Human capital shocks and innovation: Evidence from Britain's Lost Generation

Jan Bietenbeck[1]

Lund University

Davide Cipullo[2]

Università Cattolica del Sacro Cuore

Edward Pinchbeck[3]

University of Birmingham

Luca Repetto[4]

Uppsala University

Abstract

This paper investigates the long-term impact on innovation of the human-capital shock to British communities caused by World War I (WWI) soldier losses. Using a novel dataset linking parish-level WWI mortality to patent data (18951979) and inventor characteristics, we compare innovation in high- versus low-mortality areas. Differenceindifferences results show a persistent decline in patenting: a 10% increase in deaths reduces the likelihood of innovation by 0.080.11 percentage points and the likelihood of top innovations by roughly three times as much. Effects are larger when highskill individuals (elite graduates or engineer officers) are lost and are mitigated by innovationsupporting ecosystems (proximity to universities or connective infrastructure). At the inventor level, inventors in highmortality parishes patent less after the war, especially those with engineering background or working in advanced sectors. The adverse effects are partly offset by moving to less affected areas or by coauthorship, highlighting the role of complementary workers and innovation networks.

Keywords: World War; Innovation; Human Capital; Patents; Lost Generation

JEL classification: D74, O15, O31

First version: Sept. 2024. This version: February 2026

---

[1]Department of Economics, Lund University. Email: jan.bietenbeck@nek.lu.se

[2]Department of Economics, Università Cattolica del Sacro Cuore. Email: davide.cipullo@unicatt.it

[3]Department of Economics, University of Birmingham. Email: e.w.pinchbeck@bham.ac.uk.

[4]Department of Economics, Uppsala University. Email: luca.repetto@nek.uu.se.

We would like to thank Jingyuan Zeng, Gustaf Holmberg and Rebecca Wiechers for excellent research assistance, as well as Felipe Carozzi, Adrian Gortzak, Elias Papaioannou, Vincenzo Scrutinio, Marco Tabellini and participants to seminars held at ASREC Europe Conference 2023 and 2025, University of Birmingham, Uppsala University, Università Cattolica in Milan, UIL/HEUU/Urban lab workshop in Sigtuna, ASWEDE 2024, SIEP 2025, and Università di Bologna for comments and suggestions. This work was supported by BA/Leverhulme Small Research Grant SRG2021-210936, Jan Wallanders och Tom Hedelius stiftelse samt Tore Browaldhs stiftelse grant P22-0060, and the Centre for Economic Performance. This work is based on data provided through www.VisionofBritain.org.uk and uses statistical material which is copyright of the Great Britain Historical GIS Project, Humphrey Southall and the University of Portsmouth.

1. Introduction

Innovation is a key driver of economic growth (e.g. Romer, 1990; Aghion and Howitt, 1992; Glaeser et al., 1992), so understanding its determinants is crucial. Innovative activity is highly spatially concentrated and persistent over long periods (Audretsch and Feldman, 1996; Akcigit, Grigsby and Nicholas, 2017*b*; Andrews, 2023), yet history offers stark examples of rapid, long-run decline in once-leading innovation hubs. Existing research highlights the importance of local institutions and policy (Akcigit et al., 2022; Andrews, 2023) on the one hand, and individual endowments and exposure to innovative role models (Bell et al., 2019) on the other. The relative importance of these forces – and how they shape longrun innovation dynamics – remains unclear, in part because studies differ in setting, outcome measures, and identification strategies.

This paper examines how a large, sudden shock to communities' upper-tail human capital – the mortality due to World War I (WWI) military deaths – affects long-run innovation in Britain. Mortality during the War differed greatly across space and disproportionately fell on young and highly educated men – the so-called "lost generation" (Winter, 1977) – creating a substantial and plausibly exogenous shock to the human capital stock of British communities that enables us to assess both environmental and individual determinants of innovation within the same setting. Because fighting took place abroad, British communities experienced no destruction of physical capital, allowing us to focus on the consequences of human capital loss. Such a shock could reduce innovation both by lowering the productivity of existing inventors and by limiting the emergence of new ones.[1]

Our main results show that WWI mortality had large and persistent effects on innovation. A 10% increase in deaths lowers the probability that a parish produces any patent in a fiveyear period by about 0.080.11 percentage points, with effects roughly three times larger for breakthrough patents. These declines are concentrated in technically advanced fields and among engineers, and are mitigated by inventor mobility and coauthorship.

To document these effects, we construct a novel dataset that combines WWI mortality records with patent data for all of England and Wales, 1895–1979. Mortality and war participation come from linked Commonwealth War Graves Commission death records and military service data. Innovation is measured by linking PatentCity (geocoded patent filings aligned to 1911 parish boundaries) with PATSTAT, which provides inventor identifiers and patent abstracts. Together, these sources cover the universe of WWI deaths and all patents filed in England and Wales between 1895 and 1979, enabling analysis at both the parish – our community measure – and inventor levels.

The first part of our analysis examines how WWI fatalities affected patenting at the parish level. Because patenting is rare at this level – only 19 percent of parishes register a patent in any given year – we aggregate to five-year periods and use as a baseline outcome an indicator equal to one if a parish registers at least one patent. Building on Kelly et al.

[1]Evidence from several settings indicates that individuals who are more educated and those who are exposed to innovation during childhood are more likely to become inventors (see, e.g., Akcigit, Grigsby and Nicholas 2017*b*; Aghion et al. 2023; Berger and Prawitz 2024).

(2021), we also construct NLP-based indicators of "breakthrough" patents.[2] Our main explanatory variable is the log number of WWI deaths in a parish, and we estimate difference-in-differences and event-study specifications that compare trajectories across higher- versus lower-mortality parishes, with parish and period fixed effects. Our estimates indicate a significant and long-lasting negative effect: a 10 percent increase in WWI deaths reduces the probability that a parish registers at least one patent in a five-year period by about 0.080.11 percentage points. The effect appears immediately with the onset of WWI and, aside from a modest interwar recovery, remains persistent until 1979. It is much more pronounced for "breakthrough" patents – roughly three times as large at the very top of the quality distribution – suggesting that higher-quality innovation is more susceptible to the shock. Taken together, these results indicate that WWI altered both the geography and the quality distribution of British invention for decades.

Our continuous-treatment DiD relies on conditional parallel trends (Callaway, Goodman-Bacon and Sant'Anna, 2024), an assumption that would not be required under exogenous treatment assignment. Because WWI mortality is unlikely to be fully exogenous – for example, declining communities may have sent less prepared soldiers to the front – we instrument for parish deaths using a shift-share IV that follows Carozzi, Pinchbeck and Repetto (2023) and exploits plausibly exogenous battalion assignments and battalion-level death rates. The IV results confirm our main findings.

We then investigate the mechanisms behind the negative impact of WWI mortality on innovation. First, we test the "lost generation" hypothesis – that the loss of highly educated and skilled individuals reduced communities' long-run innovative capacity. Second, we explore heterogeneity in local innovation ecosystems that may mitigate or amplify the shock. Third, we assess whether the effect of WWI mortality on innovation productivity goes beyond the mechanical loss of deceased inventors.

In testing the "lost generation" hypothesis, we show that, conditional on total WWI deaths, the loss of an Oxford or Cambridge alumn leads to a large additional decline in subsequent innovation. Using detailed soldier records, we also show that the death of an officer with an engineering background has an additional negative effect, whereas the loss of other types of officers does not.[3] Second, we examine whether the effect of WWI mortality varies with features of the local innovation ecosystem, which prior work has shown to matter (see, e.g., Akcigit, Grigsby and Nicholas 2017*b*; Bell et al. 2019; Andrews 2023). The effect is mitigated in parishes with access to railway stations and post offices and in proximity to universities, suggesting that ecosystem advantages can partially offset human-capital losses. Third, we provide evidence that the mortality shock both reduced the productivity of existing inventors and made communities less attractive in the long run. The persistence of the effect through

[2]Focusing on breakthrough patents helps address the well-known issue that patents are an imperfect measure of innovation, as not all innovations are patented and not all patents represent true innovation (e.g. Griliches, 1990; Nagaoka, Motohashi and Goto, 2010).

[3]This latter finding aligns with recent evidence that engineers are pivotal to the production of innovative output (Hanlon, 2022) and underscores the role of technically oriented individuals in driving technology adoption and economic growth (Maloney and Valencia Caicedo, 2022).

the late 1970s indicates that part of the impact reflects reduced entry of new inventors, since most individuals old enough to serve in WWI would have been deceased by then. We also show that our findings hold when redefining the dependent variable to distinguish patents by new inventors – those who register their first patent after the War – from those by inventors who had already produced at least one patent prior to WWI. Together, these results rule out a purely mechanical decline from inventor deaths and motivate a closer investigation of how individual inventors are affected by the mortality shock.

To this end, we construct an inventor-level panel of roughly 18,000 patentees who registered at least one patent in 1895–1899 and follow them through 1979. Inventors in higher-mortality parishes patent less after the war, and the effect is larger for more innovative patents, consistent with the parish-level results.[4] In addition, we find that the impact of WWI mortality is not uniform across technological fields. Instead, it is larger in advanced sectors such as electricity, machinery, transportation, physics, and chemistry, while it is modest and not statistically significant in human necessities and construction. It is also larger for inventors who report engineering occupations and for those whose first patents are more complex or more important, defined using the importance metric in Kelly et al. (2021).[5] Overall, the declines are concentrated among inventors operating in more skill-intensive and frontier domains (Smil, 2005), consistent with innovation in these sectors relying on specialized teams, complementary technicians and engineers, and supporting infrastructure, so losses of highly trained workers and collaborators translate into larger and more persistent reductions in patenting output (Neffke, 2019; Jaravel, Petkova and Bell, 2018; Hanlon, 2022).

Lastly, we examine whether inventors can mitigate the adverse effect of mortality on productivity. We consider three margins: switching field, altering co-authoring strategy, and relocating. Relocation to another parish offsets the shock, whereas the negative effect is amplified for inventors who remain in their initial location. This finding aligns with prior evidence on exposure effects (Bell et al., 2019), suggesting that moving to highly innovative areas raises inventorship through improved human capital transmission, better infrastructure, enhanced networks, or a combination of these factors. By contrast, switching sectors does not mitigate the impact, possibly because retraining is costly. Co-authorship, instead, provides a buffer: while the likelihood of producing solo-authored patents drops substantially for inventors in high-mortality areas, the probability of patenting with at least one co-author is unaffected, indicating that access to co-inventors and broader collaboration networks can mitigate productivity losses following shocks (Jaravel, Petkova and Bell 2018).

Our paper contributes to the large body of research on the determinants of innovation, and its distribution in space and over time. Recent contributions to this literature have highlighted the importance of individual background as well as external factors, such as exposure to an innovative environment (Bell et al., 2019; Akcigit, Grigsby and Nicholas, 2017*a*;

[4]The effect is similar when we exclude WWI fatalities and when we restrict to inventors that produced at least one patent over the post-WWI period, indicating that our estimates are not driven by the mechanical loss of inventors.

[5]To ensure comparability across specifications, we assign to each inventor a field, an occupation, and measures of patent complexity based on the first patent that they publish in 1895–1899.

Andrews, 2023). Relatively few studies, however, have examined how modern innovation patterns relate to historical shocks.[6] Given the strong persistence of innovation clusters, (e.g., Andrews and Whalley, 2022), this represents an important gap in the literature. Our results provide new evidence on the importance of a large shock to the long-run innovative potential of communities, both in terms of quantity and of quality. Also, we are able to bridge the gap between studies analysing community-level outcomes and those focusing on individual inventors by providing evidence on both using the same setting and a long-run perspective.

We add to the literature on the relationship between labour scarcity and technological change (Acemoglu, 2007; Voth, Caprettini and Trew, 2023; Andersson, Karadja and Prawitz, 2022). Previous theoretical and empirical work has examined the conditions under which economies respond to a shortage of workers by adopting or developing labour-saving technologies. Our findings show, instead, that the loss of upper-tail human capital produces a persistent decline in innovative capacity. While this result may seem at odds with earlier studies, it can be reconciled by noting that the prior literature has focused on low-skilled labour, whose tasks were more readily substituted by machines.

Our work also contributes to the literature on the long-run economic consequences of war.[7] Prior work finds heterogeneous effects across settings and conflict types: bombings can have persistent negative impacts (Lin, 2022; Riaño and Valencia Caicedo, 2024), yet population and activity often recover (Davis and Weinstein, 2002; Brakman, Garretsen and Schramm, 2004). Chupilkin and Koczan (2022) show that long-run GDP effects vary with conflict characteristics, with on-territory wars yielding larger persistent losses than off-territory wars. Our setting isolates military human-capital losses without physical capital destruction and at a granular spatial scale, allowing us to show that these losses depress community-level innovation for decades. Besides this finding, we also advance this literature by tracing the evolution of outcomes before, during, and after the war. This allows us to validate our results by inspecting pre-trends and the dynamics of the treatment effect, mitigating concerns about time-varying shocks and the compression of history between treatment and a later outcome in persistence studies (Voth, 2021).

2. Background

2.1. The evolving patent landscape in Britain

The British Patent Office was first established in 1852. In 1883 the Patents Act significantly reduced patent fees and around the same time, Britain began to provide full patent rights to signatory countries under an international convention. From 1902 patents became

[6]A notable exception to this lack of long-term focus is the work on the impact of the Nazi regime on science. Waldinger (2016) use WW2 destruction as a shock to physical and human capital and show the dismissal of scientists in Nazi Germany's had long-lasting negative on academic productivity and the capacity of departments to attract new talent.

[7]Many papers also consider how war-related losses impact in the short run, including on marriage patterns, female labour force participation and gender norms (Abramitzky, Delavande and Vasconcelos, 2011; Boehnke and Gay, 2020; Brainerd, 2017), while Carozzi, Pinchbeck and Repetto (2023) find that WWI mortality led to the creation of civic capital in Britain. A broader body of related papers also examine other mortality shocks, for example due to infectious diseases, see e.g. Franck (2024) for an example that also looks at innovation.

subject to an investigation into the novelty of an invention prior to grant. The legislation was modernised by an Act in 1977 which underpins the current institutional framework. The number of patents registered in Britain increased steadily over the time period we consider in our analysis (1895-1979), albeit patenting activity was reduced during both world wars. There was also a marked shift towards corporate invention over the period. Until the 1920s, the vast majority of patents were registered to individual inventors, but by 1950 companies were registering as many patents as individuals.

There is no single dataset that provides comprehensive information about the age, occupation, education and income of inventors in England and Wales in the first half of the 20^{th} century - few individuals who generated patents described their occupation as inventors in the Census, and patent databases do not contain comprehensive inventor characteristics. To get some understanding of inventor characteristics, we therefore attempted to match all 36,101 inventors who generated a patent in the period 1906-1916 to individuals in the 1911 Census using name and location variables. We describe this exercise and compare inventor characteristics to those of the wider population in Appendix E. We successfully matched 23,825 inventors (66%). These inventors are overwhelmingly male, and on average older than the general adult population (means of 41 and 35 years old). They were also relatively wealthier as measured by servants and household rooms. The PatentCity data contains a self-reported occupation field, so we can also use this data to examine inventor occupations. In line with Nicholas (2011), we find that a remarkable proportion (47%) of patents in this period were registered to those who describe themselves as engineers. Unfortunately none of our data sources contain educational information. However, in a different context Akcigit, Grigsby and Nicholas (2017*b*) show that around 40% of US inventors in 1940 had a university degree (relative to 10% of non-inventors), suggesting that British inventors were also likely to be university-educated.

2.2. The British Armed Forces during WWI

Over 4.5 million men from England and Wales served in the British Army during WWI, and around 800,000 more served in the Royal Navy and what became the Royal Air Force (Winter, 1977). At the onset of the war, Britain's army comprised just over 240,000 professional soldiers, known as "regulars." Early in the war, heavy casualties among these regulars reduced their numbers dramatically: by the end of 1914, nearly one-third had been killed, with many more wounded or missing (Travers, 1994). In response, the British Army expanded rapidly. Initially, a call for volunteers provided reinforcements, and in January 1916, conscription was introduced and required all unmarried British men aged 1941 to serve, with eligibility later extended to married men and younger age groups. Although exemptions were granted to workers in "reserved occupations" deemed essential to the war effort, scientific and technical experts often found themselves serving at the front. It was only after the high-profile death of physicist Henry Moseley and advocacy from scientists like Ernest Rutherford that recruitment policies were revised to protect some prominent scientists from combat.

Of those who served, around 330,000 were commissioned officers (Winter, 1977). At the beginning of the war, the 20,000 or so officers of the serving and reserve armed services were primarily drawn from Britain's upper middle and upper classes, recruited through prestigious public schools and universities. As the demand for officers grew, however, men from outside these traditional backgrounds were promoted from the ranks, earning the titles "temporary gentlemen" or "ranker officers." Initially, these officers – commissioned only for the duration of the war – were drawn wherever possible from the public schools. However, this was soon infeasible and by 1916 a merit-based approach had been adopted. This favoured commissioning NCOs who had proven battlefield experience as well as men that had received short military training courses at Officer Cadet Battalion (OCBs) that emphasised leadership, initiative, and social skills. The sheer numbers recruited in this way – approximately 205,000 temporary officers were in post by the end of the war – means that a large proportion of officers that served came from lower-middle and working-class backgrounds.

2.3. Mortality in WWI and the Lost Generation

The war was fought overseas, so war-related deaths in Britain were very highly concentrated in the military. Exact numbers remain uncertain, but it is estimated that over 750,000 members of the British armed forces died in the course of the war or shortly afterwards, whereas less than 17,000 civilians were killed as a result of enemy action. The vast majority of war-related fatalities were young men. After the war, the popular press regularly referred to "surplus women" and the 1921 Census enumerated 1.7 million more women than men, an increase of 500,000 compared to 1911. Similarly, the age-related nature of the mortality shock is clear. Some half a million men under 40 were killed, with the highest mortality rates being for cohorts aged under 30 (Winter, 1977). The cohort-specific nature of war-related mortality gives rise to one widely used version of the phrase "the lost generation".

The extent to which the mortality shock differentially affected groups in British society is debated. The phrase "lost generation" has also been used to evoke the idea that the war robbed the country of its best and brightest young minds. Certainly, many promising young men were killed; among them, the poets Rupert Brooke and Wilfred Owen, composer George Butterworth, and Raymond Asquith, Liberal candidate for Derby and son of then-Prime Minister H. H. Asquith. The intellectual toll also extended to the scientific community, where, in addition to Henry Moseley, prominent figures like Glasgow zoologist Charles Martin and Cambridge neuroscientist Keith Lucas were also killed. An article published in *The Times* in 1915 highlighted the "waste of brains" by sending promising scientists to the front.[8]

Systematic evidence supports the notion that death rates were highest in the most elite social groups. Officers, many of whom were tasked with leading frontline units over the top, suffered significantly higher death rates than enlisted soldiers throughout the war. This was particularly pronounced in 1914-15, when the death rate for officers was approximately 14 percent compared to 6 percent for other ranks (Winter, 1977). During this period, it was still seen as desirable to recruit officers from public schools and universities where possible. The

[8]"The waste of brains: young scientists in the fighting line," *The Times*, 24 December 1915, p.3.

result was that Britain's educated elite was decimated. Among Oxford and Cambridge graduates, for example, death rates were 19 percent and 18 percent across the war respectively, with rates exceeding 25 percent for students who matriculated between 1910 and 1914.

While the educated elite of Britain were disproportionately lost in the war, the composition of deaths amongst non-elites remains uncertain. Bailey, Hatton and Inwood (2023) examine a sample of 2,400 individuals born in the 1890s that served in the ranks as non-officers, and hence were unlikely to have received advanced education. Unsurprisingly, wartime death was more likely for those that served in the infantry and for those that enlisted earlier. In their sample, there is no evidence that coming from a white-collar household (headed by a professional or managerial worker) or favourable childhood environmental conditions predict early enlistment into the armed services or the likelihood of death conditional on time served. Indeed, relative to those from skilled, semi-skilled, and unskilled manual households, servicemen in the infantry from white collar households were found to have a significantly higher likelihood of survival.

The evidence from Bailey, Hatton and Inwood (2023) runs counter to an interpretation in which the more affluent members of the middle and working classes were disproportionately killed in the war. However, it does not shed any light on whether the fallen were more likely than others to support future innovations. As we described above, the majority of patents in the period we study were generated by individuals who described themselves as engineers. A much smaller share declare commercial or managerial occupations. Although we have no data on WWI death rates by occupation, demobilisation statistics indicate that men working in the engineering industry represented a large fraction of non-officer servicemen (11%) and that engineers provided relatively few officers (BWO, 1922). Specifically, 11,389 engineers demobilised as officers and 359,948 as other ranks, a ratio of 1 officer per 32 others. This is similar to ratios for those demobilised in other non-engineering trades and in agriculture, but markedly different to other industries that provided large numbers of officers: professional men (1:1), commercial and clerical workers (1:12), and students and teachers (1:1.5).

In summary, the considerations in this section suggest a number of stylised facts that inform our later empirical work. First, in the period we study engineers generated a very considerable share of new patents and, although we have no data to verify it, it is likely that a high proportion were granted to individuals with an elite (university-level) education. Second, British men that were educated to university level were killed disproportionately in the war, at least in part because many volunteered in the rush to the colours at its outset. Third, some individuals with engineering experience fought as commissioned officers, but many more served in the other ranks, for example as privates and NCOs. Finally, the composition of officers changed throughout the war as people from wider socio-economic groups were by necessity commissioned as ranker officers, and these men largely came from industries such as teaching and clerical work that were less likely to support invention than engineering.

3. Data and descriptives

We construct a parish-level panel dataset for England and Wales covering five-year intervals from 1895 to 1979, and an auxiliary panel at the inventor level for the same period.[9]

Parishes were the primary administrative unit of local government in England and Wales during the early 20th century and closely approximate local communities. For consistency, we use parish boundaries as defined in 1911, the last census year before WWI, throughout our analysis to ensure uniformity in geographic units across time. Below, we describe the main data sources used and the procedures for constructing our sample.

3.1. Data on WWI deaths

Our primary sources of information on WWI soldier deaths are the online casualty database of the Commonwealth War Graves Commission (CWGC) (Commonwealth War Graves Commission, 2023), complemented with a separate database available from the military genealogy specialist Forces War Records (FWR) (Forces War Records, 2023). The CWGC dataset contains details recorded on the war graves of the Commonwealth military personnel who died in both World Wars, including name, age, date of death, rank, regiment, places of birth and/or residence, and service unit (usually a battalion).[10]

We next use additional data to identify members of the social elite in our CWGC-FWR data. Our primary method to do so relies on the Oxford University Roll of Service (Craig, 1920), and The War List of the University of Cambridge 1914-1918 (Carey, 1921). These documents name all faculty, enrolled students, and alumni of these pre-eminent universities that served during the war. We digitise these documents and match individuals listed to soldiers recorded in the processed CWGC-FWR data using surnames, initials, and rank. We successfully match 3305 individuals (23%) from the Oxford and 3260 (23%) from the Cambridge lists to our soldier-level fatalities dataset in this way. We also specify an alternative measure of social elites by using information on rank field to construct an indicator for a killed soldier being an officer.

Finally, to examine if the loss of specific skill-sets affects subsequent patenting, we also split these officers into two mutually exclusive and collectively exhaustive categories by using a combination of regiment and string matching for specific terms in rank titles and the additional information on gravestone commemorations. These are i) Engineers, who either served in the Royal Engineers or whose rank title or commemoration indicates an engineering background; ii) Other officers, which is a catch all for all other officers who we cannot identify as having engineering experience or skills.

After constructing the CWGC-FWR soldier-level dataset, we collapse the data to the parish level using soldiers' assigned places of origin. When doing so we create a count of the total number of WWI soldier deaths in each parish, and dummy variables for the parish

[9] We exclude Scotland from our analysis as Census data and digital maps for Scotland are unavailable for this period.

[10] More details on the construction of this dataset can be found in Appendix B. A discussion and validation of the geolocation of soldiers to parishes of origin can be found in Carozzi, Pinchbeck and Repetto (2023).

losing at least one Oxford or Cambridge graduate, at least one officer, and at least one officer of the five categories noted above.

3.2. Data on patents

Our primary source of patent data is the PatentCity database (Bergeaud and Verluise, 2024), which we restrict to all patents registered at the Great Britain patent office in the period from 1895 to 1979.[11] The database records each patent's Patent Number along with information on each of its inventors. We treat each patent-inventor pair as a separate record, which allows an invention to be spatially distributed. In total, our data comprises approximately 1.7M patents (2.8M patent-inventor pairs).

A distinctive feature of the PatentCity database is its provision of precise geographic coordinates for inventors, obtained by geocoding natural language location data. Using Geographic Information System (GIS) software, we map these inventor-patent combinations to the 1911 parish boundaries, allowing us to consistently align inventive activity with local communities. Doing so, we exclude patents registered by foreign residents (56 percent of patents) and patentees not geo-referenced at a sufficiently granular level (2 percent).[12] That is, our final sample of inventor-patent pairs encompasses 855K observations (600K patents).

Despite its usefulness for geolocation purposes, the information in PatentCity on the content of the patent is scarce. Thus, we merge the PatentCity data to the European Patent Office PATSTAT database (European Patent Office, 2023), using the unique patent identifier that is common to both source.[13] The combination of the two data sources is also instrumental to track the output of the same inventor over time as well as to gather information on co-authorship patterns since PATSTAT provides an inventor identifier. Appendix B.5 provides details on how we combine PatentCity and PATSTAT at the inventor-patent level.

We enrich the information contained in the PatentCity and PATSTAT datasets with additional variables constructed or complemented using Natural Language Processing (NLP) techniques. First, for all patents in PATSTAT with available abstracts (approximately 556,000 patents), we measure patent importance using the text-based methodology of Kelly et al. (2021) (see Appendix C for details).[14] The central idea underlying this measure is that a patent is more important when the language used to describe it departs from the vocabulary of earlier patents while simultaneously anticipating the vocabulary of subsequent innovations.

Our data contains the Cooperative Patent Classification (CPC) code, which reports the field(s) of the patent. Our definition of fields is the one-digit CPC code, so that the patents in

[11] The database also includes a selection of patents published before 1895 and after 1980, these latter merged from an alternative source. Due to incomplete data in earlier years and a structural break introduced by the merger, we restrict our analysis to patents registered in the period 1895-1979.

[12] See Appendix B.5 for details about the assignment of patents to parishes.

[13] The match is almost perfect, with only 3 patents published in 1900 in the Great Britain part of PatentCity not having a match in PATSTAT. 97K patents are recorded in PATSTAT but not in PatentCity, thus are not feasible for our analysis. Figure A.1 in the Appendix shows the discrepancies between the two datasets.

[14] Patents with missing abstracts in PATSTAT are concentrated in the earliest years of our sample (18951897) and in the final year (1979).

our main sample are divided in eight fields of knowledge. Since CPC codes are only available for patents published after 1910 and numerous patents are associated with multiple CPC codes, we estimate the field of most-likely application using the patent's title textual content and Natural Language Processing (NLP) machine learning techniques.[15]

3.3. Further data sources

We use several other data sources. We obtain the I-CeM Census microdata for England and Wales in 1911 from IPUMS International (Minnesota Population Center, 2019). This allows us to compute parish population as well as several proxies for wealth and demographic conditions in 1911, namely the share of households with no servants, the share of male workers in white collar jobs, and a proxy for the unemployment rate.

We use information from the British Army Service Records for 1914 to 1918, which we access through FamilySearch (FamilySearch, 2023), to construct parish-level measures of the total number of men mobilised.[16] To construct our instrumental variable, we combine this data with the CWGC information on deaths and apply several sample restrictions, described in Appendix B.

We compute parish populations for each of the five-yearly intervals in our final data using Census data obtained from the website "A Vision of Britain through Time" (VoB). We spatially re-weight population data to 1911 polygons and linearly interpolate between Census years because censuses are taken only every ten years and parish polygons change over time.

We use additional sources to construct proxies for regional innovation ecosystems before the start of the war, such as membership lists of four engineering trade bodies in pre WWI years; a list of pre-WWI universities; a list of rail stations in 1910; and, finally, a georeferenced list of all post offices in 1900 from the GB1900 Gazetteer. We then use these sources to construct a series of separate County-level ecosystem accessibility measures, computing the average distance to the nearest ecosystem factor in each County by taking the weighted average of the parish level distances, using as weights 1911 parish populations. We then take the natural log of this weighted average distance.

3.4. Sample selection and descriptive statistics

Our final parish-level dataset comprises 13,352 grouped parishes.[17] For patent data, we retain all patents that can be geolocated to at least the city level, aggregating counts into five-year periods from 1895-1899 to 1975-1979. The final dataset at the parish level is thefore a panel of 13,352 parishes for all 5-year periods from 1895 to 1979.

Our second dataset is an individual-level panel of inventors spanning all periods from 1900 to 1979. To construct this dataset, we start with the 18,095 patentees in our data

[15] We refer the reader to Appendix D for the details of the prediction exercise.

[16] Note that these records are incomplete because a fire destroyed part of the collection in 1940. FamilySearch's WWI collection we use includes the records that survived the fire (the "Burnt documents"), as well as additional records on WWI pension applications ("Unburnt documents"), for a total of about 4 million records. We describe how we assign mobilised men to parishes in the Appendix.

[17] For details on the parish grouping procedure, see Appendix B.

(excluding firms and other non-natural entities) who registered at least one patent in 1895-1899, the first period in our sample and trace their patenting activity over time using the inventor identifier available in PATSTAT.[18] We construct time-invaryang variables for each inventor based on their records (residence, WWI mortality exposure, field of specialization, patent importance) as measured when they produced their first patent in the 1895-1899 period.[19]

Table A.1 in the Appendix presents summary statistics for key parish characteristics. In 1911, the average parish had a population of about 2,700 inhabitants and covered 11.5 square kilometres. On average, each parish contributed around 231 servicemen to the WWI mobilization effort, implying a parish-level mobilization rate of about 5.5 percent of the population.

Figure 1 provides a visual overview of the variation in WWI losses and patenting activity across parishes in our final dataset. Panel A shows the spatial distribution of soldier deaths per parish, revealing substantial geographic variation. Panel B presents a histogram of soldier deaths, highlighting that roughly one-fifth of parishes did not experience any soldier fatalities, and that the distribution is highly skewed. On average, parishes reported 44 soldier deaths, equivalent to about 1 percent of the 1911 population, as shown in Table A.1. Additionally, 8 percent of parishes had at least one soldier who attended Oxford or Cambridge University and died in the war, while 9 percent experienced the death of a soldier with an engineering background.

In terms of patenting activity, Panel C of Figure 1 maps the average number of patents registered per five-year period across parishes over the full sample. Confirming prior findings, patenting is highly spatially concentrated in England and Wales. Panel D presents a histogram of the number of patents per five-year period, showing a highly skewed distribution: even when aggregated over five years, 80 percent of parish-period observations have zero patents. Among parishes with patenting activity, the mode is a single patent, although the distribution features a long right tail. Due to this skewness, our main analysis focuses on a binary outcome variable indicating whether at least one patent was registered in a given five-year period. Table A.1 in the Online Appendix shows that the mean of this indicator in our sample is 0.19.

4. Empirical strategy

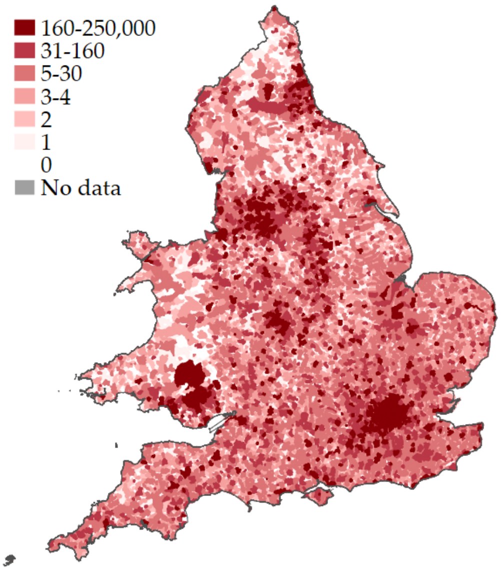
4.1. Roadmap of the empirical analysis

In the following sections, we present the empirical results of our analysis. We start by presenting the parish-level results, where we show that WWI mortality reduced local communities' innovation output, particulary for patents that are the most important and innovative. Then, we extend our analysis to the inventor-level, where we show that the negative effect of WWI mortality on innovation is not limited to inventors who were active before the

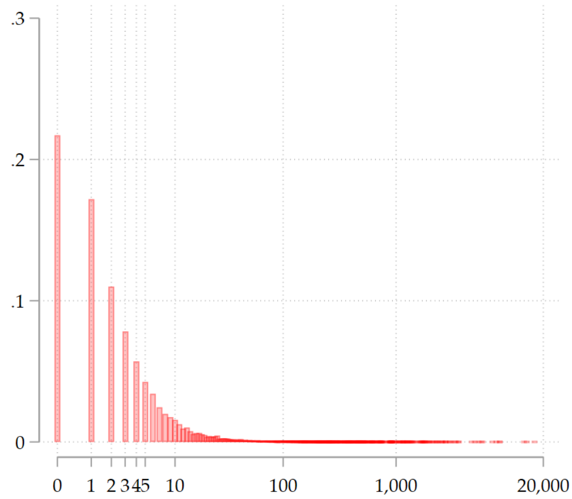
[18]See Appendix B.5 for details about the sample construction.

[19]We drop the initial 18951899 window – when, by construction, all selected inventors had a patent – so that our panel begins in 1900 and ends in 1979.

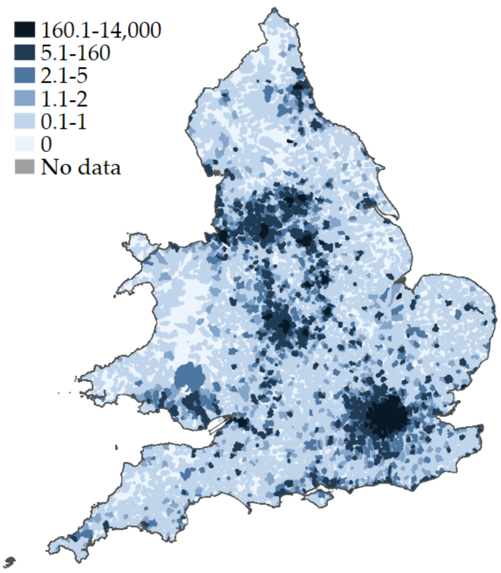
Figure 1
Patents registered and WWI deaths in England and Wales



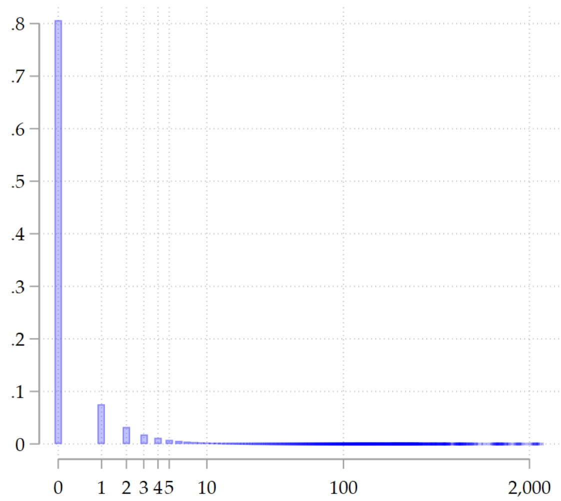
(A) N. dead WWI soldiers



(B) N. dead WWI soldiers – density



(C) N. of patents over 5-year period – average



(D) N. of patents over 5-year period – density

Notes: Historical (grouped) parishes in England and Wales. Panel A displays the number of soldiers killed in WWI originating in each parish. Panel B reports the density of the number of soldiers killed in WWI. Panel C shows the number of patents registered over a 5-year period, averaged over all periods in 1895-1979. Panel D shows the density of number of patents registered over a 5-year period (where each period counts as an observation).

war, but also affects individuals who became inventors afterwards. In Figures A.5 and A.6 in the Appendix we show the corresponding dynamic specifications. While we document that the universe of inventors shows some significant pre-trends, we argue that this evidence should not invalidate the research design because pre-war trends in patenting are comparable across inventors exposed to different levels of mortality among inventors that survived the war. We then explore the impact of WWI mortality on the production of highly innovative and complex patents, finding that these engineers and inventors working on complex and innovative projects are more affected by the human capital shock of the war. Finally, we explore the role of changing sector, co-authorship, and mobility in mitigating the impact of WWI mortality on innovation.

4.2. Empirical specifications

Our first model is a difference-in-differences (DiD) specification that compares changes in innovation in parishes more and less affected by WWI mortality. The baseline model is

$$y_{it} = \alpha + \beta log(d_i^{WWI}) \times Post\ WWI + \gamma' W_{it} + FE + \epsilon_{it}, \tag{1}$$

where y_{it} is an innovation outcome – for instance an indicator for having produced at least one patent – for parish i in 5-year period t, e.g. 1910-1914.[20] $log(d_i^{WWI})$ is the log number of servicemen from parish i killed in WWI, and $Post\ WWI$ is an indicator for all periods after and including 1915.[21] We also include controls W_{it} and different sets of fixed effects that vary by specification.[22]

To study dynamic treatment effects and to evaluate pre-trends, we also estimate the following dynamic DiD model:

$$y_{it} = \alpha + \sum_{k=1895, k\neq 1910}^{1975} \beta_k log(d_i^{WWI}) \times D_k + \gamma' W_{it} + FE + \epsilon_{it}, \tag{2}$$

where D_k is an indicator for 5-year period k. As it is customary, in estimation we exclude the pre-treatment period 1910-1914 which, therefore, becomes the baseline.

In the second part of the analysis, we use an inventor-level panel covering covering the period 19001979. The inventor-level analysis allows us to explore a much broader set of questions, such as the effect of mortality on patenting among inventors active before the War, heterogeneous impacts across sectors and patent complexity, and the role of mobility and co-authorship in mitigating the impact of the WWI human capital shock. Using this

[20]Notice that the first period in our sample is t =1895-1899, while the last period is t =1975-1979.

[21]As we take the log of WWI deaths, our baseline model drops parishes that did not experience war losses. Our results are also robust to using $log(1 + d_i^{WWI})$, as shown below in Figure A.11.

[22]These controls are the log of WWI mobilisation at the parish level, as well as variables from the 1911 Census, always interacted with linear time trends. Specifically, we include total population, the share of households in the parish with no servants (as a proxy for income), the share of white collar workers and a proxy for unemployment (to control for labour market conditions prevailing before the War). The set of controls is purposefully parsimonious because i) the difference-in-difference assumptions rely on parallel trends – for which there is evidence already in the model without controls, and ii) to avoid the inclusion of an excessive number of controls and fixed effects (that have to be interacted with linear trends or time effects).

dataset, we estimate a continuous DiD model at the inventor level as follows:

$$y_{ijt} = \alpha_j + \delta_t + \beta log(d_i^{WWI}) \times Post\ WWI + FE + \gamma' W_{it} + \epsilon_{ijt}, \tag{3}$$

where y_{ijt} is an indicator for inventor j from parish i having registered a patent in period t and the rest is as above.

4.3. Identification assumptions

One possibility to identify the causal effect of WWI mortality in this setting would be to invoke parallel trends in the outcome of treated and control units. A standard parallel trends assumption made in a continuous difference-in-difference design such as the one considered here requires that the average evolution of outcomes that units with any level of WWI mortality would have experienced without treatment is the same as the evolution of outcomes that units in the untreated group actually experienced (Callaway, Goodman-Bacon and Sant'Anna, 2024). However, under this assumption, only the difference in the outcome between "untreated" units and units with a positive "treatment" is identified, while the causal effect of having more (as opposed to less) mortality is not. To achieve identification of these "average causal responses" (Callaway, Goodman-Bacon and Sant'Anna, 2024), a stronger version of parallel trends is required, specifically that the average evolution of outcomes for the entire population if all experienced a certain mortality is equal to the path of outcomes that the group of parishes with that mortality actually experienced. Although this assumption is inherently untestable, in Section 7 we provide evidence that it appears reasonable in this context by comparing the pre-trends across all deciles of d^{WWI}. We also show that our results are widely comparable to those obtained by defining a binary treatment indicator that takes the value 1 if d^{WWI} exceeds the median.

An alternative approach to identification (which would avoid the issues highlighted above) would instead be to leverage exogenous variation in WWI mortality. Because directly invoking exogeneity of WWI mortality could be problematic, we deal with this challenge by adopting a shift-share instrumental-variable technique. Specifically, we follow Carozzi, Pinchbeck and Repetto (2023) and instrument WWI deaths with predicted deaths, constructed combining information on battalion-level mortality and the share of servicemen who served in each battalion. We discuss and validate this IV approach and show results mirror closely the DiD ones in Section 7 below. Thus, the combined approach we follow in the rest of the paper uses DiD and event-studies to document dynamics and validate pre-trends, and IV to provide a causal estimate based on quasi-random variation in battalion-level mortality risk.

5. Parish-level results

5.1. WWI human capital shock harms patenting in the long-run

Table 1 presents estimates from equation 1 of the effect of WWI casualties on parish-level patenting. Column 1 shows that higher casualties reduce the likelihood that any patent is produced by a parish resident in subsequent decades. Specifically, a 10 percent increase in deaths lowers the probability of patenting by 0.08–0.11 percentage points, a 0.3–0.5%

decline relative to the baseline mean. Columns 2–3 show that this estimate is robust to county-specific time fixed effects and to interactions of baseline controls with a linear time trend. Our preferred specification is the one reported in column 2, which limits comparisons to parishes within the same county. Column 4 addresses the possibility that innovation falls simply because war casualties reduced local population by controlling for the log of parish population, constructed from Census figures and linearly interpolated between census years.[23] While population may be an outcome and therefore a "bad control," its inclusion does not change our estimates.[24]

Figure 2 examines the dynamics of these effects by presenting estimates and 95 percent confidence intervals of the event study specification in equation 2 that includes parish and time fixed effects, and county-specific linear trends. All estimated pre-trend coefficients are small and not statistically different from zero. In the 5-year period that includes the Great War, we observe a pronounced drop in patenting activity, with a coefficient about twice as large as the one estimated using the full post-WWI period in Table 2, suggesting that the impact of war mortality was sizeable at the onset of the war and then decreased in the subsequent decades. This is indeed the case as evidenced by inspecting the figure further. After a recovery – almost to pre-war levels – in the inter-war period, patenting activity declines again before and during WW2 in parishes with high WWI mortality, and remains low for the remainder of the sample period.

The results using our instrumental variables approach for WWI deaths closely mirror those from the difference-in-differences specification discussed here. Therefore, to maintain clarity and focus, we present the IV results separately in Section 7, and concentrate on the diff-in-diffs specification throughout the main analysis.

The results in Table 1 and Figure 2 focus on a binary measure of innovation. This approach helps address the fact that only about 20% of parishes produce any patents in a given period. However, our findings are not confined to the extensive margin. First, as shown in Appendix Table A.2, when we use a pseudo-Poisson maximum likelihood model with the patent count as the dependent variable, the estimates are qualitatively similar, indicating that WWI mortality also affects the total number of patents.

Second, Figure A.2 uses a series of regressions where the dependent variable is a binary indicator for whether the parish patent rate (number of patents per capita, using the 1911 Census population) exceeds various percentiles of the overall distribution. Across almost all percentiles, higher WWI mortality reduces the probability that a parish reaches higher patenting rates.[25] Notably, the negative effects on being above the 80th or 90th percentile are even stronger than the effect at the extensive margin, suggesting that WWI mortality had a sizeable impact not just on whether parishes patented at all, but also on the intensity

[23]Although in Britain the effect of the war on demographic conditions was modest (Winter, 1977), it is still possible that innovation decreased in parishes with high mortality merely because of the population decrease.

[24]In our inventor-level analysis reported below, we provide direct evidence that observed decline in local patenting is not due to inventors losing their lives in the war.

[25]For percentiles below the 77th, these coefficients are equivalent to those in column 4 of Table 1.

of innovation (the intensive margin).[26]

Table 1
WWI deaths and patenting – OLS estimates

	(1) Any patent	(2) Any patent	(3) Any patent	(4) Any patent
$Log(d^{WWI})\times$ Post WWI	-0.009***	-0.011***	-0.008***	-0.009***
	(0.001)	(0.001)	(0.002)	(0.002)
Mean dep.var.	0.24	0.24	0.24	0.24
R2	0.49	0.50	0.50	0.50
Observations	176228	176228	176228	176228
Parish FE	Y	Y	Y	Y
Time FE	Y	Y	Y	Y
County × Time	N	Y	Y	Y
Controls × Linear trend	N	N	Y	Y
Population (log)	N	N	N	Y

Notes: OLS estimates of the effect of WWI deaths on the probability that at least one patent is registered by a resident of parish i during 5-years period t. Column 2 adds county-time fixed effects. Column 3 adds controls interacted with a linear trend. Column 4 also control for the log of population (interpolated). Standard errors clustered at the parish level. *, **, and *** represent 10%, 5%, and 1% significance levels, respectively.

5.2. Highly innovative patents are affected the most

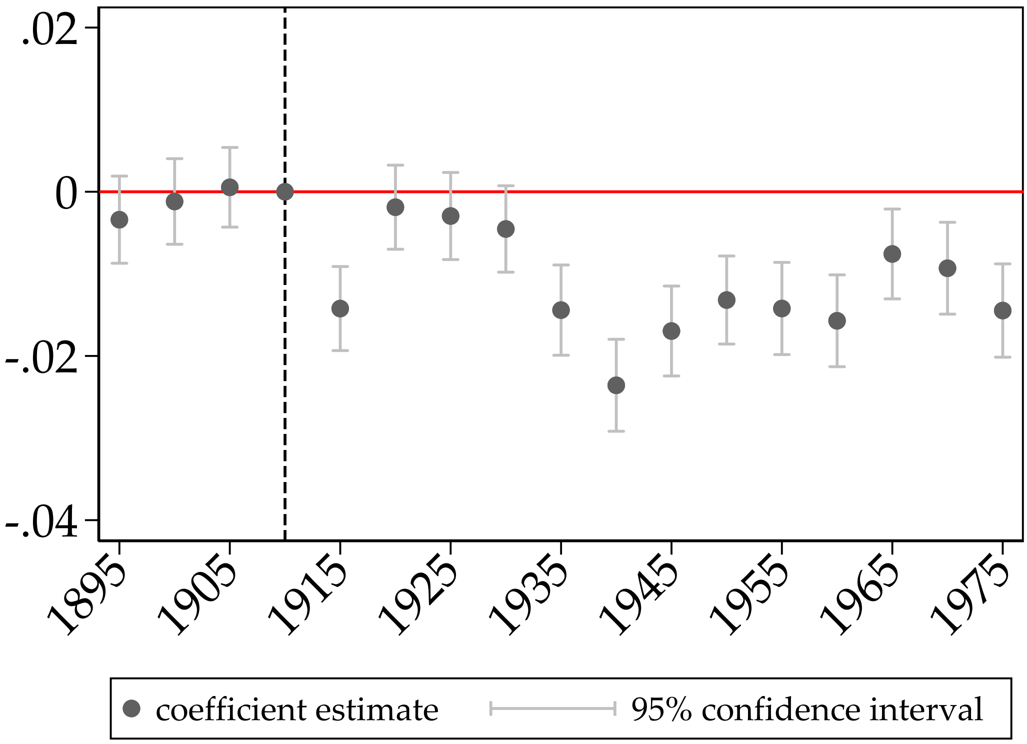
The results presented thus far demonstrate that WWI mortality led to a decline in the production of local innovation. However, this analysis centers on the *quantity* of innovation and does not consider its *quality* – that is, the technological significance and impact of each patent. It is plausible that WWI mortality altered the balance between the number and the importance of innovations, potentially reducing the overall volume of patenting without necessarily affecting the production of high-impact inventions. To investigate this question, we study whether WWI human capital losses affected not just the output but also the quality of innovations. Patent quality is measured using a text-based indicator of innovativeness, constructed according to the approach of Kelly et al. (2021) briefly outlied earlier. This allows us to assess whether WWI mortality lowered the likelihood that communities generated more influential patents, in addition to reducing patenting activity overall.

Figure 3 presents a series of estimates from repeated applications of equation 1, where the dependent variable is progressively restricted to patents above increasing importance thresholds. For each five-year window, we rank patents by their importance and construct a binary variable indicating whether parish i had at least one patent meeting or exceeding a given threshold during period t. The initial estimate includes all patents, treating any patent

[26] Appendix Table A.4 shows that the extensive-margin effect is largely confined to smaller parishes (population below 2,500 in 1911), which is expected since larger populations mean a higher likelihood of at least one patent (the probability is 14% in small parishes and 73% in large ones). Among larger parishes, where zero-patent observations are rare, we consistently find a negative effect of WWI mortality on intensive innovation outcomes (see Figure A.3, which restricts the sample to large parishes with positive patent counts).

FIGURE 2

WWI DEATHS AND PATENTING – EVENT-STUDY RESULTS



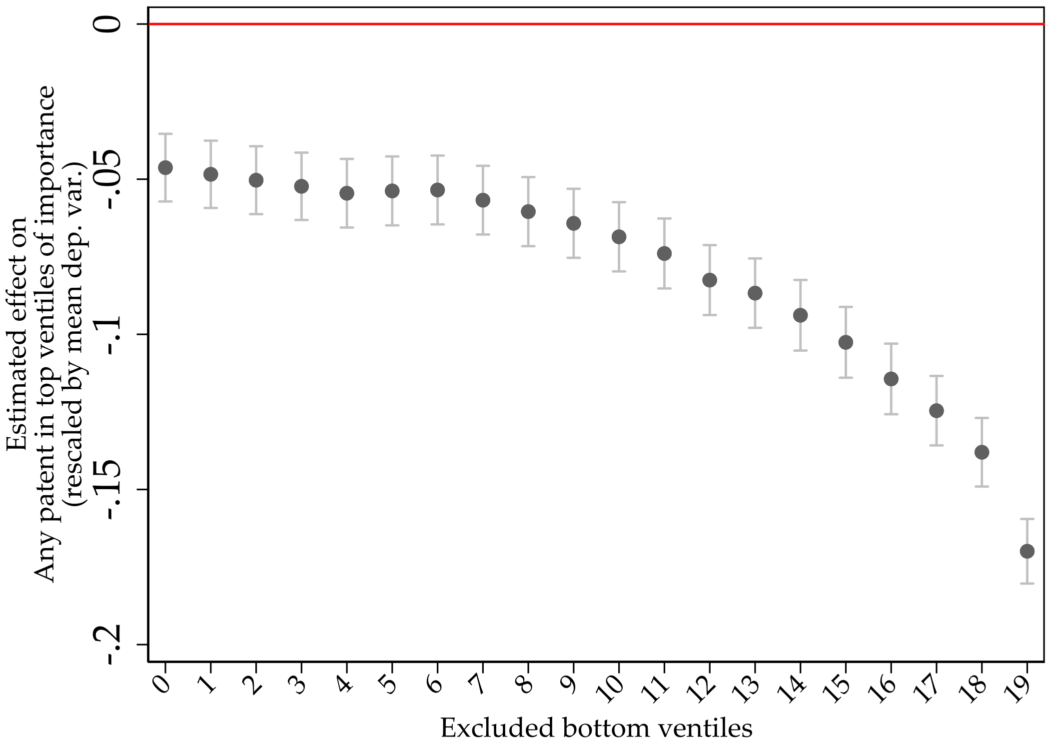
Notes: OLS estimates of the effect of WWI deaths on the probability that at least one patent is registered by a resident of parish i during 5-years period t. Estimates from equation 2, including time, parish, and county-time fixed effects. Confidence intervals constructed using s.e. clustered at the parish level.

as important. Subsequent iterations successively tighten the criterion: first to only patents in the top 95% of the importance distribution, then the top 90%, and so forth, culminating with an indicator that equals one only if a parish resident published at least one patent in the top 5% of importance within that window.

By construction, the mean of the dependent variable – critical for assessing the relative magnitude of the estimated coefficients – declines with each iteration, as the threshold for a patent being classified as important increases. To account for this, all estimates and confidence intervals are rescaled by the mean of the dependent variable in the corresponding estimation sample. The results suggest that the negative impact of the WWI death shock on innovation potential grows with patent importance. In relative terms, a higher exposure to the WWI mortality shock reduces the probability of publishing at least one patent in the top half of the importance distribution by approximately one-third more than the effect on the probability of publishing a patent of any importance. This relationship is even more pronounced in the right tail of the importance distribution: the effect of the WWI shock on the probability of producing a patent in the top 5 percent of importance is roughly three times as large as the effect on the probability of registering a patent of any importance, suggesting that WWI mortality affected both the quantity and the quality of the innovation produced

FIGURE 3

WWI DEATHS AND PATENTING – IMPORTANT PATENTS



Notes: OLS estimates of the effect of WWI deaths on the probability that at least one patent whose innovativeness (measured as in Kelly et al. (2021) is in the top ventiles of the distribution of patents published at the same time is registered by a resident of parish i during 5-years period t. Each coefficient refers to a different regression, so that the left-most coefficient reflects the estimates reported in Table 1 and the right-most coefficient is an indicator equal to 1 if at least a patent registered belongs to the 5 percent of the most important patents published during the five-year period t. Estimates include time, parish, and county-time fixed effects. Coefficients are resecaled by the mean of the corresponding dependent variable. Confidence intervals constructed using s.e. clustered at the parish level.

by British communities.

5.3. The lost generation

To confirm that the impact of war deaths on innovation, as documented in the previous section, stems directly from the loss of talented and skilled individuals, we extend our baseline model (column 2 of Table 1) as follows:

$$y_{it} = \alpha + \beta_1 log(d_i^{WWI}) \times Post + \beta_2 \mathbb{1}(HC_i) \times Post + FE + \epsilon_{it}, \quad (4)$$

where $\mathbb{1}(HC_i)$ is an indicator variable for human capital that takes the value one when the parish experiences a relatively high number of deaths of individuals that have skills and characteristics that are associated with the generation of new ideas. The coefficient of interest is β_2, which measures the effect on innovation of the loss of these skills, conditional on the overall number of deaths. We will define this indicator in different ways using the information contained in the mortality records.

Results are reported in Table 2. In columns 1-4 we report coefficients on various human capital measures without conditioning on overall WWI mortality. Columns 5-6 instead report estimates from the full model in eq. 4 that conditions on WWI deaths. Specifically, in column 1 we define $\mathbb{1}(HC_i)$ as an indicator for the parish having lost at least one individual belonging to the social elite, defined as those who studied at Oxford or Cambridge. The coefficient is negative and significant and indicates that the loss of a social elite reduces the likelihood of subsequent patenting by 3.6 percentage points. In columns 2 and 3, we examine the effect of deaths of officers. In column 2, we replace the social elite indicator with an indicator for death of any officer, finding again a negative effect. In column 3, we divide officers in two groups, engineers and other officers. Results indicate that the loss of an officer with an engineering background has much larger negative effect on subsequent patenting than the loss of other types of officers.

In column 4, we add back our indicator for losing a social elite and show that both those and engineer deaths seem to matter. Finally in column 5 we include all the human capital measures simultaneously and an indicator for WWI overall mortality being above the median. In this case the coefficient on all our human capital variables remain similar in magnitude to earlier columns, supporting the idea that the loss of social elites and men with engineering skills has additional effects on subsequent patenting over and above the loss of other men. This result is in line with recent evidence from the UK and the Americas showing that engineers are key drivers of innovation and growth and produce more patents than other types of innovators (Hanlon, 2022; Maloney and Valencia Caicedo, 2022). In column 6 we include the log number of deaths in estimation and find its large negative effect is preserved even when including all the human capital indicators, of which only the one for engineers is significant. This result suggests that engineer losses are the most harmful for innovation, and that our measure of overall WWI mortality does a generally good job of capturing the loss of skilled men that we can observe in our data, and that is also captures additional information about the loss of talented men that we cannot easily observe.

5.4. Heterogeneity and the mitigating role of institutions

Previous work demonstrates that the presence of inventor role models, transport links, communications technologies, and institutions can drive the generation of new ideas (see e.g., Akcigit, Grigsby and Nicholas 2017*b*; Bell et al. 2019; Andrews 2023). Local access to these resources in the period prior to WWI could potentially mitigate the effects of war deaths on innovation. Specifically, if these ecosystem factors can substitute for human capital inputs in the production of new ideas, we may expect to find that the effect of war deaths is smaller in places where they are more accessible. We conduct heterogeneity analysis to test for this using four county-level accessibility measures and our parish level data. In all cases, we separate out counties with above-median access to an ecosystem factor, and test for differential effects relative to the overall average effect by adding a triple interaction term

TABLE 2

WWI DEATHS AND PATENTING – OLS ESTIMATES – IMPORTANCE OF ELITES

	(1)	(2)	(3)	(4)	(5)	(6)
I(Social elite) x Post	-0.036***			-0.020**	-0.019**	-0.009
	(0.008)			(0.009)	(0.009)	(0.010)
I(Officer) x Post		-0.014***				
		(0.004)				
I(Engineer off.) x Post			-0.039***	-0.032***	-0.032***	-0.019**
			(0.007)	(0.008)	(0.008)	(0.009)
I(Non eng off) x Post			-0.011***	-0.009**	-0.007	-0.000
			(0.004)	(0.004)	(0.005)	(0.005)
$I(d^{WWI} > med)\times$ Post					-0.005	
					(0.005)	
$Log(d^{WWI})\times$ Post						-0.008***
						(0.002)
R2	0.50	0.50	0.50	0.50	0.50	0.50
Observations	177276	177276	177276	177276	177276	177276

Notes: OLS estimates of the effect of WWI deaths on the probability that at least one patent is registered by a resident of parish i during 5-years period t. Elites are defined as Oxford and Cambridge graduates. Parish, time, and county-time FE included in all specifications. Standard errors clustered at the parish level. *, **, and *** represent 10%, 5%, and 1% significance levels, respectively.

to our baseline specification.[27]

Results are reported in Table 3. In column 1, we find that the effect of war deaths is significantly smaller – approximately half as large in absolute terms – in counties with above-median access to engineering role models. In column 2, we examine proximity to rail stations in 1910. Here there is also a mitigating effect of similar magnitude. We find comparable effects with respect to access to post offices in column 3, and in places that are closer to pre WWI universities (column 4). Notably, across all columns and as expected, we find that the mean of the dependent variable is considerable larger in places with better access to ecosystem factors, consistent with the idea that these factors are positively associated with patenting. This implies that the effects of war deaths are in fact considerably smaller in these places, and in particular suggests that universities and innovation networks can mitigate the effects of shocks, at least to some degree.

5.5. Both existing and new inventors are affected

War mortality can affect communities' innovative output in several ways. It might mechanically harm innovation because of the direct effect of inventors losing their life in the war. Or it can lower the productivity of pre-existing inventors, induce them to migrate to places less exposed to the shock, or reduce future innovation by reducing the attractiveness of the location for other inventors. Finally, the loss of talented men in the war might affect

[27]The county-time effects in this specification absorb the two-way interaction between above median accessibility and post WWI.

TABLE 3

WWI DEATHS AND PATENTING – OLS ESTIMATES – HETEROGENEITY

	(1) Engineers	(2) Stations	(3) Post Off.	(4) Univer.
$Log(d^{WWI})\times$ Post WWI	-0.014***	-0.014***	-0.013***	-0.014***
	(0.002)	(0.002)	(0.002)	(0.002)
x I(Access > med)	0.005*	0.005**	0.005**	0.006**
	(0.002)	(0.002)	(0.002)	(0.002)
Mean dep.var.(Access < med)	0.15	0.14	0.18	0.17
Mean dep.var.(Access > med)	0.32	0.33	0.30	0.30
R2	0.50	0.50	0.50	0.50
Observations	176228	176228	176228	176228

Notes: OLS estimates of the effect of WWI deaths on the probability that at least one patent is registered by a resident of parish i during 5-years period t by pre-WWI characteristics defined at the district level. Parish, time, and county-time FE included in all specifications. Standard errors clustered at the parish level. *, **, and *** represent 10%, 5%, and 1% significance levels, respectively.

the production of new inventors in the next generation, for example because of the presence of fewer role models.

In Table 4, we explore these mechanisms further. We start by identifying, among all patents, those registered by a new inventor – defined as a patentee who did not appear in the sample earlier. In column 2, we observe that patenting by new inventors declines in high-mortality parishes, with a coefficient even larger than the one found using all patents (reported in column 1). In column 3, instead, we consider only patents registered by pre-existing inventors – defined as those who registered at least one patent in the first period of our sample.[28] Again we find a large and negative effect, suggesting that war mortality negatively affected both inventors who were already active and experienced, as well as reducing the probability that the eco-system of the parish encourages the decision to become an inventor or attracts from the outside individuals with the potential to become inventors. It is worth noting that the negative estimates for pre-existing inventors indicate that our main results at the parish level reported earlier cannot be driven entirely by the mechanical effect of inventors (or potential future inventors) losing their life in WWI. On the contrary, their magnitude suggest that patenting by pre-existing inventors dropped significantly more than patenting by new ones.

6. Inventor-level results

Motivated by the evidence previously shown that active inventors were also affected, in this section we move to an inventor-level analysis using the individual-level panel of inventors covering the period 19001979. Leveraging the detailed information available in this

[28]Specifically, we restrict the sample to patents granted to inventors who registered at least one patent in the period 1895-1899. We then drop this period from estimation in all specifications of Table 4.

TABLE 4

WWI DEATHS AND PATENTING – OLS ESTIMATES – NEW AND EXISTING INVENTORS

	Any patent	Any patent new inv.	Any patent existing inv.
$Log(d^{WWI})\times$ Post WWI	-0.012***	-0.015***	-0.039***
	(0.001)	(0.001)	(0.002)
Mean dep. var.	0.24	0.17	0.08
R2	0.51	0.47	0.42
Observations	167232	167232	167232

Notes: OLS estimates of the effect of WWI deaths on an indicator equal one if the inventor j from parish i registers a patent during 5-years period t. Column 1 presents results from equation 3 for the full sample. In column 2 the outcome is an indicator for the parish having registered a patent by a "new" inventor, defined as a patentees who did not yet register any patent in our sample period. Finally, in column 3, we restrict to patents by "existing" inventor, defined as a patentee who registered at least one patent in the first period, i.e. 1895-1899. For this reason we do not use this period in estimation in this specification, and in all others for comparability. Individual and time effects are included in all specifications. Standard errors clustered at the inventor level. *, **, and *** represent 10%, 5%, and 1% significance levels, respectively.

dataset, we are able to examine heterogeneity in the effects of WWI mortality on innovation across several dimensions. In particular, we are able to explore the impact of WWI mortality on the production of patents by inventors with different backgrounds, fields of specialization, co-authorship patterns and their location.

6.1. Baseline results

We start by estimating the baseline DiD model in equation 3 by OLS. Column 1 of Table 5 reports the results and shows a negative effect, in line with our parish-level estimates. In column 2, we restrict the sample to patentees who were not killed in WWI.[29] The effect is very similar in the restricted sample, which reassures us that our parish-level results are not driven by a mechanical fall in patenting due to the death of inventors. Instead, they reflect the negative impact of WWI mortality on both surviving inventors who were active before the war and on individuals who became inventors afterwards. In column 3, the outcome is an indicator for registering a "breakthrough" patent, defined as one whose importance (calculated following Kelly et al. 2021) exceeds the 80th percentile of the distribution of all patents registered in the past five years. In column 4, we use a citation-based measure of importance, defining a highly cited patent as one that receives more citations than 80 percent of all patents registered in the past five years.[30] Under both the breakthrough and citation-based measures, we find that inventors in higher-mortality areas are less likely to produce highly innovative patents.[31] Compared with the sample mean of each outcome,

[29]To identify these, we match patentees to the CWGC WWI fatalities list using the initial of the first name and the full surname. This approach is conservative, in that it may incorrectly classify some surviving inventors as having been killed when they share a first-name initial with a deceased soldier. Alternative matching methods (e.g., using more initial letters or the full first name) yield very similar results; see Appendix Table A.5.

[30]This measure is imperfect because citation counts in PATSTAT have been systematically recorded only since the 1980s, so highly influential patents that lost relevance before then may go undetected.

[31]Using one- or ten-year windows instead of five years yields similar results; we omit them for brevity.

the point estimates in columns 3 and 4 imply that the decline in high-quality patenting is proportionally larger than the decline in patenting of any type (column 1).

TABLE 5
WWI DEATHS AND PATENTING – INVENTOR-LEVEL ESTIMATES

	Any patent	Any patent	Any breakthr. patent	Any highly cited patent
$Log(d^{WWI})\times$ Post WWI	-0.003***	-0.004***	-0.001***	-0.001***
	(0.001)	(0.001)	(0.000)	(0.000)
Sample				
Mean dep.var.	0.08	0.11	0.02	0.02
R2	0.29	0.33	0.16	0.15
Observations	288000	144416	288000	288000

Notes: OLS estimates of the effect of WWI deaths on an indicator equal one if the inventor j from parish i registers a patent during 5-years period t. The sample is restricted to patentees who registered at least one patent in 1895-1899. Sample period is then restricted to 1900-1979. Column 1 presents results from equation 3 for the full sample, whereas in column 2 the sample is restricted to individuals who do not appear in the CWGC dataset of WWI fatalities, matched using the initial of the first name and surname. In column 3 and 4, the outcomes are, respectively, an indicator for the inventor registering a "breakthrough" patent or a highly-cited patent (see text for details). Controls, individual and time fixed effects are included in all specifications. Standard errors clustered at the inventor level. *, **, and *** represent 10%, 5%, and 1% significance levels, respectively.

6.2. The negative effect of mortality on highly innovative and complex patents

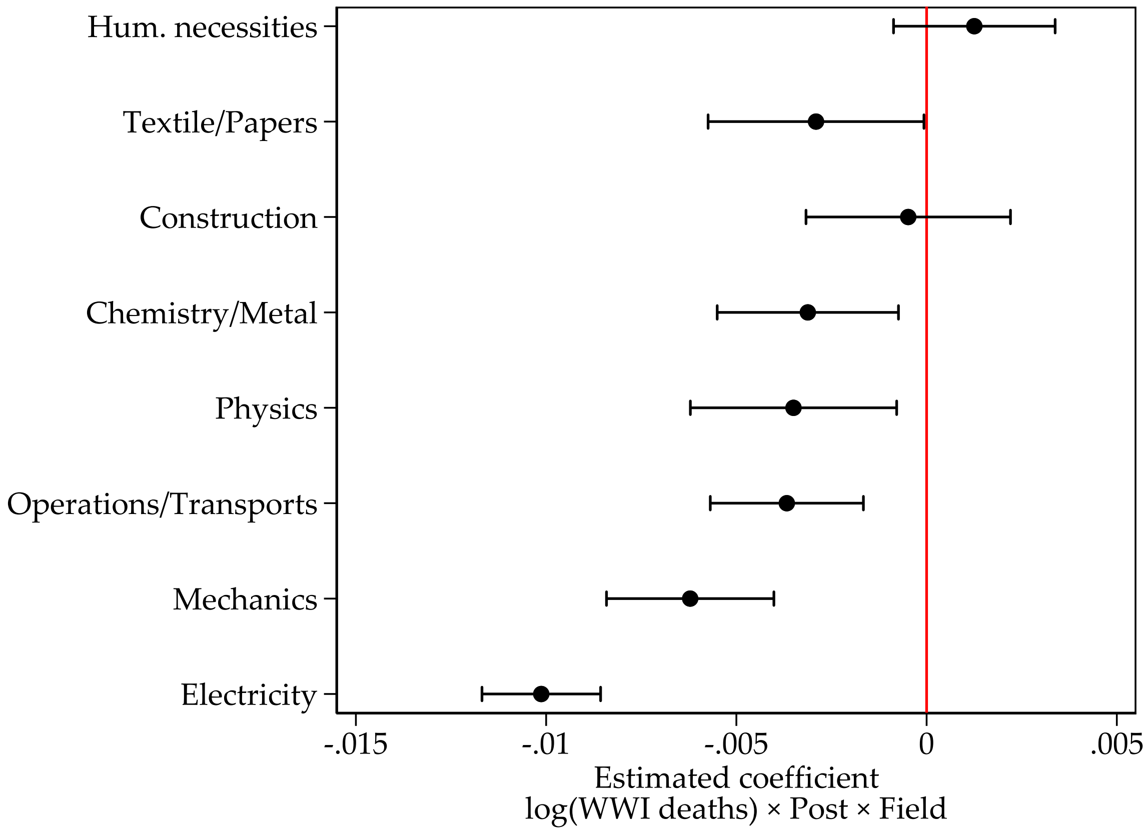
By concentrating on inventors who were already active before WWI, we can identify which fields of knowledge experienced a greater impact from the mortality shock. To do this, we estimate an extended version of equation (3) in which the $log(d_i^{WWI}) \times Post$ term is interacted with a series of indicators, one for each inventors pre-war field of specialization, measured by the one-digit CPC code of the first patent they produced during 1894–1899. This approach enables us to break down the average effect of WWI mortality by field, revealing how the shock differently affected areas that require different skills and expertise to generate innovation.

Results are presented in Figure 4, which reports the estimated coefficients for each interaction term. Results show substantial heterogeneity across fields, with the largest effects in electricity, mechanics, transportation, physics and chemistry. These science-intensive sectors, at the forefront of innovation at the time, required highly trained scientists and the execution of increasingly complex tasks. The need to commercialize scientific breakthroughs made laboratory research, specialized skills, and coordinated teams and infrastructure essential (Edgerton and Horrocks, 1994).[32]

[32]These sectors were also those where engineers were mostly active, as shown in Appendix's Figure A.7. Figure A.8 in the Appendix reports scatter plots of the share of engineers among inventors against two proxies for technological complexity: the length of the patent's abstract (panel A), and patent importance, computed following Kelly et al. (2021) (panel B). In both cases, we observe a strong positive relationship: patents that are more complex and more influential were disproportionately produced by engineers.

FIGURE 4

HETEROGENEOUS EFFECTS BY FIELD OF SPECIALIZATION



Notes: Inventor-level OLS estimates of the effect of WWI deaths on an indicator equal to one if at least one patent is registered during 5-years period t, interacted by the field of specialization of each inventor, as proxied by the field of specialization of the first patent produced by the inventor in 1894–1899 (panel a). Estimates from equation 3, including time and parish fixed effects, where the term $log(d_i^{WWI} \times Post)$ has been interacted with each field, as reported on the vertical axis. Confidence intervals constructed using s.e. clustered at the individual level. Fields of specialization are assigned according to the NLP technique explained in section 3.

Figure 5 examines heterogeneity in the effects of mortality on patenting by different measures of complexity and importance of the inventor's first patent. The figure illustrates that the negative effect of WWI deaths on patenting is not uniform: it is significantly more pronounced among inventors whose early work or training marked them as central to highly innovative or complex fields.

We start by contrasting inventors who recorded engineering as their occupation at the time of registering their first patent with those who were not. Results indicate that engineers experienced a notably larger negative impact from mortality, underscoring that the loss of human capital due to the War was particularly damaging to highly skilled inventors.

In addition, we examine heterogeneity by dividing inventors according to the complexity and importance of their first patent. Specifically, inventors whose initial patent falls above the median in either complexity (proxied by abstract length) or importance (measured using the metric from Kelly et al. 2021) experience greater declines in subsequent patenting activity.[33] The last coefficients reported in the Figure report results using the first principal component of patent complexity and importance, providing a summary index. These estimates reiterate that the negative impact of the mortality shock is most pronounced among inventors whose early work was more innovative or complex.

Taken together, these results indicate that the decline in innovation after WWI was disproportionately concentrated among inventors operating at the technological frontier or engaged in producing more complex inventions. This pattern underscores how the human capital shock induced by WWI had profound and lasting consequences for British technological progress.

6.3. The mitigating effects of mobility and co-authorship

The longrun decline in innovation following WWI mortality suggests that the human capital losses undermined both incumbent and future inventors, possibly by disrupting coauthorship networks and degrading the broader innovation ecosystem through the loss of complementary workers.

We explore this mechanism using our sample of inventors in Figure 6.[34] In column 2 we decompose our patenting indicator into filings that remain within the inventor's original field, and in column 3 into filings in a different field. Neither estimate differs meaningfully from the baseline effect in column 1, suggesting that midcareer retraining into a new technical field is neither common nor effective as a response to the human capital shock.[35]

In columns 4 and 5 we perform a similar decomposition but now for geographic mobility, contrasting patenting in the same parish as the first registered one parish with those filed while residing elsewhere. Here we observe a clear margin of adjustment: higher WWI mortality reduces the probability of patenting in the original parish but raises the probability of patenting in a new one. This pattern accords with recent evidence showing positive effects of being exposed to high-innovation areas on becoming inventors and patenting. Similarly to Bell et al. (2019) and Moretti (2021), the positive effect we document for movers could be due to gaining access to better networks, role models or mentors, as well as accessing specific infrastructures and complementary workers, who appear to be crucial for technologically advanced research.

Finally, in columns 6-7 we distinguish patenting as a single author, or co-authoring. he negative mortality effect on singleauthored patents again mirrors the baseline, but its impact on coauthored patents is statistically indistinguishable from zero. As in Jaravel, Petkova and Bell (2018), this suggests that working in teams provides a buffer against human capital losses – coauthors help sustain productivity – while solo projects are more vulnerable to negative productivity shocks and might be more likely to be de-prioritized.

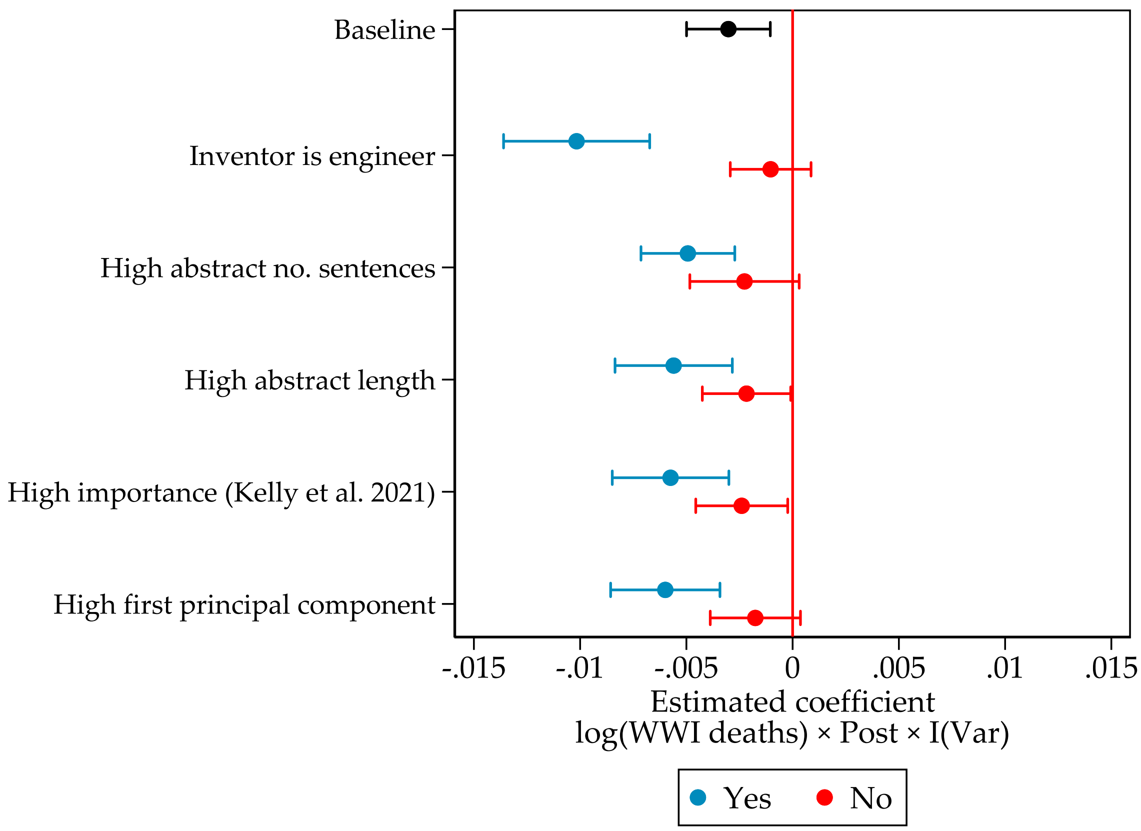
[33]Throughout our study period, the length of patent abstracts was not regulated, resulting in considerable variation. Prior to 1978, providing an abstract was optional; nevertheless, fewer than 9 percent of patents in our PATSTAT-based sample lack an abstract.

[34]See Table A.6 in the Appendix for full estimates.

[35]To facilitate interpretation across columns, all coefficients are rescaled by the corresponding sample mean of the outcome.

FIGURE 5

HETEROGENEOUS EFFECTS BY PATENT COMPLEXITY

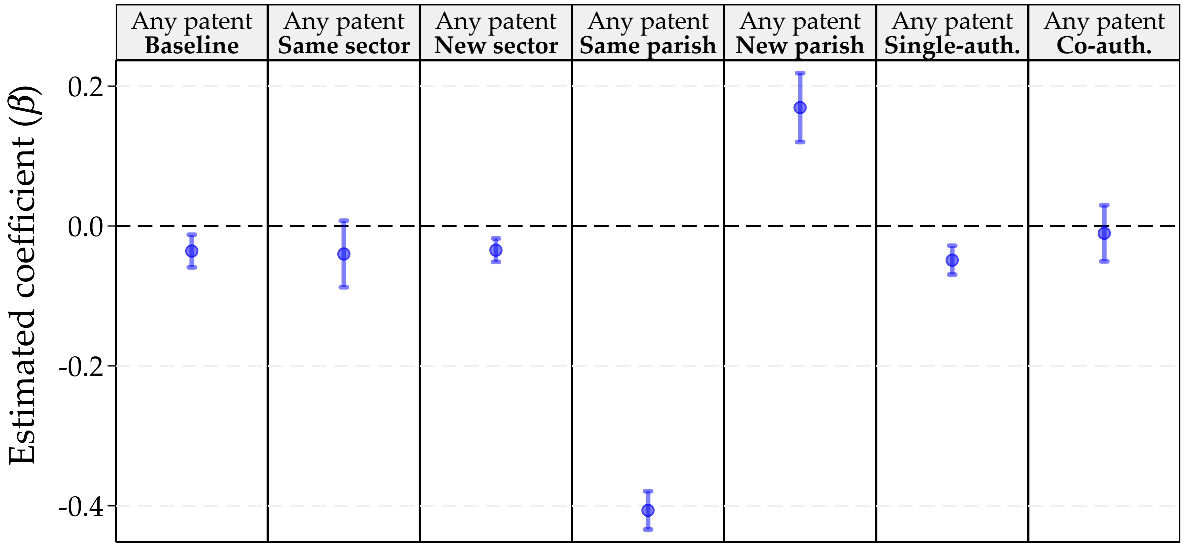


Notes: This figure presents inventor-level OLS estimates of the impact of WWI deaths on patenting outcomes, examining how effects vary according to characteristics of the inventor's first patent (see main text for details). The first row displays the overall (uninteracted) effect of WWI mortality. The following rows show estimated effects separately for inventors who reported being engineers versus non-engineers at the time of their first patent. Additional rows present effects for inventors whose first patent was of high or low complexity, defined as above or below the median based on measures such as abstract length or sentence count. The figure further distinguishes between high and low patent importance, and includes a summary using the first principal component of all measures. The outcomes are binary indicators for having patented in a given period, and coefficients are reported for each interaction term. All models include time and parish fixed effects. Confidence intervals constructed using standard errors clustered at the inventor level.

In sum, our results provide suggestive evidence consistent with recent research on exposure effects. Relocating to communities less affected by human capital losses enables inventors to offset the negative impact on their productivity, likely by accessing a larger and higher-quality pool of potential collaborators and co-workers. At the same time, we confirm the importance of team-specific innovative capital, as previously documented for the United States (Jaravel, Petkova and Bell, 2018), by showing that inventors who relied more heavily on teamwork were better able to mitigate the adverse effects of the shock.

FIGURE 6

INVENTOR-LEVEL ESTIMATES – SECTORS, MOVERS, AND CO-AUTHORSHIP



Notes: Inventor-level OLS estimates of the effect of WWI deaths on a patenting from eq. 3. All coefficients are rescaled by the corresponding sample mean of the outcome to facilitate comparability across specifications. Col. 1 shows the baseline estimate for comparison, using as outcome an indicator for registering a patent in a given period. In col. 2 the outcome is defined as patenting in the same sector as the one the inventor first patented in, while in col. 3 we focus on patents in a new sector. Col. 3-4 use as outcomes indicators for a patent registered int he same parish as the initial one, or a new one, respectively. Finally, in col. 5-6 the we use indicators for a registered patent being single- or co-authored. All specifications include time, parish, and county-time fixed effects. Confidence intervals constructed using s.e. clustered at the parish level.

7. Robustness checks

7.1. Strong parallel trends

As discussed in Section 4, difference-in-differences analysis using a continuous treatment variable require assuming a stronger version of the parallel-trends assumption (Callaway, Goodman-Bacon and Sant'Anna, 2024), which cannot be directly tested. To assess its plausibility in our setting, we construct an empirical exercise that mimics as closely as possible the theoretical comparison across all possible mortality levels, before and after the treatment. Specifically, to produce Figure A.9 in the Appendix, we divide the sample based on deciles of the distribution of $log(d_i^{WWI})$ and estimate event studies specifications in which we compare every decile of $log(d_i^{WWI})$ with each other (one comparison at the time) and using a binary treatment indicator (i.e., being in the higher decile of the two). Then, for each regression, we formally test for joint statistical significance of the pre-war interaction coefficients and save the p-values. Figure A.9 presents the distribution of p-values obtained performing all the regressions. Only one of the attempted regression (36 regressions) shows a statistically significant imbalance in pre-treatment coefficients (p=0.048). We conclude that the strong parallel trends assumption is likely to hold in our setting.

7.2. IV results

In this subsection, we discuss an alternative instrumental variable strategy that does not rely on any of the parallel trends assumptions needed for the diff-in-diffs approach. Specifically, we instrument the log of WWI deaths with the log of predicted deaths based on battalion-level mortality rates. The instrument is constructed by aggregating battalion-level mortality to the parish level, using the share of a parish's servicemen assigned to each battalion as weights. The intuition behind this approach is that during WWI, servicemen were assigned to battalions – fighting units roughly 1,000 strong – in a sequential manner based on enlistment order (Bet-El, 2009; Carozzi, Pinchbeck and Repetto, 2023). This sequential assignment is plausibly exogenous, as it is driven by the order of arrival rather than by underlying parish characteristics. Consequently, serving in one battalion rather than in the next is unlikely to be correlated with pre-existing determinants of innovation at the parish level.[36]

Specifically, indexing battalions with j and parishes with i, we instrument $log(d_i^{WWI})$ with

$$z_i = Log\left(m_i \sum_{j=1}^{J} \alpha_{ij}\tilde{\delta}_j\right),$$

where m is mobilisation, $\alpha_{ij} = m_{ij}/m_i$ is the share of mobilised in regiment j and, finally, $\tilde{\delta}_j = \frac{d_j - d_{ij}}{m_j - m_{ij}}$ is battalion j's leave-out-mean death rate.

This instrument has a shift-share structure, with shares α_{ij} and shocks δ_j, estimated by the leave-out means $\tilde{\delta}_j$. Because, in this setting, variation in mortality across battalions is likely to be driven mostly by where they were deployed and by the fortunes of war, we follow the approach that assumes shocks are exogenous (Borusyak, Hull and Jaravel, 2022). Formally, our identification assumption is that our measures of battalion-level mortalities are conditionally uncorrelated to other parish-level determinants of patenting activity. [37]

Borusyak, Hull and Jaravel (2022) show that in shift-share designs with exogenous shocks, the IV orthogonality condition can be formulated either as requiring that the instrument is uncorrelated with parish-level unobservable determinants of the outcome or, equivalently, as imposing that the shocks – here, battalion-level mortality rates – are uncorrelated with shock-level unobservables.

In Figure A.10, we provide evidence in favour of this orthogonality assumption using either approach. In panel A, we regress several parish-level pre-WWI characteristics on the instrument.[38] Most coefficient estimates are close to zero and statistically insignificant at

[36]Construction of the instrument is possible only for battalions and parishes that appear in both the CWGC-FWR data and the FamilySearch data. As a consequence, the sample size is reduced by about one third in the IV results. For additional technical details on the instrument's construction, we refer the reader to Carozzi, Pinchbeck and Repetto (2023).

[37]This assumption is weaker than requiring that random assignment of servicemen to battalions. In fact, the necessary and sufficient condition for instrument exogeneity is requiring the death rates δ_j not to be systematically related to any other pre-WWI parish-level characteristic that determine present or future innovation.

[38]All specifications include the logarithms of 1911 population and WWI mobilisation, historic county dummies

conventional levels, indicating that the instrument is not correlated with observable characteristics that could affect patenting activity. In panel B of Figure A.10, we conduct a similar after aggregating the data at the battalion level, using the shares of servicemen from each parish serving in a battalion as weights (Borusyak, Hull and Jaravel, 2022).[39] This amounts to regressing battalion-level mortality shocks (δ_j) on the economic, demographic or geographical characteristics of the community of origin of soldiers that make up these battalions. Once again, we find evidence in favour of the orthogonality assumption. Battalion-level death rates are uncorrelated with pre-determined parish level characteristics.

We confirm the relevance of the instrument for our parish-level analysis in Panel A of Table 6, which shows large first-stage coefficients across all four specifications that mirror those presented in Table 1. F-statistics are well above 1000 in all specifications, suggesting the instrument is strong throughout.

Panel B of Table 6 presents the main parish-level IV estimates. Across the four specifications, the estimated effects of WWI deaths are very similar (but slightly larger in magnitude) than the corresponding OLS estimates reported in Table 1. The event-study results, shown in Figure 7, are also qualitatively similar to the OLS counterpart in Figure 2. In particular, pre-trend coefficients are close to zero and statistically insignificant, further supporting the assumption that the instrument is uncorrelated with pre-WWI determinants of innovation. Post-WWI coefficients indicate an immediate drop in innovative activity during the war, a recovery in the interwar period, and another reduction during WW2 followed by a modest rebound. We conclude that endogeneity of WWI deaths does not appear to be a primary concern for our main results. The instrumental variable estimates for the inventor-level analysis also confirm the OLS Difference-in-Differences results and are reported in Table A.7 in the Appendix.

7.3. Additional robustness checks

We next summarize additional robustness checks that address several potential concerns. First, our baseline treatment uses the log of WWI deaths, which excludes parishes with zero fatalities. Figure A.11 and Table A.8 re-define treatment as $\log(1+\text{deaths})$, thereby retaining zero-death parishes (and inventors active there). The results are unchanged.

Second, the main analysis aggregates outcomes to five-year periods. Figure A.12 replicates the event-study specification using two-year intervals, and the qualitative pattern mirrors the baseline.

Third, Table A.9 shows that the findings are robust to substantial changes in the estimation strategy. Panel A uses alternative outcomes, including the number of patents and patents per 1,000 inhabitants (1911 Census). Panel B replaces d^{WWI} with the mortality rate, defined as the ratio of d^{WWI} to 1911 population. Panel C defines a binary exposure indicator, $I(d^{WWI} > med)$, and estimates a binary difference-in-differences model. Table A.9 also reports Poisson pseudo maximum likelihood estimates, following Silva and Tenreyro

and regiment mobilisation shares.

[39] This aggregation is performed using the *ssaggregate* command in Stata 17.

TABLE 6
WWI DEATHS AND PATENTING – PARISH-LEVEL IV ESTIMATES

	$Log(d^{WWI})$ × Post WWI	$Log(d^{WWI})$ × Post WWI	$Log(d^{WWI})$ × Post WWI	$Log(d^{WWI})$ × Post WWI
A. First-stage				
$z\times$ Post WWI	0.946***	0.925***	0.816***	0.816***
	(0.006)	(0.007)	(0.009)	(0.009)
F-stat	21409	16708	8601	8617
Observations	113473	113473	113473	113473
	Any patent	Any patent	Any patent	Any patent
B. IV (second stage)				
$Log(d^{WWI})\times$ Post WWI	-0.014***	-0.014***	-0.011***	-0.011***
	(0.002)	(0.002)	(0.003)	(0.003)
Mean dep.var.	0.30	0.30	0.30	0.30
Observations	113473	113473	113473	113473
Parish FE	Y	Y	Y	Y
Time FE	Y	Y	Y	Y
County × Time	N	Y	Y	Y
Controls × Linear trend	N	N	Y	Y
Population (log)	N	N	N	Y

Notes: Panel A: IV first-stage estimates of the log number of WWI deaths on the instrument, both interacted with an indicator for time periods after 1915-19. Panel B: IV estimates of the effect of WWI deaths on the probability that at least one patent is registered by a resident of parish i during 5-years period t. Specifications follow those in Table 1. Standard errors clustered at the parish level. *, **, and *** represent 10%, 5%, and 1% significance levels, respectively.

(2006), which are well suited to count data with many zeros (Chen and Roth, 2023). All specifications yield results consistent with our baseline dummy-outcome estimates.

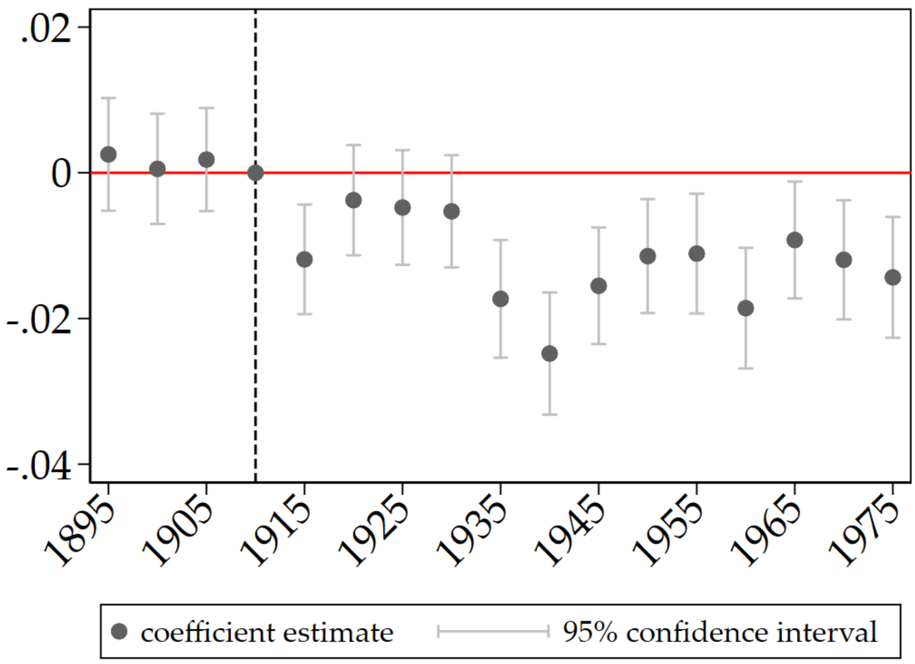
8. Conclusions

In this paper, we examine how the large population shock caused by WWI servicemen mortality affected innovation in British communities. Communities that suffered greater losses during the war experienced a decline in innovation–producing fewer and less influential patents–and remained behind for decades. This effect was particularly pronounced when the losses included highly skilled individuals, such as graduates from elite universities or engineers. By contrast, the effect was substantially smaller in parishes with access to institutions that favour innovation and the spread of knowledge, such as railway stations, post offices, or universities.

Using inventor-level data for innovators who were active before the War, we uncover several additional facts. The negative impact of WWI mortality is larger for inventors who stayed in their parish of origin, but it can be mitigated and even reversed by relocating to less-affected communities–consistent with exposure to a stronger innovation environment

FIGURE 7

WWI DEATHS AND PATENTING – EVENT-STUDY – IV



Notes: IV estimates of the effect of WWI deaths on the probability that at least one patent is registered by a resident of the parish during a given 5-years period. Estimates include time and parish fixed effects and county-time fixed effects. Confidence intervals constructed using s.e. clustered at the parish level.

raising productivity (Bell et al., 2019). The adverse effect is concentrated in advanced sectors and among inventors with an engineering background or more complex patents, whereas co-authorship provides a buffer: solo patenting drops in high-mortality areas while co-authored patenting is largely unaffected. Sector switching, by contrast, does not mitigate the shock. Together, these results point to the importance of a supportive innovation ecosystem and of access to complementary inventors for sustaining inventive activity.

Within the same setting, we document evidence for several mechanisms emphasised in the literature. Losing men – especially the highly educated – harms innovation in the long run by reducing both the output of existing inventors and the entry of new ones. At the same time, pre-war exposure to an innovation-friendly ecosystem and post-war exposure to more favourable environments through relocation can partly offset these losses. Our results thus speak to the literature on what makes places and people innovative (Bell et al., 2019; Akcigit, Grigsby and Nicholas, 2017*b*; Andrews, 2023), showing that both human-capital shocks and the local innovation environment shape long-run inventive activity – and that exposure to better environments, whether through location or collaboration, can partly restore it.

References

Abramitzky, Ran, Adeline Delavande, and Luis Vasconcelos. 2011. "Marrying up: the role of sex ratio in assortative matching." *American Economic Journal: Applied Economics*, 3(3): 124–57.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. 2012. "Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration." *American Economic Review*, 102(5): 1832–56.

Acemoglu, Daron. 2007. "Equilibrium bias of technology." *Econometrica*, 75(5): 1371–1409.

Aghion, Philippe, and Peter Howitt. 1992. "A Model of Growth Through Creative Destruction." *Econometrica*, 60(2): 323–351.

Aghion, Philippe, Ufuk Akcigit, Ari Hyytinen, and Otto Toivanen. 2023. "2022 Klein lecture parental education and invention: the finnish enigma." *International Economic Review*, 64(2): 453–490.

Akcigit, Ufuk, John Grigsby, and Tom Nicholas. 2017*a*. "Immigration and the rise of American ingenuity." *American Economic Review*, 107(5): 327–331.

Akcigit, Ufuk, John Grigsby, and Tom Nicholas. 2017*b*. "The rise of american ingenuity: Innovation and inventors of the golden age." National Bureau of Economic Research.

Akcigit, Ufuk, John Grigsby, Tom Nicholas, and Stefanie Stantcheva. 2022. "Taxation and innovation in the twentieth century." *The Quarterly Journal of Economics*, 137(1): 329–385.

Andersson, David, Mounir Karadja, and Erik Prawitz. 2022. "Mass migration and technological change." *Journal of the European Economic Association*, 20(5): 1859–1896.

Andrews, Michael J. 2023. "How do institutions of higher education affect local invention? Evidence from the establishment of US colleges." *American Economic Journal: Economic Policy*, 15(2): 1–41.

Andrews, Michael J, and Alexander Whalley. 2022. "150 years of the geography of innovation." *Regional Science and Urban Economics*, 94: 103627.

Audretsch, David B, and Maryann P Feldman. 1996. "R&D spillovers and the geography of innovation and production." *The American economic review*, 86(3): 630–640.

Bailey, Roy E, Timothy J Hatton, and Kris Inwood. 2023. "Surviving the Deluge: British servicemen in World War I." *Economics & Human Biology*, 49: 101216.

Bell, Alex, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. 2019. "Who becomes an inventor in America? The importance of exposure to innovation." *The Quarterly Journal of Economics*, 134(2): 647–713.

Bergeaud, Antonin, and Cyril Verluise. 2024. “A new dataset to study a century of innovation in Europe and in the US.” *Research Policy*, 53(1): 104903.

Berger, Thor, and Erik Prawitz. 2024. “Inventors among the "Impoverished Sophisticate".” *The Journal of Economic History*, 84(4): 1175–1207.

Bet-El, Ilana R. 2009. *Conscripts: forgotten men of the Great War.* The History Press.

Boehnke, Jörn, and Victor Gay. 2020. “The Missing Men World War I and Female Labor Force Participation.” *Journal of Human Resources*, 0419–10151R1.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel. 2022. “Quasi-experimental shift-share research designs.” *The Review of Economic Studies*, 89(1): 181–213.

Brainerd, Elizabeth. 2017. “The lasting effect of sex ratio imbalance on marriage and family: Evidence from World War II in Russia.” *Review of Economics and Statistics*, 99(2): 229–242.

Brakman, Steven, Harry Garretsen, and Marc Schramm. 2004. “The strategic bombing of German cities during World War II and its impact on city growth.” *Journal of Economic Geography*, 4(2): 201–218.

BWO. 1922. “Statistics of the Military Effort of the British Empire During the Great War, 1914-1920.”

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro HC Sant’Anna. 2024. “Difference-in-differences with a continuous treatment.” National Bureau of Economic Research.

Carey, Gordon Vero. 1921. *The war list of the University of Cambridge: 1914-1918.* CUP Archive.

Carozzi, Felipe, Edward Pinchbeck, and Luca Repetto. 2023. “Scars of War: The Legacy of WWI Deaths on Civic Capital and Combat Motivation.” CEPR, CEPR Press, Paris and London. https://cepr.org/publications/dp18343.

Chen, Jiafeng, and Jonathan Roth. 2023. “Logs with Zeros? Some Problems and Solutions*.” *The Quarterly Journal of Economics*, qjad054.

Chupilkin, Maxim, and Zsoka Koczan. 2022. *The economic consequences of war: Estimates using synthetic controls.* European Bank for Reconstruction and Development.

Commonwealth War Graves Commission. 2023. “Online Casualty Database.” [Online; accessed 1-Jan-2021].

Correia, Sergio. 2015. “Singletons, cluster-robust standard errors and fixed effects: A bad mix.” *Technical Note, Duke University*, 7.

Correia, Sergio, Paulo Guimarães, and Tom Zylkin. 2020. "Fast Poisson estimation with high-dimensional fixed effects." *The Stata Journal*, 20(1): 95–115.

Craig, Edwin Stewart. 1920. *Oxford University roll of service.* Clarendon Press.

Davis, D. R., and D. E. Weinstein. 2002. "Bones, bombs, and break points: The geography of economic activity." *American Economic Review*, 92(5): 1269–1289.

Edgerton, David EH, and Sally M Horrocks. 1994. "British industrial research and development before 1945 1." *The Economic History Review*, 47(2): 213–238.

European Patent Office. 2023. "PATSTAT Global – Autumn 2023." European Patent Office, Vienna, Austria.

FamilySearch. 2023. "United Kingdom World War I Service Records." [Online; accessed 1-Jan-2021].

Forces War Records. 2023. "Online Database." [Online; accessed 1-Jan-2021].

Franck, Raphaël. 2024. "Labor scarcity, technology adoption and innovation: evidence from the cholera pandemics in 19th century France." *Journal of Economic Growth*, 1–41.

Glaeser, Edward L, Hedi D Kallal, Jose A Scheinkman, and Andrei Shleifer. 1992. "Growth in cities." *Journal of political economy*, 100(6): 1126–1152.

Griliches, Zvi. 1990. "Patent Statistics as Economic Indicators: A Survey." *Journal of Economic Literature*, 28(4): 1661–1707.

Hanlon, W Walker. 2022. "The rise of the engineer: Inventing the professional inventor during the Industrial Revolution." National Bureau of Economic Research.

James, Brigadier EA. 2012. *British Regiments 1914-1918.* Andrews UK Limited.

Jaravel, Xavier, Neviana Petkova, and Alex Bell. 2018. "Team-specific capital and innovation." *American Economic Review*, 108(4-5): 1034–1073.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2021. "Measuring technological innovation over the long run." *American Economic Review: Insights*, 3(3): 303–320.

Lin, Erin. 2022. "How war changes land: Soil fertility, unexploded bombs, and the underdevelopment of Cambodia." *American Journal of Political Science*, 66(1): 222–237.

Maloney, William F, and Felipe Valencia Caicedo. 2022. "Engineering growth." *Journal of the European Economic Association*, 20(4): 1554–1594.

Minnesota Population Center. 2019. "Integrated Public Use Microdata Series, International: Version 7.2 [dataset]."

Moretti, Enrico. 2021. "The effect of high-tech clusters on the productivity of top inventors." *American Economic Review*, 111(10): 3328–3375.

Nagaoka, Sadao, Kazuyuki Motohashi, and Akira Goto. 2010. "Patent statistics as an innovation indicator." In *Handbook of the Economics of Innovation*. Vol. 2, 1083–1127. Elsevier.

Neffke, Frank M. H. 2019. "The value of complementary co-workers." *Science Advances*, 5(12): eaax3370.

Nicholas, Tom. 2011. "Independent invention during the rise of the corporate economy in Britain and Japan 1." *The Economic History Review*, 64(3): 995–1023.

Riaño, Juan Felipe, and Felipe Valencia Caicedo. 2024. "Collateral damage: The legacy of the secret war in Laos." *The Economic Journal*, ueae004.

Romer, Paul M. 1990. "Endogenous technological change." *Journal of political Economy*, 98(5, Part 2): S71–S102.

Schürer, Kevin, and Edward Higgs. 2014. "Integrated Census Microdata (I-CeM), 1851-1911 [data collection]." *UK Data Service. SN*, 7481.

Silva, JMC Santos, and Silvana Tenreyro. 2006. "The log of gravity." *The Review of Economics and statistics*, 88(4): 641–658.

Smil, Vaclav. 2005. *Creating the twentieth century: Technical innovations of 1867-1914 and their lasting impact.* Oxford University Press.

Travers, Tim. 1994. "The Army and the Challenge of War 1914-1918." *In D.G. Chandler and I. Beckett (eds.) The Oxford History of the British Army*, 211–234.

University of Portsmouth. 2011. "Great Britain Historical GIS Project." [Online; accessed 1-Feb-2022].

Voth, Hans-Joachim. 2021. "Persistence–myth and mystery." In *The handbook of historical economics*. 243–267. Elsevier.

Voth, Hans-Joachim, Bruno Caprettini, and Alex Trew. 2023. "Fighting for growth: Labor scarcity and technological progress during the British industrial revolution." *CEPR Discussion Papers*, , (17881).

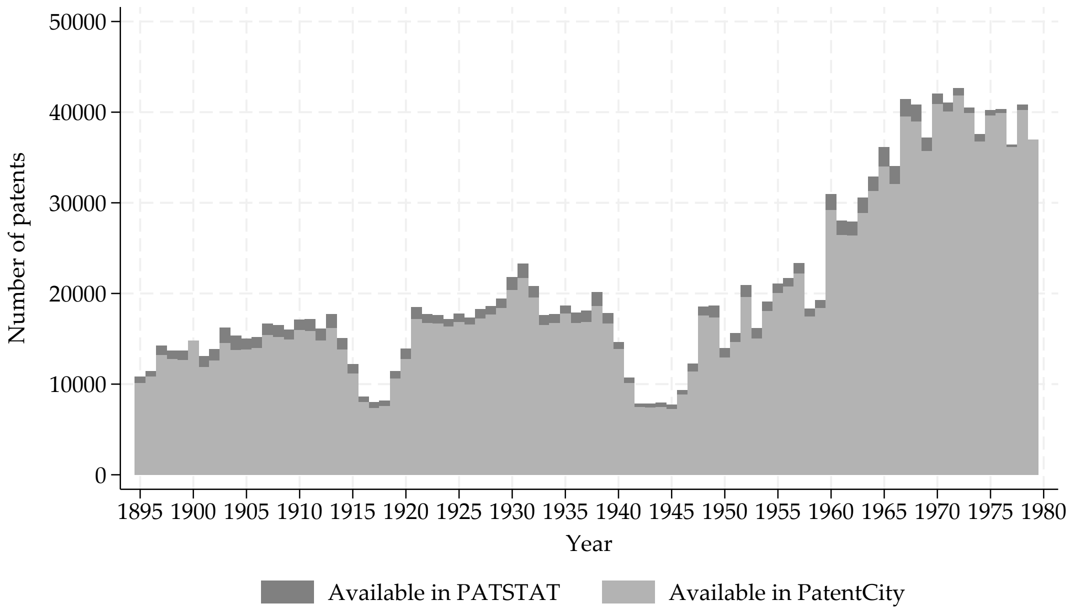
Waldinger, Fabian. 2016. "Bombs, brains, and science: The role of human and physical capital for the creation of scientific knowledge." *Review of Economics and Statistics*, 98(5): 811–831.

Winter, Jay M. 1977. "Britain's 'Lost Generation'of the First World War." *Population Studies*, 31(3): 449–466.

Online Appendix

A. Additional Figures and Tables

FIGURE A.1
PATENTCITY AND PATSTAT DATA



Notes: This figure reports, for each year in our sample, the number of patents registered at the Great Britain patent office and reported in PatentCity (light gray) and in PATSTAT (dark gray), respectively.

Table A.1
Descriptive statistics

	Mean	Std. dev.	Min	Max
A. Parish-level data				
Population 1911	2,697.14	62286.08	0	7004730
Area (sq. km)	11.54	17.97	0	1314
Population density 1911	211.56	761.36	0	23507
Share of households with servants 1911	0.15	0.12	0	1
Male ratio 1911	0.49	0.09	0	1
N. Mobilised in WWI	230.61	5877.70	0	640759
N. Mobilised/Population in WWI (%)	5.52	5.63	0	74
N. Dead in WWI	43.76	1030.10	0	113271
N. Dead/Population WWI (%)	1.04	1.88	0	78
At least one WWI dead (dummy)	0.78	0.41	0	1
At least one WWI dead had elite educ. (dummy)	0.08	0.26	0	1
At least one WWI dead was engineer (dummy)	0.09	0.28	0	1
Ever registered a patent (1894-1979)	0.60	0.49	0	1
Observations	13352			
B. Parish-level panel data (1895-1979)				
Any patent registered (over 5-year period)	0.19	0.40	0	1
N. patents registered (over 5-year period)	3.34	185.75	0	29501
N. patents reg. by existing inv. (over 5-year period)	0.79	51.51	0	12912
N. patents reg. by first-time inv. (over 5-year period)	0.90	41.04	0	7861
Observations	226984			

Notes: Panel A: Descriptive statistics at the (grouped) parish level. Data are either from the 1911 Census or from the WWI war records, as indicated. Panel B: Descriptive statistics from the parish-level panel, where each observation is a parish observed over a 5-year period. The panel covers all 5-year periods in 1895-1979. Any patent is an indicator for having at least one patent being registered by patentee residing in the parish in a given 5-year period. Similarly, n. patents registered is the average number of patents registered in the parish. Existing inventors are those who registered at least one patent in the period 1985-1899. Patents by first-time inventors are those registered by individuals who appear in the data for the first time.

TABLE A.2

WWI DEATHS AND PATENTING – PPLME ESTIMATES

	(1)	(2)	(3)	(4)
$Log(d^{WWI})\times$ Post WWI	-0.103***	-0.099***	-0.049***	-0.026***
	(0.010)	(0.009)	(0.010)	(0.009)
Mean dep.var.	1.63	1.63	1.63	1.62
Observations	109173	109077	109077	108785
Parish FE	Y	Y	Y	Y
Time FE	Y	Y	Y	Y
County × Time	N	Y	Y	Y
Controls × Linear trend	N	N	Y	Y
Population (log)	N	N	N	Y

Notes: Poisson pseudo maximum likelihood estimates of the effect of WWI deaths on the number of patents registered in the parish during a given 5-years period. Estimates obtained using the command *ppmlhdfe* in Stata (Correia, Guimarães and Zylkin, 2020). Standard errors clustered at the parish level. *, **, and *** represent 10%, 5%, and 1% significance levels, respectively. The number of observations is lower in this specification because *ppmlhdfe* drops singletons and observations separated by fixed effects (Correia, 2015).

TABLE A.3

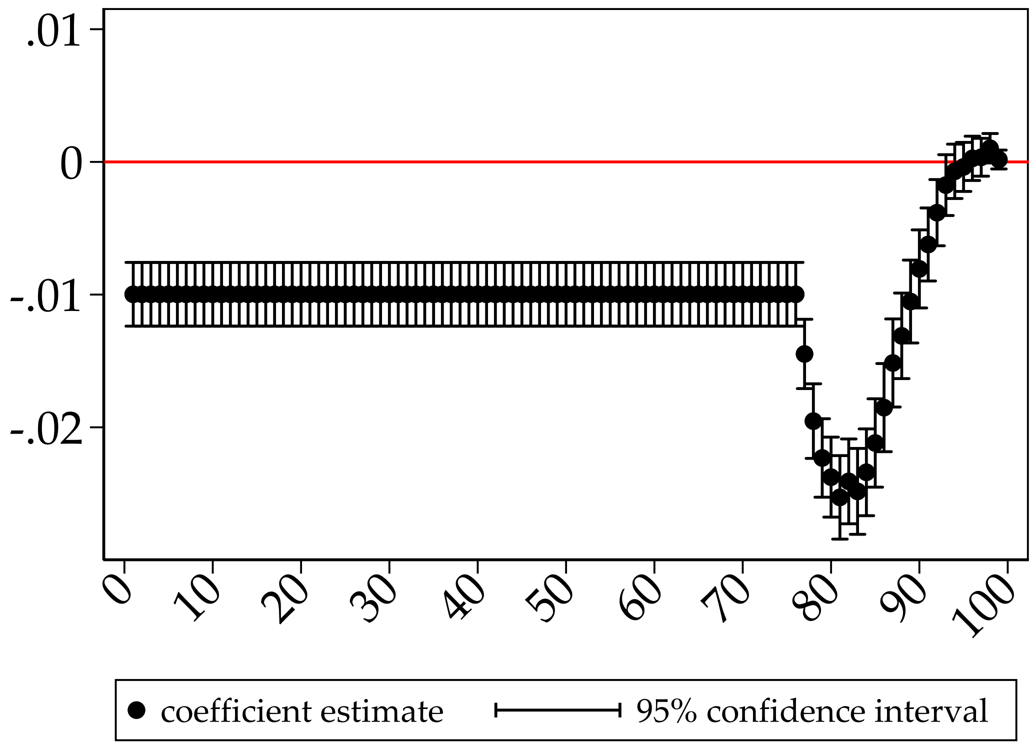
WWI DEATHS AND PATENTING – OTHER SPECIFICATIONS

	(1) Any patent	(2) N. patents	(3) Patents p.c.
A. Log deaths			
$log(d^{WWI})\times$ Post WWI	-0.01***	-0.10***	-0.02***
	(0.001)	(0.011)	(0.002)
Observations	176228	122388	176228
B. Death rate p.c.			
$d^{WW1}p.c.\times$ Post WWI	-0.27***	-10.18***	-0.39**
	(0.087)	(2.174)	(0.154)
Observations	226542	137279	225895
C. Deaths above median			
$I(d^{WW1} > med)\times$ Post WWI	-0.02***	-0.39***	-0.05***
	(0.004)	(0.063)	(0.006)
Observations	226967	137296	225985

Notes: Each cell in the Table represents a separate regression. Estimates in columns 1 and 3 are obtained by OLS. Estimates in column 2 are obtained by poisson pseudo maximum likelihood estimation using the command *ppmlhdfe* in Stata (Correia, Guimarães and Zylkin, 2020). All regressions control for parish fixed effects and County x time fixed effects. Standard errors clustered at the parish level. *, **, and *** represent 10%, 5%, and 1% significance levels, respectively. The number of observations is lower throughout column 2 because *ppmlhdfe* drops singletons and observations separated by fixed effects (Correia, 2015). Patents p.c are patents per capita * 1000. To account for outliers, values of this variable above the 95% percentile are set equal to the 95% percentile value.

Figure A.2

WWI deaths and patenting – OLS estimates – Patent rate above each percentile of distribution of patent rates



Notes: OLS estimates of the effect of WWI deaths on the probability that the number of patents registered by residents of parish i during 5-years period t, divided by the population of parish i at the 1911 census is above each percentile of the distribution of patent rates in our sample. Each coefficient is a different regression, where the dependent variable refers to the percentile specified on the horizontal axis. Estimates are based on equation 1, including time and parish fixed effects, county-time fixed effects, and controls. Confidence intervals constructed using s.e. clustered at the parish level.

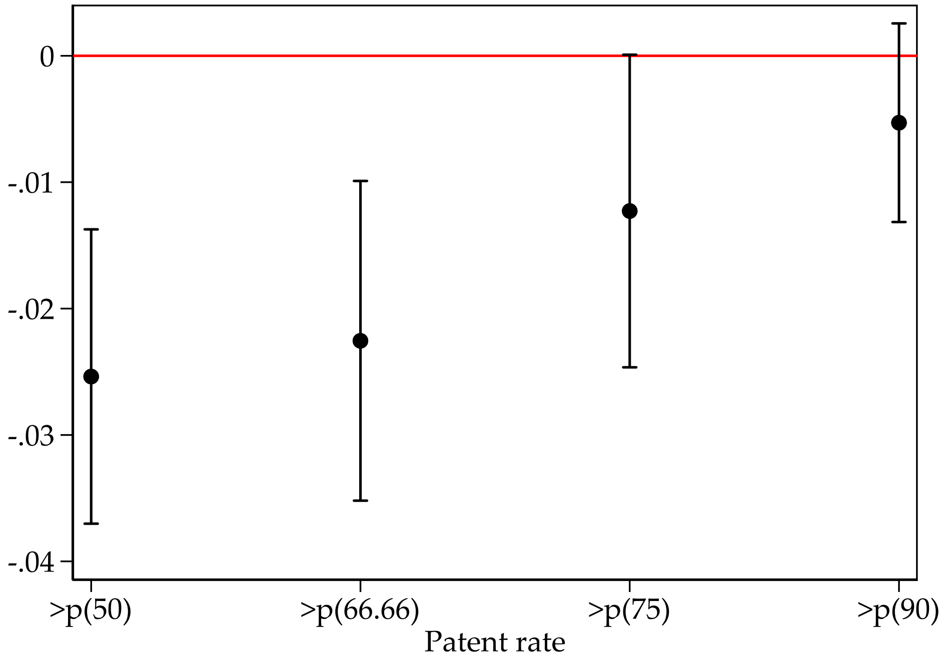
TABLE A.4

WWI DEATHS AND PATENTING – OLS ESTIMATES - LARGE AND SMALL PARISHES

	(1) Pop. $\leq$ 2,500 Any patent	(2) Pop. $\leq$ 2,500 Any patent	(3) Pop. $>$ 2,500 Any patent	(4) Pop. $>$ 2,500 Any patent
$Log(d^{WWI})\times$ Post WWI	-0.002	-0.004*	-0.002	0.004
	(0.002)	(0.002)	(0.004)	(0.004)
Mean dep.var.	0.15	0.15	0.74	0.74
R2	0.34	0.34	0.41	0.42
Observations	152059	152059	25200	25200
Parish FE	Y	Y	Y	Y
Time FE	Y	Y	Y	Y
County $\times$ Time	Y	Y	Y	Y
Controls $\times$ Linear trend	Y	Y	Y	Y
Population (log)	N	Y	N	Y

Notes: OLS estimates of the effect of WWI deaths on the probability that at least one patent is registered by a resident of parish i during 5-years period t. In Columns 1 and 2, the sample is restricted to parishes with less than 2,500 inhabitants as measured on the 1911 census. In Columns 3 and 4, the sample is restricted to parishes with more than 2,500 inhabitants as measured on the 1911 census. Standard errors clustered at the parish level. *, **, and *** represent 10%, 5%, and 1% significance levels, respectively.

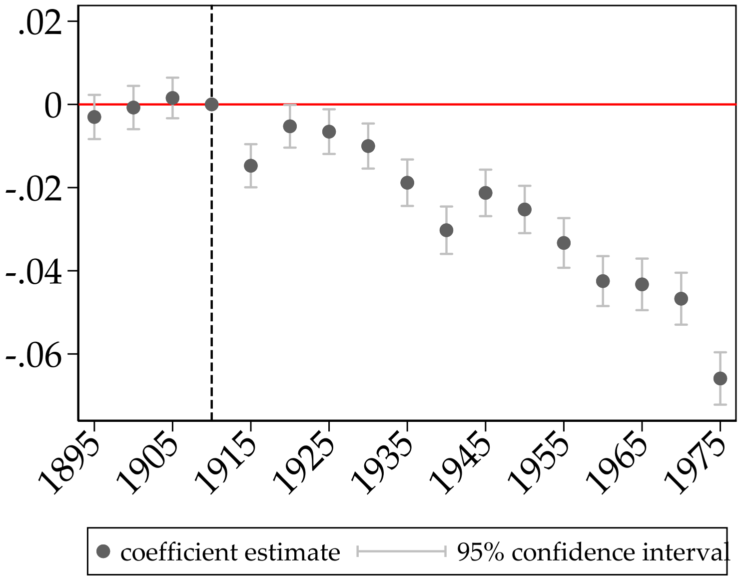
Figure A.3
WWI deaths and patenting – OLS estimates – Intensive margin in large parishes



Notes: OLS estimates of the effect of WWI deaths on the probability that the number of patents registered by residents of parish i during 5-years period t, divided by the population of parish i at the 1911 census is above each percentile of the distribution of patent rates in our sample. Each coefficient is a different regression, where the dependent variable refers to the percentile specified on the horizontal axis. Estimates are based on equation 1, including time and parish fixed effects, county-time fixed effects, and controls. Parishes with less than 2,500 inhabitants are excluded. Parishes-periods pairs for which the number of registered patents is 0 are excluded. Confidence intervals constructed using s.e. clustered at the parish level.

Figure A.4

WWI deaths and patenting – Excluding company patents



Notes: OLS estimates of the effect of WWI deaths on the probability that at least one patent is registered by an individual from the parish during a 5-years period (patents registered by companies are excluded). The mean of this outcome is 0.192. Estimates from equation 2, including time and parish fixed effects and county-time fixed effects. Confidence intervals constructed using s.e. clustered at the parish level.

TABLE A.5

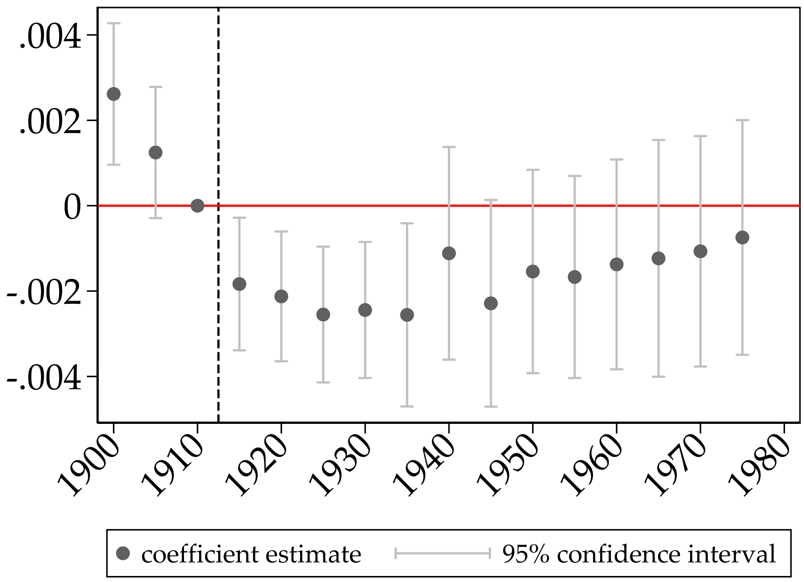
WWI DEATHS AND PATENTING – INVENTOR-LEVEL ESTIMATES – EXCLUDING WWI DEATHS

	Not killed in WWI	Not killed in WWI	Not killed in WWI	Active after WWI
$Log(d^{WWI})\times$ Post WWI	-0.004***	-0.003**	-0.002	-0.005***
	(0.001)	(0.001)	(0.002)	(0.001)
Match on				
Mean dep.var.	0.11	0.12	0.14	0.23
R2	0.33	0.34	0.35	0.29
Observations	137280	124416	103920	74800

Notes: OLS estimates of the effect of WWI deaths on an indicator equal one if the inventor j from parish i registers a patent during 5-years period t. The sample is restricted to patentees who registered at least one patent in 1895-1899. Sample period is then restricted to 1900-1979. Columns 1-3 present results from equation 3 restricting the sample to patentees who do not appear in the CWGC dataset of WWI fatalities, using different matching criteria. In column 1, we require a match on surname and the first 2 name initials; in column 2, surname and the first 3 initials; and, finally, in column 3 we require both surname and first name to match perfectly. In column 4, the sample is restricted to individuals who are active, defined as patentees who registered at least one patent in the post-WWI period. Individual and time effects are included in all specifications. Standard errors clustered at the inventor level. *, **, and *** represent 10%, 5%, and 1% significance levels, respectively.

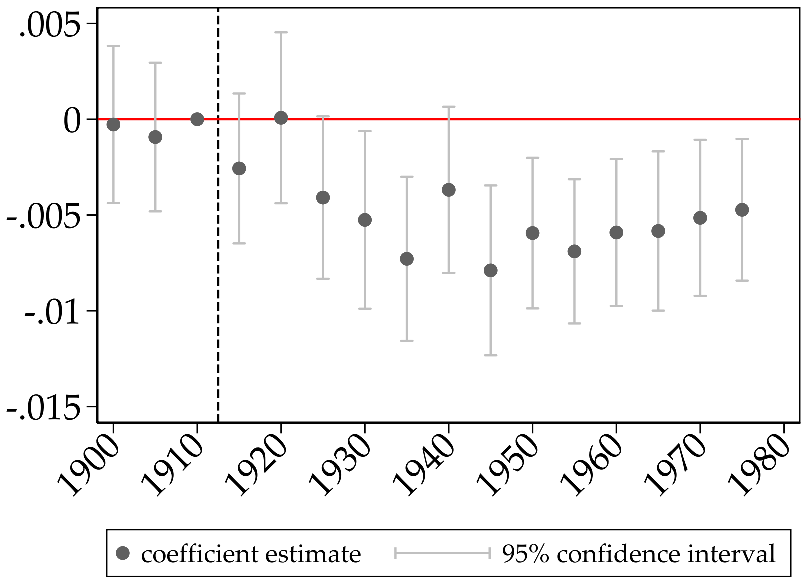
Figure A.5

Inventor-level event study



Notes: Inventor-level OLS estimates of the effect of WWI deaths on the probability that at least one patent is registered during 5-years period t. Estimates from equation 3, including time and parish fixed effects. Confidence intervals constructed using s.e. clustered at the individual level.

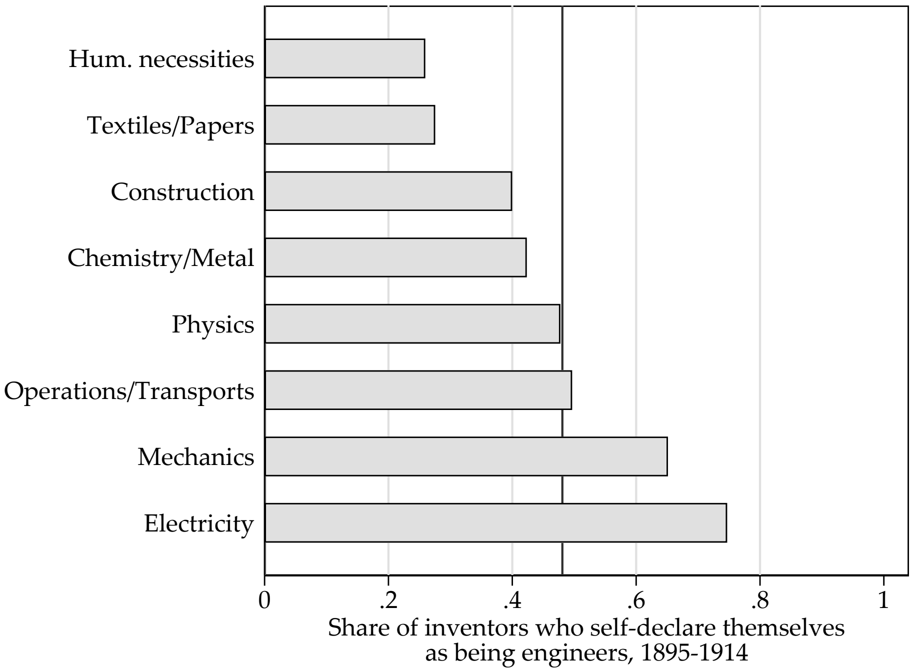
FIGURE A.6
INVENTOR-LEVEL EVENT STUDY – ONLY INDIVIDUALS ACTIVE AFTER WWI



Notes: Inventor-level OLS estimates of the effect of WWI deaths on the probability that at least one patent is registered during 5-years period t. Sample is restricted to patentees that are still active after WWI. Estimates from equation 3, including time and parish fixed effects. Confidence intervals constructed using s.e. clustered at the individual level.

Figure A.7

Predominance of engineers among inventors by field

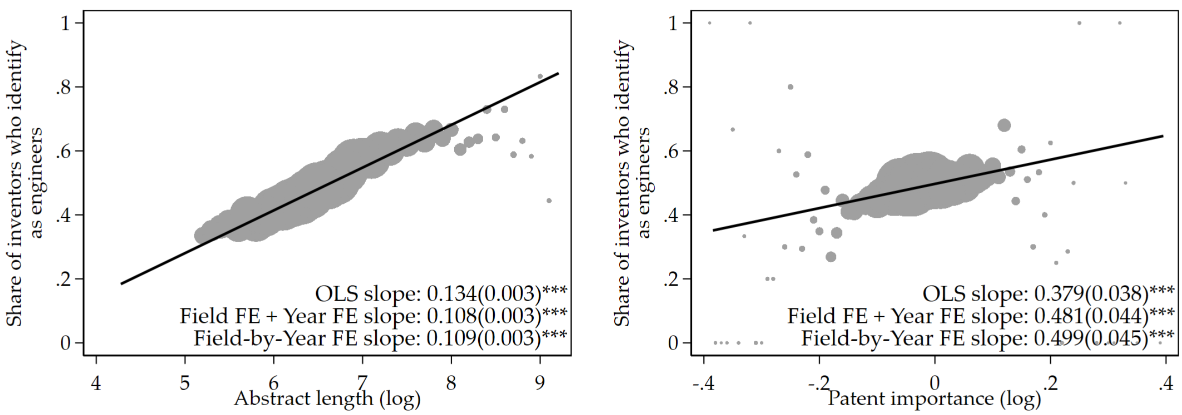


Notes: This figure reports the incidence of engineers among inventors across different fields of specialization in the pre-WWI period. For each major technological sector, we compute the share of inventors who report their occupation as being "engineer". The figure illustrates the extent to which engineering expertise was predominant among British inventors of the period across various technological domains.

Figure A.8

Engineers, patent complexity and importance

(A) Abstract length (B) Patent importance



Notes: Scatter plots of the abstract length (panel A) and patent importance (panel B – measured following Kelly et al. 2021) against the share of authors who were engineers. Estimated OLS slopes (with or without year and field $\times$ year fixed effects) are also reported, with robust s.e. in parentheses.

TABLE A.6

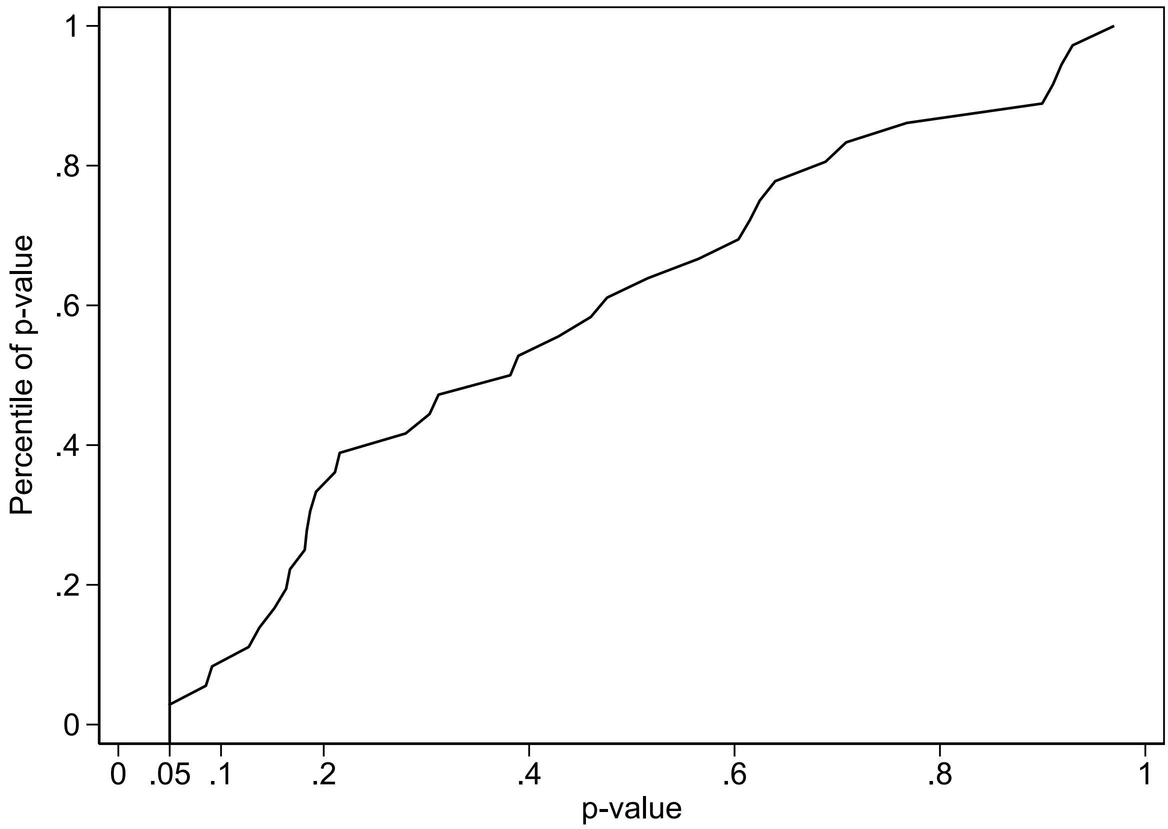
INVENTOR-LEVEL ESTIMATES – SECTORS, MOVERS, AND CO-AUTHORSHIP

		Same vs. new sector		Same vs. new parish		Authorship	
	Any patent	Any patent same sector	Any patent new sector	Any patent Stayer	Any patent Mover	Any patent Single-auth	Any patent Co-auth
$Log(d^{WWI}) \times$ Post	-0.036***	-0.040*	-0.035***	-0.406***	0.169***	-0.049***	-0.010
	(0.012)	(0.024)	(0.009)	(0.014)	(0.025)	(0.011)	(0.020)
Mean dep.var.	0.084	0.034	0.061	0.032	0.057	0.062	0.038
R2	0.29	0.21	0.25	0.24	0.26	0.25	0.20
Observations	288000	288000	288000	288000	288000	288000	288000

Notes: Inventor-level OLS estimates of the effect of WWI deaths on a patenting from eq. 3. All coefficients are rescaled by the corresponding sample mean of the outcome to facilitate comparability across specifications. Col. 1 shows the baseline estimate for comparison, using as outcome an indicator for registering a patent in a given period. In col. 2 the outcome is defined as patenting in the same sector as the one the inventor first patented in, while in col. 3 we focus on patents in a new sector. Col. 3-4 use as outcomes indicators for a patent registered int he same parish as the initial one, or a new one, respectively. Finally, in col. 5-6 the we use indicators for a registered patent being single- or co-authored. All specifications include time, parish, and county-time fixed effects. Confidence intervals constructed using s.e. clustered at the parish level. Standard errors clustered at the parish level. *, **, and *** represent 10%, 5%, and 1% significance levels, respectively.

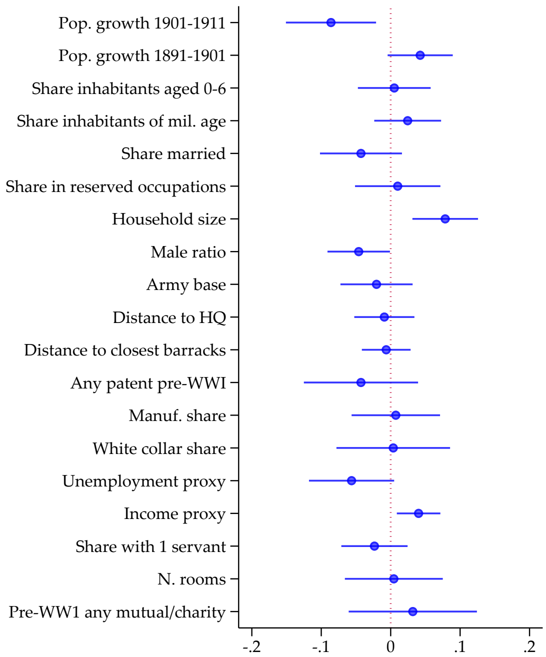
FIGURE A.9

STRONG PARALLEL TRENDS

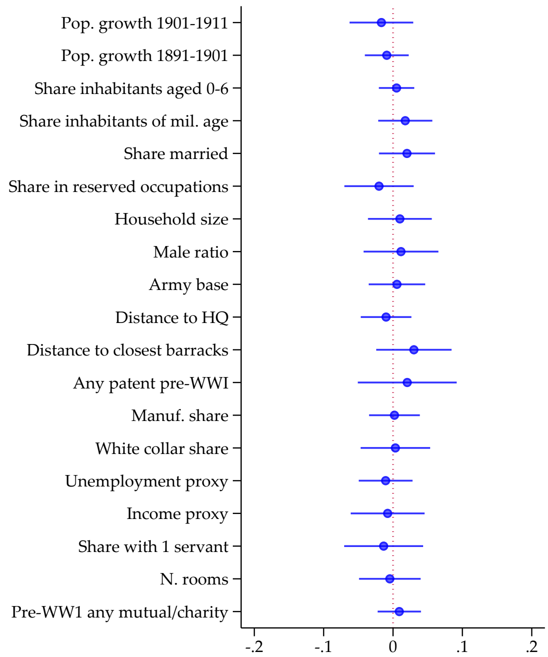


Notes: The figure plots the empirical CDF of p-values from pairwise strong parallel-trends tests across deciles of $\log(\text{number dead})$. For each admissible pair of deciles (i, z) (with $i < z$ and $i+z \leq 10$), we estimate an event-study regression of the outcome on $\mathbf{1}\{\text{decile} = z\} \times \mathbf{1}\{t\}$ indicators (omitted year 1910), with parish and county-time fixed effects and standard errors robust to clustering at the parish level. We test $H_0 : \beta_{1895} + \beta_{1900} + \beta_{1905} = 0$ and plot the distribution of the resulting p-values; the vertical line indicates $p = 0.05$.

Figure A.10
Instrumental Variable Balancing Checks



(A) Parish-level



(B) Battalion-level

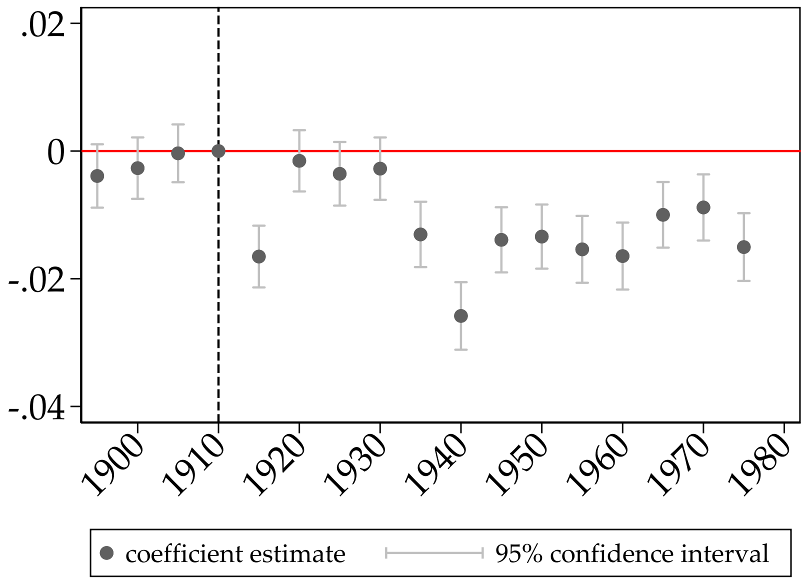
Notes: Panel A: OLS estimates of individual regressions of the instrument z_i on different variables, with 95% confidence intervals. The first coefficient shows the first-stage, i.e., the regression coefficient of the effect of the instrument on the (standardised) instrumented variable, $log(d_i^{WWI})$. All specifications control for the log of 1911 population and WWI mobilisation, regiment mobilisation shares and historic county fixed effects. Panel B: OLS estimates of individual regressions of the battalion-level death rate δ_j on different variables. All variables have been aggregated at the battalion level using the *ssaggregate* command in Stata 17 (residualized using population, mobilisation and mobilisation shares) following Borusyak, Hull and Jaravel (2022). All outcomes are standardised to have mean zero and unit standard deviation. Population growth rates are trimmed at the 1st and 99th percentile, while household size, the number of rooms per person, and area are trimmed at the 99th percentile. Share of inhabitants of military and share of married men are calculated restricting to individuals aged 14-35 at the time of the 1911 Census. Income proxy is the first principal components of white collar share, unemployment, share with 1 servant, and n. rooms per person. Standard errors clustered at the historic county level (panel A) or regiment level (panel B).

TABLE A.7
WWI DEATHS AND PATENTING – INVENTOR-LEVEL IV ESTIMATES

	Any patent	Any patent	Any breakthr. patent	Any highly cited patent
$Log(d^{WWI})\times$ Post WWI	-0.003***	-0.003**	-0.001***	-0.001***
	(0.001)	(0.001)	(0.000)	(0.000)
Sample				
Mean dep.var.	0.08	0.11	0.02	0.02
R2	0.00	0.00	0.00	0.00
Observations	279952	140576	279952	279952

Notes: IV estimates of the effect of WWI deaths on the probability that at least one patent is registered by inventor j during 5-years period t. Specifications follow those in Table 5. Standard errors clustered at the parish level. *, **, and *** represent 10%, 5%, and 1% significance levels, respectively.

FIGURE A.11
WWI DEATHS AND PATENTING – OLS ESTIMATES – INCLUDING PARISHES WITHOUT ANY WWI DEATHS



Notes: OLS estimates of the effect of WWI deaths on the probability that at least one patent is registered by a resident of parish i during 5-years period t. Estimates are based equation 2, including time and parish fixed effects and county-time fixed effects. The treatment variable in this regression is log(1+number of deaths), which implies that parishes without any WWI deaths are not dropped from the sample. Sample size: N=245,599. Confidence intervals constructed using s.e. clustered at the parish level.

TABLE A.8

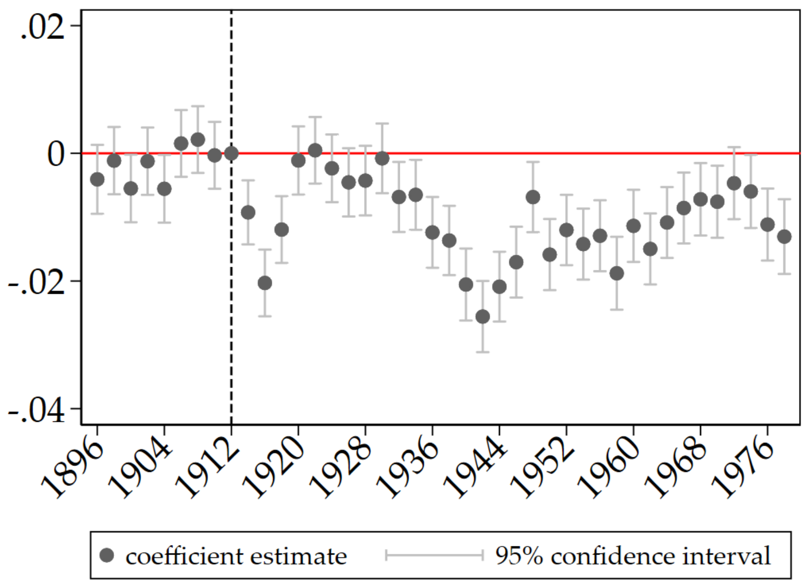
WWI DEATHS AND PATENTING – INVENTOR-LEVEL ESTIMATES INCLUDING PARISHES WITHOUT ANY WWI DEATHS

	(1) Any patent	(2) Any patent
x_post	-0.003***	-0.003***
	(0.001)	(0.001)
Mean dep.var.		
Observations	288464	288464
Parish FE	Y	Y
County × Linear trend	N	Y
Controls × Linear trend	N	Y

Notes: OLS estimates of the effect of WWI deaths on an indicator equal one if the inventor j from parish i registers a patent during 5-years period t. The sample is restricted to patentees who registered at least one patent in 1895-1899. Sample period is then restricted to 1900-1979. The treatment variable in this regression is log(1+number of deaths), which implies that inventors active in parishes without any WWI deaths are not dropped from the sample. Column 1 presents results from equation 3 for the full sample, whereas in column 2 the sample is restricted to individuals who do not appear in the CWGC dataset of WWI fatalities, matched using the initial of the first name and surname. In column 3 and 4, the outcomes are, respectively, an indicator for the inventor registering a "breakthrough" patent or a highly-cited patent (see text for details). Controls, individual and time fixed effects are included in all specifications. Standard errors clustered at the inventor level. *, **, and *** represent 10%, 5%, and 1% significance levels, respectively.

FIGURE A.12

WWI DEATHS AND PATENTING – BI-YEARLY DATA



Notes: OLS estimates of the effect of WWI deaths on the probability that at least one patent is registered by a parish resident during a 2-year period. Estimates are based on equation 2 and include time, parish, and county-time fixed effects. Confidence intervals are based on standard errors clustered at the parish level.

Table A.9

WWI deaths and patenting – other specifications

	(1) Any patent	(2) N. patents	(3) Patents p.c.
A. Log deaths			
$log(d^{WWI})\times$ Post WWI	-0.01***	-0.10***	-0.02***
	(0.001)	(0.011)	(0.002)
Observations	176228	122388	176228
B. Death rate p.c.			
$d^{WW1}p.c.\times$ Post WWI	-0.27***	-10.18***	-0.39**
	(0.087)	(2.174)	(0.154)
Observations	226542	137279	225895
C. Deaths above median			
$I(d^{WW1} > med)\times$ Post WWI	-0.02***	-0.39***	-0.05***
	(0.004)	(0.063)	(0.006)
Observations	226967	137296	225985

Notes: Each cell in the Table represents a separate regression. Estimates in columns 1 and 3 are obtained by OLS. Estimates in column 2 are obtained by poisson pseudo maximum likelihood estimation using the command *ppmlhdfe* in Stata (Correia, Guimarães and Zylkin, 2020). All regressions control for parish fixed effects and County x time fixed effects. Standard errors clustered at the parish level. *, **, and *** represent 10%, 5%, and 1% significance levels, respectively. The number of observations is lower throughout column 2 because *ppmlhdfe* drops singletons and observations separated by fixed effects (Correia, 2015). Patents p.c are patents per capita * 1000. To account for outliers, values of this variable above the 95% percentile are set equal to the 95% percentile value.

B. Data

B.1. Military records

Data on British service personnel killed in WWI was obtained from the Commonwealth War Graves Commission (CWGC) (Commonwealth War Graves Commission, 2023), an intergovernmental organisation dedicated to marking, recording and maintaining the graves, memorials and memories of the men and women of the Commonwealth forces who died in both World Wars. Open data from this organisation contains information on names, time of death, rank, fighting unit, honours (e.g., gallantry medals), age and a string from which we can extract the location of origin of dead soldiers. Data on locations is augmented using information from Forces War Records (FWR), a military genealogy specialist website (Forces War Records, 2023). To implement this merge, we searched FWR for all soldiers in CWGC, and record place of birth and residence whenever available. After merging the datasets we perform a number of cleaning operations and restrict attention to soldiers from England and Wales who were killed between 1 August 1914 and 31 December 1918. We then geolocate soldiers to 1911 parish of origin by a combination of matching location strings to parish names and batch geolocation. When assigning soldiers to parishes, we prioritise the FWR data and residential location such that we only use birthplace when residence is missing or cannot be matched to a parish. The geolocation procedure assigns parishes to 73% (585,371) of the soldiers killed.

We next use ancillary data to identify members of the social elite in our CWGC-FWR data. Our primary method to do so relies on the Oxford University Roll of Service (Craig, 1920), and The War List of the University of Cambridge 1914-1918 (Carey, 1921). These documents name all faculty, enrolled students, and alumni of these pre-eminent universities that served during the war. We digitise these documents and match individuals listed to soldiers recorded in the processed CWGC-FWR data using surnames, initials, and rank. We successfully match 3305 individuals (23%) from the Oxford and 3260 (23%) from the Cambridge lists to our soldier-level fatalities dataset in this way.[40] As this definition of elites is highly selective, we then specify an alternative measure of social elites by using the information in the rank field of the CWGC-FWR data to construct a dummy that proxies for a killed soldier being an officer (either commissioned or non-commissioned).[41]

Data on 4,135,026 war records of soldiers mobilised during WWI is obtained from FamilySearch, a non-for-profit organisation which offers on-line access to large genealogical datasets (FamilySearch, 2023). FamilySearch draws its information from the British Army Service Records for 1914 to 1920. These records contain information on enrolled soldiers including names, place of residence, birthplace, age at the time of enlistment, year and unit in which the soldier was enlisted.[42] When cleaning and processing this information, we use as reference the Table of Organisation of each regiment as detailed in James (2012).

[40] The match rate reflects the high mortality in these groups, well above the mortality in the general population of servicemen.

[41] The ranks recorded in this field are highly heterogeneous for officers but not for privates, so we define officers as those whose rank does not coincide with a set of predetermined strings (e.g. “Private”, “Rifleman").

[42] Digitised versions of these records can be consulted at `www.ancestry.co.uk`. The FamilySearch collection,

To construct our instrument, we apply several sample restrictions to the FamilySearch data. We start with the original 4.1 million war records and exclude 2.05 million records for which the battalion is missing. We then drop 57,795 duplicate entries (defined as identical across all variables) and 72,131 records dated before 1905 or after 1920. From the remaining sample, we exclude 735,768 individuals that could not be geolocated and 28,578 from regiments with zero or negligible mortality, such as the Hussars. Finally, to ensure we have sufficient observations to construct the shares serving in each battalion, we drop 38,281 soldiers from battalions with fewer than 100 servicemen. This leaves us with a final sample of just over 2.6 million geolocated servicemen with complete battalion information.

B.2. Further data sources

Individual-level information on the English and Welsh population before the Great War is obtained from the 1911 Census of population. The data we use originates from Schürer and Higgs (2014) and is distributed by IPUMS (Minnesota Population Center, 2019). We use this data both at the individual level and to construct aggregates at the parish level. From this source, we obtain information on the several income proxies including the number of servants and the number of rooms per household. We obtain aggregate area-level information for Census years 1901-1931, as well as digital maps for parishes, districts and constituencies from "A Vision of Britain through Time" (VoB), an online library of spatial data created by the Geography Department at the University of Portsmouth (University of Portsmouth, 2011). The parish-level population counts in the VoB data come from the *Census Reports* that were published following each Census. There are known to be discrepancies between the population counts in this source and the more recently published micro-data, for example because not all records have survived or there is ambiguity in the true parish in the individual level records. Consequently, in general we use the counts from the *Census Reports* where available. Further, to minimise discrepancies we also implement the corrections to assigned parishes in the 1911 micro-data using the look-up tables published on the I-CeM website.[43]

We obtain membership lists of four engineering trade bodies in pre WWI years – the Institution for Mechanical Engineers, 1896; Iron and Steel Institute, 1888; Institution of Automobile Engineers, 1910; Institution of Electrical Engineers, 1887 – from the website Grace's Guide, a list of pre-WWI universities from Wikipedia, a list of rail stations in 1910 from a GIS first constructed by Jordi-Marti Henneberg, and a georeferenced list of all post offices in 1900 from the GB1900 Gazetteer. We geocode the first two sources using batch geolocation of address information. We then use these sources to construct a series of separate County-level ecosystem accessibility measures. Counties are large – there are 52 in England

which includes the extracted data, is called "United Kingdom, World War I Service Records, 1914-1920". The original sources of this information are the "Burnt documents" (record code WO 363) and the "Unburnt collection" (record code WO 364), which are kept in the National Archives at Kew in London. The Burnt Documents are roughly 2.5 million records on WWI soldiers which survived the fire resulting from an incendiary bomb hitting the War Office Record Store in 1940. The Unburnt Collection is made of soldier information obtained from pension claims. This collection was stored separately in 1942 and, therefore, did not suffer the fate of many of the Burnt Documents.

[43]Available at https://www.essex.ac.uk/-/media/newparids11.txt?la=en, accessed on May 5, 2023.

and Wales – and nest parishes. To construct each measure, we first compute the linear distance between each parish and the nearest ecosystem factor (e.g., a university), then obtain the average distance to that factor in each County by taking the weighted average of the parish level distances, using as weights 1911 parish populations. We then take the natural log of this weighted average distance.

B.3. Spatial Units of Analysis and Reconciliation

Our main analysis is based on a 1911 parish-level dataset covering England and Wales. We take 1911 as our reference year because it was the last Census conducted before the onset of the Great War in 1914. The civil parishes we use in our analysis are administrative units corresponding to the lowest level of local government in the United Kingdom. Civil parishes evolved from ecclesiastical parishes during the 19^{th} century, but by 1880 had no religious or ecclesiastical duties. In 1911, the territories of England and Wales were divided into 14,664 parishes, of which 13,404 in England and 1,260 in Wales. We drop all parishes that had zero population in 1911 – usually parcels of empty land in remote rural areas – and 10 additional parishes that have repeated names within the same county.

Parishes are nested within local government districts, of which there were 1,861 in 1911, and in turn within 52 counties. Parish boundaries change over time and in some cases variables are only available at other (higher) levels of aggregation. In order to aggregate or re-weight information to common boundaries we use a spatial matching procedure based on the assumption of uniform population distribution within parishes. Because our main spatial units (parishes) are relatively small (10 sq. km on average) and parish boundaries are often quite stable in the 30-year period we study, we expect the measurement error induced by making this assumption to be limited. We selectively group parishes using a semi-automated approach that groups those with similar names and that lie in close proximity. For example, we group High Abbotside with Low Abbotside and we also group Ledbury Urban with Ledbury Rural. In addition to this, we group a small number of suburban parishes with the corresponding city parish. For example, we group the suburb of Sculcoates with Hull. Reflecting that soldiers from the London area often report their location as being London, we also group together London parishes based on the historical conurbation definition available from Vision of Britain. Finally, we exclude 26 parishes, of which 25 are parcels of empty land and the other is unnamed. After restrictions and grouping, our final parish set encompasses 13,352 parishes.

Our data on 1911 parishes come from two different sources: the 1911 Census micro-data from I-CeM and the *Census Reports* from VoB. These sources use different parish codes and contain a slightly different set of parishes, so we create a mapping file and reconcile the data before conducting analysis.

B.4. Geolocation Procedure, Measurement Error, and Validation

Our empirical analysis requires adequately geolocating soldiers based on information on their place of birth and residence from the sources described above. Here we provide details of the geolocation procedures used to assign soldiers to their parish of origin.

The CWGC data on soldiers killed during WWI includes 796,601 records.[44] Given that our analysis will focus on England and Wales only, we remove servicemen born in Scotland, Ireland, and abroad. We then extract information for residence or birthplace (or both) from either the birthplace and residence fields in FWR or the "additional information" string included in the CWGC source.

Geolocation of WWI dead soldiers proceeds by combining a) direct string matches with parish names based on data from FWR on historic county and location of birthplace/residence, b) direct string matching as above but based on the CWGC additional information field, and c) latitudes and longitudes obtained from a batch geolocating service to which we input the FWR locations. For the batch geolocation process, we use a service provided by the company OpenCageGeo, which is based on OpenStreetMap and is available across platforms. In order to validate the geolocation process used by this source, we randomly selected 800 individual servicemen and validated the imputed locations by hand. Only 9 observations in this sample were incorrectly imputed and 6 of these 9 were imputed to nearby areas. Hence, we conclude that the geolocation process based in this method is sufficiently reliable for our purposes, resulting in a limited amount of measurement error.

The data on parish of origin (birthplace or residence) of mobilised men in WWI – obtained from FamilySearch – has a slightly different structure and, therefore, we use a different procedure from the one used for CWGC/FWR data.[45] To match the FamilySearch records to an individual parish we combine: a) a direct string match with parish names for records that have both an historic county and a location, b) direct string matching with parish names for records that only include no county information (only match to parishes with unique names), c) hand matching of a fraction of remaining records carried out by identifying locations via GoogleMaps. We are able to geolocate just over 2.6 million of these records.

When using this data together with the CWGC information on deaths to construct our instrument, we further exclude 2.14 million records for which the battalion is missing. Finally, we drop 40,386 entries that are duplicates in terms of all variables, 68,487 records dated before 1905 or after 1920, as well as 27,335 from regiments with zero or negligible mortality, such as the Hussars. Finally, to ensure we have enough observations to construct the shares serving in each battalion, we drop 23,916 soldiers from battalions with less than 100 servicemen in the data.

Because of the measurement error deriving from the geolocation and the incompleteness of the FamilySearch records, some parishes exhibit values of mobilisation or WWI deaths that are unusually large relative to their population. To ensure that these possible outliers are not driving the results, we identify all parishes in which the number of mobilised and deaths have per-capita values above the 99^{th} value of the respective distribution. We then replace those figures with the imputed number of dead and of mobilised obtained by

[44] This number is in line with the 702,410 born in the British Isles and killed in the war, as reported by the British government (BWO, 1922) because the CWGC data also includes men from British dominions and Commonwealth countries.

[45] For example, the batch geocoding procedure that we used and validated when using FWR data on locations for killed soldiers yields very poor results when used with the FamilySearch strings.

multiplying the 1911 parish population by the district-level death or mobilisation rates, as appropriate.[46] Results are robust to not applying this correction.

In a previous paper, we report various validation checks that support our geolocation approaches. We refer interested readers to Carozzi, Pinchbeck and Repetto (2023).

B.5. Patents data

We draw on two principal sources of patent micro-data: PATSTAT and PatentCity.

PATSTAT is the European Patent Offices global patent database, providing standardized bibliographic data on over 100 million patents collected from more than 100 authorities worldwide. For our purposes, we focus on the subset of patent applications filed at the UK Intellectual Property Office between 1852 and 1979, encompassing information on inventors, applicants, technical classifications, titles and – where available – textual abstracts European Patent Office (2023).

PatentCity is a historical database of British patent applications covering the years 1894-1979 and specifically designed for spatial analysis of inventors' locations (Bergeaud and Verluise, 2024).[47] It comprises more than 3.8 million British patent records, and for about 1.2 million of these, at least one inventor can be matched to a location within the United Kingdom at the parish, district, or city level. Of these, we are able to assign a parish in Britain or Wales to 1.13 million of them. The combination of PATSTAT and PatentCity provides a comprehensive panel of British patents, allowing us to observe inventor identities and their location, as well as information on patent title, abstract, and sector.

B.5.1. Assignment of inventors to parishes

We assign each inventor-patent tuple to a parish using the latitude and longitude coordinates provided directly by PatentCity, which are the result of PatentCitys own internal geocoding of inventor addresses. We retain only observations with sufficiently precise geolocation quality, defined as address matches at the city, district, street, house-number, or postal-code level, and to patentees located in Great Britain. These coordinates are spatially matched to historical 1911 parish boundaries based on polygon intersection, discarding non-matching points.

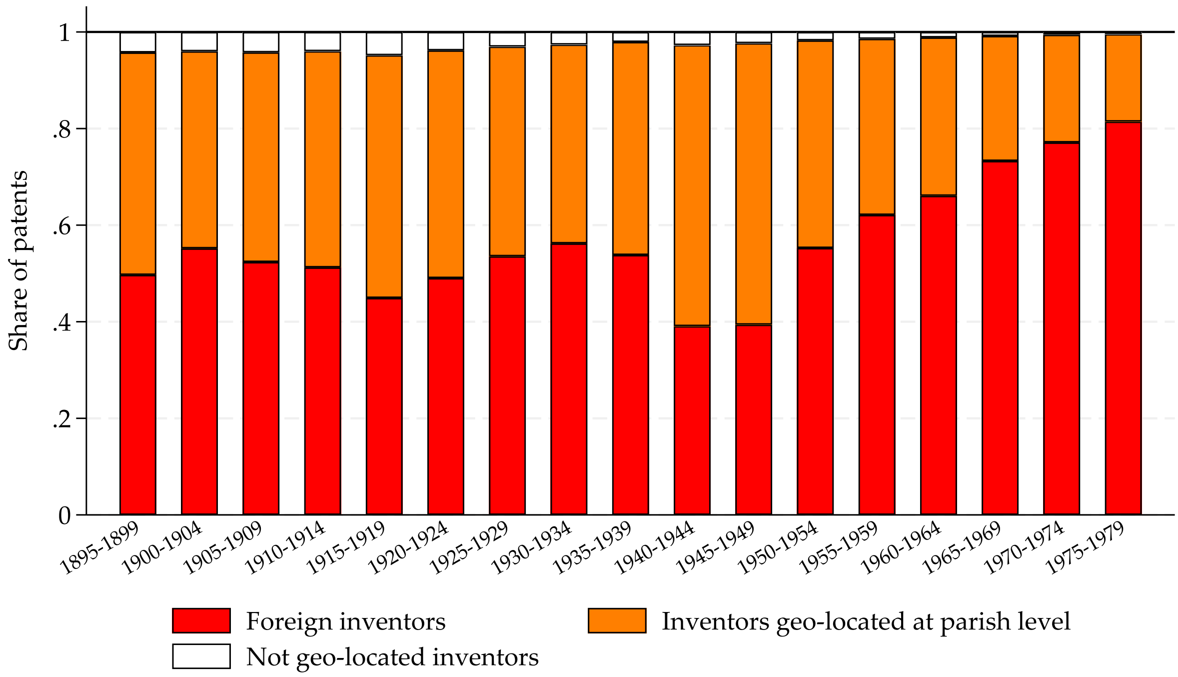
More specifically, we discard 56 percent of patents because the patentee is located outside the United Kingdom and 2 percent because the georeferenced information provided in PatentCity is not sufficiently detailed or the coordinates thereby reported do not match with the map of 1911 parishes of Great Britain. Figure B.1 reports the share of patents registered by foreign inventors, the share of patents for which we are able to assign a parish to at least one of the inventors, and the share of patents that we are not able to geolocate, for each five-year window in our sample.

[46]We use the historical county when district and grouped parish boundary coincide.

[47]In the original dataset there are patents dating as early as 1784 and as late as 2021. However, coverage before 1894 is incomplete, and the source used after 1979 is different. For these reasons, we restrict to the period 1894-1979.

Figure B.1

Inventors geo-located at parish level



Notes: This figure reports, for each five-year period in our sample, the share of patents available in PatentCity that are granted to foreign inventors (in red), the share of patents produced by inventors residing in Great Britain that we are able to geo-locate at the parish level (in orange), and the share of inventors residing in Great Britain for which the precision of the information available in PatentCity does not allow a parish-level geo-location. Our estimation samples are based on inventor-patents pairs the fall in the orange category.

B.5.2. Inventor-level link between PatentCity and PATSTAT

Because PATSTAT and PatentCity do not share inventor identifiers, we link the two sources at the inventor level in three stages. First, within each patent we generate the full cross-product of PatentCity and PATSTAT inventor entries, producing a long-format list of all admissible inventorinventor candidate pairs (the two sources share a unique patent identifier, so patents can be merged exactly). We then clean name strings and use Statas *matchit* command to compute fuzzy similarity scores between each PatentCity name and each candidate PATSTAT name within the same patent. For each PatentCity inventor, we select the PATSTAT inventor with the highest similarity score, and we require a minimum score of 0.25; lower-scoring pairs are left unmatched. For matched records, we import the cleaned PATSTAT inventor name and assign a unique inventor ID based on PATSTATs numeric identifier. Finally, when the same PATSTAT inventor is matched multiple times within a patent, we keep only the highest-scoring match to enforce a one-to-one mapping within each patent. The resulting dataset provides a harmonized inventor-level linkage that combines PatentCitys geolocation and contextual information with PATSTATs standardized names and numeric inventor identifiers. For single-authored patents, we merge inventor-level records by the exact patent identifier across the two sources.

Based on these criteria, we match PatentCity and PATSTAT inventors for approximately

700,000 inventor-patent tuples in our PatentCity main data (i.e., more than 92 percent of all inventor-patent tuples). Overall, we identify all inventors for 551,063 patents, while for the remaining 49,153 patents we miss at least one inventor.

This match is instrumental for constructing the individual-level panel of inventors active in 1895-1899, which we then follow through 1979. Based on the criteria outlined above, we identify 32,619 inventor-patent pairs (out of 34,508 available in PatentCity). We exclude 1,808 pairs identified by PATSTAT as registered by companies (i.e., non-individual inventors), leaving 30,957 inventor-patent pairs. Using the PATSTAT inventor identifier to track repeated inventors, our final sample contains 18,095 individuals.

C. Construction of the patent importance measure (Kelly et al., 2021)

C.1. Construction of the Textual Representation

The construction of the patent importance measure begins by assembling the universe of patents registered at the Great Britain patent office contained in PATSTAT (with no further restrictions in order to maximize the algorithm's predictive power) and retaining only those observations for which an abstract is available. This yields approximately 1.6 million patents. The text of each abstract is then transformed into a TF–IDF representation using the `scikit-learn` Python package.[48] Each vector is subsequently normalized to have unit L^2 norm so that cosine similarity between any two patents corresponds to the Euclidean dot product of their normalized TF–IDF vectors.

C.2. Similarity Windows and Mean Cosine Similarity

For each patent j filed in year t, we compute its similarity to two sets of comparator patents. The backward set consists of all patents filed during the five years preceding t, that is $(t-5,\ldots,t-1)$, while the forward set consists of all patents filed during the ten years after t, that is $(t+1,\ldots,t+10)$. For each of these windows, we compute the mean cosine similarity between patent j and the patents in the corresponding set. Because the TF–IDF vectors are normalized, this step reduces to computing dot products between the vector for patent k and the vectors of all patents in the relevant window.

To make this computation feasible at scale, we follow the centroid-based approach introduced by Kelly et al. (2021). For each year, we construct the sum of the normalized vectors of all patents falling in the backward window and, separately, the sum of the normalized vectors of all patents falling in the forward window. The mean similarity of patent k to a window is then obtained by taking the dot product of the vector for k with the corresponding sum vector and dividing by the number of patents in the window.

C.3. Importance Measure

Following the definition in Kelly et al. (2021), we define the technological importance of patent k as the ratio between its similarity to future patents and its similarity to prior

[48] The vocabulary is restricted to the 1,000 most frequent terms in the corpus and English stopwords are removed.

patents,

$$\text{Importance}_j = \frac{\text{ForwardSimilarity}_j}{\text{BackwardSimilarity}_j}$$

This measure is high when the textual content of patent k is more similar to language that subsequently appears in later innovations than to the terminology used in earlier ones, indicating that the ideas embodied in k anticipate future developments in the technological frontier.

D. CPC prediction algorithms

This section provides details on the Natural Language Processing (NLP) models used to predict the one-digit and two-digit Cooperative Patent Classification (CPC) code from patent titles.

We use the patent's title as the input for our NLP algorithms. For each exercise, we restrict attention to patents for which a unique one-digit (resp., two-digit) CPC code is observed in PATSTAT. We divide these patents randomly into a training sample (80%) and a held-out testing sample (20%), using a fixed random seed to guarantee replicability. The training sample is used to estimate the relationship between textual features extracted from the patent title and the CPC field, while the testing sample provides an unbiased evaluation of out-of-sample predictive accuracy.[49] As the name suggests, the training sample is used to "teach" the algorithm about the relationship between the text of each patent's title and the one-digit CPC code.

We estimate four supervised classifiers commonly used in text classification: *Naive Bayes*, *Decision Tree*, *Random Forest*, and *Support Vector Machine* (SVM). All models are trained on the same TF–IDF representation of the cleaned patent titles, where text is lowercased, punctuation is removed, English stopwords are dropped, and terms are weighted using term-frequency inverse-document-frequency with a maximum vocabulary size of 5,000 features.[50] Each model is then used to predict the CPC field for the patents in the testing sample, and its performance is evaluated by comparing predicted and observed CPC codes.

Lastly, we utilize the same algorithm to predict the one-digit CPC code also for patents published before 1910 (for which a CPC code is not available in our sources) and for patents that used to report multiple CPC codes (summing up to ≈ 1.3 million patents). Among the algorithms, the SVM achieves the highest predictive accuracy ($\approx$ 85%), and therefore we use the SVM predictions as our preferred measure of the most-likely field. We utilize the following approach to precict also the subfield of each patent (i.e., the two-digit CPC code) for a total of 29 subfields.

[49]Notice that this exercise is performed an all patents registered at the Great Britain patent office and available on PATSTAT, with no restriction about publication year and/or geo-reference to parishes in PatentCity. We follow this route in order to maximize the accuracy of the prediction algorithm. The total number of patents used is 3,420,207.

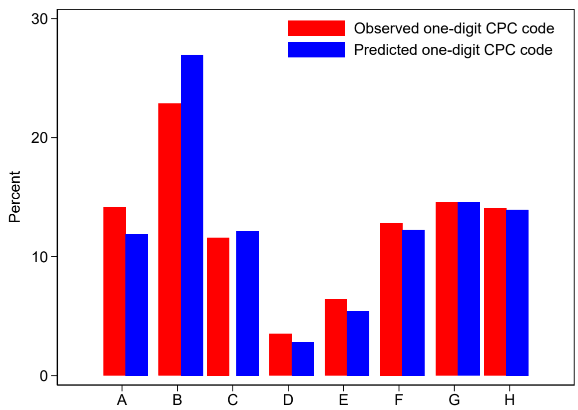
[50]These transformations are implemented in Python using the `nltk` and `scikit-learn` libraries.

For one-digit CPC codes, we present and discuss three sets of diagnostics based on the testing sample for each of the four classifiers. First, in Figure D.1, we compare the marginal distribution of predicted CPC fields with the true empirical distribution of fields in the test sample. For each classifier, Figure D.1 plots overlaid histograms showing that the predicted and observed distributions are broadly similar, although the degree of alignment varies across algorithms. In all panels, red bars reflect the distribution of one-digit CPC codes as they are observed in the training sample while blue bars report the distribution of predicted one-digit CPC codes in the same sample of patents. All classifiers perform very precisely according to this metric, perhaps with the exception of the Naive Bayes, which tends to overestimate the share of patents in sector B (Performing operations and transports) and to underestimate the majority of other fields.

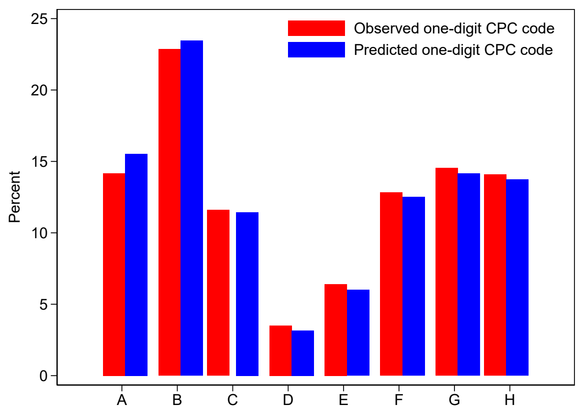
FIGURE D.1

COMPARISON OF OBSERVED AND PREDICTED CPC DISTRIBUTIONS ACROSS CLASSIFIERS

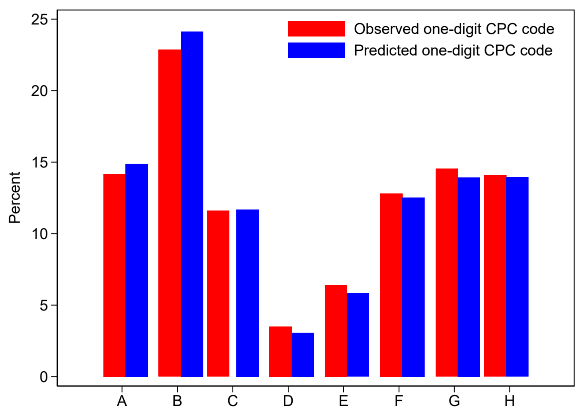
(A) NAIVE BAYES



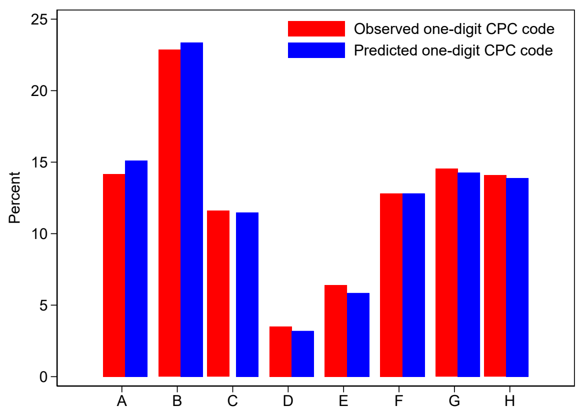
(B) DECISION TREE



(C) RANDOM FOREST



(D) SUPPORT VECTOR MACHINE



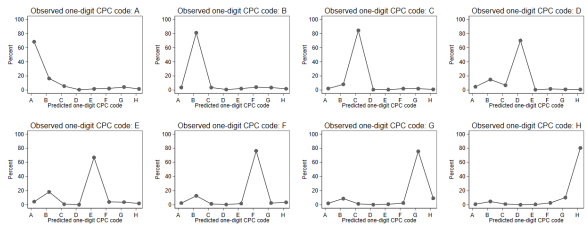
Notes: Each panel compares the empirical distribution of observed one-digit CPC codes (red) with the distribution predicted by the corresponding machine-learning classifier (blue), computed on the held-out test sample. A closer alignment indicates better predictive performance.

Second, we examine the full pattern of misclassifications by computing, for each observed field, the share of patents that are predicted into each possible field. For every algorithm, Figure D.2 reports a panel of eight plots (one per CPC field), which can be interpreted as a visual confusion matrix. In a well-performing classifier, each plot displays a dominant peak at the correct CPC category. The evidence presented in Figure D.2 is very reassuring about the validity of the prediction exercise, as for all CPC codes and classifiers we constantly see a spike exactly in concurrence with the observed field. Similarly to the results presented in Figure D.1, the Naive Bayes classifier visibly underperforms compared to the other alternatives.

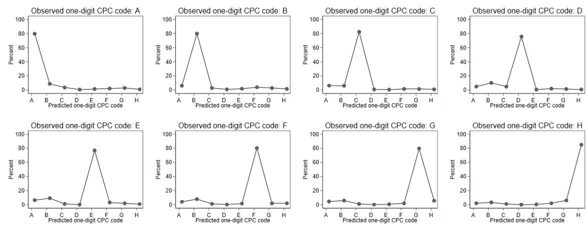
FIGURE D.2

CONDITIONAL DISTRIBUTION OF PREDICTED CPC CODES BY CLASSIFIER

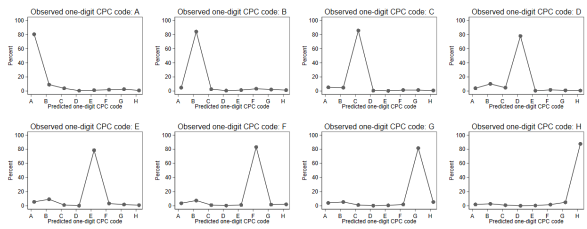
(A) NAIVE BAYES



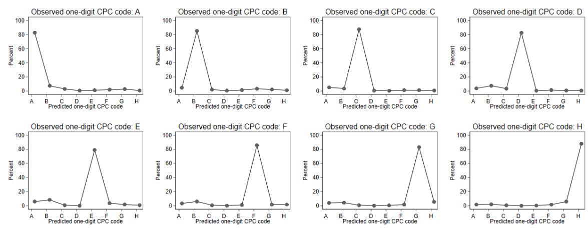
(B) DECISION TREE



(C) RANDOM FOREST



(D) SUPPORT VECTOR MACHINE



Notes: Each panel summarizes, for a given classifier, the conditional distribution of predicted one-digit CPC codes given the observed CPC field. Within each subfigure, the eight small plots correspond to patents whose true CPC is $A, \dots, H$; on the horizontal axis we report the predicted CPC code, and on the vertical axis the share (in percent) of patents in that true field that are assigned to each predicted field. Thus, for a patent with true field F, the plotted values trace out the distribution of predicted fields $hatF$. For each classifier, these shares are computed as the number of patents with true field F and predicted field $\hat{F}$, divided by the total number of patents with true field F, multiplied by 100. A well-performing classifier displays a dominant peak and relatively low mass on other entries.

Third, Tables D.1–D.4 report the detailed classification performance for each of the four supervised models considered in the validation exercise: Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine (SVM). For every one-digit CPC field (AH), we compute two standard metrics used in multi-class text classification: *recall* and *precision*. Recall is defined as $\Pr(\hat{k} \in F \mid k \in F)$ and measures the share of patents k truly belonging to field F that the algorithm correctly assigns to that field. Precision is defined as $\Pr(k \in F \mid \hat{k} \in F)$ and measures the share of patents predicted to be in field F that are in fact observed in that field. These metrics summarize, respectively, how often each model misses patents belonging to a field (low recall) and how often it over-assigns patents to a field (low precision). Each table also reports a final row showing the simple average of recall and precision across fields, which describes the models balanced performance across the CPC spectrum. Lastly, the final row reports the overall *multi-class accuracy*, computed as the share of patents in the held-out test sample for which the predicted one-digit CPC code exactly matches the observed CPC code. This metric serves as our primary criterion for algorithm selection. The Support Vector Machine achieves the highest multi-class accuracy among all classifiers.

TABLE D.1

FIELD-LEVEL RECALL AND PRECISION: NAIVE BAYES

CPC field	Recall	Precision
A	0.681	0.812
B	0.809	0.686
C	0.842	0.805
D	0.700	0.872
E	0.668	0.791
F	0.762	0.797
G	0.756	0.753
H	0.805	0.813
Average	0.753	0.791
Accuracy (multi-class)	0.767	

Notes: Rows A–H report recall and precision for each one-digit CPC field. Recall is defined as $\Pr(\hat{k} \in i \mid k \in F)$ and precision as $\Pr(k \in F \mid \hat{k} \in F)$. The Average row reports averages of recall and precision across fields. The final row reports the overall *multi-class accuracy*, defined as the share of patents in the test sample for which the predicted CPC code equals the observed one.

TABLE D.2

FIELD-LEVEL RECALL AND PRECISION: DECISION TREE

CPC field	Recall	Precision
A	0.796	0.727
B	0.799	0.779
C	0.822	0.834
D	0.756	0.839
E	0.768	0.819
F	0.803	0.822
G	0.796	0.818
H	0.847	0.869
Average	0.798	0.813
Accuracy (multi-class)	0.805	

Notes: Rows A–H report recall and precision for each one-digit CPC field. Recall is defined as $\Pr(\hat{k} \in i \mid k \in F)$ and precision as $\Pr(k \in F \mid \hat{k} \in F)$. The Average row reports averages of recall and precision across fields. The final row reports the overall *multi-class accuracy*, defined as the share of patents in the test sample for which the predicted CPC code equals the observed one.

Showing the full set of diagnostics for the prediction of two-digit CPC codes would be cumbersome due to the higher number of subfields in our data. For simplicity, Table D.5 reports the average recall, precision, and multi-class accuracy, for each classifier. Support Vector Machine results being the most accurate classifier, with a multi-class accuracy that

TABLE D.3

FIELD-LEVEL RECALL AND PRECISION: RANDOM FOREST

CPC field	Recall	Precision
A	0.802	0.764
B	0.840	0.796
C	0.855	0.849
D	0.778	0.892
E	0.785	0.862
F	0.831	0.850
G	0.815	0.851
H	0.876	0.885
Average	0.823	0.844
Accuracy (multi-class)	0.831	

Notes: Rows A–H report recall and precision for each one-digit CPC field. Recall is defined as $\Pr(\hat{k} \in i \mid k \in F)$ and precision as $\Pr(k \in F \mid \hat{k} \in F)$. The Average row reports averages of recall and precision across fields. The final row reports the overall *multi-class accuracy*, defined as the share of patents in the test sample for which the predicted CPC code equals the observed one.

TABLE D.4

FIELD-LEVEL RECALL AND PRECISION: SUPPORT VECTOR MACHINE

CPC field	Recall	Precision
A	0.826	0.774
B	0.850	0.831
C	0.873	0.882
D	0.823	0.902
E	0.789	0.863
F	0.854	0.854
G	0.830	0.846
H	0.878	0.891
Average	0.840	0.856
Accuracy (multi-class)	0.846	

Notes: Rows A–H report recall and precision for each one-digit CPC field. Recall is defined as $\Pr(\hat{k} \in i \mid k \in F)$ and precision as $\Pr(k \in F \mid \hat{k} \in F)$. The Average row reports averages of recall and precision across fields. The final row reports the overall *multi-class accuracy*, defined as the share of patents in the test sample for which the predicted CPC code equals the observed one.

again exceeds 0.8.

E. Inventor characteristics

There is no single dataset that provides comprehensive information about the characteristics - for example age, occupation, education and income - of historical inventors in England and Wales. The Census of population provides a comprehensive record of individuals' names, residential location, ages, and household characteristics but has never asked about

TABLE D.5

SUMMARY PERFORMANCE METRICS FOR TWO-DIGIT CPC CLASSIFICATION

Classifier	Average recall	Average precision	Accuracy (multi-class)
Naive Bayes	0.560	0.792	0.736
Random Forest	0.686	0.798	0.801
Decision Tree	0.660	0.731	0.768
Support Vector Machine	0.697	0.819	0.818

Notes: Each row reports summary performance statistics for one of the four classifiers used in the two-digit CPC prediction exercise. Average recall and average precision are averages across all two-digit CPC fields observed in the test sample. Recall is defined as $\Pr(\hat{k} \in F \mid k \in F)$ and precision as $\Pr(k \in F \mid \hat{k} \in F)$, where F indexes two-digit CPC codes. Multi-class Accuracy is the share of patents in the held-out test sample for which the predicted two-digit CPC code exactly matches the observed code.

income or wealth, and only began to collect educational information in 1951. However, very few individuals described themselves as inventors in historical Censuses: for example, in 1911 only 274 individuals recorded their occupation as "Inventor". Our combined PATSTAT-PatentCity database does record inventor names, location, and self-reported occupation, but does not include age or gender.

To get a sense of inventor characteristics in the period we study we therefore conduct a record linkage exercise. It is infeasible to match inventors to Census records across the full period on analysis because the Census micro-data with individual names is not yet available for research. Hence, we restrict ourselves to attempting to match inventors in the period 1906 to 1916 to the 1911 Census. Our approach to record linkage is strongly influenced by the few matching variables that are available in the two datasets. We have the first name, surname, and location for all observation in the two datasets. We also have middle names for a subset of individuals in each dataset. Although the PATSTAT-PatentCity database does not include age, we can infer a plausible lower bound on an inventor's age by assuming they were at least 12 years old at the time of their first patent publication in the database.

Given these characteristics we proceed as follows. We first clean the first name of individuals in the 1911 Census using the STATA command *abeclean* from Abramitzky, Boustan and Eriksson (2012) to strip the names of special characters and occupational titles, and construct a variable with the initial of the middle name where one is reported. For the PATSTAT-PatentCity matching sample we restrict attention to the 21,489 inventors who produced a publication between 1 January 1901 and 31 December 1916. Of these 21,392 have a unique combination of first name, surname and middle initial.

We then perform a sequential merging strategy in which we perform several rounds of merges that are usually progressively less strict in terms of matched characteristics.[51] After each round we remove any inventors and Census records that have already been matched. In

[51]We also tested using the ABE algorithm for matching the datasets using only names and parish. This produced a vast number of false matches.

all rounds of matching, we first merge inventors to multiple Census records that are exactly matched on a combination of name and location characteristics. We then eliminate matches which cannot be true, usually because either middle name initials do not match or because the age of the Census record is below the minimum age we infer the inventor to be. Finally, we keep only those records where there is a single matched record in the Census. We perform five rounds of matching in total. In the first, for example we require inventors to be exactly and uniquely matched to a Census record on first name, last name, and parish. In the final round of matching, we replace the requirement for an exact match on parish with a requirement for an exact match on county, but we also require that the occupation string exactly coincides in both datasets.

This procedure successfully matches 12,966 out of the 21,489 inventors we attempt to match, at a rate of 60%. As our matching procedure is severely limited by the paucity of common characteristics across the matching datasets, we conduct a basic test for the quality of linkages by examining the consistency of the middle initial in the two datasets. In 7,706 matched cases the middle initial is available in both datasets. In 7,242 cases (94%) the initial is the same and in only 464 cases (6%) it is not.

Table E.1 reports the mean values of Census characteristics for four groups. The first column refers to the inventors we matched to the Census. In column 2 we provide statistics for the small number of individuals who call themselves inventors in the Census. In columns 3 and 4 we respectively report characteristics of the full and adult population of England and Wales. The Table suggests that inventors in the first two decades of the 20th Century were overwhelming male and on average older than the general adult population. The data also suggests that they were richer as they were more likely to have servants and to live in larger houses than the average adult. As Table E.2 shows, these differences are statistically significant. Returning to Table E.1, we note that column 1 is broadly similar to column 2 which reassures us that our matching procedure is not introducing major biases.

Table E.1
1911 Census: Mean comparisons

	(1) Inventors		(2) Cen. Inv'rs		(3) Full pop'n		(4) Adult pop'n	
	mean	sd	mean	sd	mean	sd	mean	sd
Male	0.96	0.20	0.95	0.22	0.48	0.50	0.47	0.50
Age	39.65	12.83	47.83	13.94	27.88	19.57	35.46	16.75
No. Servants	0.65	2.05	0.50	2.45	0.28	2.32	0.32	2.51
No. Rooms	7.70	25.28	10.07	59.99	6.14	30.52	6.36	31.19
No. People	6.65	68.86	4.79	6.09	8.69	100.17	8.84	106.97
Observations	12571		274		35053905		26230697	

Notes: Inventors are defined as individuals that publish an patent between 1906 and 1916 and are matched from the PATSTAT database. Census inventors are those that record their occupation as Inventor in the 1911 Census. Full and Adult (12 yo+) population exclude those in institutions e.g. hospitals and prisons.

Finally, we can also use the same inventor dataset (pre-matching) to examine the occupations of inventors using the occupational string text in the PATSTAT-PatentCity matching

TABLE E.2

1911 CENSUS: MEAN DIFFERENCE TESTS INVENTORS AND OTHER ADULTS

	(1)					
	Inv. vs Non					
	Count	Mean	Count	Mean	Diff	T-Stat
Male	12,571	0.958	26,218,126	0.474	-0.483	-108.54
Age	12,559	39.645	26,137,768	35.456	-4.189	-28.02
No. Servants	12,571	0.645	26,218,126	0.324	-0.321	-14.34
No. Rooms	12,533	7.705	26,078,120	6.361	-1.344	-4.82
No. People	12,571	6.654	26,218,126	8.844	2.190	2.29

Notes: Inventors are defined as individuals that publish an patent between 1906 and 1916 and are matched from the PATSTAT database. Census inventors are those that record their occupation as Inventor in the 1911 Census. Full and Adult (12 yo+) population exclude those in institutions e.g. hospitals and prisons.

sample. There are 8072 individuals with non-missing data. Of these 4588 (57%) have a string that contains text suggesting that they identify as an engineer or mechanic (string contains "ENGIN" or "MECHANI"), 1362 (17%) have a string that suggests a commercial occupation ("MANAGER" "MERCHANT" "DIRECT" "CLERK"), 1347 (17%) identify as a manufacturer ("MANUF" of "FACTURE") and 459 (6%) have text which suggests gentleman or woman ("GENTLE" "LEISURE").