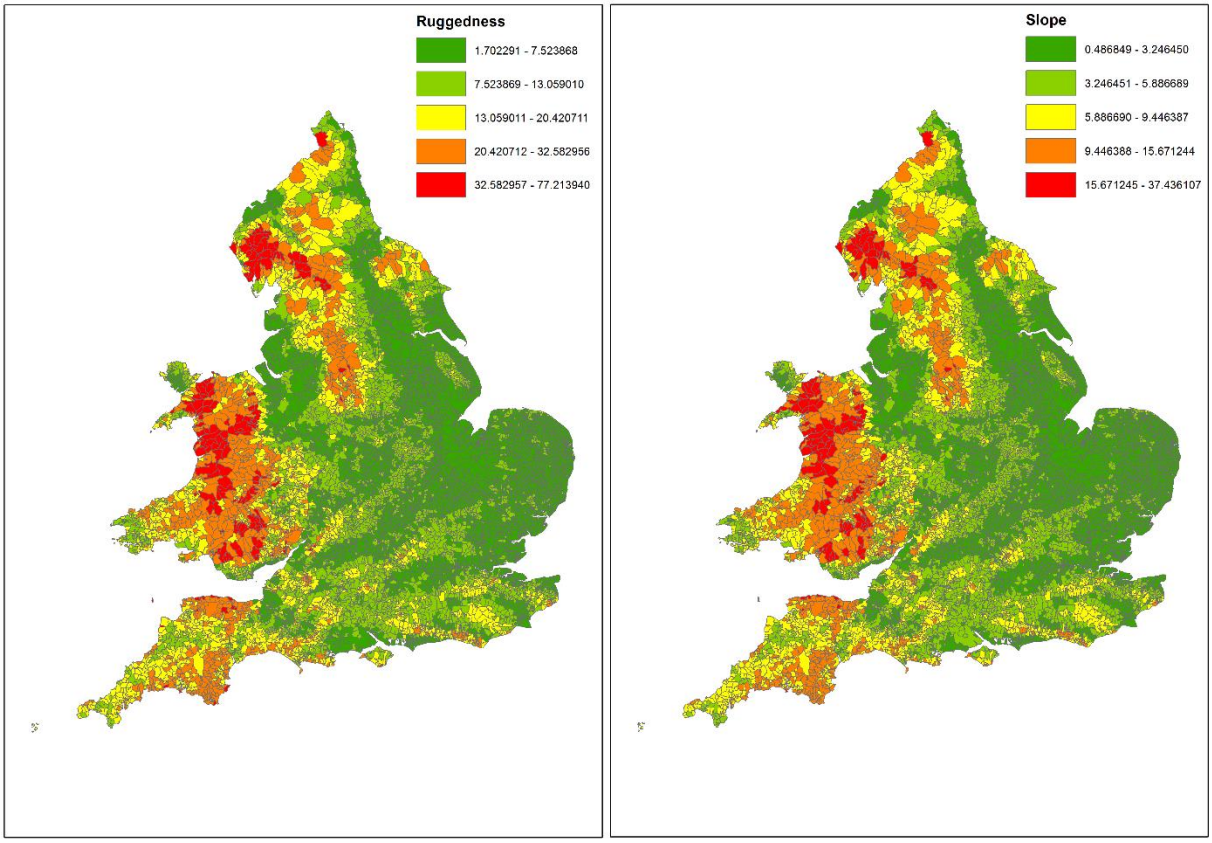
We obtained several elevation DEM rasters, preferably DTM , covering the entire England and Wales. In decreasing order in terms of accuracy, the most precise one database was LIDAR (5x5m.), Landmap Data set contained in the NEODC Landmap Archive (Centre for Environmental Data Archival). In second instance, we used EU-DEM (25x25m.) from the GMES RDA project, available in the EEA Geospatial Data Catalogue (European Environment Agency). The third dataset was the Shuttle Radar Topography Mission (SRTM 90x90m), created in 2000 from a radar system on-board the Space Shuttle Endeavor by the National Geospatial-Intelligence Agency (NGA) and NASA. And finally, we have also used GTOPO30 (1,000x1,000m) developed by a collaborative effort led by staff at the U.S. Geological Survey's Center for Earth Resources Observation and Science (EROS). All those sources have been created using satellite data, which means all of them are based in current data. The lack of historical sources of elevation data obligate us to use them. This simplification may be considered reasonable for rural places but it is more inconsistent in urban surroundings where the urbanization process

altered the original landscape. Even using DTM rasters, the construction of buildings and technical networks involved a severe change in the surface of the terrain. Several tests at a local scale were conducted with the different rasters in order to establish a balance between precision and operational time spend in the calculations. Total size of the files, time spend in different calculations and precision in relation to the finest data were some of the comparisons carried on. After these, we opted for SRTM90.

As stated in the text, the spatial units used as a basis for the present paper were civil parishes, comprising over 9000 continuous units. In this regard, we had to provide a method to obtain unique elevation variables for each unit, keeping the comparability across the country. We estimated six variables in total: elevation mean, elevation std, slope mean, slope std, ruggedness mean and ruggedness std. Before starting with the creation of the different variables, some work had to be done to prepare the data. In order to obtain fully coverage of England and Wales with SRTM data, we had to download 7 raster tiles. Those images were merged together, projected into the British National Grid and cut externally using the coastline in ArcGIS software.

Having the elevation raster of England and Wales, we proceed to calculate the first two variables: the elevation mean and its standard deviation. A python script was written to split the raster using the continuous units, to calculate the raster properties (mean and standard deviation) of all the cells in each sub-raster, and to aggregate the information obtained in a text file. These files were subsequently joined to the previous shapefile of civil parishes, offering the possibility to plot the results.

Appendix Figure 1: Slope and ruggedness measures




The second derivative of those results aimed to identify the variability of elevation between adjacent cells. In this regard, two methods were developed to measure this phenomenon: ruggedness and slope. Ruggedness is a measure of topographical heterogeneity defined by Riley et al (1999). In order to calculate the ruggedness index for each unit, a python script was written to convert each raster cell into a point keeping the elevation value, to select the adjacent values using a distance tool, to implement the stated equation to every single point, to spatially join the points to their spatial units and to calculate aggregated indicators (mean and standard deviation) per each continuous units.

In order to calculate the slope variable for each unit, a python script was written to convert the elevation into a slope raster, to split the raster using the continuous units, to calculate the raster properties (mean and standard deviation) of all the cells in each sub-raster, and to aggregate the information obtained in a text file. The obtained results for both ruggedness and slope are displayed at the end of this note. As the reader will appreciate, the scale of the indices is different (1 - 2 times) but the geographical pattern is rather similar. In this regard, we used for the paper those variables derived from slope measures because the time spend in calculations was rather lower.

## Appendix A.4: Exposed coal

The shapefile of exposed coalfields of England and Wales c. 1830 was created by Max Satchell using the Digital Geological Map Data of Great Britain 1: 625,000 bedrock produced by the British Geological Survey (BGS). Exposed coalfields can be defined as those sections of coalfields where coal-bearing strata are not concealed by geologically younger rocks. They may, however, be overlain by natural (and man-made) sediments of the Quaternary period where they would form overburden in the exposed coalfield. Quaternary deposits are often unconsolidated sediments comprising mixtures of clay, silt, sand, gravel, cobbles and boulders. Exposed coalfields are of major historical importance because they were places where coal seams crop out at or near the ground surface making coal easiest to both discover and mine. For more details

https://www.campop.geog.cam.ac.uk/research/occupations/datasets/catalogues/documentation/exposedcoalfieldsenglandandwales1830.pdf

## Appendix A.5: Additional results

Following from section 5, we show that a variable for the LCP interacted with coaching inns by 1802 is not significantly related to population growth from 1801 to 1831. See table A.5.1.

Table A.5.1: Effects of least cost path (LCP) and Coaching Inns on population growth in pre-railway era

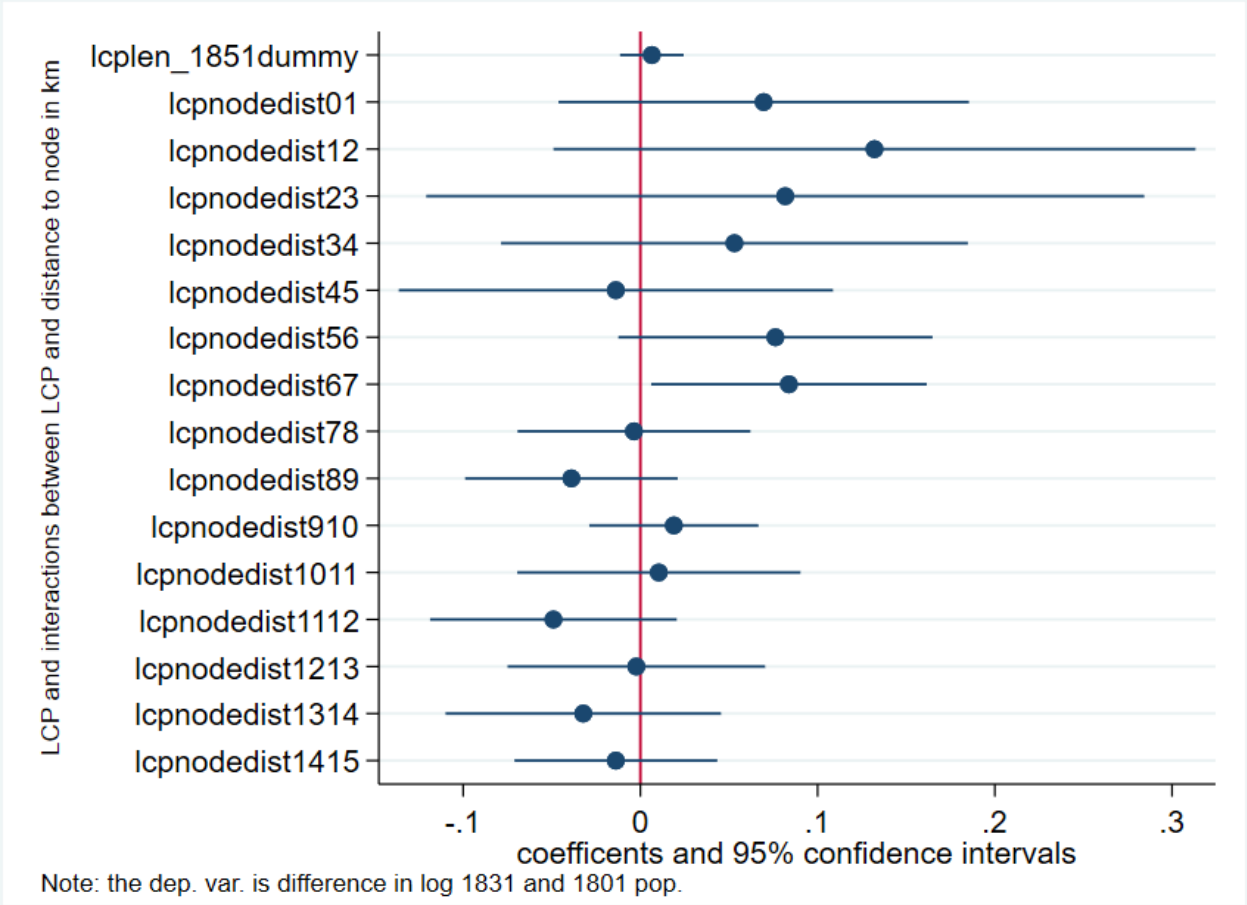
| | (1) | (2) | (3) |
|---|---|---|---|
| Estimator | OLS | OLS | OLS |
| Dependent variable: | Δ1831,1801 ln pop | | |
| LCP in unit | 0.0005 | 0.0117 | 0.0071 |
| | (0.0083) | (0.0076) | (0.0076) |
| LCP in unit* 1802 Coaching inn | -0.0319 | -0.0252 | -0.02371 |
| | (0.0224) | (0.0219) | (0.0221) |
| County FE? | N | Y | Y |
| Second Nature controls? | N | N | Y |
| Observations | 8,337 | 8,337 | 8,337 |
| R-squared | 0.072 | 0.110 | 0.1228 |

Notes: Standard errors in parentheses are clustered on counties. *** p<0.01, ** p<0.05, * p<0.1. All models include first nature variables and ln pop 1801 density as controls. For definitions of first and second nature variables see table 1. For definitions of county fixed effects see the text. All units less than 7 km from an LCP node are dropped.

Next, we show that the effects of the LCP on 1801 to 1831 population growth are sometimes positive when we do not restrict the sample to units less than 7 km from LCP nodes. To illustrate, consider specifications similar to column (3) in table 3 where we add dummy variables for being 0 to 1 km from the LCP node, 1 to 2 km from the LCP node and so on up to 14 to 15 km from the LCP node and interactions between these 15 variables and the indicator for having the LCP in the unit. The specifications also include first nature controls, second nature controls, and county fixed effects. The interactions between LCP and distances from

nodes are shown in figure A.5.1. Notice that in units less than 7 km from nodes the LCP is positively associated with population growth in the pre-railway era.

Figure A.5.1: effects of LCP and distance to LCP nodes on population growth from 1801 to 1831




Next, we show our main estimates change when excluding units at 1, 3, and 5 km from LCP nodes. We find the IV estimate gets much larger, so our sample restriction implies an under-estimate for the effects of stations. See table A.5.2.

Table A.5.2: Estimates for effect of 1851 station on population growth from 1851 to 1891 using different samples based on distance to LCP nodes

| | Dependent var.: Δ1891,1851 Ln Pop | | | | | |
|---|---|---|---|---|---|---|
| | Exclude units less 1km | | Exclude units less 3 km | | Exclude units less 5 km | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Estimator type | OLS | IV | OLS | IV | OLS | IV |
| Station in unit by 1851 | 0.172*** | 0.856*** | 0.175*** | 0.733*** | 0.176*** | 0.463*** |

| | | | | | | |
|---|---|---|---|---|---|---|
| | (0.021) | (0.232) | (0.021) | (0.223) | (0.0201) | (0.200) |
| County FEs? | Y | Y | Y | Y | Y | Y |
| Second Nature controls? | Y | Y | Y | Y | Y | Y |
| Kleibergen-Paap F stat | | 66.36 | | 66.66 | | 56.97 |
| Observations | 9,215 | 9,215 | 9,044 | 9,044 | 8,754 | 8,754 |
| R-squared | 0.333 | | 0.317 | | 0.315 | |

Notes: Standard errors in parentheses are clustered on counties. *** p<0.01, ** p<0.05, * p<0.1. All specifications include first nature variables, second nature variables, county fixed effects, and 1851, 1841, and 1831 ln pop density as controls.

Next, we report IV estimates using two instruments, the LCP and the LCP interacted with 1802 coaching inns. The IV coefficient for 1851 stations is slightly smaller at 0.300 (S.E. 0.174) but similar to our main result. See table A.5.3.

Table A.5.3: Estimates for effect of 1851 station on population growth from 1851 to 1891 using coaching inns as a second instrument.

| | Dependent var.: Δ1891,1851 Ln Pop | |
|---|---|---|
| | (1) | (2) |
| Estimator type | OLS | IV |
| Station in unit by 1851 | 0.159*** | 0.300* |
| | (0.0217) | (0.174) |
| County FEs? | Y | Y |
| Second Nature controls? | Y | Y |
| Kleibergen-Paap F stat | | 24.779 |
| Observations | 8,337 | 8,341 |
| R-squared | 0.306 | |
| First stage: Station in unit by 1851 | | |
| | | (3) |
| | | OLS |
| LCP in unit | | 0.0637*** |
| | | (0.0126) |
| LCP in unit* 1802 Coaching inn | | 0.0865** |
| | | (0.0365) |
| County FE? | | Y |
| Second Nature controls? | | Y |
| Observations | | 8,337 |
| R-squared | | 0.216 |

Notes: Standard errors in parentheses are clustered on counties. *** p<0.01, ** p<0.05, * p<0.1. All specifications include 1802 coaching inns, first nature variables, second nature variables, county fixed effects, and 1851, 1841, and 1831 ln pop density as controls. All units less than 7 km from an LCP node are dropped.

Next, we report results for our change on change specification. They show a positive and significant effect of the change in station access and the log change in population from 1821 to 1891. See table A.5.4.

Table A.5.4: Estimates for effect change in 1891 and 1821 stations on difference in log 1891 and 1821 population

| | Dependent var.: Δ1891,1821 Ln Pop | |
|---|---|---|
| | (1) | (2) |
| Estimator type | OLS | IV |
| Change in Station 1891, 1821 | 0.324*** | 0.605** |
| | (0.023) | (0.274) |
| County FEs? | Y | Y |
| Second Nature controls? | Y | Y |
| Kleibergen-Paap F stat | | 29.758 |
| Observations | 8,337 | 8,337 |
| R-squared | 0.345 | |

| First stage: Change in Station 1891, 1821 | (3) |
|---|---|
| | OLS |
| LCP in unit | 0.0720*** |
| | (0.0146) |
| County FE? | Y |
| Second Nature controls? | Y |
| Observations | 8,337 |
| R-squared | 0.223 |

Notes: Standard errors in parentheses are clustered on counties. *** p<0.01, ** p<0.05, * p<0.1. All specifications include 1802 coaching inns, first nature variables, second nature variables, county fixed effects, and 1851, 1841, and 1831 ln pop density as controls. All units less than 7 km from an LCP node are dropped.

Next, we report results using the difference in log 1881 and 1851 occupational shares. The text reports results using the difference in 1881 and 1851 occupational shares without taking logs. The results in table A.5.5 give similar conclusions.

Table A.5.5: Estimates for effect of 1851 station on difference in log male occupational shares 1881 and 1851

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Estimator | OLS | IV | OLS | IV | OLS | IV |
| Dependent variable: | Δ log male agriculture occupational share | | Δ log male secondary occupational share | | Δ log male tertiary occupational share | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Station in unit by 1851 | -0.117*** | -0.336* | 0.0645*** | 0.365** | 0.120*** | 0.3002 |
| | (0.0138) | (0.185) | (0.0139) | (0.169) | (0.0178) | (0.282) |
| Kleibergen-Paap F stat | | 47.8 | | 46.75 | | 47.48 |
| Observations | 8,333 | 8,333 | 7935 | 7935 | 8,178 | 8,178 |
| R-squared | 0.343 | | 0.143 | | 0.403 | |

Notes: Standard errors in parentheses are clustered on counties. *** p<0.01, ** p<0.05, * p<0.1. All specifications include county fixed effects, first nature variables, second nature variables, 1851, 1841, and 1831 ln pop density as controls, and 1851 male shares in agricultural, secondary, tertiary, mining, or unspecified occupations. For definitions of first and second nature variables see table 1. The instrument for station in unit by 1851 is an indicator if unit has LCP in its boundaries. All units less than 7 km from an LCP node are dropped.

Next, we use different percentiles of 1801 population density like the 50$^{th}$ or 70$^{th}$ to test for heterogenous effects. The results are very similar as table A.5.6 shows.

Table A.5.6: Heterogeneous effects of getting a station by 1851 on population growth from 1851 to 1891 using different percentiles for 1801 population density

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Estimator | OLS | IV | OLS | IV |
| Dependent variable: | | | | |
| Station by 1851 | 0.214*** | 0.497 | 0.186*** | 0.522*** |
| | (0.101) | (0.192) | (0.026) | (0.187) |
| Below 50$^{th}$ pct. pop den. 1801 | -0.006 | 0.0188 | | |
| | (0.011) | (0.016) | | |
| Station by 1851* Below 50$^{th}$ pct. pop den. 1801 | -0.131*** | -0.451*** | | |
| | (0.039) | (0.166) | | |
| Below 70$^{th}$ pct. pop den. 1801 | | | -0.061** | -0.017 |
| | | | (0.016) | (0.024) |
| Station by 1851* Below 70$^{th}$ pct. pop den. 1801 | | | -0.041 | -0.350** |
| | | | (0.037) | (0.165) |
| Kleibergen-Paap F stat | | 19.281 | | 19.95 |
| Observations | 8,377 | 8,377 | 8,337 | 8,337 |
| R-squared | 0.306 | | 0.307 | |

Notes: Standard errors in parentheses are clustered on counties. *** p<0.01, ** p<0.05, * p<0.1. All specifications include county fixed effects, first nature variables, second nature variables, 1851, 1841, and 1831 ln pop density as controls. For definitions of first and second nature variables see table 1. All units less than 7 km from an LCP node are dropped. In (2) the instruments are the indicator for LCP and the indicator for LCP interacted with dummy for unit below 50$^{th}$ percentile 1801 population. In (4) we

add the instrument indicator for LCP interacted with dummy for unit below 70$^{th}$ percentile 1801 population.

Finally, we show the IV estimates for effects of log station distance on different in tertiary shares. The IV estimates imply tertiary shares increased significantly less for all units between a log distance of 1.5 and 2.5, or 5 to 12 km.

Figure A.5.2:. IV estimates for effect of 1851 station distance on the difference in 1881 and 1851 tertiary shares