Centre for Economic Performance

Discussion Paper ISSN 2042-2695

No. 2174
April 2026

Marshall at the times of Marshall

Filippo Boeri
Olmo Silva

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

UKRI Economic and Social Research Council

Abstract
In the late nineteenth century, Alfred Marshall identified three micro-foundations of agglomeration economies: labour pooling (LP), input sharing (IO), and knowledge spillovers (KS). An extensive literature has tested the existence of the three Marshallian forces in modern economies. However, there is limited quantitative evidence on the existence of such forces at the times of Marshall. To shed light on these issues, we exploit novel geo-localised census-level data on entrepreneurs and business proprietors retrieved from six consecutive UK Censuses (1851–1911), coupled with census-level workers' data, information on historical patents and historical IO tables. We estimate co-agglomeration models to assess the relative importance of LP, IO, and KS in explaining industrial clustering during Britain's industrialisation. Our results point to a strong role for KS and LP, but only limited evidence for IO. We also show that the strength of the three forces increased over time, and that there is considerable heterogeneity across industries with different characteristics.

Keywords: agglomeration economies; Alfred Marshall; economic history; historical censuses
JEL codes: N83; N93; R12

This paper was produced as part of the Urban Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

We are grateful to the participants at the 2023 Cologne FRESH Meeting, the 2024 CEPR Annual Symposium of Economic History, the 2025 Meeting of the Urban Economics Association as well as Gabriele Cristelli, Stephan Heblich, Henry Overman, Giorgio Pietrabissa,Will Strange, Matt Turner and Yanos Zylberberg and for their comments and suggestions. We are grateful to Tongmeng Xie for his excellent research assistance. We remain responsible for any error or omission.

Filippo Boeri, Department of Economics, City St George's, University of London. Olmo Silva, Department of Geography and Environment, London School of Economics and Centre for Economic Performance at the London School of Economics.

Published by
Centre for Economic Performance
London School of Economic and Political Science
Houghton Street
London WC2A 2AE

1 Introduction

Industries are typically geographically concentrated. This is because firms gain significant benefits from locating near one another. In his pioneering work, Marshall (1890) specifically argued that economic activities cluster because firm productivity is enhanced in dense networks through three mechanisms: labour market pooling (LP), input–output linkages (IO), and knowledge spillovers (KS).[1] While early empirical work on agglomeration economies mostly focused on the distinction between urbanisation and localisation economies[2], subsequent research sought to identify the specific channels underlying agglomeration. Several studies have separately analysed Marshall's three mechanisms, providing evidence for each (Holmes, 1999, Fallick et al., 2006, Almazan et al., 2007, Arzaghi and Henderson, 2008).[3]

To pin down the relative strength of the three Marshallian forces, Ellison et al. (2010) introduced the concept of coagglomeration – the tendency of industries to locate together. Their approach focussing on industry pairs (instead of industries) has two key advantages: first, studying pairs sheds more direct evidence on the industry-to-industry flows that drive agglomeration in a way that looking at the concentration of individual industries cannot; second, studying the links between co-agglomeration and pair-wise Marshallian forces helps dealing with unobservable locational advantages that could bias the results when the unit of observation is the single industry. Their *horse race* approach shows that LP, KS and IO are all important for clustering of economic activities in the modern US economy – with LP being the dominant force. Building on this approach, Faggio et al. (2017) use modern UK data to show that the relative and absolute strength of the Marshallian agglomeration drivers varies systematically with industry and firm characteristics.

Our study contributes to this literature by identifying the relative strength of the Marshallian agglomeration forces during Marshall's own time – namely 1851-1911. While we have a good understanding of the micro-foundations of agglomerations for modern economies, much less is known about their role at the time when they were first theorised. Importantly, the British economy of the late 19th century and early 20th century provides

[1] Since then, a large literature has examined the advantages accruing to firms from clustering (Sveikauskas, 1975, Moomaw, 1981, Carlton, 1983, Nakamura, 1985, Glaeser et al., 1992, Henderson, 1994).

[2] The former, also known as Jacobs externalities (Jacobs, 1961), capture benefits from diversity and the overall size of the urban market, while the latter (Marshall–Arrow–Romer externalities) reflect industry-specific geographic concentration (Dicken and Lloyd, 1990, Glaeser et al., 1992).

[3] Duranton and Puga (2004) proposed an alternative taxonomy of sharing, matching, and learning – emphasising the role of firm and worker heterogeneity. Strange et al. (2006) linked the three mechanisms to both urbanisation and localisation economies, arguing that competitive instability and technological innovativeness are associated with city size, while skill orientation is more relevant for industry clustering. Other contributions have highlighted firm size and organisation as drivers of agglomeration (Glaeser and Kerr, 2009, Rosenthal and Strange, 2010).

us with an ideal 'lab' to study the relative importance of LP, IO and KS at a time of rapid industrialisation, fast urban growth and marked technological adoption.

To shed light on these issues, we follow the co-agglomeration approach of Ellison et al. (2010) and Faggio et al. (2017) to analyse the relative importance of LP, IO, and KS in explaining clustering. We exploit several novel datasets that have not previously been used to study historical agglomeration. First and foremost, we use individual-level geo-referenced data on entrepreneurs and business proprietors from six consecutive UK censuses (1851-1911), covering roughly 11,000 consistently delineated parishes. This allows us to study patterns of co-agglomeration based on the location choices of the entrepreneurs – as opposed to the location of workers. We combine these data with worker-level censuses (on average, 30 million individuals per census year), digitised patent records, and historical input–output tables to construct LP, IO, and KS proxies.

To implement our analysis, we map the spatial distribution of entrepreneurs and their employees across parishes, and use the continuous agglomeration index of Duranton and Overman (2005) to quantify clustering. IO linkages are constructed from historical input–output tables (Horrell et al., 1994, Meyer, 1955, Thomas, 1985). LP is instead measured by linking individual census records over time using individuals' names, addresses, and other identifiers, which allows us to trace job-to-job transitions and construct sectoral connectedness measures. Finally, to obtain a proxy for KS, we start by linking patents to census records using inventors' names and addresses to assign inventors to industries. Lacking data on citation flows (which do not exist for historical patents), we construct a novel proxy for KS by using text analysis to identify key words that represent the technologies described in the patents – and measure 'technological closeness' by looking at average text similarity across patents.

To by-pass endogeneity and omitted variable concerns, we gather extensive information on other locational advantages that could bias our analysis. These include industry-pairs' shared dependence on primary resources – e.g., agriculture or mining output – and proximity to transport hubs (ports, roads and waterways), main cities and coal deposits (whose resources were powering industrialisation). Furthermore, we construct LP, IO and KS proxies using similar data for the US economy at the time to show that these more exogenous proxies produce similar results.

Our results suggest a significant role for knowledge spillovers and labour pooling, but only limited evidence for input sharing. This is in contrast to recent evidence showing that LP and IO dominate KS in explaining clustering. We also find that all three forces – especially IO – became more important over the decades between 1851 and 1911. This is consistent with the emergence of specialised clusters in the later stages of the industrial

revolution, which more heavily rely on cross-industries flows of knowledge, workers and intermediate inputs. We also study heterogeneity in the micro-foundations of agglomeration and find that the three forces are more dominant in sectors characterised by large and more technologically advanced firms, and a higher degree of industrial agglomeration.

Our work relates to a growing literature that uses historical data to address central questions in urban economics (Hanlon and Heblich, 2022). Recent studies have examined productivity effects of agglomeration, historical drivers of industry location such as wartime shocks (Davis and Weinstein, 2008, Redding et al., 2011), and the origins of industrial clusters (Buenstorf and Klepper, 2009, Klepper, 2010, Buenstorf and Klepper, 2010). Heblich and Trew (2019) focus on the role of finance in promoting industrialisation, while Heblich et al. (2025b) focus on the role of slavery in wealth accumulation and manufacturing growth. The closest study to ours is Hanlon and Miscio (2017), who analyse the evolution of 31 English cities between 1851 and 1911 and show that industries expanded more rapidly in cities with greater local supplier presence and occupational similarity.[4] We take a different approach, focusing on industry-level co-agglomeration to identify the relative strength of LP, IO, and KS and their heterogeneity across industries.

Very recently, Heblich et al. (2025a) study how the industrial concentration of British cities that emerged during the industrial revolution gave rise to significantly negative effects on long-run productivity – contributing to the modern-day North-South divide in the UK economy. They interpret their results through the lenses of a quantitative spatial model (QSM) that emphasizes the importance of sectoral diversity (in the spirit of Jacobs) in sustaining long-term growth. Our work complements their study by shedding light on the micro-foundations of that early industrial specialisation – thus providing further insights into the mechanisms that led to cluster 'ossification'.

Leveraging the availability of historical census data, other studies have investigated urbanisation, migration, and structural change in nineteenth-century England. Heblich et al. (2020) show that the railway network led to the first large-scale separation of workplace and residence, accounting for almost half of land value and population growth in London between 1801 and 1921. Arthi et al. (2022) examine the effects of the U.S. Civil War cotton shock, showing adverse mortality impacts in cotton districts and propagation across industries via input–output linkages. Bogart et al. (2022) find that railway development between 1851 and 1891 accelerated population growth and reinforced spatial divergence. Our evidence on the Marshallian 'trinity' complements this literature by explicitly identifying the mechanisms through which such shocks propagated across space and sectors.

[4]The authors find no significant relationship between initial population and subsequent population growth in the second half of the nineteenth century, consistent with evidence from Tapia et al. (2018) for Spain and Egidi et al. (2021) for Italy.

Our paper makes three substantive and novel contributions. First, we provide the first quantification of the relative strengths of Marshall's micro-foundations at a time of industrial take-off using modern econometric techniques. Our results show that Marshall's insights on the importance LP, KS and IO not only project an important influence on our understanding of modern urban economics, but were also valid at the time of his writing. Second – and related – we provide credible historical estimates of the micro-foundations of agglomeration that are important for the growing QSM literature that investigates industrialisation and structural breaks to understand how past economic shocks – from innovation to trade and infrastructure developments – affect economic growth and the current geography of economic development. We hope our work will provide more precise guidance on how to parametrise some of the key forces underpinning such QSM frameworks – thus opening avenues to a better understanding of the historical foundations of modern economic clusters. Lastly, we make some data-related contributions. Our use of historical data on entrepreneurs is novel. More importantly, our way to identify KS from patents using text analysis is new – and provides a method to investigate knowledge sharing at a time when citation flows were not available.

The remainder of the paper is structured as follows. Section 2 presents stylised facts on the evolution of the English economy between 1851 and 1911. Section 3 describes the data sources. Section 4 discusses the construction of proxies for industrial coagglomeration and Marshallian forces. Section 5 outlines the empirical framework. Sections 6 and 7 presents the main results. Section 8 concludes.

2 Stylised Facts

In this section, we briefly discuss some stylised facts about the evolution of the British economy over the period considered in our study. These are useful for situating our study in a unique historical context of radical changes in terms of population distribution over space, and workforce distribution across sectors.

While the period 1760-1800 (before our time-window) was marked by unprecedented technological change (Mokyr, 2005), the effects of the industrial revolution started to materialise in the second half of the nineteenth century. These included: an accelerated growth rate of 2% (Crafts, 2018); population growth from 8.6 to 30 million (Hinde, 2003); and average years of schooling increasing from four to six (Morrisson and Murtin, 2009). The technological progress that led to these extraordinary results was highly unevenly distributed across sectors, and fostered the reallocation of factors across industries and regions. Moreover, the 'early start' in the processes of industrialisation led the country to develop a

highly specialised economy with a unique spatial concentration of activities.

In the next sections, we provide more details of the key transformations sweeping through the British economy of those days.

2.1 Demographic trends

Between 1851 and 1911, the English population doubled. Given the absence of large-scale foreign migration, population growth was driven primarily by rising birth rates, improvements in life expectancy and a sustained decline in infant mortality that continued throughout the twentieth century and occurred in both urban and rural areas (as documented by Gregory (2008)). At the same time as the population grew, so did the labour force. Internal mobility also increased as individuals left their places (counties) of birth to seek employment in rapidly industrialising urban economic centres. See Figure A1 for a graphical representation of these trends.

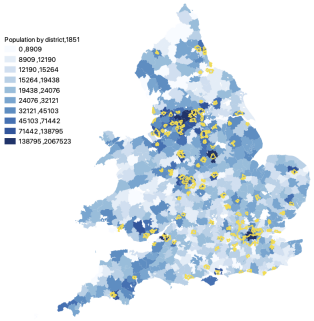
2.2 Spatial trends

During the Victorian and Edwardian eras, the share of population living in the ten largest counties increased, driven primarily by 'new' intermediate centres in Yorkshire, areas of Wales and North-Western ports. Districts in strategic trade locations – such as Middlesbrough, and Grimsby – recorded rapid growth, as did rising centres like Cardiff and districts incorporated into larger urban conglomerates – e.g. Croydon. In contrast, the population of London grew broadly in line with the rest of the country (Figure A2), while the share of the population living in the largest and most historical cities declined. This pattern is consistent with Eckert et al. (2023) and others, who document the rise of new 'factory cities' as engines of late industrialisation.

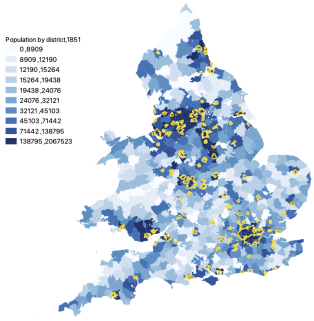
Figures 1 provides more details. Overall, population growth was concentrated in four main regions: Yorkshire, Cardiff, London, and the North Western coastal cities and ports. However, population growth in London's historic core was modest – but more than offset by the expansion of suburban districts on the city's outskirts. Map 1d further illustrates the growth of areas beyond the boundaries of modern Greater London, in line with the findings of Heblich et al. (2020).

Figure 1: A rapidly changing economic geography

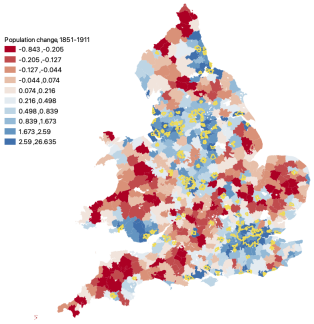
(a) Population by district, 1851



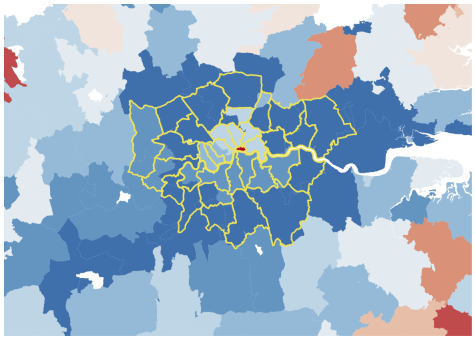
(b) Population by district, 1891



(c) Population change, 1851-1911



(d) Population change in the London area, 1851-1911



Source: Authors' calculations based on the Historical British Censuses.

2.3 Workforce reallocation across sectors

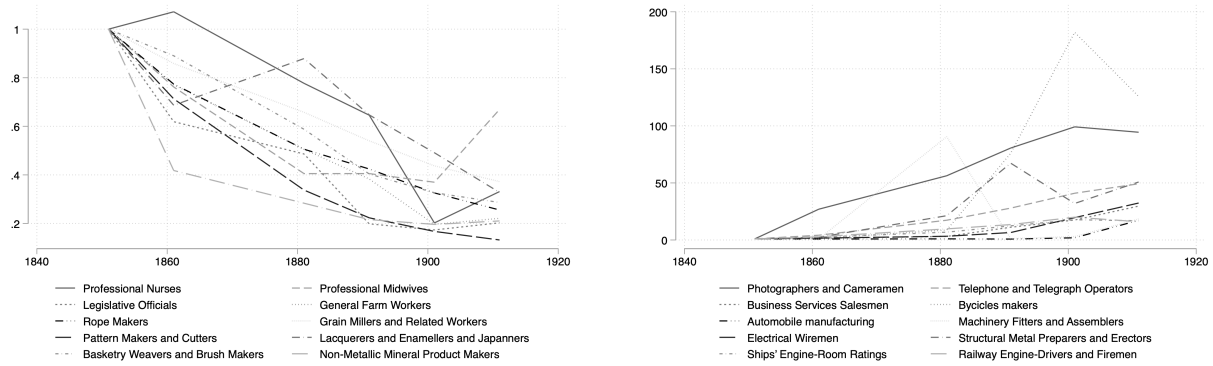
Over this period, the agricultural workforce remained roughly constant at about two million, while employment in manufacturing and services more than doubled (see Figure A3).

Each macro-sector also underwent substantial shifts in its composition. Figure 2 illustrates changes in the relative share of industries within the total workforce, net of overall population growth. New occupations that were almost absent in 1851 – such as photographers, telephone operators, and automobile manufacturers – expanded rapidly, while employment declined most sharply in agriculture, mining, and healthcare-related activities. These trends were driven by the rise of new ideas and technologies on the one hand, and the mechanisation of certain industries/occupations on the other (e.g., agriculture), which allowed for a significant redistribution of workers across the economy.

Figure 2: Best and worst performing industries/occupations

(a) Declining occupations

(b) Rising occupations



Source: Authors' calculations based on the Historical British Censuses.

2.4 Industrial concentration trends

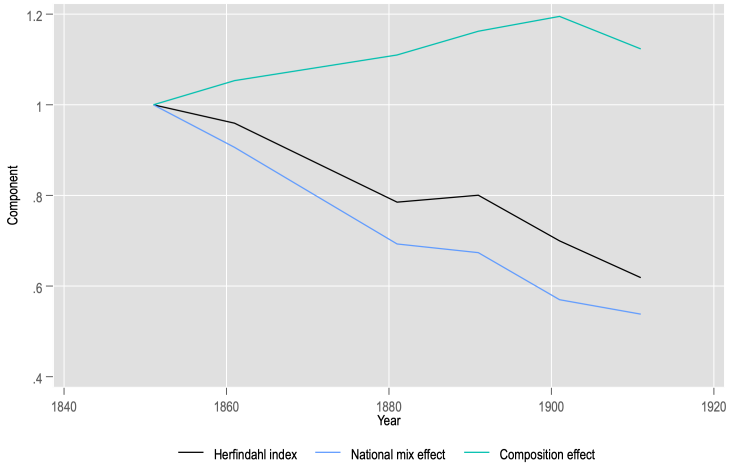
The rise of new sectors documented above reduced industrial concentration at the aggregate level. However, behind this macro trend lies a more nuanced story of industrial geographical concentration. To provide evidence on this issue, in Figure 3 we decompose the evolution of local industrial concentration into two components: changes in the national industry mix and changes in the local composition of economic activities. To do so, we use 'Shapley values' following the procedure suggested by Shapley et al. (1953).[5]

We find that the national Herfindahl index – constructed as the employment-weighted mean of regional Herfindahl indices – declined significantly over the period. However, our decomposition reveals that this aggregate decrease was *entirely* driven by changes in the national industry mix: as new sectors grew and the economy diversified, no single industry dominated employment to the same extent as before. On the other hard, holding the industry mix constant, we observe an *increase* in the local concentration of economic activities. That is, within individual regions, employment became more concentrated in fewer industries over time – consistent with the emergence of specialised industrial clusters as the industrial revolution was unfolding.

This pattern provides an important motivation for the coagglomeration analysis that follows: while the national economy was diversifying, local economies were becoming more specialised, pointing to the presence of agglomeration forces that drew related industries together.

[5]We describe the decomposition in detail in Section C.1 in the Appendix.

Figure 3: Local industrial concentration trends



Source: Authors' calculations based on the Historical British Censuses.

3 Data

3.1 Main data sources

In this study, we exploit four main data sources. The first source is the *I-CeM (Integrated Census Microdata)* dataset, which provides standardised individual-level data from the British censuses conducted between 1851 and 1911 for England, Wales, and Scotland. These decennial censuses recorded information on all individuals, including names, ages, sex, marital status, household relationships, occupations, birthplaces, and employer details. The original data were collected on handwritten household schedules, completed either by the head of household or by enumerators, and preserved in manuscript form by the General Register Office. The I-CeM project drew on large-scale transcriptions undertaken by commercial genealogy companies as part of digitisation initiatives for family history users. The Cambridge Group for the History of Population and Social Structure subsequently transformed these raw transcriptions into a coherent research resource. The resulting dataset integrates individual-level records across multiple censuses, creating a machine-readable archive of unparalleled scale for the nineteenth and early twentieth centuries (Higgs et al., 2013).

A central challenge in constructing I-CeM was the harmonisation of occupational and locational information. The original records contained transcriptions with many spelling variants, abbreviations, and local job titles. These were standardised through dictionaries that consolidated common variants, and then mapped onto the HISCO system (Historical

International Standard Classification of Occupations) developed by Van Leeuwen et al. (2002). To enable longitudinal analysis, I-CeM also introduced consistent parish identifiers, which map changing administrative units onto a stable framework. In total, there are 11,527 parishes, with an average area of 13 square kilometres – roughly the size of Bath or the Royal Borough of Kensington and Chelsea in modern-days Britain. The median parish covers only 7 square kilometres, while the third quartile reaches 14 square kilometres. The few outliers that significantly increase the mean area are due to specific geographical features. For instance, the three largest parishes correspond to the Black Mountains in Wales, the North Pennines Hills in northern England, and the Peak District region.

The second source is the *British Business Census of Entrepreneurs (BBCE)*, produced by the Cambridge Group for the History of Population and Social Structure. Constructed directly from I-CeM, it draws on the same census transcriptions for England and Wales between 1851 and 1911 (Bennett et al., 2019). Whereas I-CeM provides harmonised records for the entire population, the BBCE focuses specifically on entrepreneurs and business proprietors, systematically extracting and coding those recorded as employers or self-employed. Building on the I-CeM framework, the BBCE offers a unique resource for analysing the scale, distribution, and attributes of entrepreneurial activity in nineteenth- and early twentieth-century Britain, while remaining fully compatible with the broader population micro-data. At present, these data have been hardly explored (Hannah and Bennett, 2022, Bennett et al., 2022). Their use to investigate agglomeration patterns is novel – and one of the contributions of our work.

The third source is data from the *UK Intellectual Property Office*. For each patent, we observe information on inventors, abstracts, and full descriptions. An extensive cleaning process was required to correct typographical errors and digitisation problems. For this study, we restrict our attention to patents registered by inventors residing in England and Wales at the date of publication. To facilitate this restriction and link patents to census records (Section 4.2), we extracted and cleaned unstructured information on inventors' residences. The final dataset includes more than 40,000 patents recorded between 1880 and 1911. We use this dataset linked to historical census data to first identify inventors' sector of employment, and then construct a proxy for knowledge spillovers using a quantitative text analysis approach that measures patents' technological similarities. This is also a novel contribution of our study. More details are provided in the next section.

We also draw on historical input–output tables from multiple sources. Horrell et al. (1994) constructed a table for 1841 from a variety of historical sources, arguing that the large number of empty cells reflects the relatively simple production technologies of the period rather than omissions of inter-industry linkages. Meyer (1955), later extended by

Conrad and Meyer (1964), instead produced a table for 1907 based on the first British Census of Production. Similarly, Thomas (1985) constructed a 41-sector table for 1907, supplementing the Census with company histories, industrial handbooks, and other sources. In our work, we experiment with information derived from all three data sources as they trade off 'precision' (more industries) for 'exogeneity' (early measurement).

3.2 Additional data

We also gather similar information for the United States to construct proxies that are less likely to be affected by unobservable locational attributes that could drive the co-location of industries and bias our analysis. The first data source is the *1880 United States Census*, which has been digitised and made accessible through the Integrated Public Use Microdata Series (IPUMS). IPUMS provides harmonised samples that enable micro-level analyses of individuals, households, and communities. Next, we use the 1899 U.S. input–output table of Whitney (1967), which was reconstructed from the Census of Manufactures, agricultural and mining censuses, and other official reports, balancing gross outputs, intermediate uses, and value added in a manner comparable to modern national accounts. Finally, we use data on US patents obtained from Moody's Orbis Intellectual Property. The database compiles comprehensive information on patent filings and grants from multiple jurisdictions. These records are subjected to extensive manual cleaning and validation procedures to ensure consistency, accuracy, and reliable firm-level attribution.

In addition to the data used in the main analysis, we draw on several datasets to construct an extensive set of proxies for local productive amenities. These are discussed in Section A.1 the Appendix.

3.3 Industry classification

The UK and US historical censuses we use assign two different classifications to individuals: the Historical International Standard Classification of Occupations (OCCHISCO); and the 1950 US Census Bureau industry classification (IND95US).

We conduct most of the empirical analysis at the three-digit OCCHISCO level rather than IND95US. As disccused by IPUMS, for nineteenth-century data, the OCCHISCO classification captures occupational nuances by preserving detail from historically vague titles (e.g., 'works in a cotton mill'), re-organising manufacturing and transport groups, and aligning with the hierarchical HISCO framework (Van Leeuwen et al., 2002). The third digit ('unit group') is well suited to nineteenth-century occupational statements and to distinguishing work tasks within early-industrial settings, allowing us to assign workers

to 82 manufacturing groups.

By contrast, IND95US harmonises industries to the 1950 Census scheme to maximise comparability across years. These 1950-based records are not intended for highly detailed industry tracking and are often too coarse for applications – such as our co-agglomeration analysis – that require fine within-manufacturing variation. Moreover, before 1930, the census did not clearly separate 'industry' from 'occupation', further limiting the interpretability of industry-based groupings for nineteenth-century contexts. Consequently, we prefer the three-digit OCCHISCO structure as it provides the finer, historically consistent granularity needed to capture clustering patterns in the nineteenth-century economy. Nevertheless, in Section A.2 we replicate our main analysis using data aggregated to the IND95US level (40 categories). The results are broadly consistent with our baseline. We return to this point below when we discuss our findings.

4 Coagglomeration and agglomeration forces

4.1 Co-agglomeration

In this section, we define how we measure co-agglomeration and briefly present some related descriptive evidence.

4.1.1 Methods

We analyse spatial co-agglomeration defined as the degree of co-location of individual plants. Previous work on co-agglomeration patterns has mostly relied on discrete indices (Ellison et al., 2010, Faggio et al., 2017). Here, we adopt the continuous agglomeration index proposed by Duranton and Overman (2005) (DO).

Ideally, the construction of the DO index requires plant-level data to carry out the following steps: *i)* identify the coordinates of each plant; *ii)* compute pairwise distances; and *iii)* map the industry-pair distribution of plants. Such data are not available for nineteenth-century England and Wales as there is no historical database of firms, their location and their sector of operation. We therefore use the information contained in the BBCE to identify entrepreneurs and business owners, their sector of occupation and their location. Unfortunately, we do not have the exact address of their employment so we proxy workplace location with place of residence of entrepreneurs (i.e., one of approximately 11,000 parishes, consistently defined over time). Given the limited transport technologies available in the nineteenth century (at least compared to modern days), work tended to be

relatively local – and we do not see this as a major limitation of our data (we return to this point below).

With this information, we construct a co-agglomeration index following these steps:

1. Identify the centroid of each parish.

2. Compute the number of entrepreneurs and workers by parish–industry–year.

3. Assign to each entrepreneur–industry–year observation the average number of workers per entrepreneur in that parish.

4. Estimate the Duranton and Overman (2008) co-agglomeration index:

$$\bar{K}_{ji}^{emp}(d) = \frac{1}{h \sum_{r=1}^{n_i} \sum_{s=1}^{n_j} e(r)e(s)} \sum_{r=1}^{n_i} \sum_{s=1}^{n_j} e(r)e(s) f\left(\frac{d - d_{r,s}}{h}\right),$$

   where $d_{r,s}$ is the Euclidean distance between plants r and s, f is a Gaussian kernel density function with bandwidth h, and n_i and n_j are the number of plants in industries i and j respectively.

5. Define distance thresholds (15, 21, 30, 60, 90, 120, 150, 180 km) and, following Behrens (2016), compute the integral of the function up to each $\bar{d}$. This yields the probability that a pair of firms is located within distance $\bar{d}$ of each other.

Given the approximation involved in Step 1, our index should be regarded as a quasi-continuous version of the fully-fledged continuous approach of Duranton and Overman (2008) – and a concern is that place of work and place of residence may differ. However, commuting distances in nineteenth-century England and Wales were severely limited by transport constraints, implying that such discrepancies are likely to be relatively small. Before the advent of steam railways, which enabled the first large-scale separation of workplace and residence, most individuals travelled on foot (Heblich et al., 2020). Combined with working days that commonly exceeded twelve hours, individuals had little scope for lengthy commutes – and place of residence was, for most, a close proxy for place of work.[6] Notably, commuting costs at the time were substantially higher relative to trade costs than is the case today. As noted by Glaeser and Kohlhase (2004), by the late nineteenth century canals, railways, and coastal shipping had already dramatically reduced the cost of moving manufactured goods and raw materials. On the other hand, the dominant mode of

[6] Contemporary accounts and subsequent historical research confirm that most employees lived very near their workplace throughout the nineteenth century, as public transport was either unavailable or unaffordable for most workers (Seltzer and Wadsworth, 2024, Dyos, 1953, Booth, 1893).

'personal transportation' – walking – limited the effective commuting range to one or two miles. This asymmetry between the costs of moving goods and the costs of moving people – formalised in recent quantitative spatial models as the distinction between inter-regional trade costs and intra-urban commuting costs (Redding and Turner, 2015, Ahlfeldt et al., 2015, Monte et al., 2018) – implies that the limited separation of home and workplace in this period is not merely an approximation but a defining feature of the nineteenth-century economy.

A second dimension of approximation involved in Step 1 relates to our inability to capture the within-parish variation in clustering of activities – as we assign entrepreneurs to the parish centroid. Given that we subdivide the territory into more than 11,000 units, this is likely to be of second-order importance. Moreover, any measurement error is likely to be attenuated in the construction of the continuous index: entrepreneurs (and workers) mis-allocated to their residence parish are likely to work (or be employed) in neighbouring parishes given the limitation on travel distance imposed by the historical transport infrastructure network and technologies. This should imply only minor displacement of centroids within our preferred 30km buffer used to compute the index.[7] Nonetheless, we exploit different thresholds to define our co-agglomeration metric and test the robustness of our results when using these alternative definitions.

Another dimension of approximation relates to Step 3. Information on the number of workers employed by entrepreneurs and business proprietors in the BBCE is only available for the years 1851–1871. To bypass this limitation, in our main analysis we assign the same employment to all entrepreneurs in a given sector, in a given year, in a given parish. However, to assess robustness, we construct two alternative versions of the DO index:

- Method 2: A 'standard' DO function that assumes each firm employs one worker.
- Method 3: A DO function using self-reported employment data available for 1851–1871.

As reported in Table A5 in the Appendix, the alternative measures are highly correlated with our baseline index. We also studied co-agglomeration models that use these alternative DO proxies, reaching similar conclusions (we return to this point below).

[7] In the Appendix, we report the results from the following simple exercise. We randomly assign approximately 200,000 points across the territory; then measure the distance from each point to the centroid of the parish where it is located; and finally repeat the process 100 times. Following this procedure, we obtain a measurement error distribution that shows how 'far off' from the centroid of a parish a random location within its boundaries would be. Our results are shown in Figure A4. As can be seen, the mean error is approximately 2.4 km, the median error is about 1.8 km, and the 95th percentile is 6.3 km. This suggests that our DO metrics calculated at 15km to 180km thresholds are unlikely to be significantly affected by measurement error.

4.1.2 Descriptive facts

We briefly discuss some descriptive facts on the co-agglomeration patterns pinned down by our methods. Our findings are tabulated in Appendix B.

Table B1 lists the 15 most co-agglomerated industry pairs identified using our DO metric with a distance threshold at 30km. Most seemed reasonable. These include: lacquerers with locksmiths; locksmiths with gunsmiths; musical instrument makers and bookbinders; dairy product processors with cookers, roasters and related heat treaters. Our classification also pins down some 'historical' sector pairs such as textile printers with fibre preparers; and spinners with fibre preparers. Table B2 presents the 10 most agglomerated industries (i.e., *own* coagglomeration). These include locksmiths, potters and clay activities, musical instruments, fibre preparers, textile printers, lacqueres, and spinners. These are 'classic' examples of nineteenth British industrial clusters – well-known to economic historians.

4.2 Agglomeration forces

In this section, we present the way we construct proxies for the three agglomeration forces defined by Marshall (1890) and discuss some relevant descriptive facts.

4.2.1 Labour pooling

We proxy the strength of labour pooling between industries by tracking job-to-job transitions across consecutive censuses. To this end, we perform large-scale record linkage of census manuscripts, matching individuals on a combination of time-invariant characteristics and family information. Core identifiers include metaphone-encoded first and last names, exact date of birth, and place of birth, with additional reinforcement from intra-household relations (spouses, parents, children, siblings) recorded in the schedules.

The linkage algorithm proceeds iteratively. Candidate matches are first restricted via blocking on names and geography, combined with demographic filters such as age and gender. Potential pairs are then compared using similarity metrics and ranked by match probability, with duplicates removed. High-probability matches are retained, and the inferred family links are reintroduced into the comparison process until convergence. On average, this procedure yields around ten million matches per census pair, corresponding to 45–50% of the surviving population between censuses. This can be considered a good result in the context of historical record linkage, where data incompleteness, name variation, and migration often constrain matching accuracy.[8]

[8] In the historical census–linkage literature, match rates are often much lower (e.g. 20–30%, (Fu et al., 2014). Even the most ambitious full-count linkage projects typically aim for linkage rates around 50% when meth-

From these matched individuals, we construct an occupational mobility index (OMI) that quantifies excess worker flows between sectors. Following Neffke et al. (2018), for each ordered industry pair (i, j) we compute

$$OMI_{ij} = \frac{F^s_{ij} F^s_{\cdot\cdot}}{F^s_{i\cdot} F^s_{\cdot j}}, \tag{1}$$

where F^s_{ij} is the observed number of transitions from i to j, $F^s_{i\cdot}$ the total number of outflows from i, $F^s_{\cdot j}$ the total number of inflows into j, and $F^s_{\cdot\cdot}$ the total number of transitions across all industries. By construction, OMI_{ij} captures whether mobility from i to j is greater or smaller than expected given industry sizes, and thus measures the extent to which industries share a pooled labour market through observed historical worker transitions.

Table B3 in Appendix B report the fifteen most-connected sectors when considering our LP. All the top fifteen most closely related sectors capture sensible associations. To mention a few, these include gilders (workers who apply a thin layer of gold onto furniture) with cabinet makers or gilders with woodworkers; textile printers with printers and related workers; metal rolling mill workers with metal processors; jewelers with watch and precision instrument makers; and wheelwrights with coach and carriage makers.

4.2.2 Input sharing

Proxies for input sharing are typically derived from input–output tables. Since industry-level data are unavailable for the nineteenth century, we rely on reconstructions produced *ex post*. Our baseline analysis uses the table of Conrad and Meyer (1964), while we also replicate results with the alternative matrices of Thomas (1985) and Horrell et al. (1994). The pros and cons of these different tables were discussed above in the Data section.[9]

Using these input–output tables, we quantify the intensity of linkages between industries by examining bilateral intermediate flows. For each ordered pair of industries (i, j), we first calculate the share of inputs that i buys from j, $\text{IN}_{ij} = \frac{z_{ij}}{\sum_k z_{ik}}$, and symmetrically the share of inputs that j buys from i, IN_{ji} (z_{ij} denotes the intermediate input purchased by sector i from sector j). Similarly, we compute the share of outputs that i sells to j, $\text{OUT}_{ij} = \frac{z_{ji}}{\sum_k z_{ki}}$, and the corresponding share for j's sales to i, OUT_{ji}, where the denomina-

ods are carefully tuned and metadata are rich. For example, the 'Record Linking Glass Ceiling' project in Derbyshire, reports achieving around 40–45% match rates between 1881 and 1911 (Diduch et al., 2024). Abramitzky et al. (2021) show that, even when linking duplicate transcriptions of the same census, maximal recovery rarely exceeds 60–65%, and that inter-census linkages typically achieve 20–40% coverage under conservative assumptions.

[9] Appendix A reports binscatter plots showing the correlation between the input-sharing measures obtained from the three matrices. These show a relatively good alignment.

tors exclude sales to final demand.

From these quantities, we compute the maximum input share and the the maximum output share, which proxy, respectively, upstream and downstream dependence:

$$U_{ij} = \max\{\text{IN}_{ij}, \text{IN}_{ji}\} \tag{2}$$

$$D_{ij} = \max\{\text{OUT}_{ij}, \text{OUT}_{ji}\} \tag{3}$$

Finally, we define a synthetic linkage index as:

$$L_{ij} = \max\{U_{ij}, D_{ij}\} \tag{4}$$

Applied to Conrad and Meyer (1964)'s table, these indices provide a concise measure of the extent to which nineteenth-century industries were connected through intermediate input and output flows. Table B4 in Appendix B presents some evidence. The patterns we detect capture meaningful associations – including well drillers and borers with miners and quarrymen; weavers with spinners, twisters and winders; and butchers and meat preparers with brewers and beverage makers and with food preserves.

We also construct a set of pairwise dissimilarity indices to capture industry differences in their reliance on primary resources. Specifically, for each industry i, we compute the share s_i^r of its total intermediate consumption sourced from a primary resource sector r – namely, agriculture, mining, and gas/electricity/water. Following Faggio et al. (2017), the dissimilarity between industries i and j in their reliance on resource r is defined as:

$$D_{ij}^r = \frac{1}{2}\left|s_i^r - s_j^r\right|, \tag{5}$$

where $s_i^r = v_i^r / \sum_k v_i^k$ and v_i^r denotes the value of intermediate inputs that industry i sources from resource sector r. The index takes a value of zero when two industries draw on a given resource in identical proportions, and increases as their input profiles diverge.

4.2.3 Knowledge spillovers

Several studies have exploited patent citation data to proxy knowledge spillovers (Jaffe et al., 1993, Yang and Lin, 2012, Faggio et al., 2017). This approach is not feasible for historical patents, as nineteenth-century inventors were not required to cite prior work.

To capture knowledge spillovers between industries, we instead construct a novel proxy based on historical *patent text*. To begin with, we retrieve the full text of all patents recorded

in Espacenet between 1880 and 1911 and implemented a multi-step procedure to link these innovations to the industries of their inventors (this is similar to the approach we used above to match workers across census decades). Specifically, we first apply a named-entity recognition algorithm, fine-tuned on a synthetic training sample, to extract inventor names, addresses, and reported occupations. Second, inventors are matched to census records through a sequence of plausibility filters, threshold adjustments, sorting, and de-duplication. Third, each patent is assigned to the inventor's industry, using census-based information whenever available (or self-reported industry otherwise). This procedure yielded industry assignments for 36,412 patents out of 70,000.

We consider this a 'good outcome', given that recent studies linking patents to individuals or firms in historical settings report successful linkage rates ranging from 40% to 65%, depending on the richness of available metadata, the accuracy of name disambiguation, and the use of auxiliary sources such as census or company directories.[10] Moreover, we deliberately adopted a conservative matching strategy, prioritising the reduction of false positives at the expense of overall coverage.

With these assignments in place, we proceed to construct a measure of technological proximity across industries through a multi-step text-based approach. First, we apply quantitative text analysis to the full body of each patent to identify the technological domains in which innovative activity occurred. The raw text is pre-processed to remove punctuation, non-alphabetic characters, and standard stop-words, as well as high-frequency terms that are generic to the patenting process (e.g. 'invention' 'apparatus', 'method', 'system'). We then implement term-frequency and relevance filtering to isolate the most salient technology-related keywords, which capture the substantive content of each innovation. The resulting vocabulary provides a compact yet informative representation of the technological space, allowing us to compare industries based on the semantic similarity of the innovations attributed to their inventors.[11]

With these data, we construct a similarity index across all pairs of industries as follows. Let $T_i = (T_{i1}, T_{i2}, \ldots, T_{iN})$ denote the distribution of patents from industry i across N technology classes, where $T_{i\tau}$ is the share of patents of industry i in class τ. For each industry

[10] For example, Petralia et al. (2017) use patent text and inventor metadata to map technologies to industries, achieving comparable coverage levels. Catalini et al. (2019) also highlight that even modern patent–firm linkage projects face substantial name and address ambiguities, with recall rates rarely exceeding 60%.

[11] Table B5 lists the 30 most frequent technical terms used for this purpose. Although individually these terms may appear commonplace, it is their joint frequency profile — i.e., the vector of term weights across patents — that encodes technological content. Two patents concentrating on the same subset of terms (e.g. *pressure*, *valve*, and *pipe* versus *gear*, *shaft*, and *pivot*) will receive a high cosine similarity, capturing proximity in technological space.

pair (i, j), we define

$$Tech_{ij} = \frac{T_i T_j'}{\sqrt{T_i T_i'}\sqrt{T_j T_j'}}, \tag{6}$$

which corresponds to the cosine similarity between the two patent distributions. This index captures the extent to which industries drew on similar technological knowledge bases, and thus provides a proxy for potential knowledge spillovers in the late nineteenth and early twentieth century.

As shown in Table B6 our KS proxy also identifies close connections between sectors that seem closely related in terms of technologies. These include: automobile manufacturing with machine-tool operators; gunsmiths with machine tool operators; locksmiths with gunsmiths; and watch, clock and precision instruments with machine-tool operators – just to mention a few.

To sum up, the descriptive evidence provided in Appendix B suggests that our measure of co-agglomeration and proxies for the three Marshallian forces provide a relatively accurate and credible snapshot of the industrial structure of industrialising Britain in the second half of the nineteenth century. We next use these proxies to estimate the relative importance of LP, KS and IO in explaining co-agglomeration of industries.

5 Empirical strategy

5.1 Horse-race model

Our main specification evaluates the relative importance of the Marshallian forces in shaping agglomeration economies by jointly estimating the effect of LP, IO, and KS on observed co-agglomeration between industry pairs. Formally, we estimate:

$$DO_{ijt} = \beta_1 LP_{ij} + \beta_2 IO_{ij} + \beta_3 KS_{ij} + X_{ij}'\gamma + \psi_i + \phi_j + \nu_t + \varepsilon_{ijt}, \tag{7}$$

where DO_{ijt} denotes the co-agglomeration index between sectors i and j at time t, measured following Duranton and Overman (2005). The variables LP_{ij}, IO_{ij}, and KS_{ij} are, respectively, our proxies for labour pooling, input–output linkages, and knowledge spillovers between sectors i and j, averaged over the relevant census intervals. The vector X_{ij} collects controls for dissimilarity in access to primary resources across industries, as well as a set of proxies for local productive amenities (see Section A.1). We also include sector fixed effects ψ_i and ϕ_j, and time fixed effects ν_t. This framework allows us to assess the explanatory

power of the three Marshallian channels while absorbing sector- and period-specific heterogeneity. Focusing on industry-pair co-agglomeration – as opposed to a single-industry agglomeration – helps deal with potentially unobservable locational advantages that could bias our results (as argued in Ellison et al. (2010) and Faggio et al. (2017)). Nevertheless, identification concerns may remain – and we next discuss the measures we take to alleviate potential problems.

5.2 Endogeneity concerns

Two sources of bias are salient in our setting. First, sorting: firms may self-select into locations based on unobserved characteristics, which could confound estimates of the contribution of Marshallian forces. Second, reverse causality: firms may cluster in particular locations because of Marshallian linkages, but such linkages may also arise endogenously after clustering has occurred for other reasons.

To address potential sorting based on local amenities, we rely on the data discussed above that characterise potential access to infrastructures and resources that could have lead industry to simultaneously cluster in a set of locations – irrespective of their LP, IO and KS connections. The maps in Figure A7 show the locations of ports, canals, train stations and coalfields. We also consider the location of main towns and historical Roman roads (overlaid to the other maps). For each of these factors, we proceed as follows. Consider ports as an example: we start by computing the employment-weighted average distance between each plant and the nearest port, and then rank industries by the inverse of this measure. An analogous approach is used to measure proximity to train stations, urban centres, historical roads and navigable waterways, while to construct our proxy for coal access, we consider the share of firms located within 10 km of an exposed coalfields. Using this information, we then either control in our regressions for a dissimilarity index for industry-pairs in access to these resources and infrastructures; or iteratively exclude the 10% of sector pairs with the greatest access to each factor.

To address both reverse causality and omitted variables at once, we take an alternative approach and replicate our main analysis using 'exogenous' proxies for Marshallian forces derived from the United States. These are unlikely to be affected by local unobservables and other contemporaneous forces driving industrial clustering in Britain. For input sharing, we use the 1899 U.S. input–output table reconstructed by Whitney (1967). For LP, we exploit U.S. census microdata for 1850–1910 and construct a measure of job-to-job transitions across census decades, using the same record-linkage approach as for

Britain.[12] For KS, we apply the methodology we used for British data to U.S. patents and census data, linking inventors to census records and computing a technology-closeness index across industries. These exogenous variables provide an external benchmark that should not be correlated with British co-agglomeration unobservables, thereby mitigating concerns about reverse causality and endogenous sorting.

6 Results

6.1 Baseline analysis

Our first estimation results are reported in Table 1. Columns (1)–(3) present univariate regressions of co-agglomeration on each Marshallian force separately. The coefficients are positive and statistically significant in all cases. A one-standard-deviation increase in input–output linkages is associated with a 13% increase in industrial co-agglomeration, while the corresponding effects for labour pooling and knowledge spillovers are 30% and 35%, respectively.

Columns (4)–(6) report multivariate regressions including all three forces jointly. Column (4) does not include any control besides industry and year fixed effects. In this specification, the coefficient on input sharing declines to 4.6%, and that on knowledge spillovers to 11%, while the effect of labour pooling decreases only modestly. Column (5) and (6) progressively add control that take into account the possible effects of the industry-pair's joint dependence on inputs from primary sectors transport infrastructures and other productive amenities. The coefficients remain broadly stable when input–output dissimilarity indices are added in column (5). Including dissimilarity controls for local amenities (Column 6) leads to a more pronounced reduction in magnitudes, but all three coefficients remain positive and statistically significant.[13] Specifically, the effect size of IO lies between 6.2% and 3.1%, the impact of LP is estimated to be between 27% and 21% while KS has a standardized effect of 12%-7.5%.

[12] This exercise requires a crosswalk between U.S. and British occupational and industrial classifications, as the OCCHISCO variable is available only in the 1860 Census. We begin by using all four occupational and industrial variables available in the 1860 U.S. Census. For each combination of the 1880 occupational code (used by IPUMS for the 1850–1900 Censuses), the 1950 occupational code, and the 1950 industrial code, we identify the OCCHISCO code with the highest frequency. In almost 30% of cases, the combination of the three variables corresponds to a single OCCHISCO code, and the median share of observations covered by the most frequent OCCHISCO is 0.9. We the adopt a conservative approach to deal with the other cases by excluding all combinations of the three variables for which the most frequent OCCHISCO accounts for less than 80% of observations. This procedure results in dropping 25% of the combinations, which, however, correspond to less than 7% of individuals in the Census population.

[13] Note that dropping the dissimilarity index based on the use of mining products when we include our proxy for coalfield dependence as in Column (6) does not change the results.

Compared with Faggio et al. (2017) study of the modern British economy, our estimates for LP are similar – or slightly larger; our coefficients for IO are markedly smaller; while our estimates for KS are notably larger (the authors find effects sizes of 16.5%, 8.2% 2.4% for LP, IO and KS, respectively). A similar pattern emerges when comparing our evidence to the estimates by Ellison et al. (2010) for the US economy. These results should be interpreted in the context of an economy at the early stages of industrialisation, with short value chains and many sectors linked to natural resources (thus lowering the importance of IO). Similarly, the heightened relevance of knowledge spillovers can be explained by the limited ability to codify production processes – and the importance of 'knowledge in the air' as originally proposed by Marshall.

Table 1: Coagglomeration and Marshallian forces

VARIABLES	(1) DO	(2) DO	(3) DO	(4) DO	(5) DO	(6) DO
Input-Output	0.130***			0.0460**	0.0623***	0.0318**
	(0.0219)			(0.0179)	(0.0204)	(0.0144)
Labour Pooling		0.302***		0.276***	0.277***	0.214***
		(0.0185)		(0.0170)	(0.0176)	(0.0146)
Knowledge Spillovers			0.350***	0.112***	0.120***	0.0773**
			(0.0487)	(0.0365)	(0.0370)	(0.0310)
Diss. agriculture					0.0306	0.0329**
					(0.0204)	(0.0142)
Diss. Gas & electricity					0.0613***	0.0757***
					(0.0209)	(0.0175)
Diss. mines					0.00698	0.00759
					(0.0148)	(0.0116)
Diss. ports						-0.317***
						(0.0278)
Diss. road network						-0.238***
						(0.0149)
Diss. waterways						-0.0740***
						(0.0176)
Diss. main cities						0.0762***
						(0.0122)
Diss. rail network						0.0302
						(0.0210)
Diss. coalfields						-0.314***
						(0.0483)
Observations	19,926	19,926	19,926	19,926	19,926	19,926
R^2	0.033	0.152	0.026	0.159	0.161	0.373
Year FE	✓	✓	✓	✓	✓	✓
i industry FE	✓	✓	✓	✓	✓	✓
j industry FE	✓	✓	✓	✓	✓	✓
IO Table	Meyer	-	-	Meyer	Meyer	Meyer

Robust standard errors are clustered at the industry pair level and reported in parenthesis. ***, ** and * respectively indicate 0.01, 0.05 and 0.1 levels of significance.

To lend credibility to our results, we test their robustness along several dimensions. To begin with, in Table A1 in the Appendix A we replicate the analysis using the IO tables produced by Thomas (1985) and Horrell et al. (1994) (see Section 4.2). The results are broadly similar – although the IO proxies are now insignificant and close to zero in our multivariate regressions. While the input-output table by Horrell et al. (1994) has the advantage of

predating our period of analysis – thus attenuating reverse causality concerns – it has a very rough partitioning of the economy. The table by Thomas (1985) instead extends the work of Meyer (1955) by adjusting the input-output patterns based on additional information extracted from company histories, industrial records and other sources. This risks building an element of endogeneity in our results. On balance, we see the results using the Meyer (1955) proxy as our favourite as they strike the right balance between precision and exogeneity.

In Tables A2, we replicate our analysis using alternative distance thresholds to calculate our DO co-agglomeration variable. The estimates for IO and KS are fairly stable over the range of 30km to 90km while they are smaller, but still significant at 21km and 15km. Conversely the coefficient estimates for labour pooling are incredibly steady between 15 and 120km, then progressively decline, while remaining positive and significant, beyond that threshold. At the top end, these patterns are likely to reflect the actual geographical distance over which these Marshallian forces exerted an effect. At the short distance, instead, the significant attenuation is likely to be driven by some measurement error in our co-agglomeration proxy which is based on the centroids of the businesses' parishes as opposed to their exact address.[14]

In Table A3, we address potential measurement issues arising from the imputation of employment masses for businesses in the Census of Entrepreneurs, which are used to construct the DO co-agglomeration proxy (as discussed in Section 4.1). In this table, we replicate the results under the assumption that all entrepreneurs employ exactly one worker. Although the magnitude and statistical significance of the estimates vary across specifications, the evidence remains broadly consistent with the findings discussed above.[15]

Next, in Table A4 in the Appendix we replicate the analysis at the IND95US-level, rather than at the HISCO-level. We find broadly consistent patterns, with LP having the strongest effects, followed by KS and IO – the latter being small and only significant in univariate regressions. As discussed above, we see the IND95US-based industrial classification as less accurate and relevant than the HISCO-based alternative we use in our main analysis. Nonetheless, it is reassuring that our results survive with this classification.

Table 2 instead examines the evolution of Marshallian forces over the study period. All three increase in magnitude over time. The effect of labour pooling increases by roughly two-thirds, while knowledge spillovers remain relatively stable before rising in the final

[14] We find similar patterns if we add the additional dissimilarity controls as in Columns (6) of Table 1.

[15] We do not replicate our analysis using the proxy that relies on self-reported employment for BBCE entrepreneurs as this is available only for two decades. This approach would 'confound' any evidence on the impact of using a different dependent variable with changes in the relative Marshallian forces over time – something we discuss later in this section.

two decades. Most strikingly, input–output linkages are insignificant in the early censuses but become positive, sizeable and significant by the end of the period. We take this as evidence of a maturing economy – with extending supply chains and increasingly integrated production processes. This is also reminiscent of the patterns described in the 'nursery city' framework by Duranton and Puga (2001). We return to this point below where we test the heterogeneity of our findings across industries.

Table 2: Agglomeration trends

VARIABLES	(1) DO	(2) DO	(3) DO	(4) DO	(5) DO	(6) DO
Input-Output	0.0108	0.00585	0.0185	0.0386**	0.0490***	0.0681***
	(0.0126)	(0.0127)	(0.0137)	(0.0166)	(0.0180)	(0.0219)
Labour Pooling	0.141***	0.184***	0.207***	0.249***	0.248***	0.258***
	(0.0145)	(0.0151)	(0.0154)	(0.0175)	(0.0186)	(0.0210)
Knowledge Spillovers	0.0644	0.0745**	0.0645*	0.0585	0.0902**	0.112**
	(0.0397)	(0.0336)	(0.0351)	(0.0389)	(0.0417)	(0.0444)
Diss. agriculture	0.0115	-0.000740	0.0213	0.0421**	0.0542***	0.0692***
	(0.0203)	(0.0205)	(0.0176)	(0.0187)	(0.0207)	(0.0222)
Diss. Gas & electricity	0.0623*	0.0784***	0.0746***	0.0978***	0.0795***	0.0617*
	(0.0320)	(0.0281)	(0.0225)	(0.0290)	(0.0220)	(0.0334)
Diss. mines	-0.0102	0.0167	0.00780	0.0121	0.00727	0.0119
	(0.0128)	(0.0122)	(0.0122)	(0.0134)	(0.0140)	(0.0149)
Diss. ports	-0.418***	-0.369***	-0.326***	-0.288***	-0.248***	-0.253***
	(0.0302)	(0.0284)	(0.0281)	(0.0297)	(0.0311)	(0.0355)
Diss. road network	-0.265***	-0.265***	-0.274***	-0.221***	-0.206***	-0.198***
	(0.0184)	(0.0175)	(0.0176)	(0.0176)	(0.0192)	(0.0202)
Diss. waterways	-0.0515**	-0.0770***	-0.0850***	-0.0717***	-0.0680***	-0.0909***
	(0.0211)	(0.0182)	(0.0187)	(0.0194)	(0.0202)	(0.0232)
Diss. main cities	0.0875***	0.0909***	0.0938***	0.0810***	0.0603***	0.0437***
	(0.0144)	(0.0138)	(0.0132)	(0.0143)	(0.0148)	(0.0163)
Diss. rail network	-0.0422	-0.00549	0.0303	0.0280	0.0674***	0.103***
	(0.0257)	(0.0240)	(0.0229)	(0.0246)	(0.0251)	(0.0286)
Diss. coalfields	-0.255***	-0.305***	-0.343***	-0.354***	-0.334***	-0.290***
	(0.0471)	(0.0427)	(0.0430)	(0.0577)	(0.0598)	(0.0819)
Observations	3,321	3,321	3,321	3,321	3,321	3,321
R^2	0.421	0.428	0.433	0.391	0.334	0.307
Year FE	✓	✓	✓	✓	✓	✓
i industry FE	✓	✓	✓	✓	✓	✓
j industry FE	✓	✓	✓	✓	✓	✓
IO Table	Meyer	Meyer	Meyer	Meyer	Meyer	Meyer
Year	1851	1861	1881	1891	1901	1911

Robust standard errors are clustered at the industry pair level and reported in parenthesis.
***, ** and * respectively indicate 0.01, 0.05 and 0.1 levels of significance.

Finally, in Table A5 we study a related point: how the three Marshallian forces relate to the dynamics of co-agglomeration over time. To do so, we regress changes in coagglomeration of industry pairs between 1851 and 1911 – i.e., over a sixty-year period – on LP, KS and IO. Univariate regressions in Columns (1)–(3) show that all three Marshallian forces are positively and significantly associated to long-term changes in coagglomeration. However, the *horse race* specifications of Columns (4)–(6) show that only LP and IO are positive and significant – while KS has a positive but not significant effect. The columns also show that LP still remains the strongest of the three Marshallian forces, while the point estimates of IO and KS are now much more similar. If anything, IO dominates KS in explaining the

strengthening of coagglomeration of industry pairs. Again, we interpret this pattern as evidence of a maturing economy in which supply chains progressively extend and for which sharing of input becomes more important in explaining colocation patterns.[16]

6.2 Dealing with endogeneity concerns

As discussed above, there are two potential sources of bias in our work: sorting on unobservable locational (productive) amenities and reverse causality.

Our 'augmented' specification of Column (6), Table 1 addresses the problem of unobservable local factors by appending controls that partial out the possible co-dependence of industry pairs on infrastructures, travel networks and natural resources that could drive co-location. In Table 3, we take a more 'aggressive' approach and present our findings from specifications that exclude all industry pairs where sector i and/or j falls within the top 10% of the distribution of dependence on a specific local factor (sequentially, one factor at the time across the different columns). For example, in Column (2) we exclude the top 10% pairs for which either one or both industries locate closest to ports.

Results remain largely unaffected when we consider ports, railways, waterways, and urban centres. The only significant change emerges in column (7), where we exclude the top 10% of sectors with the highest dependence on coal, measured by their proximity to parishes located in coal-rich areas – defined as those where at least 5% of the population were coal miners. We find that excluding the most coal-dependent sectors leads to a marked reduction in coefficient magnitudes – especially for IO, which is now insignificant, very small and slightly negative. This is perhaps not surprising as it is well known that coal availability played an important role in shaping agglomeration patterns during this period of rapid industrialisation given the prevailing technologies.

[16] We also investigated whether LP, KS and IO explain changes in coagglomeration patterns over 30 years – i.e., 1851–1881 and 1881-1911. We found similar patterns. Results are not presented for space reasons.

Table 3: Selection on local amenities

VARIABLES	(1) DO	(2) DO	(3) DO	(4) DO	(5) DO	(6) DO	(7) DO
Input-Output	0.0623***	0.0711***	0.0722***	0.0753***	0.0801***	0.0669***	-0.0124
	(0.0204)	(0.0234)	(0.0249)	(0.0243)	(0.0243)	(0.0246)	(0.00850)
Labour Pooling	0.277***	0.286***	0.302***	0.233***	0.263***	0.305***	0.208***
	(0.0176)	(0.0200)	(0.0216)	(0.0180)	(0.0183)	(0.0220)	(0.0166)
Knowledge Spillovers	0.120***	0.174***	0.233***	0.0962**	0.174***	0.169***	0.0687*
	(0.0370)	(0.0428)	(0.0461)	(0.0425)	(0.0436)	(0.0484)	(0.0359)
Diss. agriculture	0.0306	0.0436*	0.0436*	0.0523**	0.0383	0.0401*	0.00775
	(0.0204)	(0.0231)	(0.0257)	(0.0218)	(0.0240)	(0.0243)	(0.0164)
Diss. mines	0.00698	0.0186	0.0104	0.0202	0.00355	0.00592	-0.0332**
	(0.0148)	(0.0171)	(0.0171)	(0.0128)	(0.0150)	(0.0173)	(0.0162)
Diss. Gas & electricity	0.0613***	0.0544**	0.0470*	0.0562**	0.0594**	0.0482*	0.0257
	(0.0209)	(0.0238)	(0.0265)	(0.0235)	(0.0243)	(0.0259)	(0.0173)
Observations	19,926	17,556	16,650	16,206	16,206	16,650	16,650
R^2	0.161	0.172	0.197	0.132	0.167	0.184	0.099
Year FE	✓	✓	✓	✓	✓	✓	✓
i industry FE	✓	✓	✓	✓	✓	✓	✓
j industry FE	✓	✓	✓	✓	✓	✓	✓
Selection	none	Ports	Roads	Waterways	Cities	Railway	Coal

Robust standard errors are clustered at the industry pair level and reported in parenthesis. ***, ** and * respectively indicate 0.01, 0.05 and 0.1 levels of significance.

In Table 4 instead, we take an alternative approach to bypass possible endogeneity concerns and replicate our main analysis using 'exogenous' proxies for Marshallian forces derived from the United States.[17] These are unlikely to be affected by reverse causality and local unobservables driving industrial clustering in Britain. The estimated coefficients are smaller in magnitude but broadly consistent with our baseline evidence. We still find that LP has the most sizeable effect of the three Marshallian forces, followed by KS – while IO is not always significant or sizeable.[18] In Table 5 we further examine the evolution of the effects over time using these 'exogenous' proxies . We still find that all three Marshallian forces substantially increase in strength across the period, consistent with the results from the British data. All in all, the evidence provided in this section provides support to our causal interpretation of the evidence so far.

[17] Note that we include slightly different controls for dissimilarity in industry-pairs primary resource dependence. This is driven by the structure of the US IO tables that we use.

[18] As an additional exercise, we tested each UK agglomeration variable individually, instrumenting it with the corresponding US variable. Using this approach, we obtain a coefficient of 0.75 (s.e. 0.33) for KS, with a first-stage F-statistic of 16.4; a coefficient of 0.34 (s.e. 0.044) for LP, with a first-stage F-statistic of 194; and a coefficient of 0.13 (s.e. 0.056) for IO, though with a very low first-stage F-statistic of 2.16.

Table 4: Exogenous variables

VARIABLES	(1) DO	(2) DO	(3) DO	(4) DO	(5) DO
Input-Output	0.0212*			0.0153	0.0121
	(0.0117)			(0.00937)	(0.00738)
Labour Pooling		0.101***		0.0989***	0.0936***
		(0.0151)		(0.0156)	(0.0151)
Knowledge Spillovers			0.0792**	0.0622*	0.0593*
			(0.0344)	(0.0345)	(0.0340)
Diss. agriculture					-0.0594**
					(0.0251)
Diss. fisheries					0.00612
					(0.0130)
Diss. mines					-0.100
					(0.0723)
Diss. gasoil					0.0875
					(0.0557)
Diss. transport					-0.0190
					(0.0205)
Observations	19,440	19,926	19,440	18,960	18,960
R^2	0.001	0.013	0.002	0.016	0.023
Year FE	✓	✓	✓	✓	✓
i industry FE	✓	✓	✓	✓	✓
j industry FE	✓	✓	✓	✓	✓
IO Table	Whitney	Whitney	Whitney	Whitney	Whitney
Sample	All	All	All	All	All

Robust standard errors are clustered at the industry pair level and reported in parenthesis. ***, ** and * respectively indicate 0.01, 0.05 and 0.1 levels of significance.

Table 5: Exogenous variables - trend

VARIABLES	(1) DO	(2) DO	(3) DO	(4) DO	(5) DO	(6) DO
Input-Output	0.00228	0.0100	0.00978	0.0145	0.0182*	0.0175*
	(0.00899)	(0.00710)	(0.00769)	(0.00956)	(0.0106)	(0.0103)
Labour Pooling	0.0699***	0.0814***	0.0932***	0.101***	0.107***	0.109***
	(0.0161)	(0.0157)	(0.0155)	(0.0172)	(0.0177)	(0.0195)
Knowledge Spillovers	0.0515	0.0242	0.0404	0.0671*	0.0750*	0.0976**
	(0.0402)	(0.0393)	(0.0355)	(0.0395)	(0.0385)	(0.0470)
Diss. agriculture	-0.0516*	-0.0493*	-0.0518**	-0.0549**	-0.0638**	-0.0848***
	(0.0265)	(0.0265)	(0.0245)	(0.0274)	(0.0290)	(0.0321)
Diss. fisheries	0.00299	0.00500	0.00955	0.0121	0.00301	0.00411
	(0.0172)	(0.0178)	(0.0146)	(0.0134)	(0.0142)	(0.0142)
Diss. mines	-0.0798	-0.0904	-0.0988	-0.0996	-0.110	-0.122
	(0.0694)	(0.0702)	(0.0751)	(0.0765)	(0.0738)	(0.0789)
Diss. gasoil	0.0335	0.0689	0.0583	0.0976	0.131	0.136
	(0.0953)	(0.0665)	(0.0620)	(0.0593)	(0.0876)	(0.0864)
Diss. transport	-0.0154	0.000385	-0.00680	-0.0373	-0.0296	-0.0256
	(0.0237)	(0.0233)	(0.0237)	(0.0230)	(0.0242)	(0.0256)
Observations	3,160	3,160	3,160	3,160	3,160	3,160
R^2	0.015	0.016	0.022	0.028	0.030	0.031
Year FE	✓	✓	✓	✓	✓	✓
i industry FE	✓	✓	✓	✓	✓	✓
j industry FE	✓	✓	✓	✓	✓	✓
IO Table	Whitney	Whitney	Whitney	Whitney	Whitney	Whitney
Sample	1851	1861	1881	1891	1901	1911

Robust standard errors are clustered at the industry pair level and reported in parenthesis. ***, ** and * respectively indicate 0.01, 0.05 and 0.1 levels of significance.

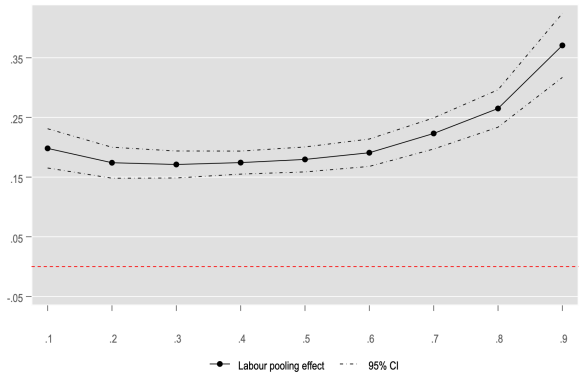
7 Heterogeneity in micro-foundations

We next explore heterogeneity in strengths of the micro-foundations of agglomeration economies across industries following the approach by Faggio et al. (2017).

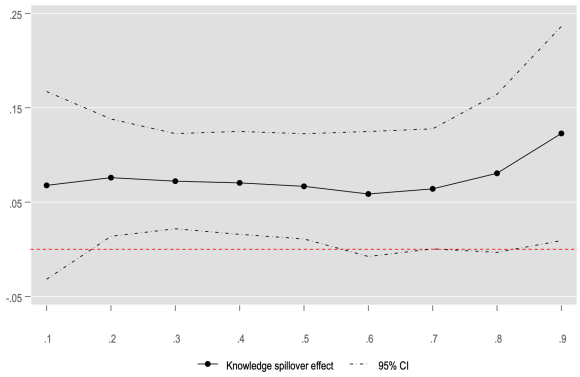
To begin with, we report estimates of Equation 7 by quantiles of the co-agglomeration distribution. This approach sheds light on whether LP, IO and KS are stronger in instances where co-location is more or less prevalent – possibly reflecting the planned or unplanned nature of such interactions (a la Jacobs vs. Marshall). Our evidence is presented in Figure 4. We find that knowledge spillovers and input sharing display relatively flat patterns across quantiles – with a slight uptick for the highest quantiles. Labour pooling instead is positively associated with the degree of co-agglomeration: its effect is strongest among industry pairs that co-locate most extensively. These patterns are in sharp contrast with Faggio et al. (2017) who find flat KS effects, increasing IO impacts and decreasing importance of LP over the co-agglomeration quantiles.

All in all, this suggests that Marshall's micro-foundations have their strongest effects when interactions are repeated and of a planned, predictable nature. Stated differently, Marshall – not Jacobs – was right at the time of Marshall.

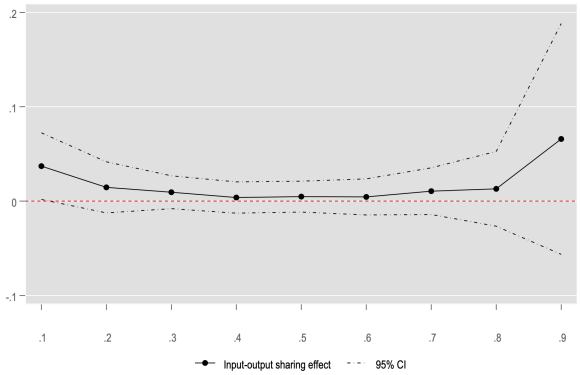
Figure 4: The effect of Marshallian forces at different quantiles of agglomeration – Multivariate



(a) Labour pooling



(b) Knowledge spillovers



(c) IO – Meyer

Note: Variables are transformed to have unit standard deviation for interpretation. The figure plots regression coefficients and 95% confidence intervals from quantile regressions that simultaneously include all three Marshallian forces. Confidence intervals from bootstrapped standard errors clustered on industry pairs. All regressions control for dissimilarity in use of resources.

Next, we examine heterogeneity in the strength of the Marshallian forces across subsamples of industry pairs defined by firm size, workforce age, innovativeness, and industry-specific agglomeration. Specifically, we rank industries based on the median of the distribution of a given characteristic – for example, firm size – and classify each pair according to whether both industries lie above the median, both below, or one on each side. Our results are presented in Table 6.

Columns (1)–(3) suggest that all three Marshallian forces are concentrated in industry pairs comprising larger firms. The IO and KS coefficients are not statistically significant for mixed and small-firm pairs, while the LP coefficient is roughly three times as large for pairs of large-firm industries as for small-firm ones. This contrasts with the results for modern clusters presented in Faggio et al. (2017), but is consistent with the nineteenth-century

economy, where many smaller industries tended to operate in relative autarky. With respect to workforce age (columns (4)–(6)), industries employing younger workers exhibit stronger associations with labour pooling, input sharing, and knowledge spillovers. While this may relate to industry dynamism and the 'nursery city' hypothesis of Duranton and Puga (2001), we lack data on the age of the industry itself to more thoroughly investigate this channel.

Columns (7)–(9) analyse heterogeneity with respect to the degree of innovativeness measured as the share of 'breakthrough innovations' in each sector's total patent output (following the methodology proposed by Kelly et al. (2021)).[19] Interestingly, the coefficients on the three forces are significant only for pairs of innovative industries, whereas for pairs of non-innovative industries only labour pooling appears to play a role.[20] Finally, columns (10)–(12) show that industry-specific agglomeration amplifies all three forces, confirming that the micro-foundations of coagglomeration are most pronounced in sectors already characterised by strong within-industry clustering.

[19] See Section D for details on the construction of this measure.

[20] We also investigated whether our results differ depending on the skill-set of the workforce. Results were mixed: IO had a stronger effect among low-skill intensive industry pairs, while LP did not vary along this dimension. Surprisingly, KS seem to be more important for industry-pairs with a low-skilled workforce. This might suggest a degree of substitutability between the skills embodied in workers and the knowledge flows that can be absorbed from spillovers from other firms in a cluster. However, our measure of skills is very crude and closer to a socio-economic background taxonomy. This might explain some of the puzzling patterns, so we do not tabulate our results – which remain available upon request.

Table 6: Heterogeneous agglomeration

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
		size			age			Innovative			Agglomeration	
VARIABLES	high	mixed	low	high	mixed	low	high	mixed	low	high	mixed	low
Input-Ouput	0.132***	0.0166	-0.000780	-0.0160*	0.0112	0.155***	0.105***	0.0333	-0.0301	0.200***	-0.0325***	-0.0123***
	(0.0389)	(0.0161)	(0.00764)	(0.00932)	(0.0112)	(0.0377)	(0.0342)	(0.0212)	(0.0229)	(0.0398)	(0.00995)	(0.00454)
Labor Pooling	0.372***	0.249***	0.133***	0.131***	0.238***	0.414***	0.363***	0.311***	0.177***	0.414***	0.170***	0.124***
	(0.0372)	(0.0256)	(0.0202)	(0.0224)	(0.0252)	(0.0400)	(0.0306)	(0.0282)	(0.0263)	(0.0417)	(0.0128)	(0.0127)
Knowledge Spillovers	0.337***	0.0193	-0.00278	-0.0152	0.0437	0.389***	0.158**	0.109**	0.0310	0.227**	0.0553*	-0.0355*
	(0.0930)	(0.0467)	(0.0475)	(0.0411)	(0.0431)	(0.107)	(0.0695)	(0.0506)	(0.0690)	(0.0983)	(0.0299)	(0.0206)
Diss. agriculture	-0.159	-0.0117	0.0236	0.0252	0.0225	-0.0932	0.134***	0.00997	-0.0541	-0.0115	-0.00584	0.0273***
	(0.108)	(0.0308)	(0.0198)	(0.0183)	(0.0262)	(0.0599)	(0.0487)	(0.0203)	(0.0533)	(0.0621)	(0.0152)	(0.0104)
Diss. mines	0.0815***	-0.0150	-0.0146	-0.0831*	-0.0276	0.136***	0.0199	0.00938	0.0466	0.0362	-0.0427**	-0.0249***
	(0.0234)	(0.0238)	(0.0301)	(0.0439)	(0.0230)	(0.0221)	(0.0208)	(0.0234)	(0.0358)	(0.0645)	(0.0168)	(0.00881)
Diss. Gas												
& electricity	0.0830	0.0155	0.0282	0.0179	0.0185	-0.00308	0.000714	0.0636**	0.0685*	0.184***	-0.0158	-0.0136
	(0.0850)	(0.0254)	(0.0223)	(0.0149)	(0.0269)	(0.0655)	(0.0538)	(0.0257)	(0.0350)	(0.0555)	(0.0138)	(0.00905)
Constant	0.0150	-0.0342***	0.0482***	0.0171	-0.000477	-0.0296	0.0143	-0.0150	0.0183	0.00695	-0.0269***	0.0325***
	(0.0209)	(0.0128)	(0.0129)	(0.0123)	(0.0134)	(0.0199)	(0.0168)	(0.0130)	(0.0182)	(0.0282)	(0.00914)	(0.00499)
Observations	5,166	10,080	4,680	4,680	10,080	5,166	5,166	10,080	4,680	4,920	10,086	4,920
R-squared	0.298	0.111	0.053	0.060	0.112	0.342	0.305	0.172	0.060	0.272	0.099	0.152
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
i industry FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
j industry FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Robust standard errors are clustered at the industry pair level and reported in parenthesis. ***, ** and * respectively indicate 0.01, 0.05 and 0.1 levels of significance.

8 Conclusion

This paper has examined the role of industrial agglomeration forces during the Victorian and Edwardian eras, when they were first theorised by Alfred Marshall. We show that Marshallian forces were already shaping the spatial distribution of firms, broadly in line with Marshall's predictions, and that their importance increased between 1851 and 1911.

Relative to modern economies, labour pooling exhibited a similar magnitude, input sharing was less relevant, and knowledge spillovers played a much stronger role. This pattern is consistent with an economy in the early stages of industrialisation: value chains were short and closely tied to natural resources, limiting the scope for input sharing; limited codification of production processes heightened the need for proximity in learning and knowledge transfer. The effects were most pronounced in sectors characterised by larger and technologically advanced firms, younger workforces, and higher degrees of industrial agglomeration.

Why should we care? First and foremost, because understanding the past helps dealing with the present. The processes that lead to the rapid industrialisation of certain regions of the British economy in the nineteenth century also sowed the seeds of later ossification and industrial decline. Forming a better understanding of the forces that brought about such hyper-specialisation helps identifying the micro-foundations of such long-term dynamics. It also helps appreciating that agglomerative forces might underpin a possible inter-temporal trade-off – with rapid industrial take-off on the back of specialised industrial clustering followed by decline and backwardness a century and a half later. In this sense, while Marshall was right in describing what he observed, Jacobs might have had 'the upper hand' in describing the economic recipe for long-term sustainable growth (as argued by Heblich et al. (2025a) for the British historical context and Glaeser et al. (1992) for the US economy).

Our evidence is also important for a growing literature that uses quantitative spatial models (QSM) to understand the long-term distribution of economic activities and historic structural economic breaks vis-a-vis global shocks – such as opening to trade, technological innovations and changes in means of production. Our results show that modern estimates of the Marshallian agglomerative forces do not closely mimic those that underpinned economic clusters during the industrial revolutions. Our estimates provide more precise guidance on how to parameterise such QSM frameworks to better understand the causes and long-term consequences of early industrial clustering.

References

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., and Pérez, S. (2021). Automated linking of historical data. *Journal of Economic Literature*, 59(3):865–918.

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2015). The economics of density: Evidence from the berlin wall. *Econometrica*, 83(6):2127–2189.

Almazan, A., De Motta, A., and Titman, S. (2007). Firm location and the creation and utilization of human capital. *The Review of Economic Studies*, 74(4):1305–1327.

Alvarez-Palau, E. J. and Dunn, O. (2019). Database of historic ports and coastal sailing routes in england and wales. *Data in brief*, 25:104188.

Arthi, V., Beach, B., and Hanlon, W. W. (2022). Recessions, mortality, and migration bias: Evidence from the lancashire cotton famine. *American Economic Journal: Applied Economics*, 14(2):228–55.

Arzaghi, M. and Henderson, J. V. (2008). Networking off madison avenue. *The Review of Economic Studies*, 75(4):1011–1038.

Behrens, K. (2016). Agglomeration and clusters: Tools and insights from coagglomeration patterns. *Canadian Journal of Economics/Revue canadienne d'économique*, 49(4):1293–1339.

Bennett, R. J., Montebruno, P., Van Lieshout, C., and Smith, H. (2022). Business entry and exit: Career changes of proprietors in england and wales (1851–81) using record-linkage. *Social Science History*, 46(2):255–289.

Bennett, R. J., Smith, H., Van Lieshout, C., Montebruno, P., and Newton, G. (2019). *The age of entrepreneurship: Business proprietors, self-employment and corporations since 1851*. Routledge.

Bogart, D., Satchell, M., Alvarez-Palau, E. J., You, X., and Taylor, L. S. (2017). Turnpikes, canals, and economic growth in england and wales, 1800-1850. *Dan Bogart: Research: Research on Transport and the English Industrial Revolution*.

Bogart, D., You, X., Alvarez-Palau, E. J., Satchell, M., and Shaw-Taylor, L. (2022). Railways, divergence, and structural change in 19th century england and wales. *Journal of Urban Economics*, 128:103390.

Booth, C. (1893). *Life and Labour of the People in London*, volume 4. Macmillan and Company.

Buenstorf, G. and Klepper, S. (2009). Heritage and agglomeration: the akron tyre cluster revisited. *The Economic Journal*, 119(537):705–733.

Buenstorf, G. and Klepper, S. (2010). Why does entry cluster geographically? evidence from the us tire industry. *Journal of Urban Economics*, 68(2):103–114.

Carlton, D. W. (1983). The location and employment choices of new firms: An econometric model with discrete and continuous endogenous variables. *The Review of Economics and Statistics*, pages 440–449.

Catalini, C., Guzman, J., and Stern, S. (2019). Hidden in plain sight: venture growth with or without venture capital. Technical report, National Bureau of Economic Research.

Codrington, T. (1919). *Roman roads in Britain*. Society for promoting Christian knowledge.

Conrad, A. H. and Meyer, J. R. (1964). The economics of slavery and other. *Studies in Econometric History (Aldine Publishing Co., Chicago, 1964)*.

Crafts, N. (2018). *Forging ahead, falling behind and fighting back: British economic growth from the Industrial Revolution to the Financial Crisis*. Cambridge University Press.

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

Dicken, P. and Lloyd, P. (1990). Location in space: Theoretical perspectives. *Economic Geography*.

Diduch, E. et al. (2024). The record linking glass ceiling. applying automated methods to the census and women's marriage records, 1881–1911. *Historical Life Course Studies*, 14:126–143.

Duranton, G. and Overman, H. G. (2005). Testing for localization using micro-geographic data. *The Review of Economic Studies*, 72(4):1077–1106.

Duranton, G. and Overman, H. G. (2008). Exploring the detailed location patterns of uk manufacturing industries using microgeographic data. *Journal of Regional Science*, 48(1):213–243.

Duranton, G. and Puga, D. (2001). Nursery cities: Urban diversity, process innovation, and the life cycle of products. *American Economic Review*, 91(5):1454–1477.

Duranton, G. and Puga, D. (2004). Micro-foundations of urban agglomeration economies. In *Handbook of regional and urban economics*, volume 4, pages 2063–2117. Elsevier.

Dyos, H. J. (1953). Workmen's fares in south london, 1860–1914. *The Journal of Transport History*, (1):3–19.

Eckert, F., Juneau, J., and Peters, M. (2023). Sprouting cities: How rural america industrialized. In *AEA Papers and Proceedings*, volume 113, pages 87–92. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Egidi, G., Quaranta, G., Salvati, L., Salvia, R., and Antonio, G. M. (2021). Investigating density-dependent patterns of population growth in southern italy, 1861–2019. *Letters in Spatial and Resource Sciences*, 14(1):11–30.

Ellison, G., Glaeser, E. L., and Kerr, W. R. (2010). What causes industry agglomeration? evidence from coagglomeration patterns. *American Economic Review*, 100(3):1195–1213.

Faggio, G., Silva, O., and Strange, W. C. (2017). Heterogeneous agglomeration. *Review of Economics and Statistics*, 99(1):80–94.

Fallick, B., Fleischman, C. A., and Rebitzer, J. B. (2006). Job-hopping in silicon valley: some evidence concerning the microfoundations of a high-technology cluster. *The review of economics and statistics*, 88(3):472–481.

Fu, Z., Boot, H., Christen, P., and Zhou, J. (2014). Automatic record linkage of individuals and households in historical census data. *International Journal of Humanities and Arts Computing*, 8(2):204–225.

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A., and Shleifer, A. (1992). Growth in cities. *Journal of political economy*, 100(6):1126–1152.

Glaeser, E. L. and Kerr, W. R. (2009). Local industrial conditions and entrepreneurship: how much of the spatial distribution can we explain? *Journal of Economics & Management Strategy*, 18(3):623–663.

Glaeser, E. L. and Kohlhase, J. E. (2004). Cities, regions and the decline of transport costs. *Papers in regional Science*, 83(1):197–228.

Gregory, I. N. (2008). Different places, different stories: Infant mortality decline in england and wales, 1851–1911. *Annals of the Association of American Geographers*, 98(4):773–794.

Hanlon, W. W. and Heblich, S. (2022). History and urban economics. *Regional Science and Urban Economics*, 94:103751.

Hanlon, W. W. and Miscio, A. (2017). Agglomeration: A long-run panel data approach. *Journal of Urban Economics*, 99:1–14.

Hannah, L. and Bennett, R. (2022). Large-scale victorian manufacturers: Reconstructing the lost 1881 uk employer census. *The Economic History Review*, 75(3):830–856.

Heblich, S., Nagy, D. K., Trew, A., and Zylberberg, Y. (2025a). The death and life of great british cities. *NBER Discussion Paper 34029*.

Heblich, S., Redding, S., Trew, A., and Voth, H.-J. (2025b). Slavery and the british industrial industrialization. *NBER Discussion Paper 30451*.

Heblich, S., Redding, S. J., and Sturm, D. M. (2020). The making of the modern metropolis: evidence from london. *The Quarterly Journal of Economics*, 135(4):2059–2133.

Heblich, S. and Trew, A. (2019). Banking and industrialization. *Journal of the European Economic Association*, 17(6):1753–1793.

Henderson, I. V. (1994). Where does an industry locate? *Journal of Urban Economics*, 35(1):83–104.

Higgs, E., Jones, C., Schürer, K., and Wilkinson, A. (2013). The integrated census microdata (i-cem) guide.

Hinde, A. (2003). *England's population: a history since the Domesday survey*. Hodder Arnold.

Holmes, T. J. (1999). How industries migrate when agglomeration economies are important. *Journal of Urban Economics*, 45(2):240–263.

Horrell, S., Humphries, J., and Weale, M. (1994). An input-output table for 1841. *Economic History Review*, pages 545–566.

Jacobs, J. (1961). *Death and Life of Great American Cities*. Random House, New York City.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *the Quarterly journal of Economics*, 108(3):577–598.

Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3):303–320.

Klepper, S. (2010). The origin and growth of industry clusters: The making of silicon valley and detroit. *Journal of Urban Economics*, 67(1):15–32.

Margary, I. D. (1955). *Roman Roads in Britain: South of the Foss Way. Bristol Channel*, volume 1. Phoenix House.

Margary, I. D. (1973). *Roman Roads in Britain (third ed.)*, volume 1. John Baker.

Marshall, A. (1890). Principles of economics: An introductory volume.

Meyer, J. R. (1955). An input-output approach to evaluating the influence of exports on british industrial production in the late 19th century. *Explorations in Economic History*, 8(1):12.

Mokyr, J. (2005). The intellectual origins of modern economic growth. *The Journal of Economic History*, 65(2):285–351.

Monte, F., Redding, S. J., and Rossi-Hansberg, E. (2018). Commuting, migration, and local employment elasticities. *American Economic Review*, 108(12):3855–3890.

Moomaw, R. L. (1981). Productivity and city size: a critique of the evidence. *The Quarterly Journal of Economics*, 96(4):675–688.

Morrisson, C. and Murtin, F. (2009). The century of education. *Journal of Human capital*, 3(1):1–42.

Nakamura, R. (1985). Agglomeration economies in urban manufacturing industries: a case of japanese cities. *Journal of Urban economics*, 17(1):108–124.

Neffke, F. M., Otto, A., and Hidalgo, C. (2018). The mobility of displaced workers: How the local industry mix affects job search. *Journal of Urban Economics*, 108:124–140.

Petralia, S., Balland, P.-A., and Morrison, A. (2017). Climbing the ladder of technological development. *Research Policy*, 46(5):956–969.

Redding, S. J., Sturm, D. M., and Wolf, N. (2011). History and industry location: evidence from german airports. *Review of Economics and Statistics*, 93(3):814–831.

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

Rosenthal, S. S. and Strange, W. C. (2010). Small establishments/big effects: Agglomeration, industrial organization and entrepreneurship. In *Agglomeration economics*, pages 277–302. University of Chicago Press.

Satchell, M. (2017). Exposed coalfields of england and wales. GIS shapefile derived from BGS 1:625,000 digital geological maps.

Satchell, M. and Bogart, D. (2017). Settlement points of england and wales, c. 1563?1911. GIS point shapefile of 4,010 settlement locations, derived from historical town lists and texts.

Satchell, M., Wrigley, E. A., Shaw-Taylor, L., You, X., and Henneberg, J. (2023). 1851 england, wales and scotland rail lines. ArcGIS shapefile dataset.

Seltzer, A. J. and Wadsworth, J. (2024). The impact of public transportation and commuting on urban labor markets: Evidence from the new survey of london life and labour, 1929–1932. *Explorations in Economic History*, 91:101553.

Shapley, L. S. et al. (1953). A value for n-person games.

Shorrocks, A. F. et al. (2013). Decomposition procedures for distributional analysis: a unified framework based on the shapley value. *Journal of Economic Inequality*, 11(1):99–126.

Strange, W., Hejazi, W., and Tang, J. (2006). The uncertain city: Competitive instability, skills, innovation and the strategy of agglomeration. *Journal of Urban Economics*, 59(3):331–351.

Sveikauskas, L. (1975). The productivity of cities. *The Quarterly Journal of Economics*, 89(3):393–413.

Tapia, F. J. B., Díez-Minguela, A., and Martinez-Galarraga, J. (2018). Tracing the evolution of agglomeration economies: Spain, 1860–1991. *The Journal of Economic History*, 78(1):81–117.

Thomas, M. (1985). An input-output approach to the british economy, 1890–1914. *The Journal of Economic History*, 45(2):460–463.

Van Leeuwen, M., Maas, I., Miles, A., Edvinsson, S., Karlsson, J., Jarnaes-Erikstad, M., Pelissier, J., Rébaudo, D., de Sève, M., van de Putte, B., et al. (2002). *HISCO. Historical international standard classification of occupations*. Leuven University Press.

Whitney, W. (1967). The structure of the american economy in the late nineteenth century. *PhD Dissertation*.

Yang, C.-H. and Lin, H.-L. (2012). Openness, absorptive capacity, and regional innovation in china. *Environment and Planning A*, 44(2):333–355.

A Additional Tables, Charts and Data Descriptions

A.1 Additional data sources

In this section, we discuss several other data sources that we use to construct an extensive set of proxies for local productive amenities.

The dataset on *Historical Ports and Sailing Shipping Routes in England and Wales* (Alvarez-Palau and Dunn, 2019) provides geo-referenced information on 479 ports and landing places between 1540 and 1914, together with reconstructed coastal sailing routes based on historical charts, bathy-metric data, and visibility analysis. Supplied in GIS-ready format, it enables analysis of maritime connectivity, transport costs, and the spatial evolution of coastal infrastructure. For this study, we use information on ports identified up to 1842.

The dataset *1830 England and Wales Navigable Waterways shapefile* (Bogart et al., 2017) provides a GIS representation of navigable rivers and canals in 1830, as part of a time-dynamic database spanning 1600–1948. Digitised primarily from first-edition Ordnance Survey maps and supplemented by historical sources, it includes attributes on opening, closing, and commercial-use dates. This dataset offers a consistent spatial framework for analysing inland navigation within the broader transport network.

Roman roads in Britain were first systematically described by Codrington (1919), based on topographical observation and antique reports, and later refined by Margary (1955, 1973), who introduced a standardised numbering system and distinguished between certain, probable, and conjectural routes. Their combined work established the canonical mapping of the Roman road network still used today. For this study, we use a GIS shapefile digitising their mapped network, providing a consistent spatial framework for analysis.

The dataset *1851 England, Wales and Scotland Rail Lines*, constructed by Satchell et al. (2023), comprises an ArcGIS shapefile of approximately 6,336 miles of railway lines open to passenger or freight services in 1851, covering England, Wales, and Scotland. Derived from a broader time-varying GIS spanning 1807–1998, it enables the extraction of railway network snapshots at any given year. For our analysis, we use data on train stations active by 1851.

The dataset *Settlement Points of England and Wales, c. 1563–1911* (Satchell and Bogart, 2017) provides a GIS point shapefile of 4,010 settlements, derived from historical town lists (1563–1911) and 22 historic texts (1612–1888), geo-referenced in the British National Grid. For this study, we focus on settlements recorded in 1861.

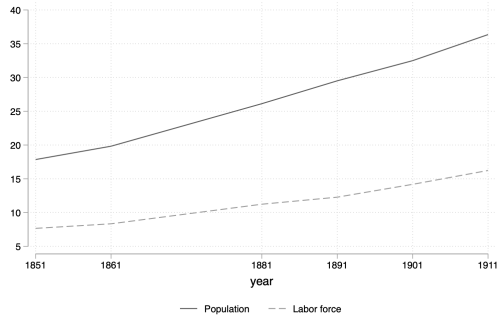
Finally, the dataset *Exposed Coalfields of England and Wales* (Satchell, 2017) provides a GIS representation of coal-bearing strata outcropping at or near the surface circa 1830. Derived from the British Geological Survey's digital geological maps, it identifies areas where coal

was most accessible before the introduction deep of mining. The shapefile, clipped to the England and Wales coastline, is a key spatial layer for analysing the geography of early industrialisation.

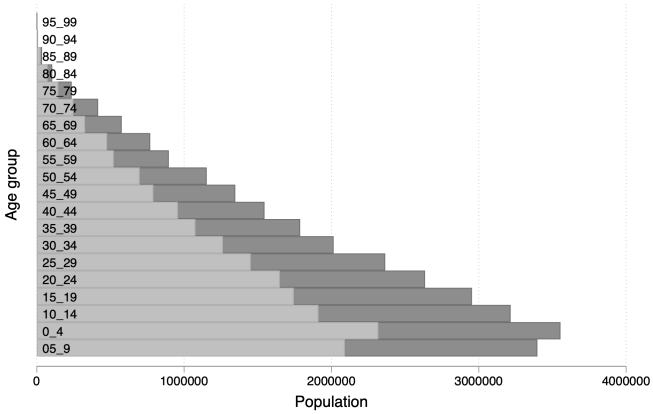
A.2 Additional Tables and Charts

Figure A1: Demographic trends

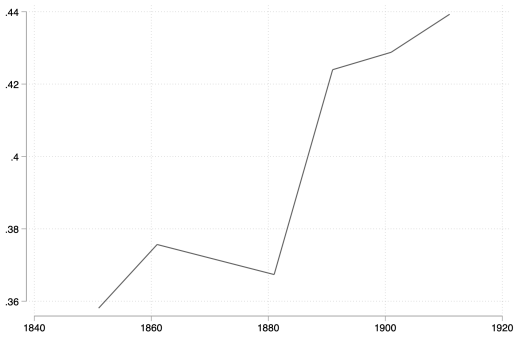
(a) Overall trend)



(b) Demographic pyramid



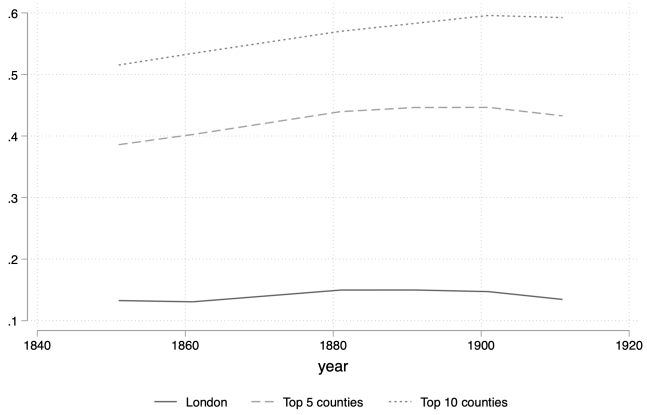
(c) Share of population born in a different county



(d) Share of foreign migrants

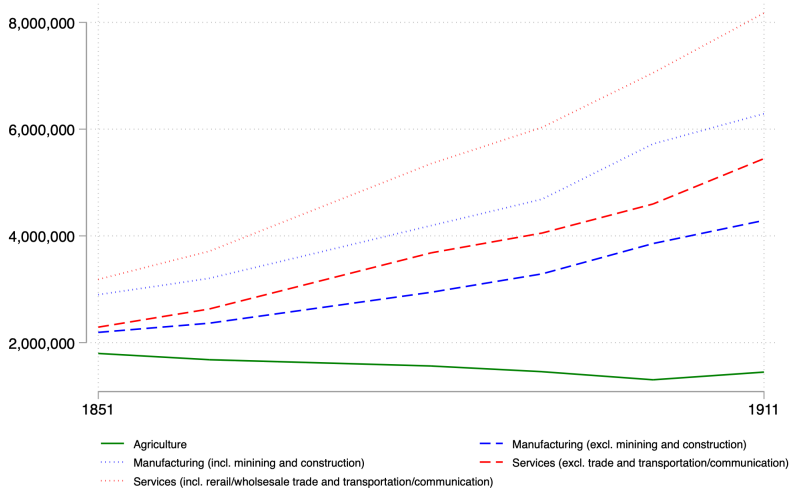
Source: Authors' calculations based on the Historical British Censuses.

Figure A2: Not just a London story



Source: Authors' calculations based on the Historical British Censuses.

Figure A3: Sectoral trends



Source: Authors' calculations based on the Historical British Censuses.

Figure A4: Distribution of measurement errors

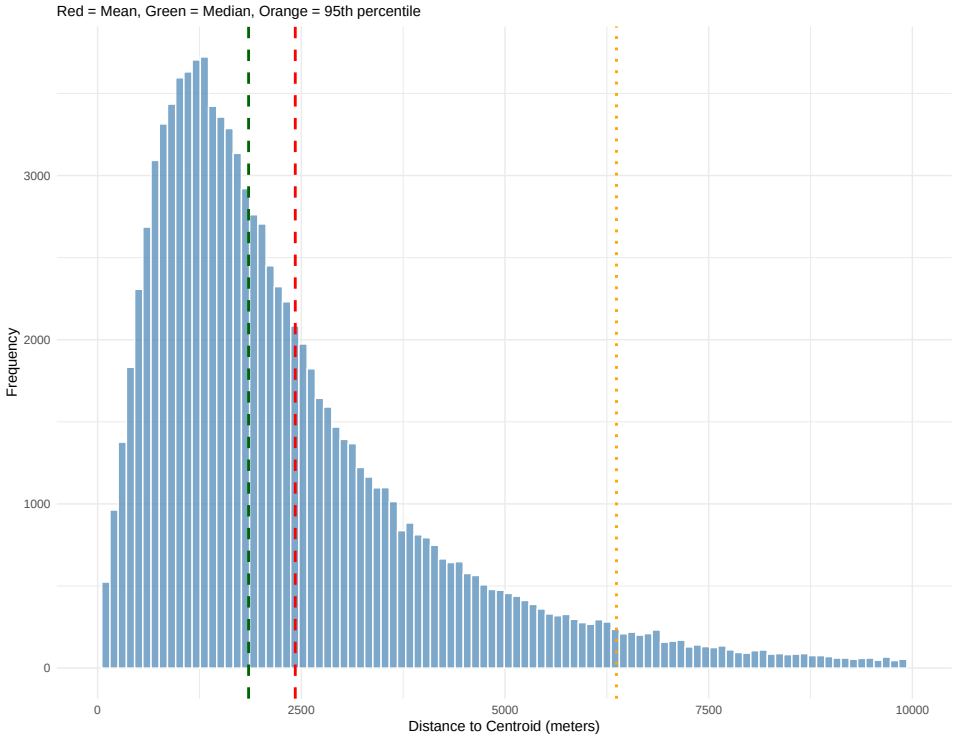


Figure A5: DO Metric: The three methods

(a) Method 1 - Method 2 (b) Method 1 - Method 3 (c) Method 2 - Method 3

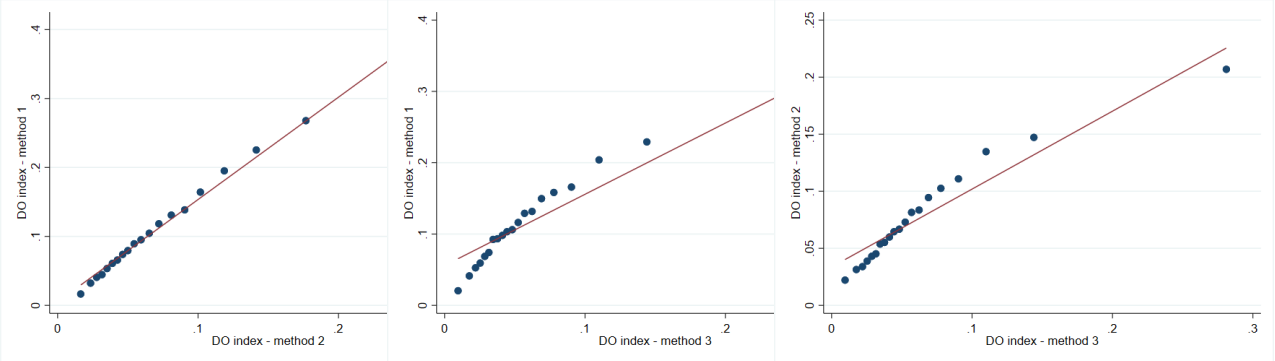


Figure A6: Input sharing, occupational classification

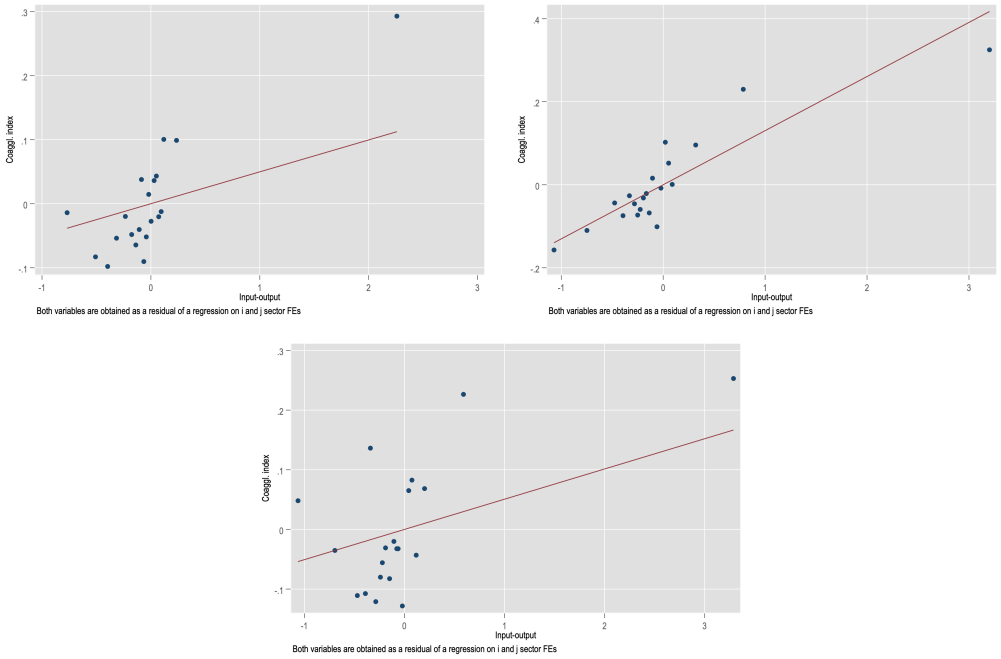


Figure A7: Amenities

(a) Open coalfields and Roman roads

(b) Train stations

(c) Ports, waterways and main towns

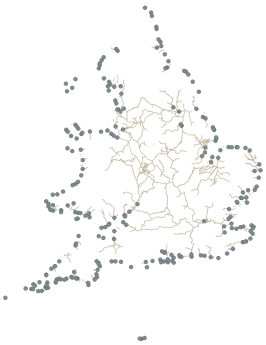


Table A1: Alternative input sharing proxies

VARIABLES	(1) DO	(2) DO	(3) DO	(4) DO	(5) DO	(6) DO	(7) DO	(8) DO
Input-Output	0.0507***	0.00291	0.00187	8.28e-05	0.0498**	-0.00139	-0.000626	-0.00878
	(0.00951)	(0.00729)	(0.00746)	(0.00667)	(0.0202)	(0.0137)	(0.0137)	(0.0118)
Labour Pooling		0.289***	0.291***	0.220***		0.290***	0.292***	0.220***
		(0.0175)	(0.0179)	(0.0144)		(0.0175)	(0.0177)	(0.0143)
Knowledge Spillovers		0.128***	0.123***	0.0820***		0.128***	0.133***	0.0721**
		(0.0380)	(0.0387)	(0.0310)		(0.0379)	(0.0383)	(0.0307)
Diss. agriculture			-0.0161	0.0275**			0.0235	0.0249**
			(0.0138)	(0.0115)			(0.0180)	(0.0124)
Diss. Gas & electricity			0.0180	0.00720			0.0465**	0.0180
			(0.0196)	(0.0147)			(0.0232)	(0.0243)
Diss. mines			0.0236**	0.00705			0.0153	0.000347
			(0.0101)	(0.00857)			(0.0212)	(0.0216)
Diss. ports				-0.316***				-0.312***
				(0.0281)				(0.0276)
Diss. road network				-0.240***				-0.239***
				(0.0150)				(0.0150)
Diss. waterways				-0.0740***				-0.0753***
				(0.0176)				(0.0177)
Diss. main cities				0.0776***				0.0762***
				(0.0122)				(0.0122)
Diss. rail network				0.0280				0.0335
				(0.0210)				(0.0211)
Diss. coalfields				-0.326***				-0.330***
				(0.0509)				(0.0509)
Observations	19,926	19,926	19,926	19,926	19,926	19,926	19,926	19,926
R^2	0.005	0.155	0.156	0.370	0.005	0.155	0.157	0.370
Year FE	✓	✓	✓	✓	✓	✓	✓	✓
i industry FE	✓	✓	✓	✓	✓	✓	✓	✓
j industry FE	✓	✓	✓	✓	✓	✓	✓	✓
IO Table	Horrell	Horrell	Horrell	Horrell	Thomas	Thomas	Thomas	Thomas

Robust standard errors are clustered at the industry pair level and reported in parenthesis. ***, ** and * respectively indicate 0.01, 0.05 and 0.1 levels of significance.

Table A2: Alternative distance thresholds (v1)

VARIABLES	(1) DO	(2) DO	(3) DO	(4) DO	(5) DO	(6) DO	(7) DO	(8) DO
Input-Output	0.0344***	0.0461***	0.0623***	0.0856***	0.0773***	0.0819***	0.0543***	0.0610***
	(0.0123)	(0.0155)	(0.0204)	(0.0294)	(0.0273)	(0.0250)	(0.0192)	(0.0189)
Labour Pooling	0.235***	0.254***	0.277***	0.277***	0.241***	0.225***	0.163***	0.159***
	(0.0140)	(0.0155)	(0.0176)	(0.0209)	(0.0205)	(0.0193)	(0.0167)	(0.0160)
Knowledge Spillovers	0.0605**	0.0834***	0.120***	0.167***	0.134***	0.0712	0.0148	0.00901
	(0.0290)	(0.0321)	(0.0370)	(0.0466)	(0.0461)	(0.0458)	(0.0433)	(0.0397)
Diss. agriculture	0.0309*	0.0322*	0.0306	0.0160	-0.00285	-0.0150	-0.0328	-0.0300
	(0.0165)	(0.0180)	(0.0204)	(0.0258)	(0.0252)	(0.0263)	(0.0235)	(0.0216)
Diss. mines	-0.00245	-0.000975	0.00698	0.0277*	0.0401**	0.0180	0.0151	0.0512***
	(0.0124)	(0.0136)	(0.0148)	(0.0162)	(0.0164)	(0.0163)	(0.0157)	(0.0140)
Diss. Gas & electricity	0.0393**	0.0491***	0.0613***	0.0782***	0.0695***	0.0653**	0.0286	0.0420*
	(0.0163)	(0.0181)	(0.0209)	(0.0260)	(0.0258)	(0.0260)	(0.0232)	(0.0228)
Observations	19,926	19,926	19,926	19,926	19,926	19,926	19,926	19,926
R^2	0.143	0.153	0.161	0.124	0.085	0.073	0.041	0.046
Year FE	✓	✓	✓	✓	✓	✓	✓	✓
i industry FE	✓	✓	✓	✓	✓	✓	✓	✓
j industry FE	✓	✓	✓	✓	✓	✓	✓	✓
Distance threshold	15	21	30	60	90	120	150	180

Robust standard errors are clustered at the industry pair level and reported in parenthesis. ***, ** and * respectively indicate 0.01, 0.05 and 0.1 levels of significance.

Table A3: No employment weights

VARIABLES	(1) DO	(2) DO	(3) DO	(4) DO	(5) DO	(6) DO
Input-Output	0.0921***			0.0184	0.0253	0.00926
	(0.0185)			(0.0153)	(0.0181)	(0.0158)
Labour Pooling		0.251***		0.230***	0.222***	0.168***
		(0.0175)		(0.0160)	(0.0160)	(0.0133)
Knowledge Spillovers			0.334***	0.147***	0.142***	0.119***
			(0.0436)	(0.0350)	(0.0353)	(0.0298)
Diss. agriculture					0.0245	0.0240*
					(0.0178)	(0.0136)
Diss. mines					-0.0407**	-0.0442***
					(0.0165)	(0.0143)
Diss. Gas & electricity					0.0425**	0.0583***
					(0.0171)	(0.0157)
Diss. ports						-0.232***
						(0.0296)
Diss. road network						-0.121***
						(0.0143)
Diss. waterways						-0.130***
						(0.0211)
Diss. main cities						0.0240*
						(0.0123)
Diss. rail network						-0.00144
						(0.0218)
Diss. coalfields						-0.156**
						(0.0661)
Observations	19,926	19,926	19,926	19,926	19,926	19,926
R^2	0.018	0.111	0.025	0.117	0.120	0.242
Year FE	✓	✓	✓	✓	✓	✓
i industry FE	✓	✓	✓	✓	✓	✓
j industry FE	✓	✓	✓	✓	✓	✓

Robust standard errors are clustered at the industry pair level and reported in parenthesis. ***, ** and * respectively indicate 0.01, 0.05 and 0.1 levels of significance.

Table A4: Coagglomeration and Marshallian forces (IND95US-level)

VARIABLES	(1) DO	(2) DO	(3) DO	(4) DO	(5) DO	(6) DO
Input-Output	0.100***			0.0377	0.0415	0.0306
	(0.0365)			(0.0347)	(0.0388)	(0.0358)
Labour Pooling		0.220***		0.202***	0.197***	0.130***
		(0.0242)		(0.0229)	(0.0232)	(0.0170)
Knowledge Spillovers			0.279***	0.0701*	0.0665*	0.0434
			(0.0517)	(0.0400)	(0.0401)	(0.0365)
Diss. agriculture					-0.0130	0.0277
					(0.0265)	(0.0194)
Diss. Gas & electricity					0.0499	0.0294
					(0.0498)	(0.0418)
Diss. mines					-0.0243	-0.0414**
					(0.0213)	(0.0170)
Diss. ports						-0.277***
						(0.0317)
Diss. road network						-0.232***
						(0.0286)
Diss. waterways						-0.0457*
						(0.0250)
Diss. main cities						0.0555**
						(0.0236)
Diss. rail network						-0.00244
						(0.0235)
Diss. coalfields						-0.152***
						(0.0553)
Observations	4,680	4,680	4,680	4,680	4,680	4,680
R^2	0.027	0.119	0.020	0.125	0.126	0.356
Year FE	✓	✓	✓	✓	✓	✓
i industry FE	✓	✓	✓	✓	✓	✓
j industry FE	✓	✓	✓	✓	✓	✓
IO Table	Meyer	-	-	Meyer	Meyer	Meyer

Robust standard errors are clustered at the industry pair level and reported in parenthesis. ***, ** and * respectively indicate 0.01, 0.05 and 0.1 levels of significance.

Table A5: Coagglomeration and Marshallian forces (60-year changes)

VARIABLES	(1) Residuals	(2) Residuals	(3) Residuals	(4) Residuals	(5) Residuals	(6) Residuals
Input-Output	0.0672***			0.0410***	0.0547***	0.0573***
	(0.0153)			(0.0144)	(0.0153)	(0.0141)
Labour Pooling		0.102***		0.0849***	0.0898***	0.116***
		(0.0152)		(0.0152)	(0.0158)	(0.0158)
Knowledge Spillovers			0.127***	0.0382	0.0448	0.0471
			(0.0412)	(0.0385)	(0.0391)	(0.0363)
Diss. agriculture					0.0554**	0.0577***
					(0.0231)	(0.0194)
Diss. mines					0.0246**	0.0221*
					(0.0121)	(0.0114)
Diss. Gas & electricity					0.0196	-0.000545
					(0.0300)	(0.0294)
Diss. ports						0.165***
						(0.0239)
Diss. road network						0.0664***
						(0.0161)
Diss. waterways						-0.0394**
						(0.0178)
Diss. main cities						-0.0439***
						(0.0139)
Diss. rail network						0.145***
						(0.0240)
Diss. coalfields						-0.0343
						(0.0635)
Observations	3,321	3,321	3,321	3,321	3,321	3,321
R^2	0.015	0.030	0.006	0.036	0.040	0.111
Year FE	✓	✓	✓	✓	✓	✓
i industry FE	✓	✓	✓	✓	✓	✓
j industry FE	✓	✓	✓	✓	✓	✓
IO Table	Meyer	-	-	Meyer	Meyer	Meyer

Robust standard errors are clustered at the industry pair level and reported in parenthesis. ***, ** and * respectively indicate 0.01, 0.05 and 0.1 levels of significance.

B Descriptive statistics

Table B1: Fifteen most co-agglomerated industry pairs (within a 30km radius)

DO_30km	i	i_name	j	j_name
0.610603	933	Lacquerers and Enamellers and Japanners	837	Locksmiths
0.599434	941	Musical Instrument Makers and Tuners	775	Dairy Product Processors
0.566352	837	Locksmiths	836	Gunsmiths
0.549998	941	Musical Instrument Makers and Tuners	926	Bookbinders and Related Workers
0.535618	933	Lacquerers and Enamellers and Japanners	836	Gunsmiths
0.531237	941	Musical Instrument Makers and Tuners	742	Cookers, Roasters and Related Heat Treaters
0.524353	926	Bookbinders and Related Workers	775	Dairy Product Processors
0.514231	928	Textile printers	751	Fibre Preparers
0.512703	775	Dairy Product Processors	742	Cookers, Roasters and Related Heat Treaters
0.501151	752	Spinners, doublers, twisters and winders	751	Fibre Preparers
0.485555	941	Musical Instrument Makers and Tuners	781	Tobacco Preparers
0.481279	926	Bookbinders and Related Workers	742	Cookers, Roasters and Related Heat Treaters
0.478517	959	Construction Workers Not Elsewhere Classified	941	Musical Instrument Makers and Tuners
0.475218	941	Musical Instrument Makers and Tuners	772	Sugar Processors and Refiners
0.466843	781	Tobacco Preparers	775	Dairy Product Processors

Table B2: Most agglomerated industries (within a 30km radius)

Code	Title	Agglomeration
837	Locksmiths	0.713144
892	Potters and Related Clay and Abrasive Formers	0.624576
941	Musical Instrument Makers and Tuners	0.620596
775	Dairy Product Processors	0.568008
751	Fibre Preparers	0.557975
836	Gunsmiths	0.533006
928	Textile printers	0.512283
933	Lacquerers and Enammellers and Japanners	0.496948
926	Bookbinders and Related Workers	0.481448
752	Spinners, doublers, twisters and winders	0.46294

Table B3: Fifteen Industry pairs reporting the highest score of labour pooling

i	j	i_name	j_name	LP
934	819	Gilders	Cabinetmakers and Related Woodworkers Not Elsewhere Classified	0.93754
746	745	Ink, Paint and Dye Makers	Petroleum-Refining Workers	0.924931
924	921	Printing Engravers (except Photo-Engravers)	Compositors and Type-Setters	0.923809
929	921	Printers and Related Workers Not Elsewhere Classified	Compositors and Type-Setters	0.92168
934	812	Gilders	Woodworkers	0.903701
929	928	Printers and Related Workers Not Elsewhere Classified	Textile printers	0.901215
910	733	Paper and Paperboard Products Makers	Paper Pulp Preparers	0.867366
880	842	Jewellery and Precious Metal Workers	Watch, Clock and Precision Instrument Makers	0.866437
722	720	Metal Rolling-Mill Workers	Metal Processors, Specialisation Unknown	0.864895
819	817	Cabinetmakers and Related Woodworkers Not Elsewhere Classified	Box makers	0.86299
814	813	Wheelwrights	Coach, carriage and wagon makers	0.860299
819	812	Cabinetmakers and Related Woodworkers Not Elsewhere Classified	Woodworkers	0.852971
873	722	Sheet-metal Workers	Metal Rolling-Mill Workers	0.822349
811	796	Cabinetmakers	Upholsterers and Related Workers	0.820248
722	721	Metal Rolling-Mill Workers	Metal Smelting, Converting and Refining Furnacemen	0.815119

Table B4: Fifteen Industry pairs reporting the highest score of input sharing (Meyer)

i	j	i_name	j_name	IO_Meyer
949	942	Other Production and Related Workers	Basketry Weavers and Brush Makers	0.328087
949	941	Other Production and Related Workers	Musical Instrument Makers and Tuners	0.260629
942	941	Basketry Weavers and Brush Makers	Musical Instrument Makers and Tuners	0.257722
713	711	Well-Drillers, Borers and Related Workers	Miners and Quarrymen	0.244503
754	752	Weavers and Related Workers	Spinners, doublers, twisters and winders	0.19892
803	761	Leather Goods Makers	Tanners and Fellmongers	0.18804
778	773	Brewers, Wine and Beverage Makers	Butchers and Meat Preparers	0.186298
871	711	Plumbers and Pipe Fitters	Miners and Quarrymen	0.176847
851	711	Electrical Fitters	Miners and Quarrymen	0.176396
931	921	Painters, Construction	Compositors and Type-Setters	0.174646
773	771	Butchers and Meat Preparers	Grain Millers and Related Workers	0.169799
781	773	Tobacco Preparers	Butchers and Meat Preparers	0.160402
896	803	Lime and Cement Makers	Leather Goods Makers	0.15656
901	803	Rubber and Plastics Product Makers (except Tire Makers and Tire Vulcanisers)	Leather Goods Makers	0.154423
774	773	Food Preservers	Butchers and Meat Preparers	0.154264

Table B5: Most frequent technical terms

Rank	Term	Frequency
1	elevation	11,789
2	screw	8,752
3	spring	8,066
4	plate	7,823
5	metal	7,596
6	edge	7,307
7	slide	6,225
8	frame	6,156
9	sheet	6,151
10	wheel	5,882
11	rod	5,815
12	slot	5,773
13	lever	5,762
14	pressure	5,617
15	arm	5,605
16	angle	5,470
17	force	5,362
18	press	5,337
19	bar	5,097
20	shaft	4,897
21	tube	4,739
22	head	4,643
23	pivot	4,578
24	motion	4,504
25	groove	4,500
26	bolt	4,375
27	projection	4,291
28	nut	4,048
29	ring	4,030
30	joint	4,027

The 30 terms listed are the most frequent technical words in the patent corpus retained to construct pairwise Jaffe similarity indices. Patent-level indices are then aggregated at the industry-pair level to yield our knowledge spillover measure; see Section 4.2.3 for a full description.

Table B6: Fifteen Industry pairs reporting the highest score of knowledge spillovers

i	j	i_name	j_name	KS
844	834	Automobile manufacturing	Machine-Tool Operators	0.1080309
834	757	Machine-Tool Operators	Rope Makers	0.1073503
757	713	Rope Makers	Well-Drillers, Borers and Related Workers	0.1056166
836	834	Gunsmiths	Machine-Tool Operators	0.1045727
834	722	Machine-Tool Operators	Metal Rolling-Mill Workers	0.103649
834	832	Machine-Tool Operators	Toolmakers, Metal Pattern Makers and Metal Markers	0.1032663
834	724	Machine-Tool Operators	Metal Casters	0.1025112
836	832	Gunsmiths	Toolmakers, Metal Pattern Makers and Metal Markers	0.1025088
781	713	Tobacco Preparers	Well-Drillers, Borers and Related Workers	0.1024904
841	834	Machinery Fitters and Machine Assemblers	Machine-Tool Operators	0.1022783
814	757	Wheelwrights [and cartwrights]	Rope Makers	0.1021877
929	722	Printers and Related Workers Not Elsewhere Classified	Metal Rolling-Mill Workers	0.1021106
837	836	Locksmiths	Gunsmiths	0.102054
842	834	Watch, Clock and Precision Instrument Makers	Machine-Tool Operators	0.1018753
844	836	Automobile manufacturing	Gunsmiths	0.1017581

C Local industrial agglomeration

In Section 2.4, we documented that the aggregate decline in local industrial concentration over the period 1851–1911 masks two opposing forces: a diversification of the national industry mix and an increase in within-region specialisation. This appendix formalises that decomposition. We first define a measure of regional industrial concentration based on the Herfindahl index (Section C.1), and then use a two-factor Shapley decomposition to separate the observed change into a national mix effect and a local allocation effect (Section C.2).

C.1 Regional industrial concentration

Let regions be indexed by $r = 1, \ldots, R$, sectors by $k = 1, \ldots, K$, and time by $t \in \{0, 1\}$. Let e_{rkt} denote employment in sector k and region r at time t, and define the regional employment share of sector k in region r as:

$$s_{rkt} = \frac{e_{rkt}}{E_{rt}}, \tag{B1}$$

where $E_{rt} = \sum_{k=1}^{K} e_{rkt}$ is total employment in region r. The regional Herfindahl index is then:

$$H_{rt} = \sum_{k=1}^{K} s_{rkt}^2. \tag{B2}$$

A higher value of H_{rt} indicates that employment in region r is concentrated in fewer sectors. To obtain a single national summary measure, we compute the employment-weighted mean of regional indices:

$$H_t = \sum_{r=1}^{R} w_{rt} \, H_{rt}, \tag{B3}$$

where $w_{rt} = E_{rt}/E_t$ and $E_t = \sum_{r=1}^{R} E_{rt}$. A decline in H_t over time could reflect either a more diversified national industry mix – with employment spreading across a larger number of sectors – or a more even spatial allocation of industries across regions, or both. To disentangle these two channels, we turn to a Shapley-based decomposition.

C.2 Decomposition via two-factor Shapley values

Shapley values (Shapley et al., 1953) provide a method to decompose aggregate outcomes into the contributions of individual factors. In contrast to sequential decompositions,

which are sensitive to the order in which factors are varied, the Shapley approach averages marginal contributions across all possible factor orderings, yielding an exact and order-independent decomposition. As shown by Shorrocks et al. (2013), this framework offers a unified approach to distributional analysis, and it has since been widely applied in studies of income inequality and poverty dynamics – among other topics.

We apply this framework to decompose the change in H_t into two components: one driven by the evolution of the national industry mix, and one driven by changes in the local allocation of industries across regions. To do so, we write employment as the product of two factors:

$$e_{rkt} = E_{kt}\, a_{rkt}, \tag{B4}$$

where $E_{kt} = \sum_{r=1}^{R} e_{rkt}$ is total national employment in sector k, and

$$a_{rkt} := \frac{e_{rkt}}{E_{kt}}, \qquad \text{so that} \qquad \sum_{r=1}^{R} a_{rkt} = 1 \;\; \forall\, k, t, \tag{B5}$$

is the share of sector k's national employment located in region r. The vector $E_t = (E_{1t}, \ldots, E_{Kt})$ captures the *national industry mix*, while the matrix $a_t = \{a_{rkt}\}_{r,k}$ captures the *spatial allocation* of industries across regions.

For any pair of factor values (E, a), we can compute the implied regional employment shares, Herfindahl indices, and weights:

$$s_{rk}(E, a) = \frac{E_k\, a_{rk}}{\sum_{j=1}^{K} E_j\, a_{rj}}, \tag{B6}$$

$$H_r(E, a) = \sum_{k=1}^{K} s_{rk}(E, a)^2, \tag{B7}$$

$$H(E, a) = \sum_{r=1}^{R} w_r(E, a)\, H_r(E, a), \tag{B8}$$

where

$$w_r(E, a) = \frac{\sum_{k=1}^{K} E_k\, a_{rk}}{\sum_{r'=1}^{R} \sum_{k=1}^{K} E_k\, a_{r'k}}. \tag{B9}$$

The observed change in the national Herfindahl index between $t = 0$ and $t = 1$ is:

$$\Delta H = H(E^1, a^1) - H(E^0, a^0). \tag{B10}$$

The two-factor Shapley decomposition splits ΔH exactly into two components. The *national mix effect* captures the change attributable to shifts in the sectoral composition of

national employment:

$$\Delta H^{\text{mix}} = \frac{1}{2}\Big[H(E^1, a^0) - H(E^0, a^0) \;+\; H(E^1, a^1) - H(E^0, a^1)\Big]. \tag{B11}$$

The *local allocation effect* captures the change attributable to the spatial redistribution of industries across regions, holding the national mix fixed:

$$\Delta H^{\text{local}} = \frac{1}{2}\Big[H(E^0, a^1) - H(E^0, a^0) \;+\; H(E^1, a^1) - H(E^1, a^0)\Big]. \tag{B12}$$

By construction, the decomposition is exact:

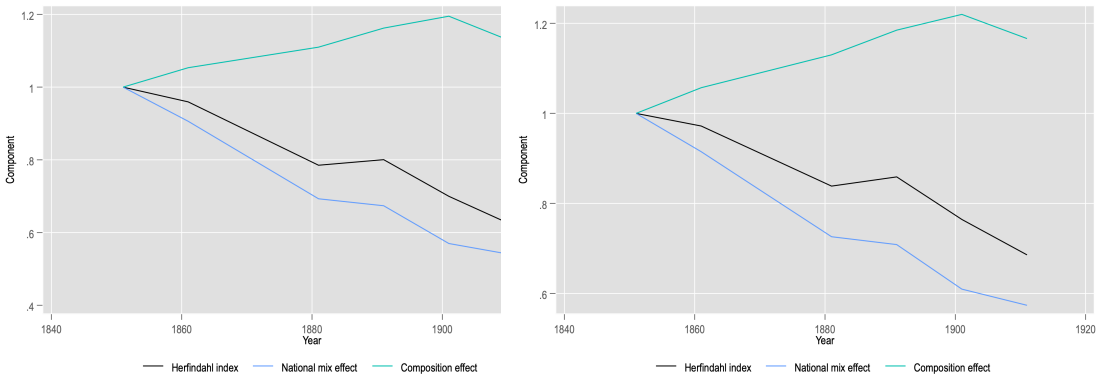
$$\Delta H = \Delta H^{\text{mix}} + \Delta H^{\text{local}}. \tag{B13}$$

As documented in Section 2.4, we find that $\Delta H < 0$ – meaning that aggregate local concentration declined over the period – but this was entirely driven by $\Delta H^{\text{mix}} < 0$, reflecting the diversification of the national economy as new sectors emerged. The local allocation effect, ΔH^{local}, is positive: holding the industry mix constant, regions became *more* specialised over time, consistent with the strengthening of agglomeration forces during the later stages of the industrial revolution.

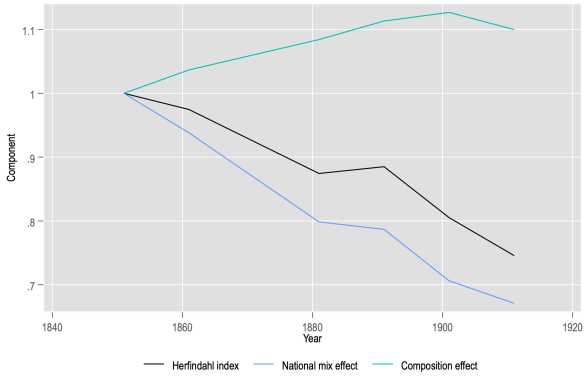
Figure B1: Shapley decomposition of ΔH

(a) Parish

(b) District

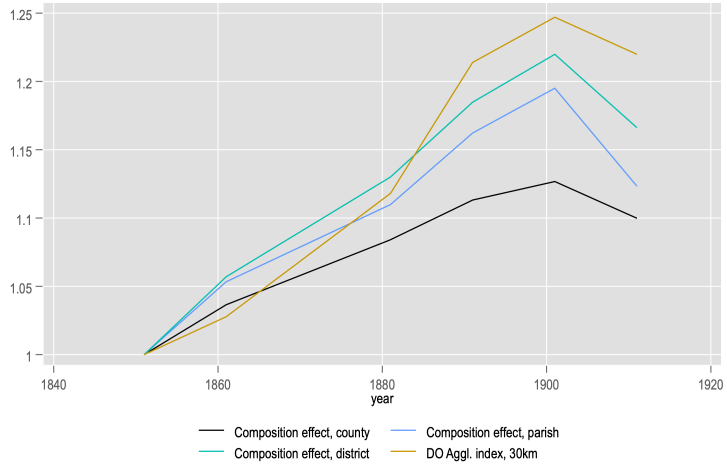


(c) County



Source: Authors' calculations based on the Historical British Censuses.

Figure B2: Agglomeration trend



Source: Authors' calculations based on the Historical British Censuses.

D Measuring innovativeness

To measure the degree of innovativeness of each industry, we follow Kelly et al. (2021), who propose a text-based measure of patent importance that captures both the novelty and the impact of an invention. The key idea is that a breakthrough patent is one that is textually dissimilar to what came before it (*novel*) but textually similar to what comes after it (*impactful*). We implement this approach in five steps.

1. **Term weighting.** For each term w in patent i filed at time t, we compute the term frequency–backward inverse document frequency:

$$TFBIDF_{w,i,t} = TF_{w,i} \times BIDF_{w,t}, \tag{B14}$$

   where $TF_{w,i}$ is the frequency of word w in patent i and $BIDF_{w,t}$ is the backward inverse document frequency of term w at time t, which down-weights terms that are common in the existing patent corpus up to time t.

2. **Normalisation and cosine similarity.** We normalise each patent's TFBIDF vector to unit length and compute pairwise cosine similarity:

$$V_{i,t} = \frac{TFBIDF_{i,t}}{\|TFBIDF_{i,t}\|}, \tag{B15}$$

$$\rho_{i,j} = V_{i,t} \cdot V_{j,t}, \tag{B16}$$

   where $\rho_{i,j} \in [0,1]$ is the cosine similarity between patents i and j, computed as the dot product of their normalised TFBIDF vectors, and $t \equiv \min(i,j)$.

3. **Backward similarity.** We measure how similar patent j is to the prior art:

$$BS_j^\tau = \sum_{i \in \mathcal{B}_{j,\tau}} \rho_{j,i}, \tag{B17}$$

   where $\mathcal{B}_{j,\tau}$ denotes the set of patents filed in the τ calendar years preceding patent j's filing date. A low value of BS_j^τ indicates that the patent is textually distant from its predecessors – i.e., that it is novel.

4. **Forward similarity.** We measure the subsequent influence of patent j:

$$FS_j^\tau = \sum_{i \in \mathcal{F}_{j,\tau}} \rho_{j,i}, \tag{B18}$$

where $\mathcal{F}_{j,\tau}$ denotes the set of patents filed in the τ calendar years following patent j's filing date. A high value of FS_j^τ indicates that the patent's language was widely adopted by subsequent inventions – i.e., that it was impactful.

5. **Patent importance.** We combine novelty and impact into a single measure:

$$q_j^\tau = \frac{FS_j^\tau}{BS_j^\tau}, \tag{B19}$$

where q_j^τ captures the importance of patent j over a τ-year horizon. Patents with high q_j^τ are those that are both novel (low backward similarity) and impactful (high forward similarity). Following Kelly et al. (2021), we classify a patent as a breakthrough innovation if q_j^τ exceeds a given threshold, and compute the share of breakthrough patents in each sector's total patent output as our industry-level measure of innovativeness.