THE UNIVERSITY OF CHICAGO

ESSAYS ON THE ECONOMICS OF INNOVATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE IRVING B. HARRIS
GRADUATE SCHOOL OF PUBLIC POLICY STUDIES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
TERENCE CHAU

CHICAGO, ILLINOIS
AUGUST 2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

For setting me on the path to this degree, I'd like to thank my parents, Lam, Kevin, Thomas, and Gilma, my grandparents Gloria and Orlando, and my siblings Orlando and Anthony.

For helping me walk this path, I extend my gratitude to all fellow Ph.D. students at the Harris School past and present—you have contributed through countless discussions to this dissertation and are coauthors of this work. I'd specially like to thank my cohort-mates: Mythili Vinnakota, Devika Lakhote, Emileigh Harrison, Goya Razavi, Jenna Allard, Afia Khan, Angela Wyse, Lucas Mation, Scott Loring, Chinmaya Kumar, and Kailash Rajah. I have thoroughly enjoyed our time together, relied on your support beyond the academic, and am forever indebted to you.

For the academic, I would like to thank my committee, Jeff Grogger, Rick Hornbeck, Dan Black, and Anders Humlum.

ABSTRACT

This dissertation examines several topics in the economics of innovation in historical settings. The first chapter examines how government investment in technology creates spillover effects in the production of knowledge, using the 1960s Space Race as an empirical setting. Combining the known universe of patent records with information on their federal reliance, along with a difference in differences design, I estimate that NASA-exposed fields increased their patenting relative to non-exposed fields, and that these patents were more impactful by citation metrics. To study the degree to which these results are driven by the reallocation of scientists and engineers, I use the inventor information in the patent documents to show that NASA-affiliated inventors obtained their first ever patents after joining NASA or obtaining a NASA contract, and not before.

The second chapter contributes to data methods in the economics of innovation by comparing traditional rules-based data linkage to machine learning methods. I apply a supervised learning strategy to link patents issued between 1840 and 1900 to individual establishment microdata in the 1870 Census of Manufactures, and conclude that a simple rules-based approach combined with manual verification plausibly yields higher confidence links. I contribute a novel dataset for future researchers by performing this higher confidence linkage.

The final chapter provides an overview of the historical patent datasets used throughout the dissertation, with a focus on their accuracy, coverage, and overlap.

CHAPTER 1

SPILLOVERS OF PUBLIC R&D: EVIDENCE FROM THE SPACE RACE

Geopolitical rivalry during the 1960s Space Race drove the world's two superpowers to make massive investments in spaceflight related technologies. This paper examines the impact of NASA's research and development efforts on the quantity and quality of space-related innovation. Combining the known universe of patent records with information on their federal reliance, along with a difference in differences design, I estimate that NASA-exposed fields increased their patenting relative to non-exposed fields, and that these patents were more impactful by citation metrics. These results are robust to removing NASA-related fields strongly related to concurrent defense spending. To explicitly account for the fact that technologies might have been selected by reasons other than achieving a Moon landing, I also produce estimates only using spaceflight-essential classes. To study the degree to which these results are driven by the reallocation of scientists and engineers, I use the inventor information in the patent documents to show that NASA-affiliated inventors obtained their first ever patents after joining NASA or obtaining a NASA contract, and not before.

1.1 INTRODUCTION

Technological innovation is one of the main drivers of increasing living standards and productivity growth. However, economic theory predicts that innovation will be undersupplied in a free market. A common policy lever to ensure the optimal supply

of research and development (R&D) is government funding (Arrow, 1962). Despite this, there are few empirical studies of large public investments in R&D, especially in applied research (Gross and Sampat, 2022; Kantor and Whalley, 2022). A particularly understudied area about the impacts of these investments is their spillover potential in technology space. Despite not having ex-ante broad applicability, applied R&D investments in risky or untested fields can lead to novel uses of incipient technologies or create new influential fields altogether.

In this paper, I study one of the largest examples of government R&D funding and measure the extent to which it also generates externalities in technology space. In particular, I answer the following question: Did the creation of the National Aeronautics and Space Administration (NASA) increase the quantity, and more importantly, the quality of inventive output of the fields it was involved with?

I leverage the creation of NASA and its sizeable R&D funding during the Space Race—0.7% of GDP in 1966 and 35.9% of all federal R&D outlays (Office of Management and Budget, 2021)—as a source of variation. Using a combination of administrative patent records and a novel historical patent dataset (Berkes, 2018), which includes the full text content of each patent, I compare changes in the quantity and quality of patenting in NASA-involved technology classes with changes in non NASA-involved technology classes, before and after NASA's creation. Relative to other government-funded fields, I find that NASA led to a 59.90% increase in spaceflight-related patenting, citations to these fields experienced a relative increase in citations of 72.27%, and that this impact extended beyond to non-spaceflight related fields.

Estimating the spillover effects of applied public R&D presents a causal identification challenge. A social planner will optimally invest in technologies that will yield the largest welfare returns, and can typically provide much larger funding than a specific private entity can.[1] The former will lead to selection bias, because treated fields are higher quality on average than control technologies. The latter will also bias estimates, because even within selected fields, higher R&D funding will typically yield a higher quantity and quality of patents.[2]

To circumvent these identification issues, I use a difference in differences design at the technology subclass level, where I compare NASA funded technologies to other government funded technologies. This setting and design have several desirable properties for identification. First, the timing of this funding was driven by the USSR's successful launch of Sputnik 1 in 1957, and the timing and decision for funding the Moonshot was largely a response to Yuri Gagarin's successful Earth orbit flight in 1961. Second, the selection of technologies invested in by NASA as the Mercury, Gemini, and Apollo programs developed was mission-driven, not driven by economic or spillover concerns. In essence, I argue that all comparison classes have ex-ante government interest, which is correlated with many sources of selection bias,

1. Despite the fact that corporate patents, foreign and domestic, have represented the largest share of US patents since the mid 1930s until today (Nicholas, 2010), at least 25% of US patents granted each year since 2005 have used some form of federal funding, with a peak of 30% in 2011. Corporate entities represent the majority of assignees in these federally funded patents (Fleming et al., 2019).

2. This discussion omits a third source of bias: governments differentially over invest in basic science, which by definition has larger spillovers (Williams & Bryan, 2021). Therefore any estimate of public versus non-public innovation will overstate the difference. My main specifications only use publicly funded innovation, and given NASA's strong emphasis on applied science during Apollo, my estimates will, if anything, be attenuated by this bias.

except that treated classes experience a large excess funding that is driven by external geopolitical factors. To the degree that NASA invested in non-spaceflight related technologies, I also produce estimates using only fields that were ex-ante known to be directly related to achieving the Moon landing and find similar event study results. Due to the overlap between spaceflight technology and contemporaneous Department of Defense investments in rocketry, I show that my results are qualitatively similar when omitting ordnance and rocket related classes.

To understand the degree to which these effects are driven by the reallocation of inventive human capital, I use the inventor information in the patents to create an inventor-level panel dataset to observe NASA-affiliated scientists and engineers. Using this data, I observe whether they had generated patents before joining NASA or not, along with any changes in their pre and post-NASA fields of work. I find that the majority of NASA-affiliated engineers and scientists obtained their first patent after joining NASA, and not before, which is congruent with the idea that the growth in spaceflight innovation during Apollo was not driven by the reallocation of existing inventors from other fields.

This article contributes to several strands of the innovation economics literature. First, it contributes to the broad literature on the drivers of technological innovation, with an emphasis on government investment as a policy lever (Arrow, 1962; Williams and Bryan, 2021). I contribute to a growing body of empirical case studies on government R&D programs using causal inference methods (Jacob and Lefgren, 2011; Howell, 2017; Azoulay et al., 2018; Gross and Sampat, 2022; Kantor and Whalley, 2022; Moretti et al., forthcoming) by studying one of the largest and most sudden

government drives in technological spending in American history. While previous work has mostly focused on the effects of investing in basic science, this article adds to our understanding of how applied innovations can also create spillovers across technological space, even when their application is narrowly targeted.

By looking at non-NASA increases in patenting and subsequent citation patterns from non-government patents, this paper adds to the literature on private-sector responses to the government funding of innovation. Whether public R&D is a complement or a substitute to private R&D has been a longstanding empirical question in the economics of innovation literature (David et al., 2000). Theoretically, firm responses to public innovation efforts are ambiguous. On the one hand, firms can contract their supply of innovation, i.e., a crowding out effect. This could be the case if public demand exerts upwards pressure on the prices of R&D inputs (e.g., scientists and engineers, specialized facilities). In this instance, I find the opposite: government interest in spaceflight technologies resulted in further private sector innovation, a crowding in effect.

This result is congruent with recent work—Slavtchev and Wiederhold (2016) show that theoretically, a public demand shift towards technologically advanced goods increases private R&D by increasing the returns to innovation for the whole economy. Empirically, recent studies find that different forms of government investment in R&D spur further innovation from the private sector. Azoulay et al. (2018) link privately generated patents to scientific publications funded by the National Institutes of Health, and find that basic science funding leads to increased private patenting. Howell (2017) and Myers and Lanahan (2022) study Department of Energy research

grants to small firms and conclude that the funding leads to increased patenting by grantees, and that these patents induce further downstream patenting by other firms. I complement this body of knowledge by showing that these crowding in effects also occur when the government invests in applied science, that they occur when the government specifies and directs the research agenda entirely, and that noticeably larger public R&D initiatives still exhibit these positive spillovers.

In related work, Gross and Sampat 2022 show that World War II government R&D investments in military technology through the Office of Scientific Research and Development helped shape the post-war direction and geographic distribution of innovation. By showing Space Race innovation effects persist even when omitting explicitly military classes, this article expands our understanding of the effects of these large innovation pushes beyond investments in military purposes (Moretti et al., forthcoming). Another advantage of the setting I study is that military inventions typically face high disclosure restrictions, which limit their spillover potential. As the civilian branch of the space effort, NASA's technology fields, while correlated with those of the Department of Defense, would have faced less restrictions and provide a setting where spillovers are allowed to develop naturally.

This paper complements previous and contemporaneous work on the effects of NASA on innovation and the economy more broadly. First, this paper is most similar, and can be seen as a direct successor to Jaffe et al. (1998), who look at federal agency patenting patterns through the years with an emphasis on NASA. They describe the spillovers of public R&D by looking at NASA's patenting behavior and the citations NASA patents have received over the decades. This paper builds on

this conceptual question by exploiting the comprehensiveness, scope, and quality of patent data achieved in recent years due to computational advances, along with the development of causal inference methods in econometrics. In particular, their sample only allows for the study of patents assigned directly to NASA, and not those developed under contracts from NASA. Second, they can only observe these assignments starting in 1969, the year the Moon landing happened. The available data at the time only allowed researchers to observe citations made after 1977, the starting year for computerized records at the National Bureau of Economic Research (Hall et al., 2001), and well after the decline of NASA's budget post-Apollo. By combining datasets that leverage modern optical character recognition to identify citations and federal reliance for all years, along with a difference in differences design, this article can provide causal answers to the question Jaffe et al. posed two decades ago.

My work also complements Kantor and Whalley (2022), who study the geographic manufacturing growth effects of the Moonshot. Using Census of Manufactures data, they create a continuous measure of county-level spaceflight specialization and produce difference in differences estimates of manufacturing value added, employment, and other outcomes on local specialization. Using a market access approach derived from Donaldson and Hornbeck's (2016) county to county trade model, they argue that NASA also produced market-wide effects on manufacturing outcomes during the Space Race. They posit that local productivity spillovers are a possible driver of positive market-wide effects. By looking directly at innovation outcomes and looking at comparisons in technological as opposed to geographic space, the present article

allows us to dissect one channel through which spillovers of the Space Race happened in ways beyond localized effects—through knowledge production directly.

Section 1.2 discusses the timeline of the Space Race, NASA's creation and funding, technology selection, and its contracting and patenting policies. Section 1.3 details data sources and key variables, Section 1.4 discusses the empirical framework and identification. Section 1.5 discusses the results and Section 1.6 concludes.

1.2 CONTEXT

1.2.1 Towards a Moon Landing

On October 4th 1957, the Soviet Union successfully launched the first artificial satellite, Sputnik I. A month later on November 3rd, the USSR launched Sputnik II, delivering a satellite weighing a hundred times more[3] than Sputnik I and containing the first living being in orbit, the dog Laika. In an attempt to respond, the United States launched the Vanguard TV-3 satellite, which immediately failed and exploded. These events over perceived American technological inferiority triggered the Sputnik Crisis, and led to the Eisenhower administration to create the National Aeronautics and Space Administration in 1958 (U.S. House of Representatives, 1958).

As early as 1959, discussions around achieving a lunar landing were taking place at various NASA facilities, particularly within the Research Steering Committee on Manned Space Flight, also known as the Goett Committee. Tasked with conceptualizing NASA's long term mission plans, members Maxime Faget from the Langley

3. Sputnik I weighed 184 pounds, while Sputnik II weighed around 17,200 pounds. Vanguard TV-3 weighed 3.3 pounds (Murray & Cox, 2004).

Research Center's Space Task Group[4] and George M. Low from the Lewis Research Center[5] urged the committee to conclude that a Moon landing should be the agency's post-Mercury goal. However, predicting that the political support for such vast spending wouldn't be there,[6] Abe Silverstein, Chief of Space Flight Programs, and T. Keith Glennan, NASA Administrator, concluded that NASA could not commit to any long term plans beyond Mercury (Brooks et al., 1979; Murray and Cox, 2004). Eventually, the Apollo program was announced in July 1960, with a reduced goal of a manned flight around the moon.

By late 1960, Low recommended the creation of a committee at the Lewis Research Center that would carry out preliminary feasibility studies for Apollo, with a particular goal of devising the requirements and options for achieving a lunar landing.[7] The resulting Manned Lunar Landing Task Group (or Low Committee)

4. NASA Langley, located in Hampton, Virginia, was the home of NASA's precursor agency, the National Advisory Committee for Aeronautics (N.A.C.A.) and the headquarters of the manned space program until November of 1961, when the Manned Spacecraft Center (MSC, now the Johnson Space Center) was established in Houston, Texas (Uri, 2021). The Space Task Group at Langley was in charge of managing the manned spaceflight program, starting with Project Mercury. Faget is credited with designing the Mercury capsule (US Patents 3,001,739; 3,093,346; and 3,270,908). He served as Director of Engineering and Development at the Manned Spacecraft Center from 1962 to 1981 (Allen, 2017).

5. Lewis, in Brook Park, Ohio, now known was the John H. Glenn Research Center at Lewis Field, was another inherited facility from the N.A.C.A. (Keeter, 2017). Low was an aeronautical engineer at Lewis, Deputy Center Director at the MSC, and then NASA Deputy and Acting Administrator (Arrighi, 2019).

6. "At 10:00 o'clock, I talked with Dr. Kistiakowsky about our budget. I found him resigned to the inevitable - that President Eisenhower is going to balance the budget, come hell or high water." (Glennan, 1993, Chapter 12).

7. "This group will endeavor to establish ground rules for manned lunar landing missions, to determine reasonable spacecraft weights, to specify launch vehicle requirements, and to prepare an integrated development plan including the spacecraft, lunar landing and take-off systems, and launch vehicles. This plan should include a time phasing and funding picture and should identify areas requiring early studies by field organizations." - Memorandum for Director of Space Flight

concluded in a February 7th, 1961 report that, given the required financial support, the agency had the technical capacity to achieve an Earth orbital flight by 1965, lunar orbit by 1967, and a Moon landing between 1968 and 1971 (Low, 1961; Low, 1999; Arrighi, 2019). These technical assessments aligned with opinions surveyed by the House Select Committee on Astronautics and Space Exploration in 1959, whose interviewed experts, including Wernher von Braun, predicted that circumlunar flights would be technically possible by the end of the next decade, with manned landings "a few years thereafter" (U.S. House of Representatives, 1959, p.4).

President John F. Kennedy, who had been inaugurated a week before the Low Committee's report was presented, commissioned an advisory committee led by Jerome B. Wiesner[8] to assess the space program. The resulting report suggested a potential cancellation of Project Mercury, or at a minimum, to cease advertising Mercury as the United States' major space objective, as any failures would be blamed on the incoming administration, and to focus on unmanned space activities (Wiesner Committee, 1961). The resulting appointment of Wiesner as President Kennedy's Special Assistant for Science and Technology cast doubt on the administration's overall stance regarding the manned space program.

On Wednesday, April 12, 1961, the Soviet Union achieved its second major milestone over the American space program by placing cosmonaut Yuri Alekseyevich Gagarin in orbit. Gagarin's approximately two hour flight launched from the

---

Programs Subject: Manned Lunar Landing Program, George M. Low (Low, 1999, Chapter 13).

8. Wiesner was the head of the Department of Electrical Engineering at the Massachusetts Institute of Technology (M.I.T.) and a member of the President's Science Advisory Committee during the Eisenhower administration (M.I.T. Libraries, 2005).

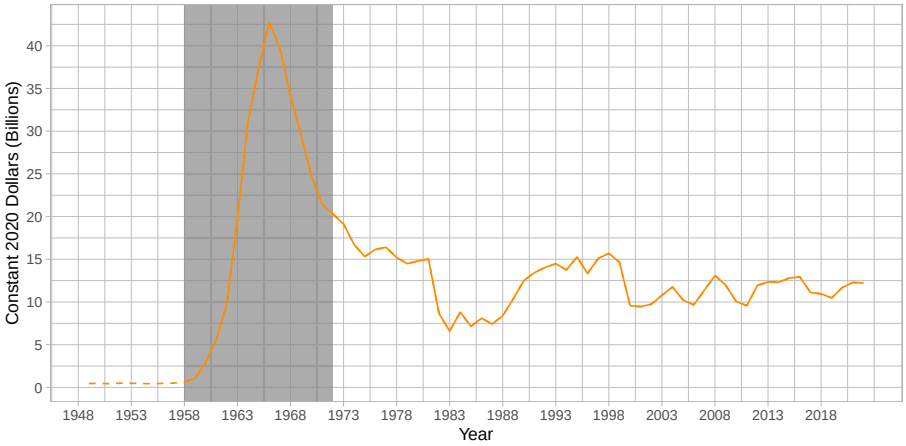
Baikonur Cosmodrome in Kazakhstan, passed over Kamchatka, Russia, the Pacific Ocean north of Hawaii, southeast over the tip of South America, the Atlantic, then entered Africa over Angola and left above Egypt, finally landing near the city of Engels, Saratov Oblast, Russia (European Space Agency, 2011).[9] Two days later, Kennedy, along with speechwriter and adviser Theodore Sorensen, summoned Jerome Wiesner, NASA Administrator James Webb, Deputy Administrator Hugh Dryden, and Director of the Office of Management and Budget David E. Bell, to discuss where could the United States achieve a "first" in space. By NASA's projections, the USSR would be first in crewed orbital flights, orbital space stations, and circumlunar flights. Given Low's previous feasibility studies, Dryden remarked that manned lunar landings would require further technology developments that could potentially be achieved first, but it would take a scientific and funding effort akin to the Manhattan Project. After the meeting, Sorensen confided off the record to Life Magazine correspondent Hugh Sidey, who was present, that they were going to the Moon (Sidey, 1994, Murray and Cox, 2004).

Soon after, President Kennedy addressed a Joint Session of Congress, publicly stating that "this nation should commit itself to achieving the goal, before this decade is out, of landing a man on the Moon and returning him safely to the Earth" (Kennedy, 1961). What followed was a sharp increase in NASA's total budget and R&D outlays (Figure 1.1), and a rapid expansion in planning, contracting, and fa-

9. The United States achieved its first orbital flight on the Mercury-Atlas 6 mission nearly a year later on February 20th, 1962, where astronaut John H. Glenn completed three orbits in under five hours. During his three passes on Friendship 7, Glenn flew over most of Africa, Australia, Hawaii, northern Mexico, the southern United States, and landed near the Bahamas (Uri, 2022; NASA, 1962).

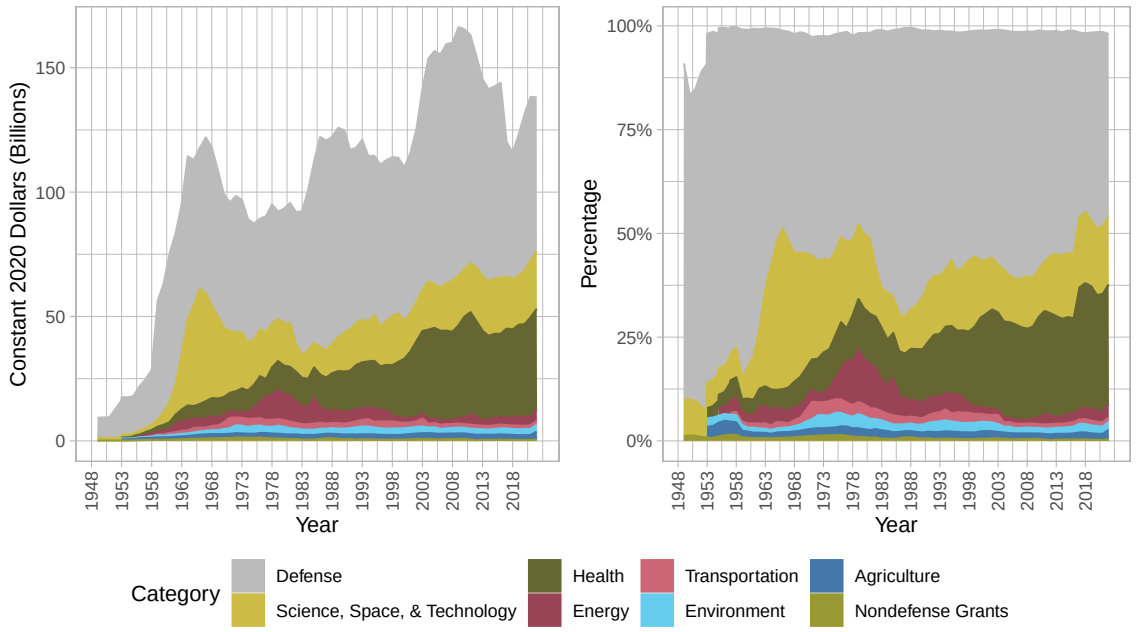
cility construction. Detailed hardware specifications and requests for contractor proposals had been completed by July (Murray & Cox, 2004). The first contract, for the design and prototyping of the Apollo Guidance Computer, was issued to the Instrumentation Laboratory at M.I.T. on August 9th, 1961 (Brooks et al., 1979). By September, land acquisition and planning for the Kennedy Space Center at Merritt Island, Florida, the Michoud Assembly Facility in Michoud, Louisiana, and the Manned Spacecraft Center in Houston (Dunbar, 2017; Mohon, 2008; Uri, 2021) were underway. By the mid-1960s, NASA's R&D outlays represented 35.9% of all federal R&D expenditure (Figure 1.2).

Figure 1.1: NASA R&D Outlays, 1948-2022



Source: Calculated from the Office of Management and Budget's (2021) Historical Tables and deflated using the CPI-U (Bureau of Labor Statistics, 2023). Dashed pre-1958 values represent the N.A.C.A. budget.

Figure 1.2: Federal R&D Outlays, 1948-2022



Source: Calculated from the Office of Management and Budget's (2021) Historical Tables. Outlays under "Other" category not plotted. The following lists the agencies included in each category by the Office of Management and Budget. Defense: Department of Defense & Other; Health: Science, Space and Technology: NASA, NSF, AEC General Science; Health: NIH & Other; Energy: DOE; Transportation: NASA Transportation, DOT: Environment, Agriculture and Nondefense Grants not specified.

1.2.2 NASA's Contracting and Patenting Policy

In order to obtain the necessary ingredients for a Moon landing within the decade, a key aspect of NASA's approach was to leverage private research and development and manufacturing capacity where possible. For example, the Saturn V rocket involved contracts with North American Aviation and its Rocketdyne division, Douglas Aviation, Boeing, and International Business Machines (IBM).[10]

10. For a listing of major system contractors for Project Mercury, see Grimwood (1963), Appendix 9. For Project Gemini, see Grimwood et al. (1968), Appendix 7. For the Apollo program, see Ertel and Morse (1969), Appendix 3 for principal contractors from August 1961 to November 1962, Morse

This public-private relationship in procurement and contracting for R&D has roots in the Department of Defense's (DOD) source evaluation procedure, which NASA adopted and modified, initially working with cost-plus-fixed-fee contracts and fixed-cost contracts post-1963 (Rosholt, 1966). Unlike the DOD, however, Section 305[11] of the Space Act allowed NASA to take title to all inventions produced while performing an R&D contract.[12] This right could be waived by the NASA Administrator and the Inventions and Contributions Board, leaving the contractor as the assignee of the resulting patents and giving the U.S. Government a royalty-free license to use the invention.[13] Watson and Holman (1966) look at all NASA contractor waiver applications up to December of 1965 and estimate that 88% of waiver applications were granted. Even when not waived, industrial contractors typically had royalty-free, non-exclusive licenses upon coming up with an invention, conditional on bringing the invention to market (Kraemer, 1999).

These title (NASA owned) and waiver (contractor owned) patent policies did

---

and Bays (1973), Appendix 3 for contractors from November 1962 to September 1964, Brooks and Ertel (1973), Appendix 3 from October 1964 to January 1966, and Ertel et al. (1978), Appendix 2 for those from January 1966 to July 1974. These listings do not account for subcontracts or smaller direct contracts.

11. For a detailed legislative history of NASA patent policy and debates over Space Act sections relating to patenting (i.e., Section 203, Section 305, and Section 306), see Watson and Holman (1966) with an emphasis on Appendix A (written by Aaronson, 1966), Rosholt (1966), and Kraemer (1999).

12. Department of Defense contracts followed this second policy, while Section 305 was modeled after Atomic Energy Commission policy since NASA was deemed to be more similar in nature to the AEC than the DOD. However, in practice the NACA's patent policies were more similar to the DOD's and NASA had a larger contractor and technological overlap with the DOD (Rosholt, 1966).

13. Inventions from NASA employees follow a similar policy, where the U.S. Government by default claims title to all of their inventions, but employees can request title to their inventions (Watson & Holman, 1966).

not preclude third parties from leveraging these technologies commercially. NASA owned patents can be licensed from its technology transfer programs for private sector use, and if waiver patent owners have not used their invention commercially, NASA can use its march-in rights[14] to force patent owners to license an agency-funded technology (Watson & Holman, 1966).

1.2.3 NASA Innovation & Spinoff

Given the cutting edge nature of NASA's mission—Apollo represented some of the earliest applications of integrated circuits and electronic and digital fly-by-wire systems—, and the sizable funding it received during the Apollo era, quantifying the spillovers of this technology has been an open question since the agency's inception (Watson and Holman, 1966; Ginzberg et al., 1976; Jaffe et al., 1998). Since 1976, the Technology Utilization Office has published a yearly report detailing commercially licensed products that have spunoff from NASA-funded technologies. In earlier reports, NASA claimed successful spillovers in diverse fields including satellite weather forecasting and communications, integrated circuits and computing, water purification, medical imaging, food processing, and materials such as memory foam and blow rubber molding[15] (Ruzic, 1976).

14. March-in rights allow the government to require the patent owning contractors to grant licenses to third parties, even when the current patent assignee refuses. NASA already had march-in rights to its funded inventions held by contractors, but the Bayh-Dole Act of 1980 extended this power to all federal agencies (Thomas, 2016).

15. Marion Franklin Rudy, who worked on NASA projects for Lockheed and Rockwell International in the 1960s, adapted this technology to create small air membranes in shoe soles (US Patent Nos. 4,183,156 and 4,219,945). He partnered with Nike to commercialize it, creating the Nike Air shoe.

Of note is NASA's co-occurrence with the birth of the integrated circuit. The Apollo Guidance Computer (AGC), developed at M.I.T.'s Instrumentation Lab, was one of the first real life, high stakes applications of integrated circuits at a time transistors were the main computational technology.[16] In the early 1960s, aerospace hardware ranging from navigation electronics to military radar was mainly analog. While the AGC is archaic by modern standards, "the exclusive use of integrated circuits in the processor ushered in a new era of computing, the novel memory design stored large amounts of data in a small space, and the human interface allowed real-time interaction with software" (O'Brien, 2010, p.xiii)—all features familiar to the modern computer user.

Ginzberg et al. (1976) argue that one of the drivers of the growth in integrated circuits and semiconductors was assured demand from the government, as space and defense accounted for between 25% to 48% of their production between 1955 and 1968. Between 1954 and 1963, this sector also consumed between 32% to 100% of US computer and computer service production. Given the emphasis on performance over cost in NASA and Department of Defense requests for integrated circuits and their derivatives, this assured demand is argued to have not only decreased production costs but also increased the reliability of these unproven components over time. Table 1.1 shows the total production and average prices of integrated circuits throughout the 1960s.

The 1960s growth in computing is also reflected in educational supply and human capital decisions. The first computer science department in the United States was

16. For a painstaking walkthrough of the inner workings of the AGC and its development history from a computer science perspective, see O'Brien (2010).

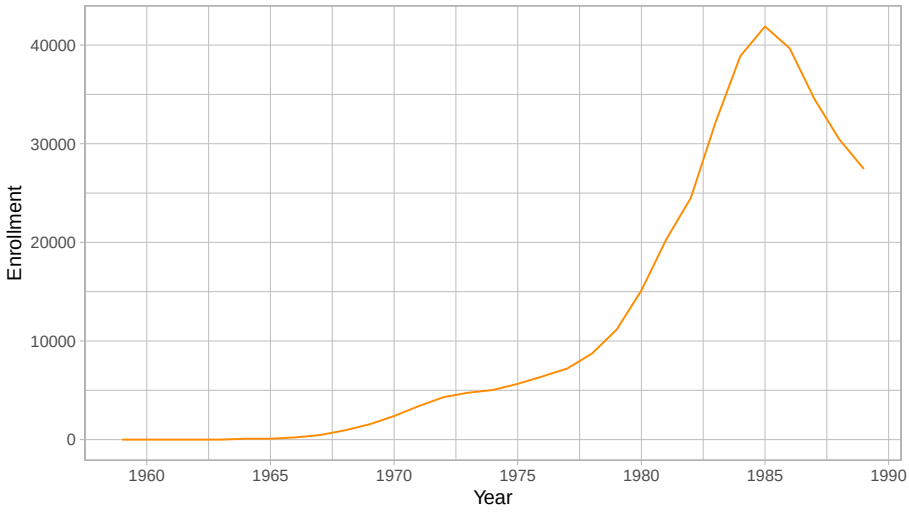
Table 1.1: US Integrated Circuit Production and Prices

Year	Production (Millions of Dollars)	Avg. Price per IC (Dollars)
1962	4	50.00
1963	16	31.00
1964	41	18.50
1965	79	8.33
1966	148	5.05
1967	228	3.32
1968	312	2.33

Note: Reproduced from Ginzberg et al. (1976), Table 2-3, p.59.

created at Purdue University in 1962, and enrollments in computer science departments soon grew thereafter as shown in Figure 1.3.

Figure 1.3: US Computer Science Bachelor's Enrollments, 1959-1990



Note: Data from Snyder (1993).

1.3 DATA

This study draws data from several sources. First, I use the United States Patent and Trademark Office's (USPTO) administrative file on the universe of recorded patents (Marco et al., 2015). This dataset contains identifiers for all registered patents from 1836 onwards, along with technology classes and associated dates. Because this dataset does not contain many other variables one would require to study patenting behavior, I also use the Comprehensive Universe of U.S. Patents dataset (CUSP), a novel and private dataset constructed by Berkes (2018) which is currently considered to be the gold standard historical patent dataset in terms of scope of variables included and completeness (Andrews, 2021). These records are constructed from high quality, optical character recognized scans of the patents themselves, from which key variables such as inventor, assignee, location, previous patents cited, and more are drawn from. The dataset also contains the full text of each patent's claims. Figure 1.4 shows an example patent, along with the information that can be gleaned from this file.

To identify NASA-involved inventions, I combine this information with federal agency reliance data from Fleming et al. (2019), who also use the patent full text to infer the degree of federal agency involvement in the production of all patents from 1926 to 2017. Specifically, they define two sources of direct federal reliance. If the patent was produced by a NASA employee, the patent would be assigned to NASA[17] (see Figure 1.4 for an example). Further, if work from a NASA contract resulted in a

17. Patents always have inventors, who are individuals (35 U.S.C. 100(f), 2021), but they can also have assignees, which can be companies or other organizations, to whom ownership of the patent is transferred.

patent, the patent's claims would start with a reference to said contract work, even if assigned to a private firm (Figure 1.5). The data further identifies patents that indirectly rely on NASA funding by determining whether the patent cites non-patent literature (e.g., scientific journal articles) that was funded by NASA. For purposes of this paper, a patent's NASA reliance is defined by being produced directly by NASA or originating from a NASA contract.[18]

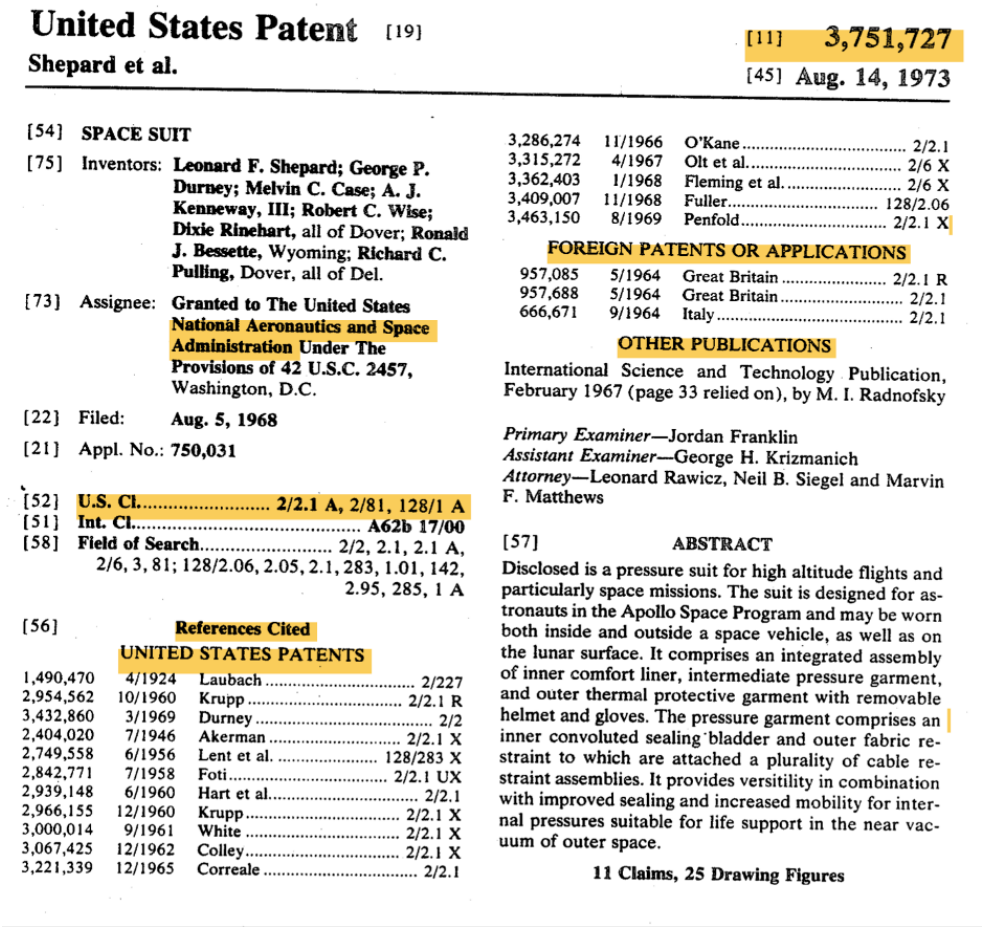
The main unit of most analyses in this paper is the USPC technology subclass. The USPC classification scheme[19] is one of many systems used for patent classification. It has a hierarchical construction, with around 500 classes at the broadest level and 150,000 narrow classes or subclasses. For the purposes of this paper, "broad class" will refer to the former, while "narrow class" or "subclass" will refer to the latter[20]. A patent can have several classes, however, the first code is chosen to capture the main contribution of the invention. Figure 1.4, for example, shows Patent 3,751,727 as being in class 2, subclass 2.1 A, class 2 subclass 81, and class 128 subclass 1 A, with 2/2.1 representing "Apparel; astronaut's body cover", while 2/81 represents "Apparel; heat resistant". Classes are assigned by the patent examiner, who is a domain expert at the USPTO, and patents are reassigned as new classes are created. This implies that class and subclass assignments are internally consistent for all patents for data queried from internal USPTO records at a given point in time.

18. To verify the extent of mistaken NASA attribution in the Fleming et al. (2019) data, I hand inspect the original patent scans for 200 randomly drawn patents and find they are all funded by NASA.

19. USPC codes are constructed similarly to Journal of Economic Literature codes. There are other patent classification schemes, such as the Cooperative Patent Classification (CPC).

20. "Class" will be used more generally, or where the distinction between broad and narrow class does not affect interpretation.

Figure 1.4: Example U.S. Patent



United States Patent [19]
Shepard et al.

[11] **3,751,727**
[45] **Aug. 14, 1973**

[54] **SPACE SUIT**

[75] Inventors: **Leonard F. Shepard; George P. Durney; Melvin C. Case; A. J. Kenneway, III; Robert C. Wise; Dixie Rinehart,** all of Dover; **Ronald J. Bessette,** Wyoming; **Richard C. Pulling,** Dover, all of Del.

[73] Assignee: **Granted to The United States National Aeronautics and Space Administration Under The Provisions of 42 U.S.C. 2457,** Washington, D.C.

[22] Filed: **Aug. 5, 1968**

[21] Appl. No.: **750,031**

[52] **U.S. Cl.** ..... **2/2.1 A, 2/81, 128/1 A**
[51] **Int. Cl.** **A62b 17/00**
[58] **Field of Search** ..... 2/2, 2.1, 2.1 A, 2/6, 3, 81; 128/2.06, 2.05, 2.1, 283, 1.01, 142, 2.95, 285, 1 A

[56] **References Cited**

UNITED STATES PATENTS

1,490,470	4/1924	Laubach	2/227
2,954,562	10/1960	Krupp	2/2.1 R
3,432,860	3/1969	Durney	2/2
2,404,020	7/1946	Akerman	2/2.1 X
2,749,558	6/1956	Lent et al.	128/283 X
2,842,771	7/1958	Foti	2/2.1 UX
2,939,148	6/1960	Hart et al.	2/2.1
2,966,155	12/1960	Krupp	2/2.1 X
3,000,014	9/1961	White	2/2.1 X
3,067,425	12/1962	Colley	2/2.1 X
3,221,339	12/1965	Correale	2/2.1
3,286,274	11/1966	O'Kane	2/2.1
3,315,272	4/1967	Olt et al.	2/6 X
3,362,403	1/1968	Fleming et al.	2/6 X
3,409,007	11/1968	Fuller	128/2.06
3,463,150	8/1969	Penfold	2/2.1 X

FOREIGN PATENTS OR APPLICATIONS

957,085	5/1964	Great Britain	2/2.1 R
957,688	5/1964	Great Britain	2/2.1
666,671	9/1964	Italy	2/2.1

OTHER PUBLICATIONS

International Science and Technology Publication, February 1967 (page 33 relied on), by M. I. Radnofsky

Primary Examiner—Jordan Franklin
Assistant Examiner—George H. Krizmanich
Attorney—Leonard Rawicz, Neil B. Siegel and Marvin F. Matthews

[57] **ABSTRACT**

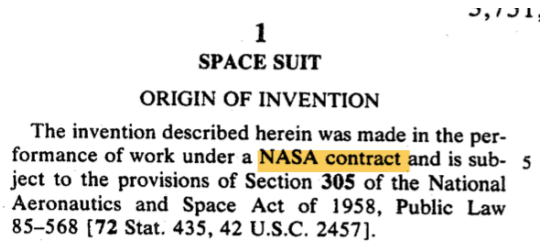
Disclosed is a pressure suit for high altitude flights and particularly space missions. The suit is designed for astronauts in the Apollo Space Program and may be worn both inside and outside a space vehicle, as well as on the lunar surface. It comprises an integrated assembly of inner comfort liner, intermediate pressure garment, and outer thermal protective garment with removable helmet and gloves. The pressure garment comprises an inner convoluted sealing bladder and outer fabric restraint to which are attached a plurality of cable restraint assemblies. It provides versitility in combination with improved sealing and increased mobility for internal pressures suitable for life support in the near vacuum of outer space.

11 Claims, 25 Drawing Figures

Figure 1.6 shows NASA and non-NASA patent counts at the individual patent level, where the non-NASA counts have been scaled. NASA patents increased sharply from its inception through the Mercury, Gemini, and Apollo eras, largely mimicking the budget shares in Figure 1.1 in that early period.

As expected, NASA's technology portfolio largely consists of patents in spaceflight-related fields. Among broad classes, measuring and testing and aeronautics and astronautics are the main fields of invention (Figure 1.7), while at the

Figure 1.5: NASA Contract Mention in Patent Text



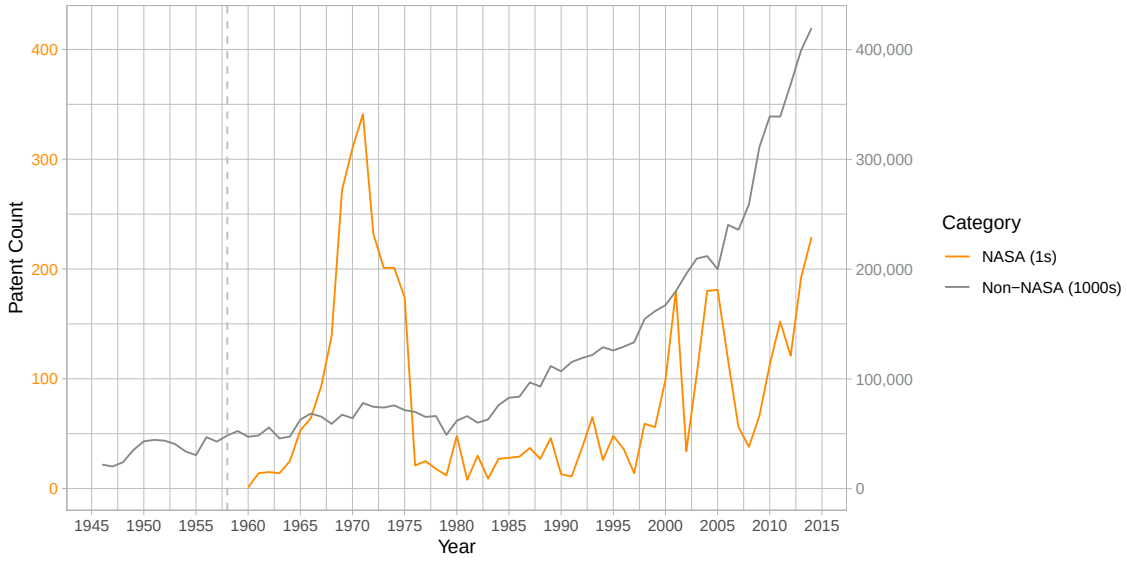
1

SPACE SUIT

ORIGIN OF INVENTION

The invention described herein was made in the performance of work under a NASA contract and is subject to the provisions of Section 305 of the National Aeronautics and Space Act of 1958, Public Law 85–568 [72 Stat. 435, 42 U.S.C. 2457].

Figure 1.6: NASA and Non-NASA Patent Counts, 1960-2019

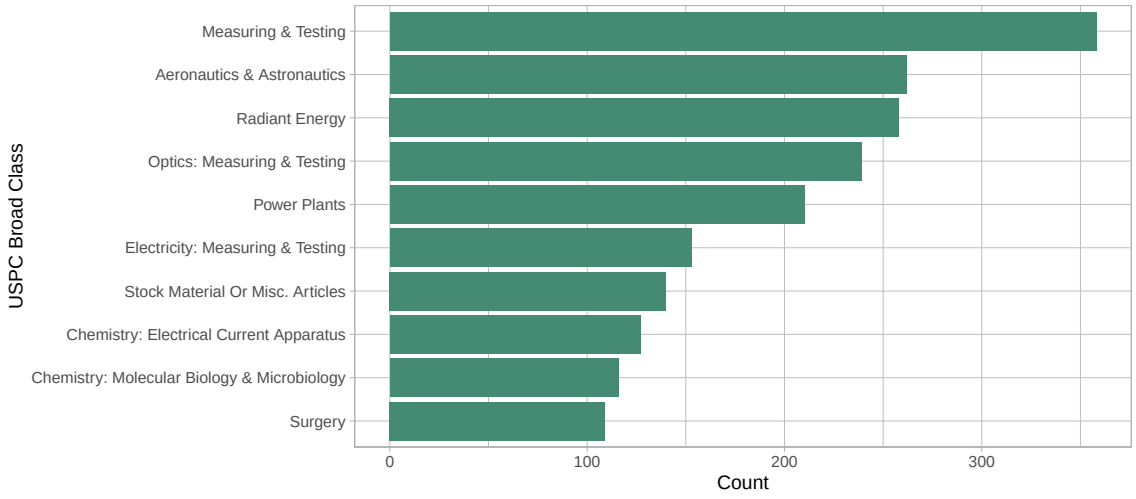


narrow class level five out of NASA's top subclasses are related to spacecraft or aerodynamics (Figure 1.8).

The main empirical analysis focuses on a subclass-by-year panel dataset from 1948, starting ten years prior to NASA's creation and after World War II,[21] and ending in 1980, around a decade after the end of the Apollo program. The main outcomes of interest are patent counts and patent citations at the subclass-year

21. The post-war period appears to be a natural point to start the analysis, given the existence of both wartime disruptions and accelerations of innovation (Gross & Sampat, 2022).

Figure 1.7: Top NASA Broad Classes

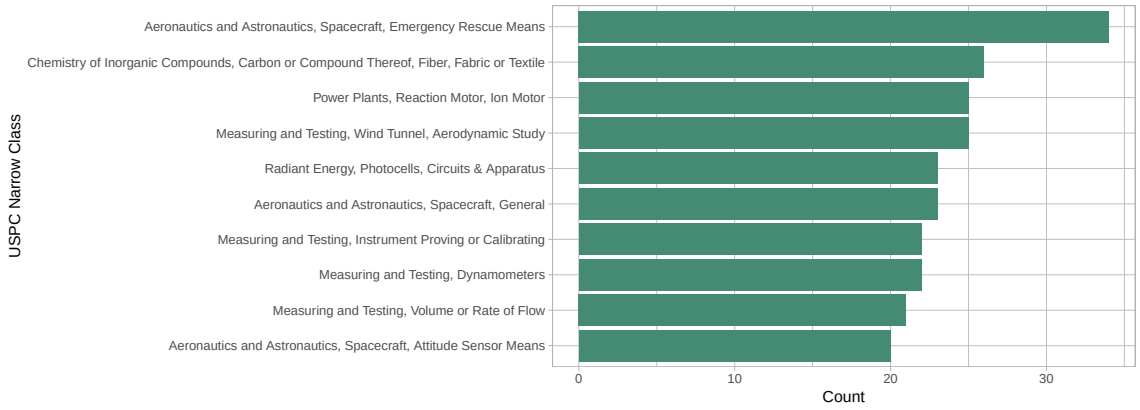


level. Citations to a class can be measured in different ways, which merits discussion. Broadly, they can be construed as a measure of impact or influence, much like academic article citations (Jaffe et al., 1993). First, prior art citations are furnished by the inventor and their attorney at the moment of application to the USPTO. Afterwards, the patent examiner will observe the list and assign further prior patents deemed necessary.

Patent forward citations[22] have been found to correlate strongly to alternative measures of novelty, both scientific and economic. For the former, patents correlate strongly to natural language processing-based metrics where the full text of a patent's claims is compared to patents past and patents future. In this framework, patents are defined to be breakthroughs when they have a large text dissimilarity with previous patents and a large text similarity with future patents (Kelly et al., 2021). For the

22. For a given pair of patents A and B where B cites A, we can define the forward citation as A having a forward citation from B. Alternatively, B backwardly cites A.

Figure 1.8: Top NASA Narrow Classes



latter, (Kogan et al., 2017) develop a novel measure of stock market responses for publicly traded firms that receive new patents and find that this measure correlates with that patent's forward citations as well.

Using the citation network embedded in the patent-level data, I construct different measures of citations that reflect different aspects of a given subclass' future impact. First, I define a measure of yearly citations received, which I refer to as "citations per year". This is the number of citations made towards a given subclass in each year. Second, I define the sum of all citations made to patents issued in a given subclass in a specific year as "lifetime citations". Broadly, the former reflects overall interest in a given subclass in a given year. If citations per year are high for a subclass-year combination, it implies that inventive activity in that year expressed reliance on that given subclass. Lifetime citations on the other hand reflect the fact that inventions generated in some years are disproportionately influential.

Additionally, I create leave-one-out measures of these metrics by omitting citations made from one subclass to the same subclass to observe broader impact of a

technology, and omit NASA patent to NASA patent citations to account for self-citing behavior.

Because the number of citations to a subclass is correlated with both the age and number of patents in it, and patenting behavior changes over the years, I also create versions of these measures where I only count citations within a fixed 20-year window of time, and I use time fixed effects throughout my estimations.

To set a baseline comparison, Table 1.2 shows a difference in means between NASA and non-NASA patents' lifetime citations for all patents issued after 1958. While the average post-1958 non-NASA patent has 10.816 citations over their lifetime, NASA patents have around 3.946 more lifetime citations. This holds when controlling for technology subclass and issue year fixed effects—the former adjusting for the fact that some technology classes have higher impact regardless of year, and the latter for the fact that citation behavior changes over time for all classes.

Table 1.2: Difference in Lifetime Citations, NASA & Non-NASA Patents

	Lifetime Citations		
	(1)	(2)	(3)
$\mathbb{I}$(NASA)	4.650***	3.891***	3.946***
	(0.359)	(0.343)	(0.329)
Constant	10.816***		
	(0.011)		
Technology FE	N	Y	Y
Issue Year FE	N	N	Y
NASA/Non-NASA Obs.	6,770 / 7,489,012	6,770 / 7,489,012	6,770 / 7,489,012

Note: Robust s.e. in parentheses. *p<0.1; **p<0.05; ***p<0.01

However, even within the same technology class, NASA funded patents plausibly receive higher funding than non-NASA funding, confounding the true estimated effect of publicly funded innovation on citations. Given this potential selection bias, we proceed to detail our empirical design.

1.4 EMPIRICAL FRAMEWORK

The main empirical analysis estimates changes in NASA-exposed technology subclasses, against changes in non-exposed technology subclasses, relative to a base year of 1957. The analysis is carried out at the subclass level instead of the patent-level for one main reason—patents are not observed until they are produced. Therefore, treated patents are always treated. On the other hand, one can observe outcomes for technology classes over time, and can estimate how these outcomes vary before

and after NASA's creation. The main outcomes of interest are patent counts and patent citations at the subclass-year level.

Formally, the outcome y_{ct} in a given subclass c in a given year t is regressed on a treatment indicator that takes a value of 1 when the year is 1958 or later and the subclass is treated, along with its leads and lags, a subclass fixed effect γ_c, a year fixed effect δ_t, and an error term ε_{ct}. This dynamic two-way fixed effects estimator allows us to obtain difference in differences estimates for each pre-treatment and post-treatment year minus the base year of 1957, and provides for a test of parallel pre-treatment trends in the same estimation.[23]

$$y_{ct} = \sum_{\tau=-10\setminus\{-1\}}^{22} \beta_\tau(\mathbb{I}\{t - 1958 = \tau\} \times \ T_c) + \gamma_c + \delta_t + \varepsilon_{ct} \tag{1.1}$$

These estimates will also be summarized with a static two-way fixed effects specification without leads and lags as follows:

$$y_{ct} = \beta(\mathbb{I}\{t \geq 1958\} \times \ T_c) + \gamma_c + \delta_t + \varepsilon_{ct} \tag{1.2}$$

In this setting, I define treated subclasses T_c as narrow classes in which NASA-

23. This design avoids some of the concerns highlighted by recent literature on two-way fixed effects estimation: treatment is defined as starting at the same time for all treated units, avoiding using earlier treatments as controls for later ones (Goodman-Bacon, 2021). However, concerns regarding treatment effect heterogeneity are still valid in this context (de Chaisemartin & D'Haultfœuille, 2020), therefore I produce estimates using Callaway and Sant'Anna's (2021) methodology in Appendix 1.7.7 and find nearly identical event study estimates.

exposed patents were produced during the Apollo era, that is:

$$T_c = \mathbb{I}\{\#\text{ 1958-1972 NASA Patents in Subclass} \geq 1\} \tag{1.3}$$

Appendix Tables 1.8 and 1.9 show the treatment and control classes with the most patents in the sample period. Because the difference in differences estimator requires pre-treatment and post-treatment observations, I do not utilize subclasses that only exist in the post-period. One concern is that if NASA is seeding many subclasses by creating entirely new technologies, this research design would not capture their influence. Using the USPTO administrative records, I find that only 34 out of all NASA patents were the first in their subclass, although 25 of these originated in the Apollo years.[24]

The identifying assumption in such a research design is that absent NASA's creation and funding, the quantity and quality of treated technologies would have evolved similarly to those NASA did not work on. Given the large possible number of control subclasses, it is hard to think ex-ante that they are as a whole a valid comparison group to NASA-exposed technologies. For my main specification, I will instead use subclasses that other federal agencies worked on before 1958 as a comparison group. In Appendix 1.7.8, I show however that results using both a) all untreated subclasses as controls, and b) only untreated subclasses within broad classes that were treated, yield qualitatively similar results to the main specification.

24. These subclass seed patents are in a number of different classes including aeronautics, nuclear measurement, fuel cells, medical diagnostics equipment, semiconductor manufacturing, optics, television signal processing, and arithmetic and calculation methods using electrical computers.

Standard errors are clustered at the subclass level to adjust for heteroskedasticity and subclass serial correlation throughout. Event study estimates also include sup-t confidence bands to account for multiple hypothesis testing (Callaway and Sant'Anna, 2021; Freyaldenhoven et al., forthcoming).

1.5 RESULTS

Table 1.3 shows summary statistics of the main outcomes in the baseline year, 1957, for both treatment and control subclasses. On average, treatment classes in the baseline year had about double the number of patents issued, those patents were cited about twice as much over their lifetime, and those classes were themselves cited twice as much in 1957. These imbalances further motivate using a differences in differences design as opposed to the naive comparison of treated and control outcomes.

Table 1.3: Summary Statistics, 1957

Covariate	Control Mean	Treat. Mean	Diff. in Means	p-value
Patents Issued	1.094	2.032	0.938	0.000
Citations	4.731	8.751	4.021	0.000
Citations (Leave-one-out)	3.777	6.196	2.419	0.000
Lifetime Citations	7.953	15.634	7.682	0.000
Lifetime Citations (LOO)	6.157	11.142	4.985	0.000
Narrow USPC Count	6,468	869		

1.5.1 Baseline Difference in Differences Results

Figure 1.9 reports the estimated coefficients from the main equation. During the Apollo years and the following decade, patenting in spaceflight related subclasses increased by a statistically significant 0.974 per year on average over other subclasses, with a peak of 1.922 in 1971. Figure 1.10 only counts non-NASA funded patents.

Table 1.4 summarizes these results using the static two-way fixed effects specification, with similar results. On average, patent counts increase by a statistically significant 1.217 within each subclass per year, and do so by 1.152 after excluding NASA-owned or contracted patents.

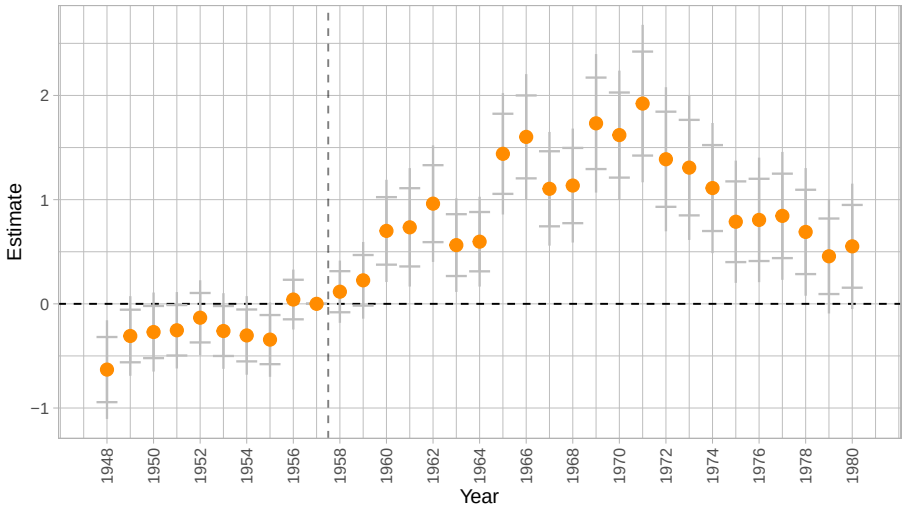
Table 1.4: Difference in Differences Estimates, Patent Counts

	Patent Issues	Patent Issues (Excl. NASA)
	(1)	(2)
$\mathbb{I}$(NASA)	1.217***	1.152***
	(0.154)	(0.153)
Subclass FE	Y	Y
Year FE	Y	Y
Observations	249,803	249,803
Note:	Subclass clustered s.e. *p<0.1; **p<0.05; ***p<0.01	

Taking the event study and static difference in differences estimates at face value, the estimates are largely unchanged when excluding NASA patents, hinting that these increases are not mostly driven by NASA's direct efforts, but rather by spillovers from other related inventions. However, another potential explanation is

that NASA funding was underreported in the original patent documents,[25] or that the text analysis carried out by Fleming et al. (2019) was not able to identify all NASA reliant patents, resulting in false negatives. To address the second point, I manually inspect the original documents for a sample of two hundred post-1958 patents in treated subclasses that appear to have no NASA funding in the data, and only find two NASA funded patents.

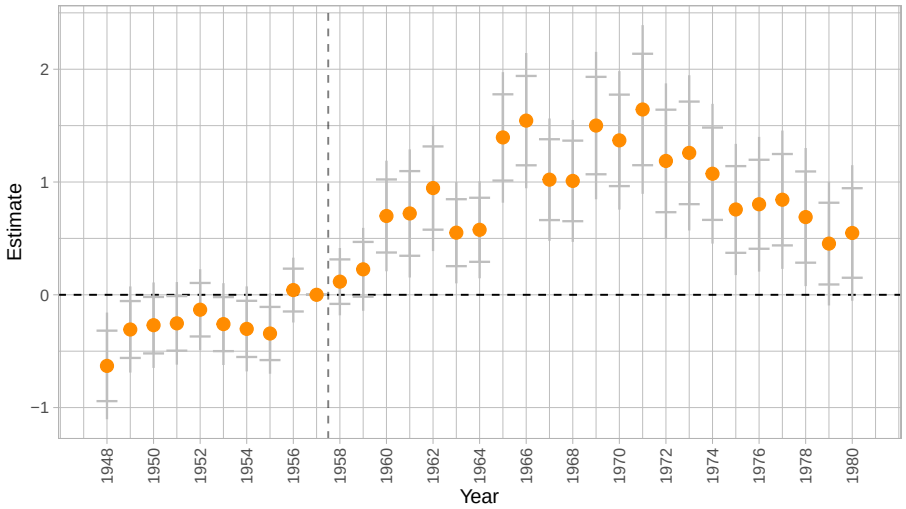
Figure 1.9: Patent Issue DID Estimates, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

25. Empirically demonstrating this would require knowledge on all 1960s NASA contract specifications, and identifying all possible patents issued to contractors originating from the contracted work. However, there is anecdotal evidence that there was an underreporting of contractor inventions to NASA (Kraemer, 1999).

Figure 1.10: Patent Issue DID Estimates, Excluding NASA Patents



Note: S.E. clustered at subclass level, ⌶: point-wise 95% CI, |: sup-t 95% confidence band.

Table 1.5: Difference in Differences Estimates, Patent Citations

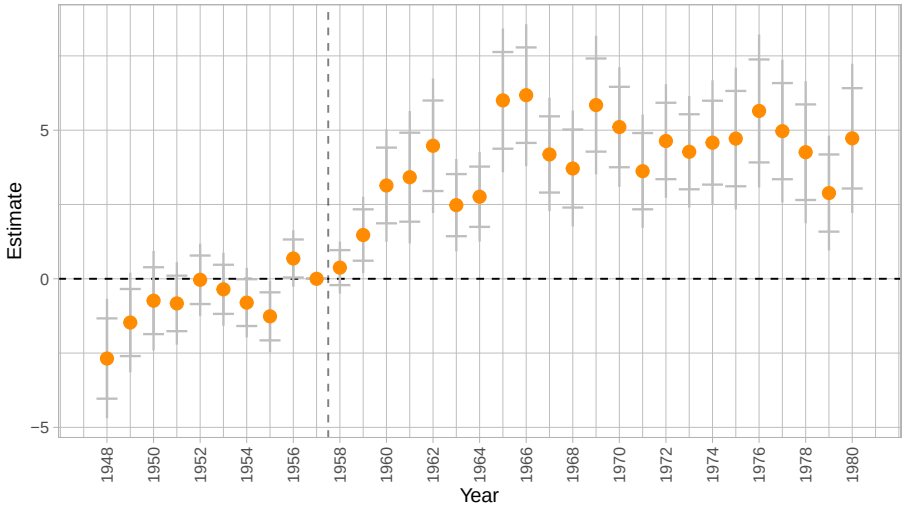
	Yearly	Yearly (Excl. NASA)	Lifetime	Lifetime (Excl. NASA)
	(1)	(2)	(3)	(4)
𝕀(NASA)	4.778***	4.771***	11.614***	11.604***
	(0.624)	(0.624)	(1.560)	(1.560)
Subclass FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Observations	249,803	249,803	249,803	249,803

Note: Subclass clustered s.e. *p<0.1; **p<0.05; ***p<0.01. Columns (1) and (3) include citations made to NASA patents, while Columns (2) and (4) exclude them.

Figures 1.11 and 1.12 show the estimates on citations by year and lifetime cita-

tions.[26] Citations per year increased by an average of 4.064 each year, and lifetime citations for patents issued in the Apollo years were 10.258 higher, similar to the static estimates in Table 1.5. These estimates imply increases in patenting and in-year citations of around 59.90% and 54.60%, and an increase in lifetime citations of around 72.27% over the 1957 treatment group mean. While patenting decreased over the years after Apollo and subsequent reductions in NASA funding, citation behavior and the lifetime citations towards treated classes remained constant in the decade after.

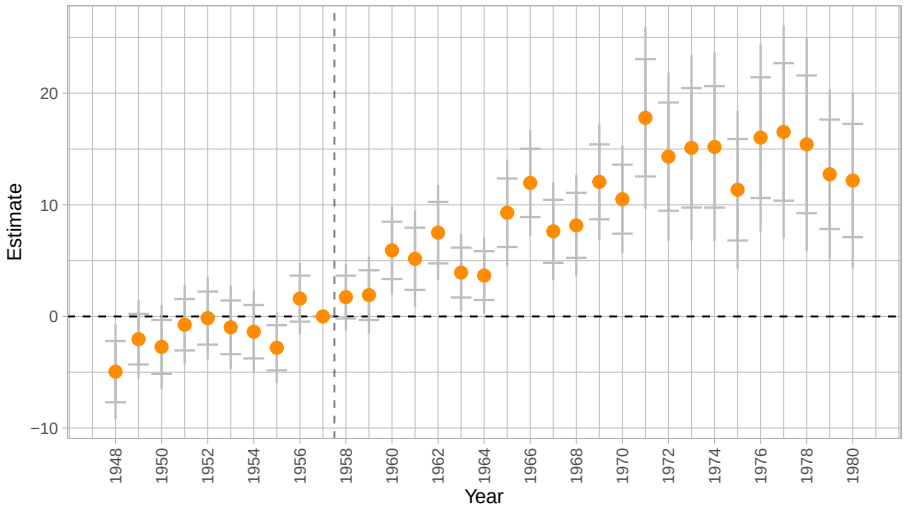
Figure 1.11: Citations by Year DID Estimates, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

26. Estimates using only a fixed window of 20 years instead of lifetime citations can be found in Appendix Table 1.10 and Figures 1.23-1.25.

Figure 1.12: Lifetime Citation DID Estimates, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

1.5.2 Citation Breadth and Length

Increased patenting activity within a set of technologies can drive increases in citations to those technologies due to self-citations—many gradual improvements over the same base technology, or high idiosyncratic citation rates can both drive citation counts among treated subclasses.

To study if the results are driven by either phenomenon as opposed to these classes having broader impact, I estimate the main regressions using leave-one-out citation outcomes as the number of citations originating from subclasses other than the subclass of interest (Table 1.6 and Figures 1.13 & 1.14) and find the results are qualitatively unchanged—most citations to these subclasses originate from other subclasses. I repeat this estimation using only citations from patents in entirely

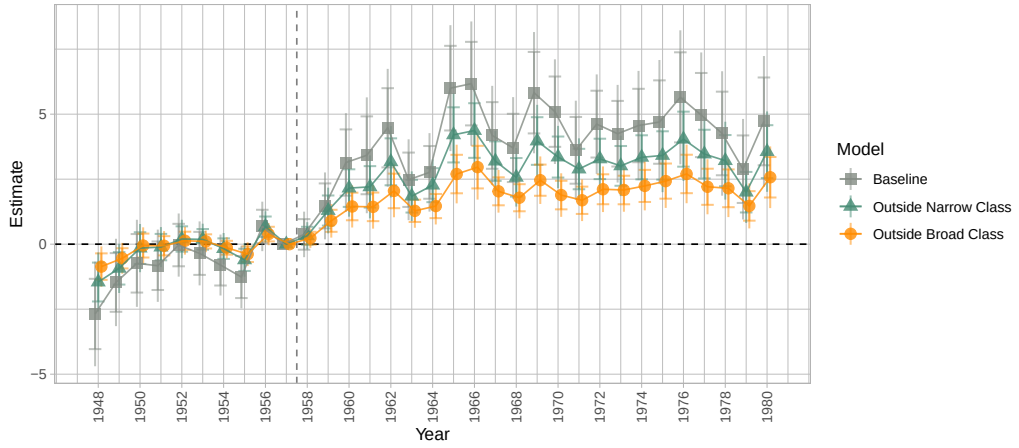
different broad classes (Columns (2) and (4) of Table 1.6, and Appendix Figures 1.26 and 1.27) and find diminished but statistically significant spillovers, implying this effect is not mostly driven by technology subclasses that are narrowly related, but rather, by classes that are in entirely different fields of invention.

Table 1.6: Difference in Differences Estimates, Leave-One-Out Citations

	Yearly	Yearly, Broad	Lifetime	Lifetime, Broad
	(1)	(2)	(3)	(4)
$\mathbb{I}$(NASA)	3.134***	2.055***	8.014***	5.860***
	(0.377)	(0.267)	(0.978)	(0.771)
Subclass FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Observations	249,803	249,803	249,803	249,803

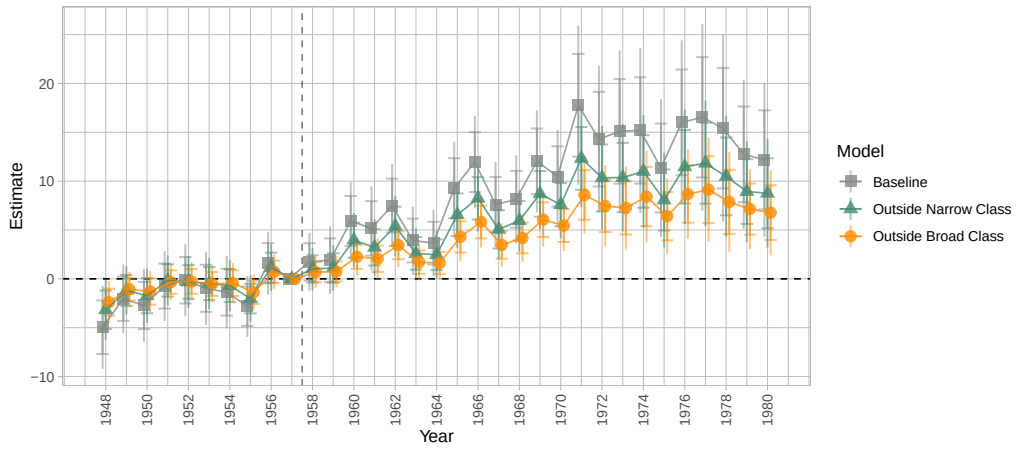
Note: Subclass clustered s.e. *p<0.1; **p<0.05; ***p<0.01. Columns (1) and (3) exclude within-narrow subclass citations, while (2) and (4) exclude within-broad class citations. All columns exclude NASA to NASA citations.

Figure 1.13: Citations by Year (Leave-One-Out) DID Estimates, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.14: Lifetime Citation (Leave-One-Out) DID Estimates, 1948-1980



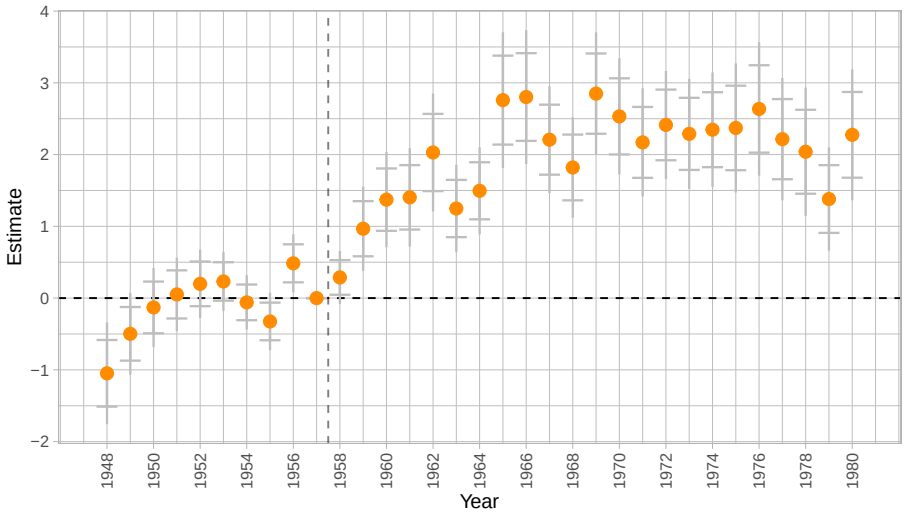
Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

To provide alternative measures of a technology's spillover breadth and length, I follow the literature on general purpose technologies (Moser and Nicholas, 2004; Rosenberg and Trajtenberg, 2004; Jovanovic and Rousseau, 2005), and calculate

alternative measures of a patent class' generality and longevity as outcomes.

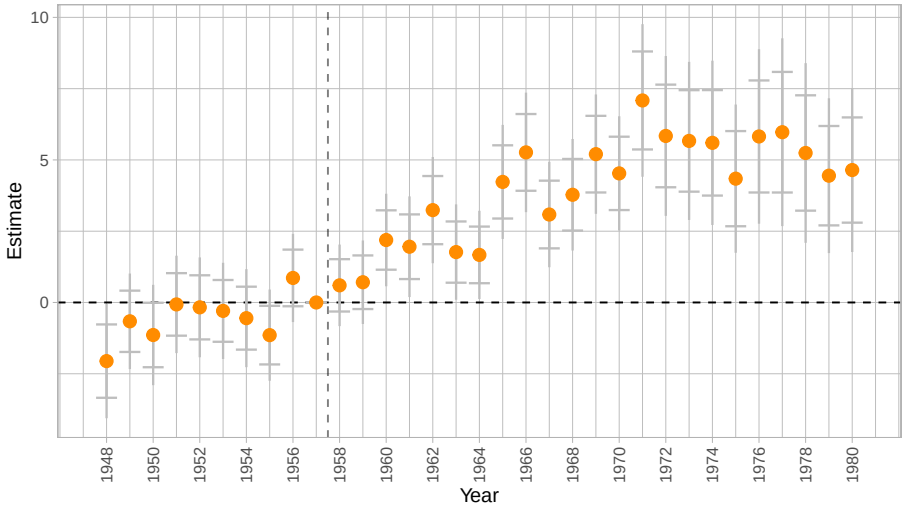
First, I estimate my main event study using a count of narrow classes that cite the treated class. While the previous set of regressions estimate the number of citations coming from other fields, this count abstracts from citation volume to calculate whether NASA's involvement changed the broadness of the impact of these technologies (Figures 1.15 and 1.16 and static estimates in Appendix Table 1.11). On average, the number of citing subclasses differentially increased for treated technology fields after treatment, with between 2.090 to 4.536 more classes citing treated classes over the control group.

Figure 1.15: Number of Citing Classes, Yearly Citations, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.16: Number of Citing Classes, Lifetime Citations, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

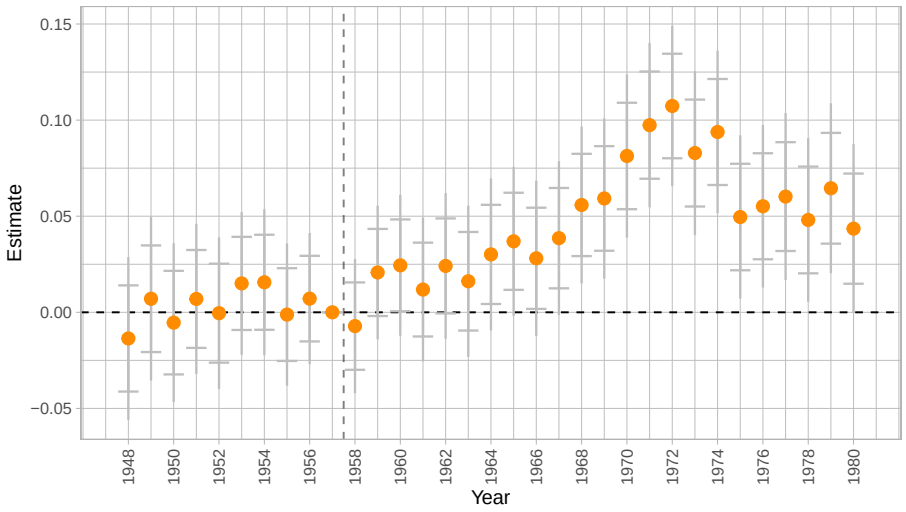
I also follow Moser and Nicholas (2004) and calculate a generality measure based on a Herfindahl-Hirschman index (HHI) for a given class and year's citations over all other J classes, defined as:

$$1 - \sum_{j=1}^{J} \left(\frac{C_j}{C} \right)^2 \tag{1.4}$$

As the second term measures the concentration of a given class' citations, one minus the concentration term implies that for a value of one, a subclass-year's citations are spread over many classes, while a value of zero implies that all citations were concentrated in one technology field. The estimates in Figures 1.17 and 1.18, while noisier than the simple counts of citing classes, are mostly positive and statistically significant. Static two-way fixed effects estimates are presented in Appendix Table

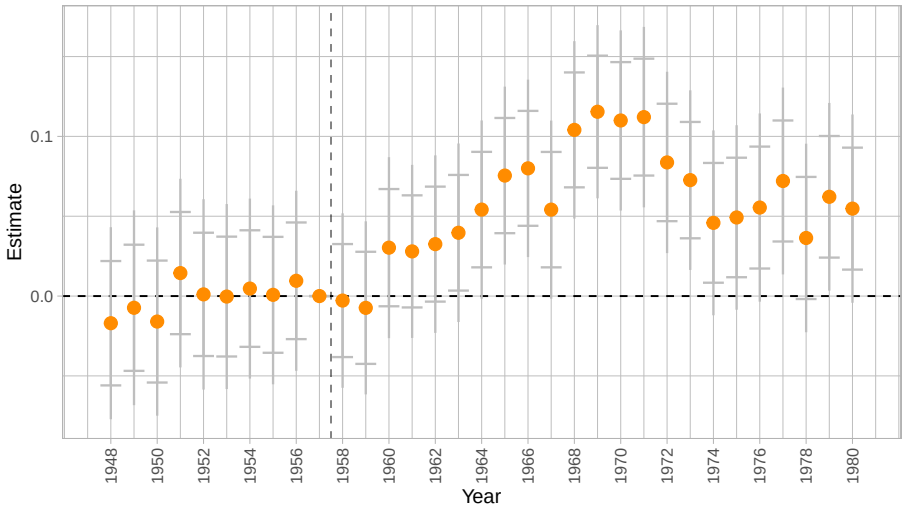
1.12. On average, the HHI is higher by 0.046 to 0.060 for treated subclasses, and it is statistically significant to the 1% level.

Figure 1.17: Herfindahl-Hirschman Based Generality, Yearly Citations, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.18: Herfindahl-Hirschman Based Generality, Lifetime Citations, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

To complement the estimates on breadth, Appendix Figures 1.28 and 1.29 estimate measures of longevity, that is, the speed of obsolescence for the inventions in a subclass-year cell. For patents issued in each subclass-year, I look at the mean and maximum lags in citations—the time difference between the original patent and the typical patent that cites it, and the gap between the original patent and the newest patent that cites it.

While the dynamic difference in differences specifications have noisier estimates, Appendix Table 1.13 shows that on average, treated subclasses have a longer citation lag of between 0.976 and 1.737 years, and this is statistically significant at the 1% level.

1.5.3 Blockbuster Patenting

The preceding sections indicate that knowledge spillovers stemming from treated subclasses increased on average. However, they do not specify if these fields generated individual patents that were particularly influential. In this following section, I devise a measure of breakthrough or blockbuster patenting based on the distribution of citation counts at the patent level.

First, I partial out subclass and year fixed effects from each individual patent's (i) lifetime citation count using Equation 1.5 to obtain a citation count adjusted for the fact that different fields and years receive varying amounts of citations.

$$y_{ict} = \gamma_c + \delta_t + \varepsilon_{ict} \tag{1.5}$$

Then, I define blockbuster patents as those that are above the 90th, 95th, and 99th percentiles of the residual citation distribution. Next, I re-estimate the static and dynamic two-way effects regressions at the subclass-year, where the outcome is an indicator that takes the value of one if the subclass-year contains at least one blockbuster patent. Difference in differences results are shown in Table 1.7 and Appendix Figures 1.30 through 1.32. On average, treated classes increased their 90th percentile blockbuster patenting by 0.041 percentage points, over a treated base year average of 0.127. At the 99th percentile, there was a statistically significant increase of 0.012 percentage points over a base year treated average of 0.0104.

Table 1.7: Difference in Differences Estimates, Blockbuster Patents

	90th Percentile	95th Percentile	99th Percentile
	(1)	(2)	(3)
$\mathbb{I}$(NASA)	0.041***	0.031***	0.012***
	(0.006)	(0.005)	(0.002)
Subclass FE	Y	Y	Y
Year FE	Y	Y	Y
Observations	249,803	249,803	249,803

Note: Subclass clustered s.e. *p<0.1; **p<0.05; ***p<0.01

1.5.4 Space-Essential Classes

Treated and control subclasses appear to display mostly parallel pre-treatment trends in the decade before NASA's creation, which appears to support the argument that NASA's technology portfolio was mostly mission-driven. However, to the degree that selection of technologies deviated from NASA's mandate, I re-estimate my main regressions using technologies that were ex-ante known to be essential to winning the Space Race: spacecraft capable of withstanding travel in the vacuum of space and re-entry into Earth, heavy-lift rockets able to carry said spacecraft beyond Earth orbit, and life support systems that can ensure the survival of humans under the stresses of space, such as heavy radiation and extreme temperatures.

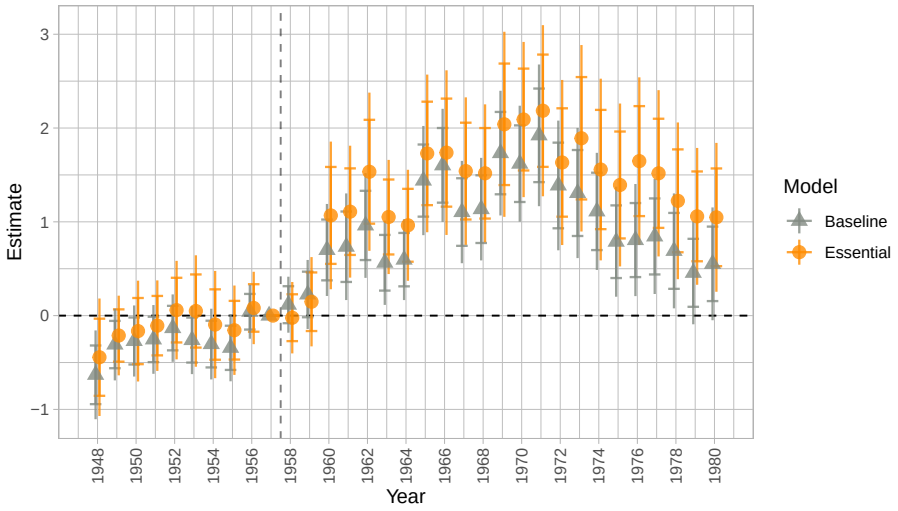
As evidenced in its first fiscal year budget, NASA had already commenced testing and development on all of these categories, ranging from life-support and restraint systems for manned spaceflight, heat and shock-resistant metals and ceramics for capsules, advanced digital guidance, control, and communications technologies, solid

and liquid fuel rockets, solar cells, and batteries (U.S. House of Representatives. Committee on Appropriations, 1960).[27]

The following figures show event study estimates of this redefined treatment, following Equation 1.1, where treated subclasses are those in technology classes matching patent broad classes in rocketry, aeronautics and aerospace, batteries, digital communications, radiant energy, advanced alloys and coatings, and computer-aided calculation. These estimates are on average similar to the estimates in Figures 1.9-1.12 (overlaid for convenience), with larger standard errors, larger estimates, and flatter pre-treatment estimates. Leave-one-out estimates are added to the Appendix (Figures 1.33 & 1.36) for brevity, but are significant and similar to their baseline versions as well.

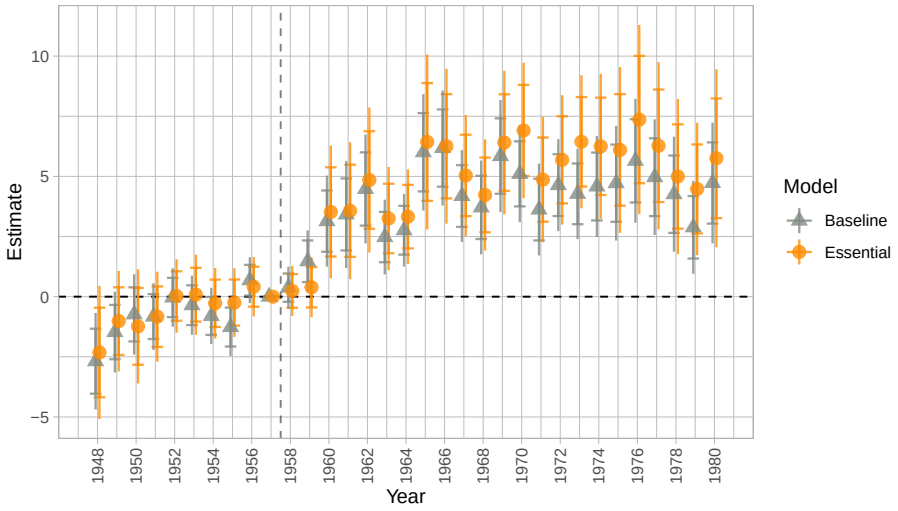
27. This is evidenced in several other points early in NASA's history. President John F. Kennedy mentions some of these necessary advances in his first speeches advocating for a Moon landing: the Address to the Joint Session of Congress on May 25, 1961, where he first proposed a landing before the end of the decade, and his Address at Rice University on the Nation's Space Effort on September 12, 1962, colloquially known as his "We choose to go to the Moon" speech. Despite lack of political and budgetary support for a Moon landing prior to this, NASA had already formed 12 committees dedicated to feasibility and planning for the lunar landing, starting as early as February 1959 (Brooks et al., 1979; Hansen, 1995).

Figure 1.19: Patent Issue Estimates, Space-Essential Classes



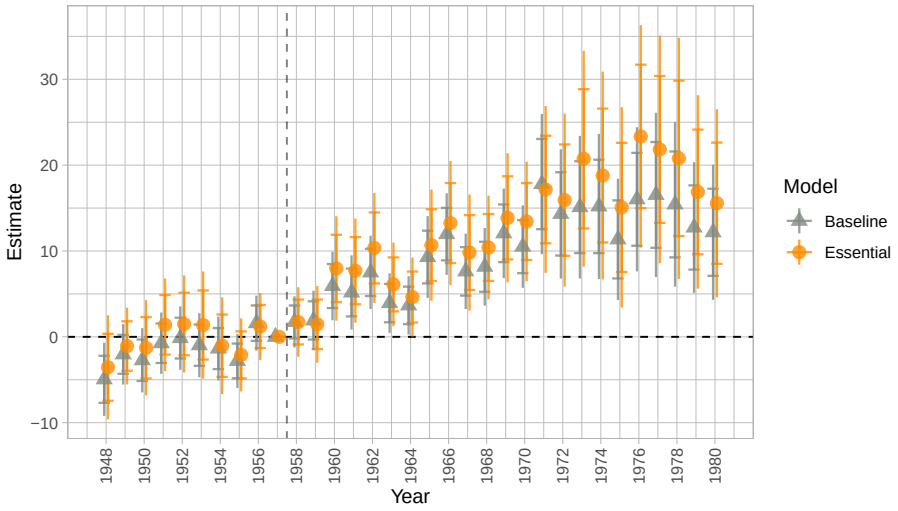
Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.20: Citations by Year Estimates, Space-Essential Classes



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

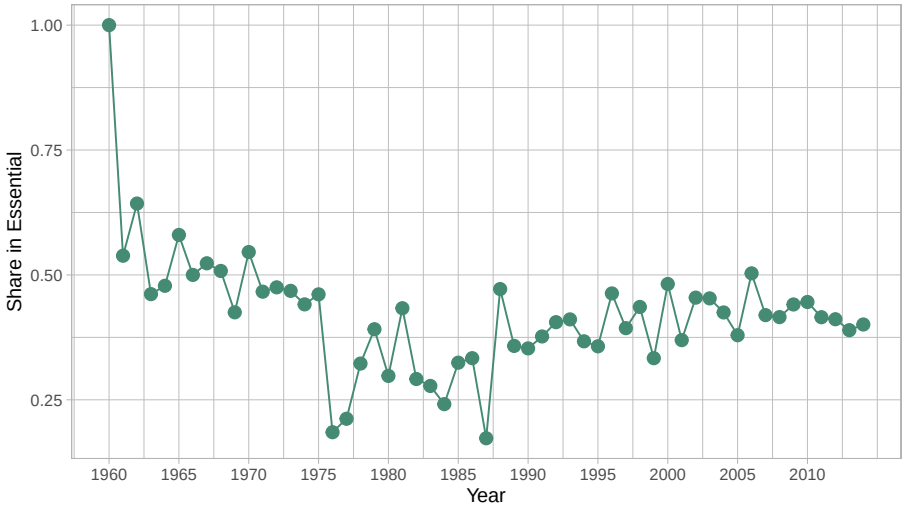
Figure 1.21: Lifetime Citations Estimates, Space-Essential Classes



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

To assess the degree to which NASA deviated from these space-essential classes throughout the years, I plot the share of treated classes within this essential subset for each year in the post-treatment sample in Figure 1.22. Over the Apollo years, this share averaged 53.692%, while over the whole sample this averaged to 42.053%.

Figure 1.22: Share of Essential Classes in Realized NASA Classes, 1958-2014



Note: First year in sample contains one single patent.

1.5.5 Non-Disclosure and Military Spending

Given that NASA was born of Cold War geopolitical competition, there are two natural questions to raise about NASA-involved inventions. First, were they affected by non-disclosure? That is, the government would have restricted their publication and potential to influence other downstream innovations. Second, are estimates of treated technologies simply capturing the effect of concurrent military spending on NASA-related technologies, such as innovations on intercontinental ballistic missiles (ICBMs)?

NASA innovations could have feasibly been affected by non-disclosure. However, given NASA's origins as the civilian branch of the United States' space effort, these inventions would have been subject to less restrictions than military efforts such as

those from the Department of Defense. To the degree that non-disclosure affected NASA related technologies, it is likely that they would also be more restricted than untreated technologies, therefore restricting the potential influence of the treatment group. Given that previous results show a positive knowledge spillover from treated technologies, non-disclosure would make these estimates lower bounds on the average treatment effect on the treated.

To analyze the degree to which the effects are being driven by concurrent military spending, I re-estimate my baseline results after omitting highly-overlapping technology categories, such as ordnance and rocket-related classes[28] in Appendix Figures 1.37-1.43. I find my main results largely unchanged. While defense spending has consistently taken the largest share of federal R&D across the years (Figure 1.2), the redirection of government technology efforts in the Apollo years coincided with some of the lowest of these shares in defense R&D spending in US history.

1.5.6 Inventor-level Reallocation

I investigate the extent to which the above results are plausibly driven by a relative reallocation of patenting as opposed to a shift in aggregate patenting. Using the inventor names in the patents, I identify all engineers and inventors that were ever issued a NASA-affiliated patent and manually match them to all other patents they held before joining NASA as employees or as contractors.[29]

28. Specifically, I omit subclasses within ordnance, ammunition and explosives, ammunition and explosive making, firearms, mechanical guns and projectors, and explosive and thermic compositions.

29. The matching procedure is discussed in Appendix 1.7.9.

NASA-affiliated inventors typically received their first patent after joining NASA, and not before. From 1940 to 1980, out of 2,276 affiliated inventors, only 28.16% held a patent before their first NASA patent. This would tentatively imply that NASA-affiliated inventive output was not mostly driven by the reallocation of existing scientists or engineers. This is not to say that NASA-affiliated inventors would not have counterfactually issued patents had they not joined NASA, but that the bulk of NASA inventors were not already producing patents elsewhere. [30] Given the discussion of non-disclosure above, however, one would expect patenting propensity to be lowest in treated fields and not in control fields.

However, conditional on holding a patent in the pre-treatment period, only 25.60% of inventors were working in at least one of their post-treatment fields in their pre-NASA patents. This implies that there was some degree of reallocation between fields for approximately 7.03% of NASA affiliated inventors.

1.6 CONCLUSION

There is growing evidence that public research and development efforts can crowd in private sector innovation efforts, and that large scale government research programs affect the direction of future research. I estimate that NASA's creation and sizable funding during the Space Race of the 1960s increased the innovation output

30. This could however also be a result of heterogeneous patenting propensity across fields. Consider an engineer who works in non-NASA field A, which has a low propensity to patent. In the post-treatment period, she joins NASA and begins working in NASA field B, which has a high propensity to patent. Then, even though her movement implied a reallocation of inventive capital from field A to B, the patent data is less likely to show that she was already inventing in the pre-period.

of spaceflight related fields, and that this increase in innovation did not originate entirely from patents that NASA originated or contracted. The patents that originated from this innovation rush had larger impact on future innovation by various citation metrics, and these impacts extended to technology fields beyond their own.

These results are robust to measuring citations and their breadth in multiple ways, to removing technology fields that are explicitly defense related, and when only looking at ex-ante spaceflight relevant technology classes.

To assess the degree to which this innovation came from reallocating scientists and engineers from other fields to spaceflight technologies, I match inventors in patent records over time, and find that most engineers that ever held a NASA patent had not received patents beforehand.

These empirical results altogether support the literature that large public R&D efforts can shape the direction and intensity of technological growth, and do so not only through their direct output, but through the crowding in of private sector innovation efforts. They also suggest that unlike commonly theorized, these causal effects are not only driven through the public sector's basic science efforts, but also through its applied innovation.

1.7 APPENDIX

1.7.1 Treatment and Control Classes

The following tables show the top 10 treatment and control classes by patent count. Columns include total patents and accrued lifetime citations for 1948-1980 patents.

Table 1.8: Top 10 Treated Classes, by Patent Count, 1948-1980

USPC Subclass	Patents	Citations
Fluid Handling, Multiway Valve Unit	2,356	23,258
Power Plants, Combustion Products Used as Motive Fluid	2,296	21,362
Electricity: Circuit Makers and Breakers, Incubator	2,056	13,025
Measuring and Testing, Volume or Rate of Flow	1,746	17,385
Measuring and Testing, Dynamometers	1,391	10,771
Machine Element or Mechanism, Gyroscopes	1,348	6,215
Ordnance	1,331	10,193
Communications: Electrical, Continuously Variable Indicating	1,108	12,635
Compositions, Organic Luminescent Material	1,051	6,672
Metal Working, Catalytic Device Making	993	8,800

1.7.2 Fixed Window Citations

The following table and figures show results for the baseline difference in differences estimates using a fixed window of forward citations within 20 years of issuance instead of lifetime citations.

Table 1.9: Top 10 Control Classes, by Patent Count, 1948-1980

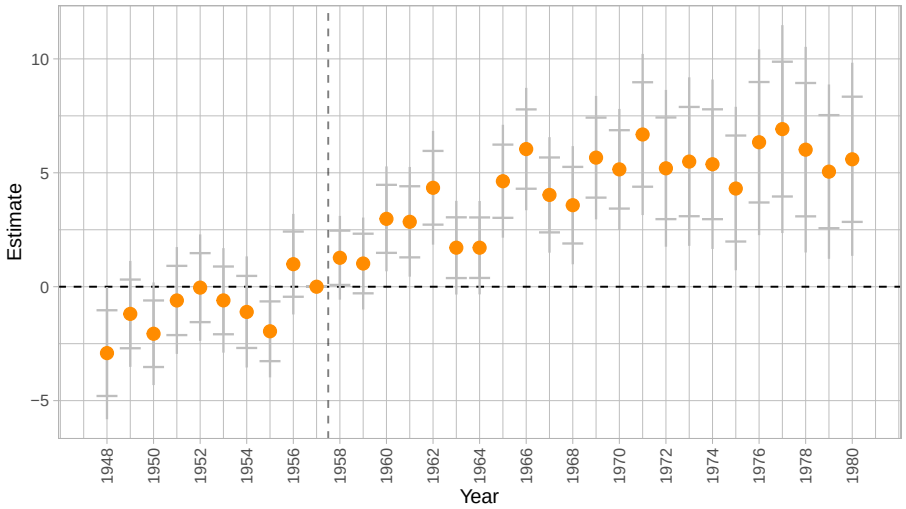
USPC Subclass	Patents	Citations
Fishing, Trapping, Vermin Destroying, Artificial Bait	1,540	12,347
Electric Lamp/Discharge Devices, Cathode Ray Tube Circuit	1,020	6,828
Electric Lamp/Discharge Devices, with Transmission Line	985	5,229
Internal Combustion Engines, Reversible	963	7,079
Printing, Bed and Platen Machines	939	6,750
Chemistry: Electrical and Wave Energy, Treating Materials	914	4,489
Land Vehicles, Suspension Arrangement	891	7,314
Brushing, Scrubbing, Cleaning, Implements	862	8,295
Specialized Metallurgical Processes, Electrothermic	835	5,368
Lubrication, Systems	687	5,150

Table 1.10: Difference in Differences Estimates, 20-Year Window Citations

	Citations	Excl. NASA	Leave-One-Out, Narrow	LOO, Broad
	(1)	(2)	(3)	(4)
$\mathbb{I}$(NASA)	5.337***	5.330***	3.704***	2.567***
	(0.789)	(0.789)	(0.489)	(0.364)
Subclass FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Observations	249,803	249,803	249,803	249,803

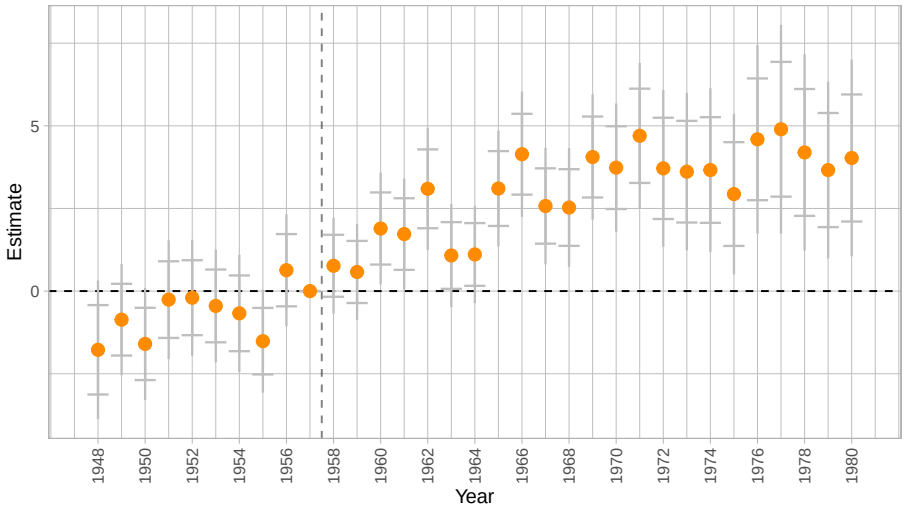
Note: Subclass clustered s.e. *p<0.1; **p<0.05; ***p<0.01

Figure 1.23: 20-Year Window Citation DID Estimates, 1948-1980



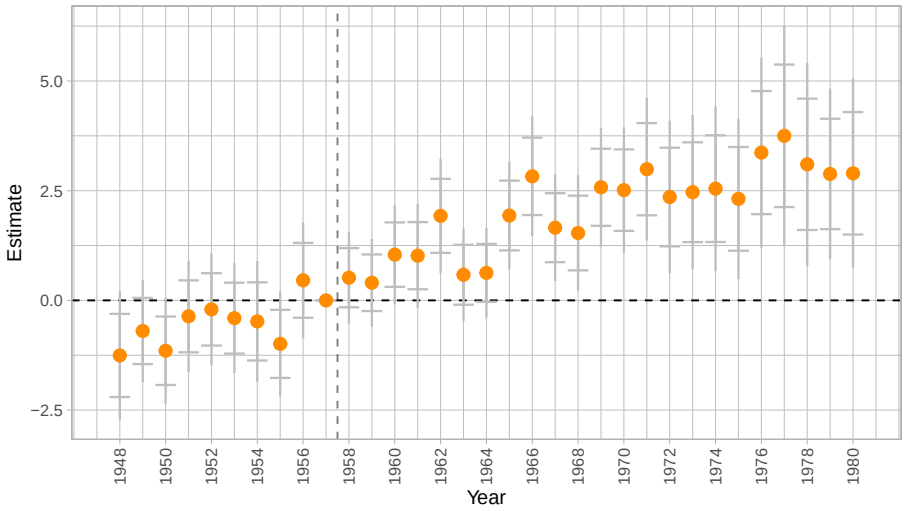
Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.24: 20-Year Window Citation (Leave-One-Out) DID Estimates, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

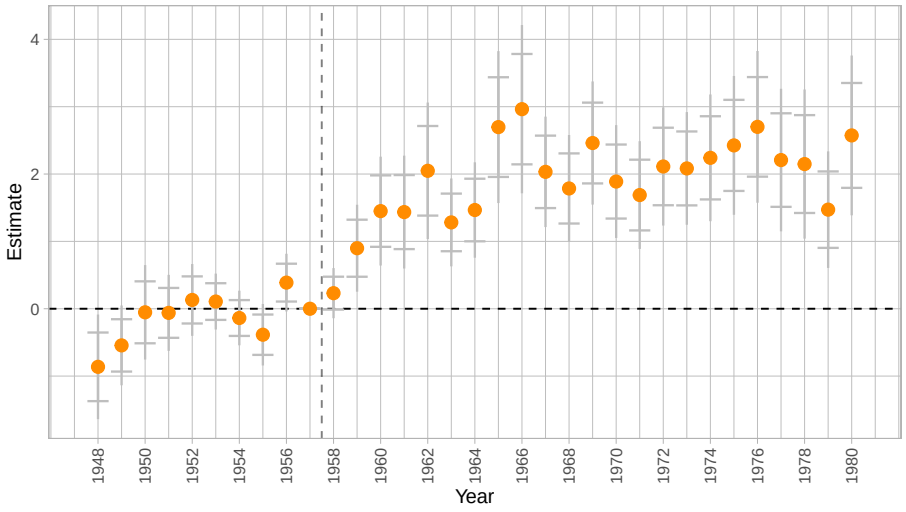
Figure 1.25: 20-Year Window Citation (Broad Leave-One-Out) DID Estimates, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

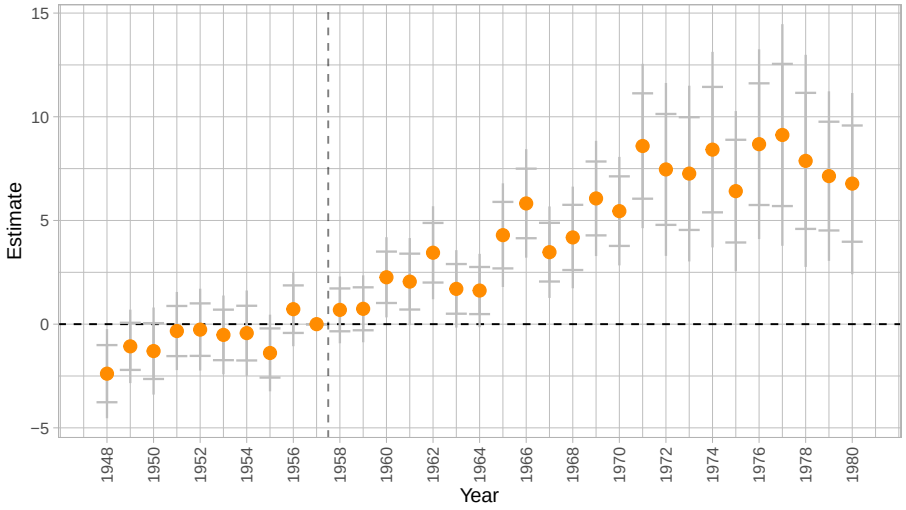
1.7.3 Additional Difference in Differences Estimates for Citation Breadth and Length

Figure 1.26: Citations by Year (Broad Leave-One-Out) DID Estimates, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.27: Lifetime Citation (Broad Leave-One-Out) DID Estimates, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Table 1.11: Difference in Differences Estimates, Number of Citing Classes

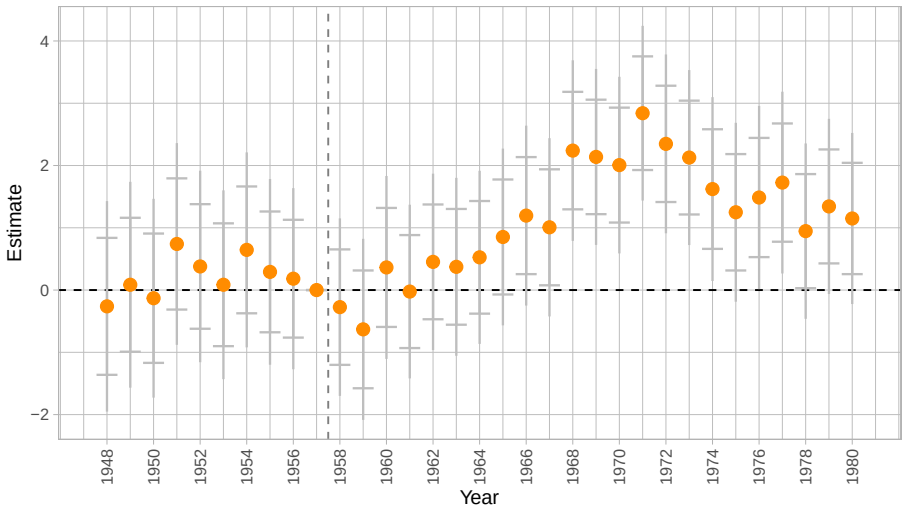
	No. Citing Classes, Yearly	No. Citing Classes, Lifetime
	(1)	(2)
$\mathbb{I}$(NASA)	2.090***	4.536***
	(0.223)	(0.534)
Subclass FE	Y	Y
Year FE	Y	Y
Observations	249,803	249,803

Note: Subclass clustered s.e. *p<0.1; **p<0.05; ***p<0.01

Table 1.12: Difference in Differences Estimates, Citation HHI

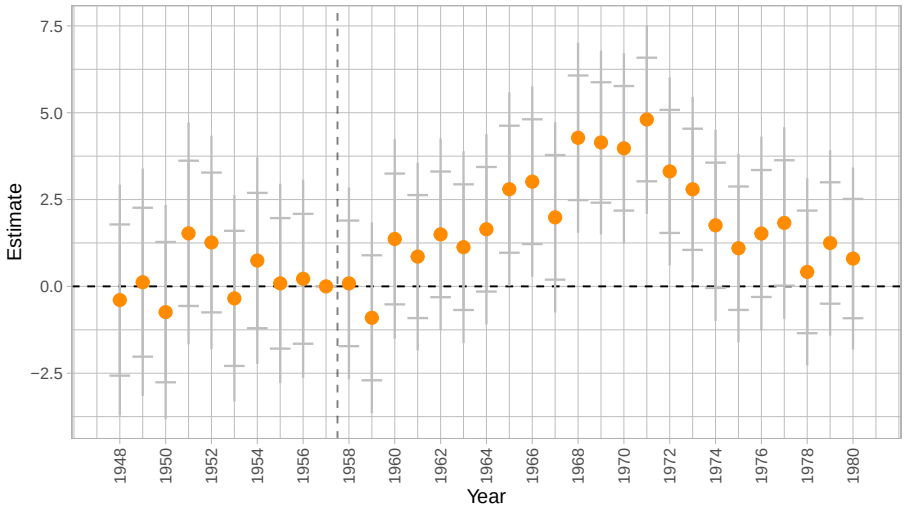
	HHI, Yearly	HHI, Lifetime
	(1)	(2)
$\mathbb{I}$(NASA)	0.046***	0.060***
	(0.008)	(0.008)
Subclass FE	Y	Y
Year FE	Y	Y
Observations	249,803	249,803
Note:	Subclass clustered s.e. *p<0.1; **p<0.05; ***p<0.01	

Figure 1.28: Mean Lag in Citations, 1948-1980



Note: S.E. clustered at subclass level, $\mathbb{I}$: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.29: Maximum Lag in Citations, 1948-1980



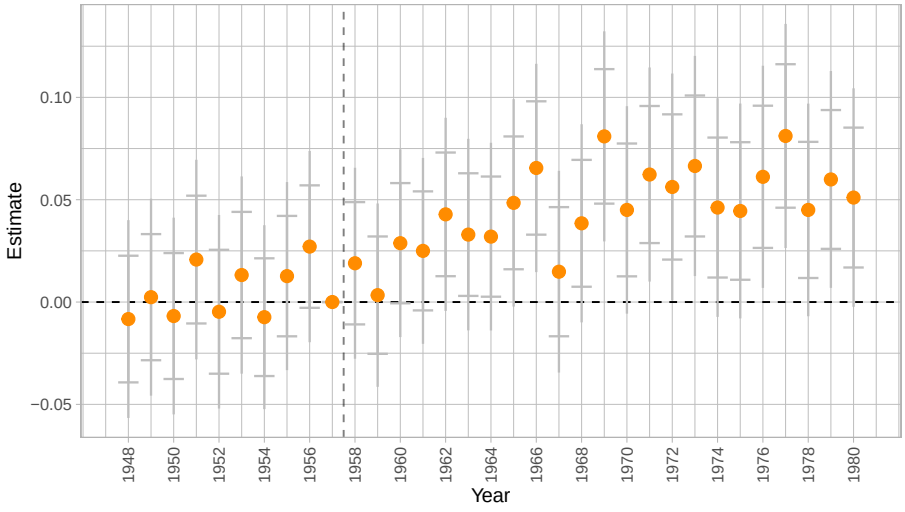
Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Table 1.13: Difference in Differences Estimates, Mean and Maximum Citation Lag

	Mean Citation Lag	Maximum Citation Lag
	(1)	(2)
𝕀(NASA)	0.976***	1.737***
	(0.175)	(0.404)
Subclass FE	Y	Y
Year FE	Y	Y
Observations	249,803	249,803
Note:	Subclass clustered s.e. *p<0.1; **p<0.05; ***p<0.01	

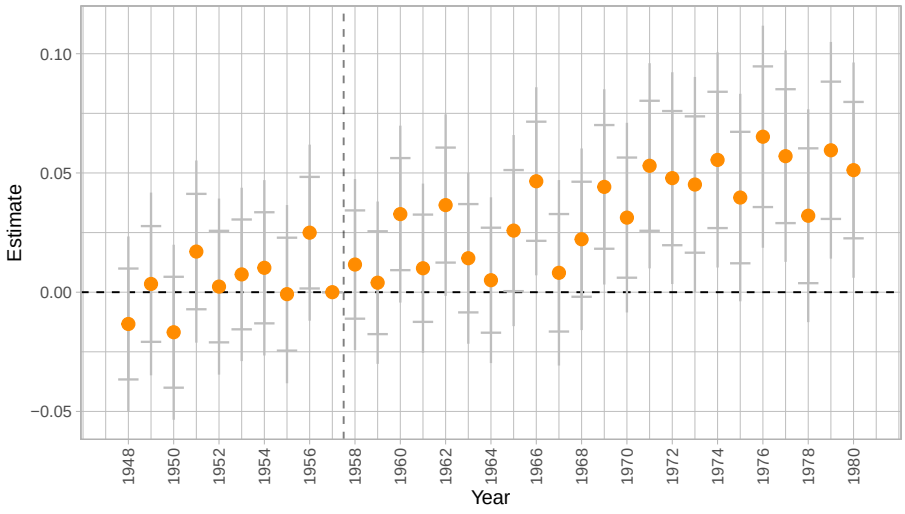
1.7.4 Additional Blockbuster Patenting Event Studies

Figure 1.30: Blockbuster Patenting, 90th Percentile, 1948-1980



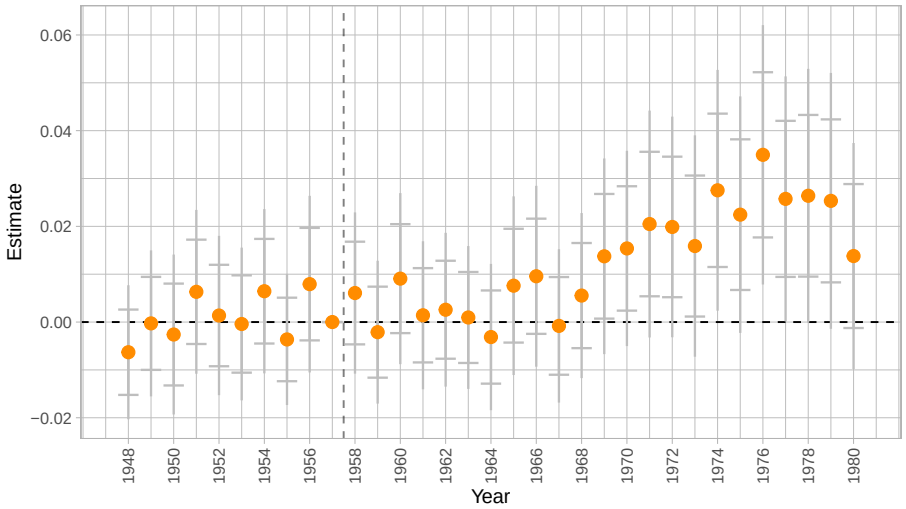
Note: S.E. clustered at subclass level, ⌶: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.31: Blockbuster Patenting, 95th Percentile, 1948-1980



Note: S.E. clustered at subclass level, ⌶: point-wise 95% CI, |: sup-t 95% confidence band.

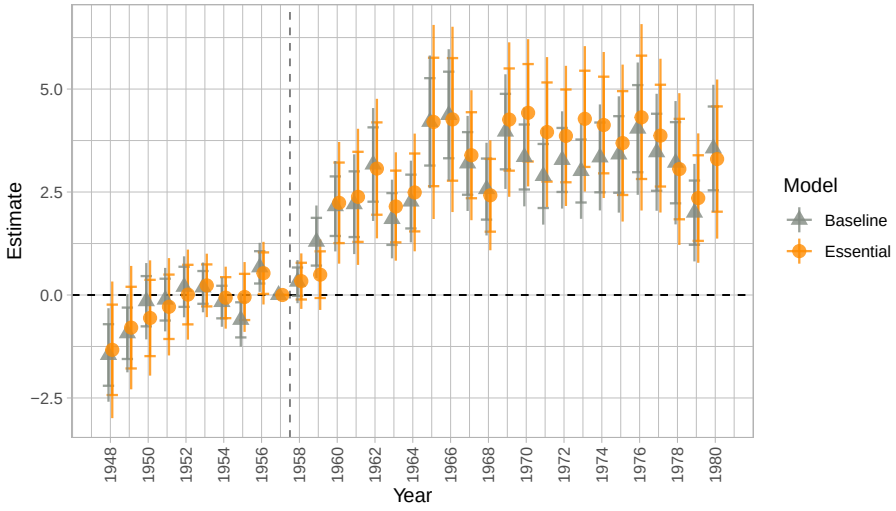
Figure 1.32: Blockbuster Patenting, 99th Percentile, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

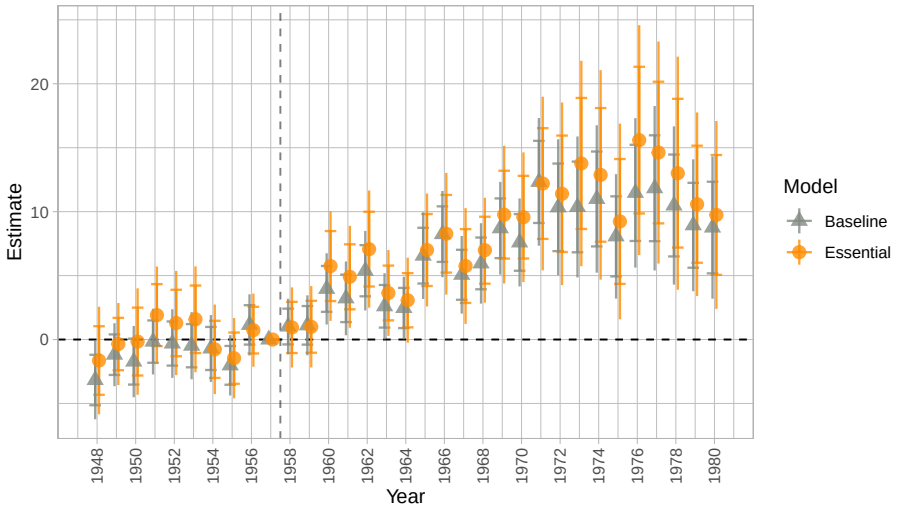
1.7.5 Additional Space Essential Class Event Studies

Figure 1.33: Citations by Year Leave-One-Out Estimates, 1948-1980



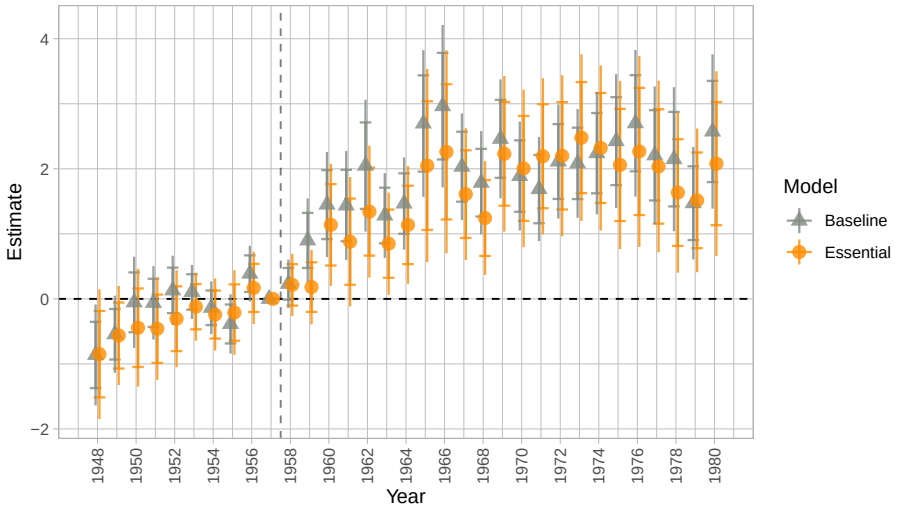
Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.34: Lifetime Citations Leave-One-Out Estimates, 1948-1980



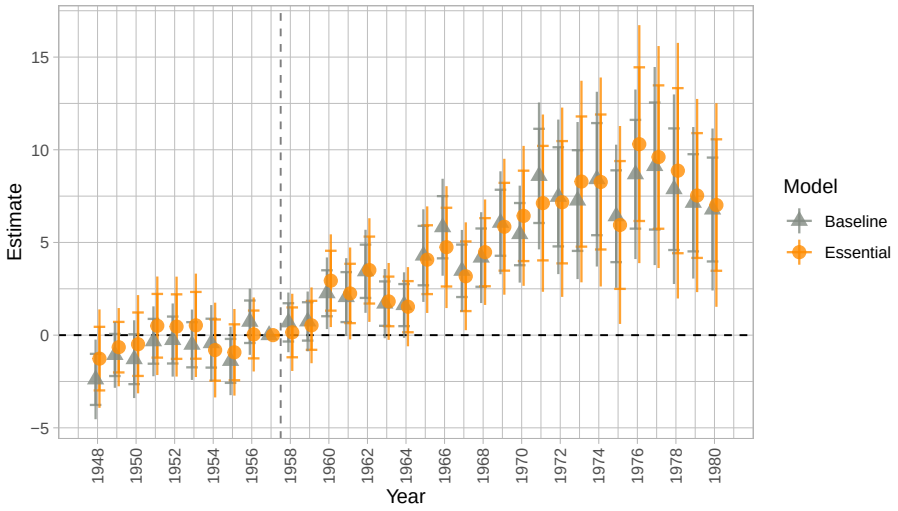
Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.35: Citations by Year Leave-One-Out Estimates, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

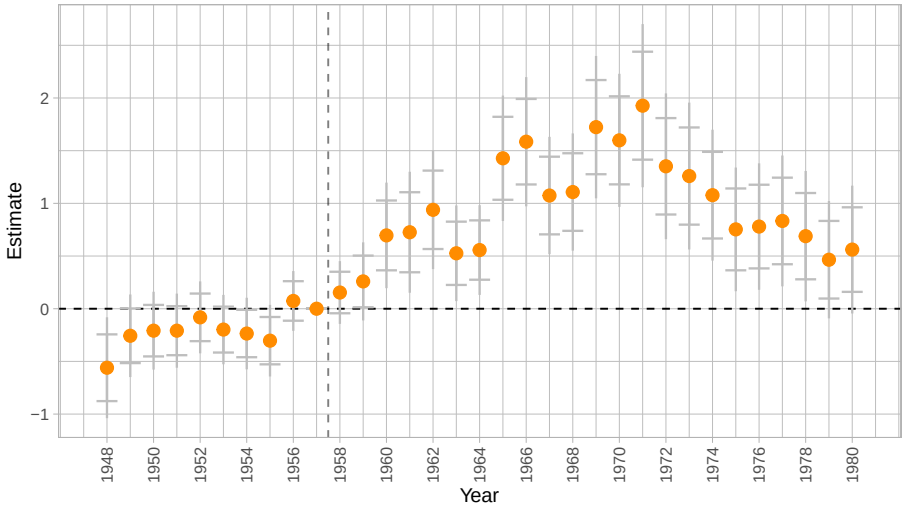
Figure 1.36: Lifetime Citations Leave-One-Out Estimates, 1948-1980



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

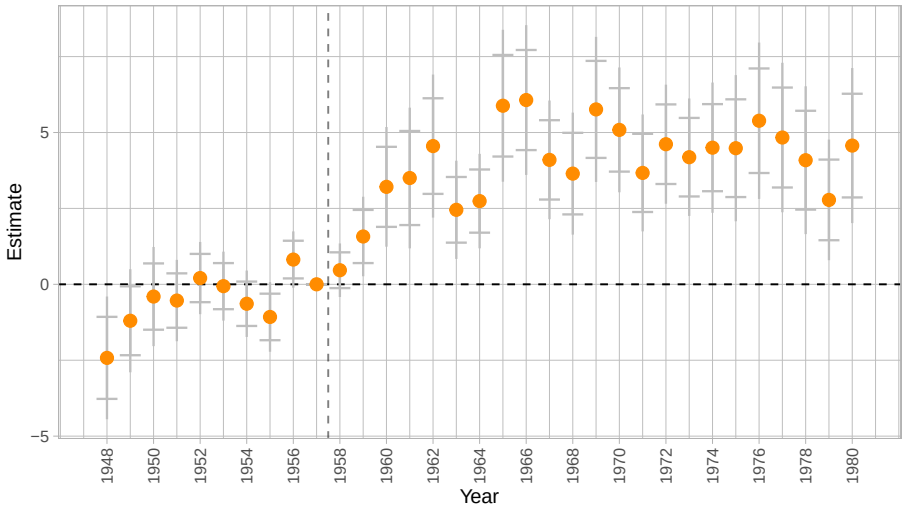
1.7.6 Estimates Excluding Military-Related Classes

Figure 1.37: Patent Issue DID Estimates, Excluding Military Classes



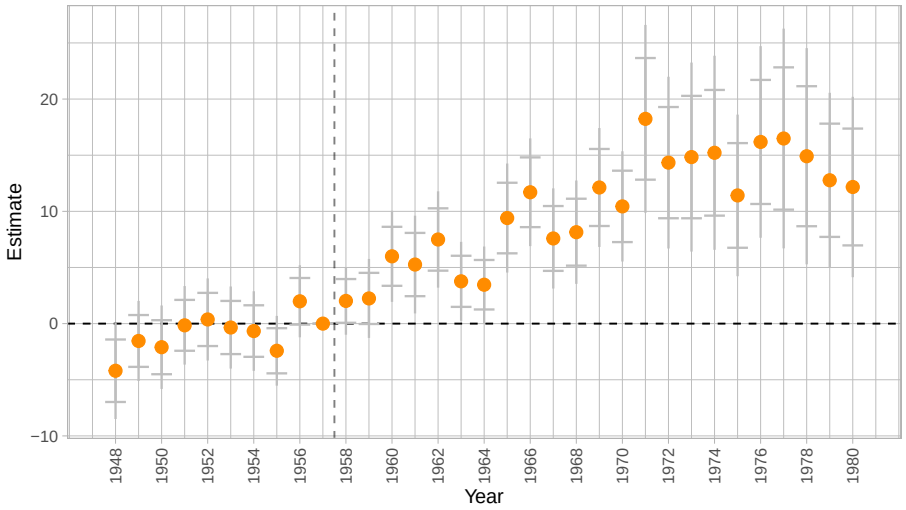
Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.38: Citations by Year DID Estimates, Excluding Military Classes



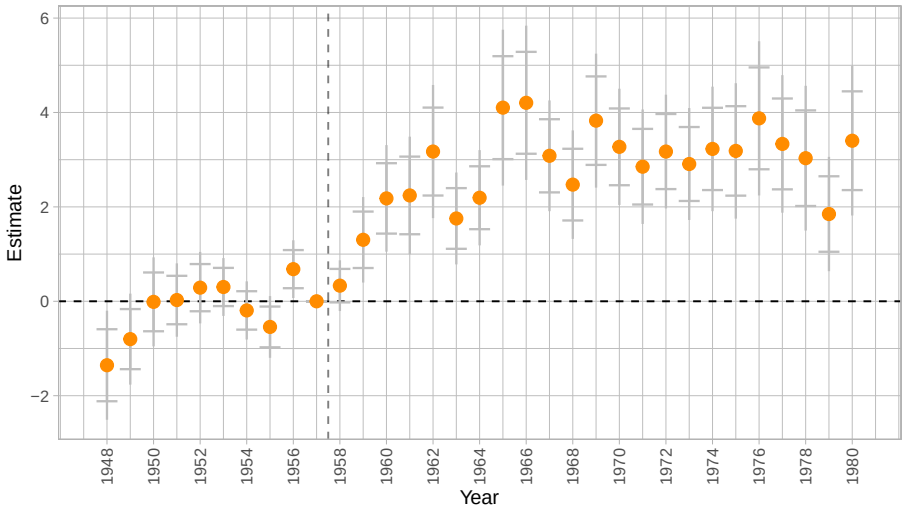
Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.39: Lifetime Citations DID Estimates, Excluding Military Classes



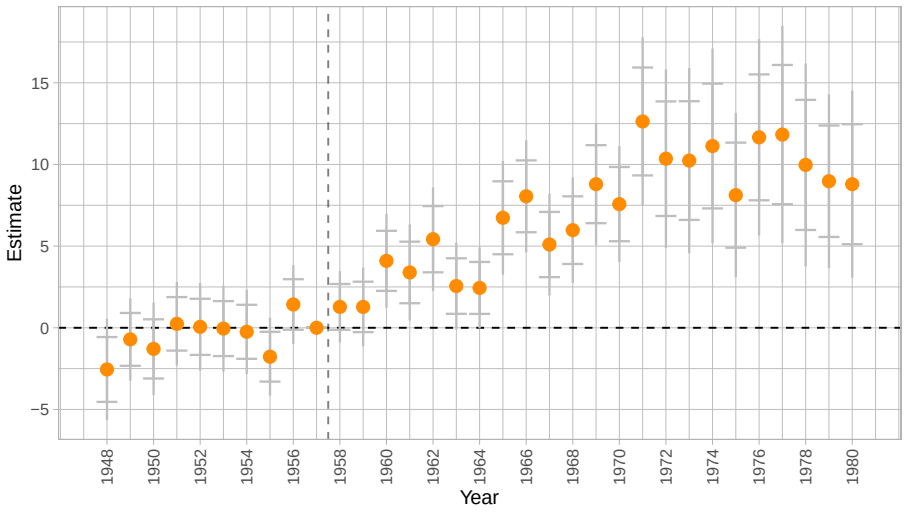
Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.40: Citations by Year (Leave-One-Out) DID Estimates, Excluding Military Classes



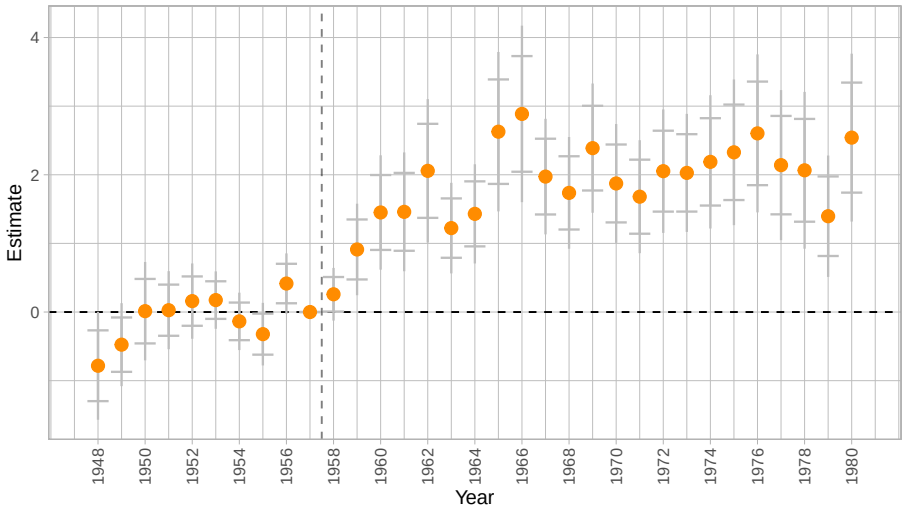
Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.41: Lifetime Citations (Leave-One-Out) DID Estimates, Excluding Military Classes



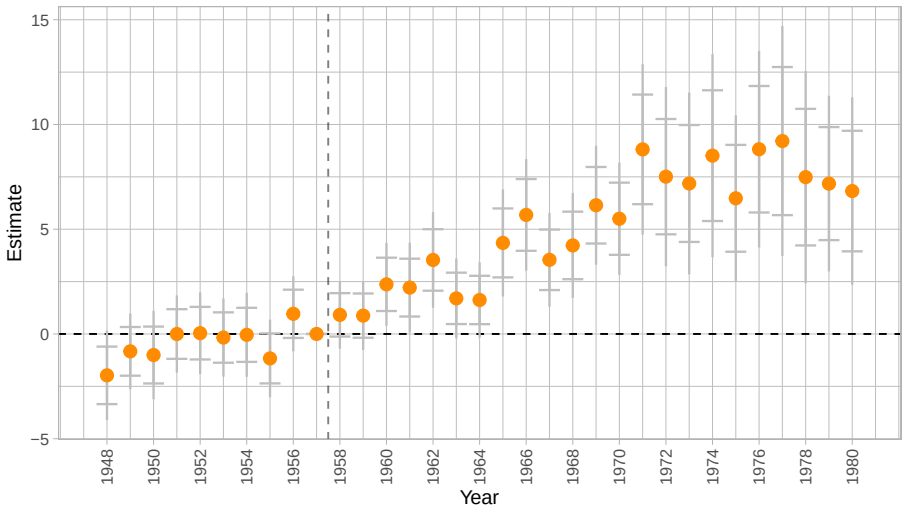
Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.42: Citations by Year (Broad Leave-One-Out) DID Estimates, Excluding Military Classes



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.43: Lifetime Citations (Broad Leave-One-Out) DID Estimates, Excluding Military Classes

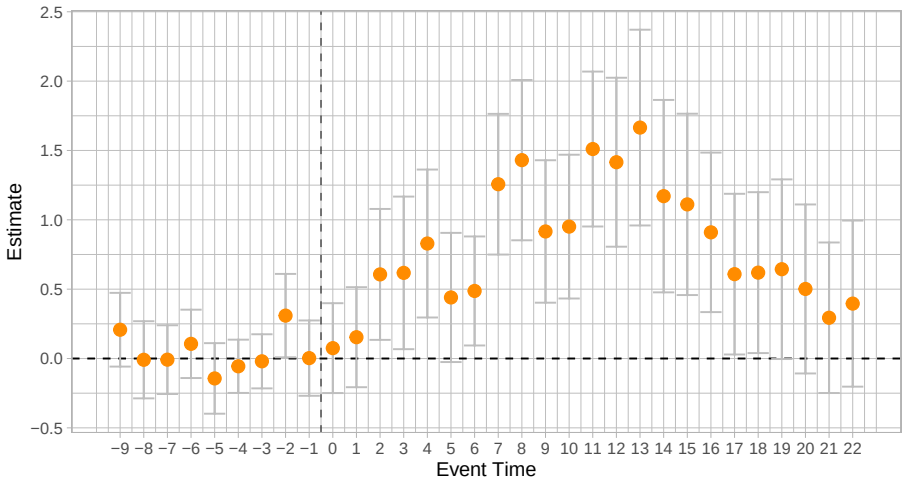


Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

1.7.7 Alternate Event Study Estimation Methods

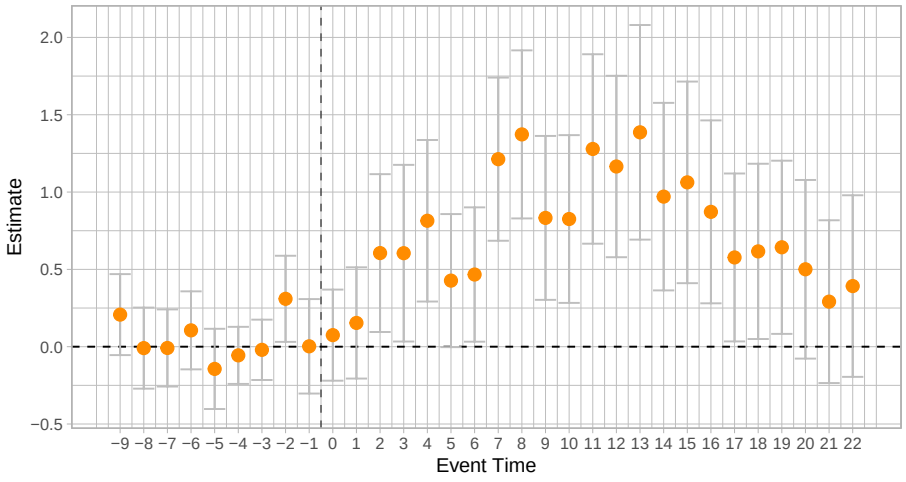
A recent strand of the econometrics literature has focused on the potential pitfalls of estimating static and dynamic two-way fixed effects regressions. In particular, bias can arise when treatments are staggered over time, when there is treatment effect heterogeneity, or when dynamic regressions are not fully saturated (de Chaisemartin and D'Haultfœuille, 2020, Goodman-Bacon, 2021, Callaway and Sant'Anna, 2021, Sun and Abraham, 2021). While my estimates have treatments happening at the same time, 1958, heterogeneous treatment effects are likely to exist in my setting, as certain technology subfields are likely higher impact on average, likelier to be treated, and have a potentially heterogenous response to different levels of R&D funding. Due to this, I re-estimate my main estimates using Callaway and Sant'Anna's (2021) method of estimating group-wise treatment effects individually and aggregating them via a group-size weighted average per year. The resulting event study estimates are mostly unchanged from my main estimates.

Figure 1.44: Patent Issue DID Estimates, Callaway-Sant'Anna



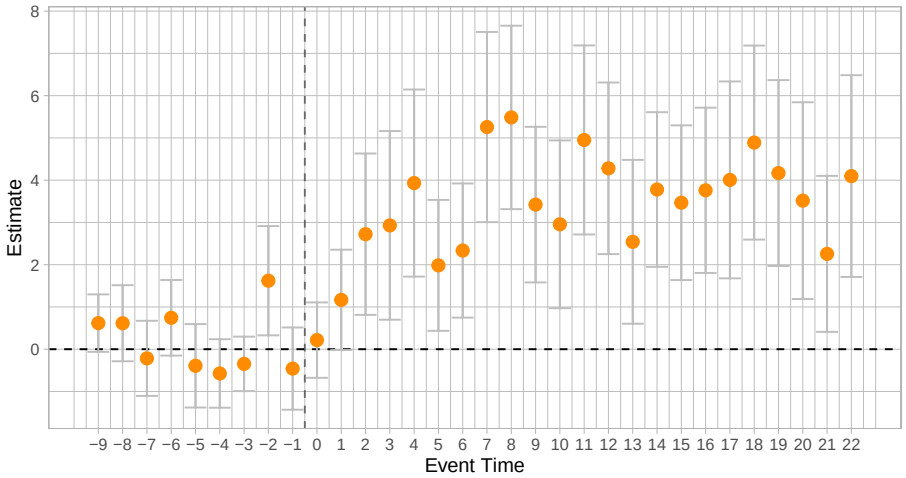
Note: S.E. clustered at subclass level, ⌶: sup-t 95% confidence band.

Figure 1.45: Patent Issue DID Estimates, Excl. NASA Patents, Callaway-Sant'Anna



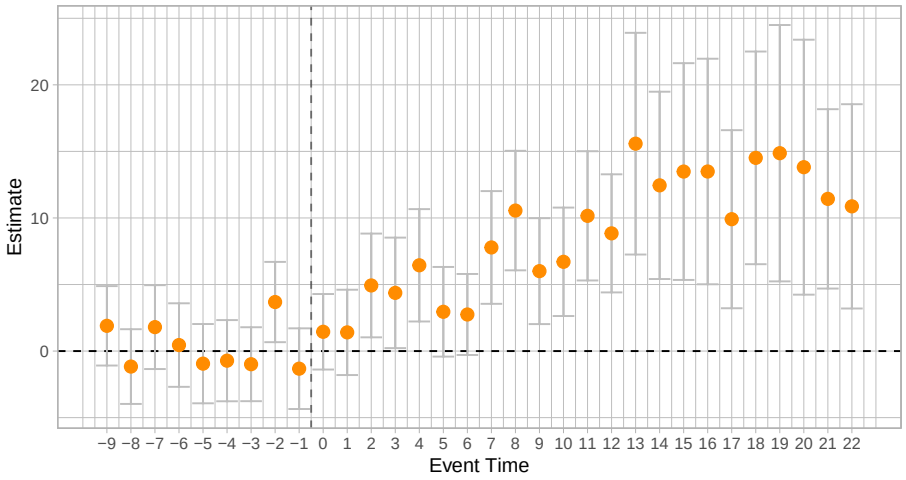
Note: S.E. clustered at subclass level, ⌶: sup-t 95% confidence band.

Figure 1.46: Citations by Year DID Estimates, Callaway-Sant'Anna



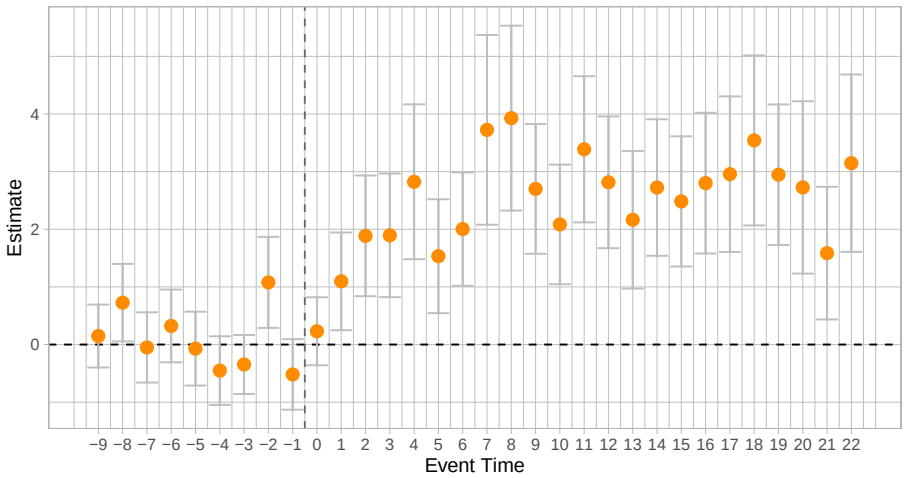
Note: S.E. clustered at subclass level, ⌶: sup-t 95% confidence band.

Figure 1.47: Lifetime Citation DID Estimates, Callaway-Sant'Anna



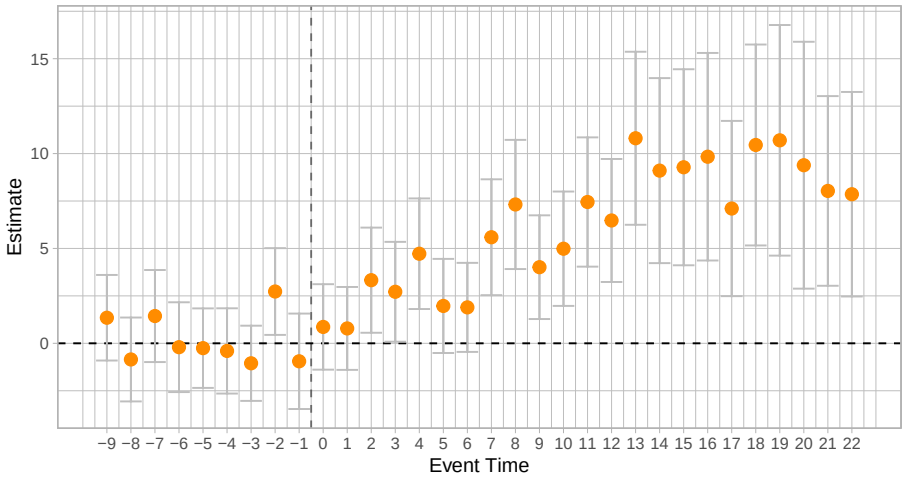
Note: S.E. clustered at subclass level, ⌶: sup-t 95% confidence band.

Figure 1.48: Citations by Year (LOO) DID Estimates, Callaway-Sant'Anna



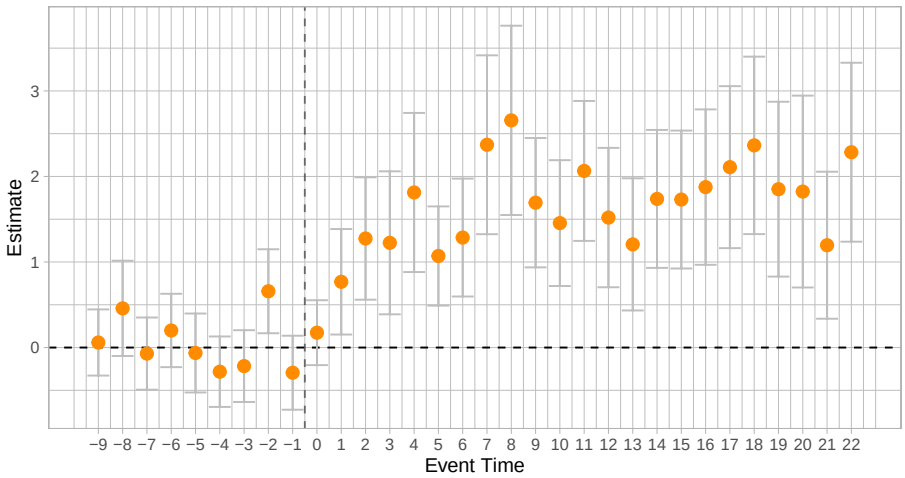
Note: S.E. clustered at subclass level, ⌶: sup-t 95% confidence band.

Figure 1.49: Lifetime Citation (LOO) DID Estimates, Callaway-Sant'Anna



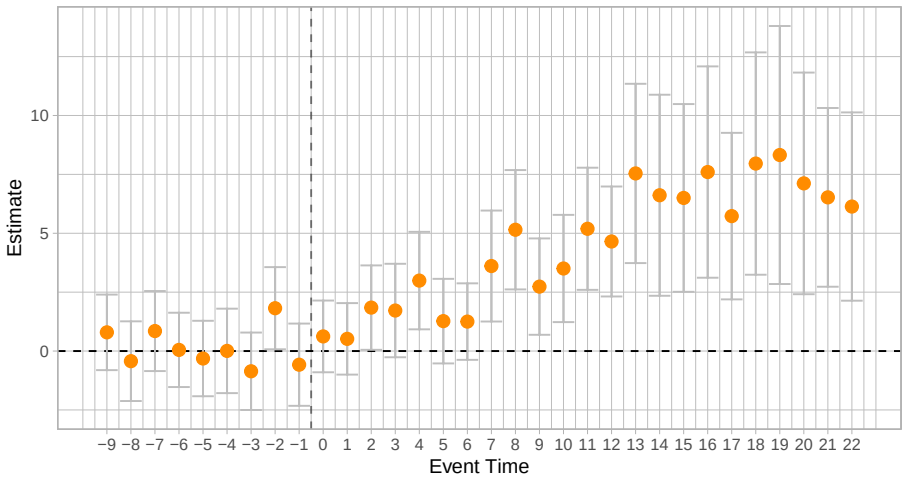
Note: S.E. clustered at subclass level, ⌶: sup-t 95% confidence band.

Figure 1.50: Citations by Year (Broad LOO) DID Estimates, Callaway-Sant'Anna



Note: S.E. clustered at subclass level, ⌶: sup-t 95% confidence band.

Figure 1.51: Lifetime Citation (Broad LOO) DID Estimates, Callaway-Sant'Anna



Note: S.E. clustered at subclass level, ⌶: sup-t 95% confidence band.

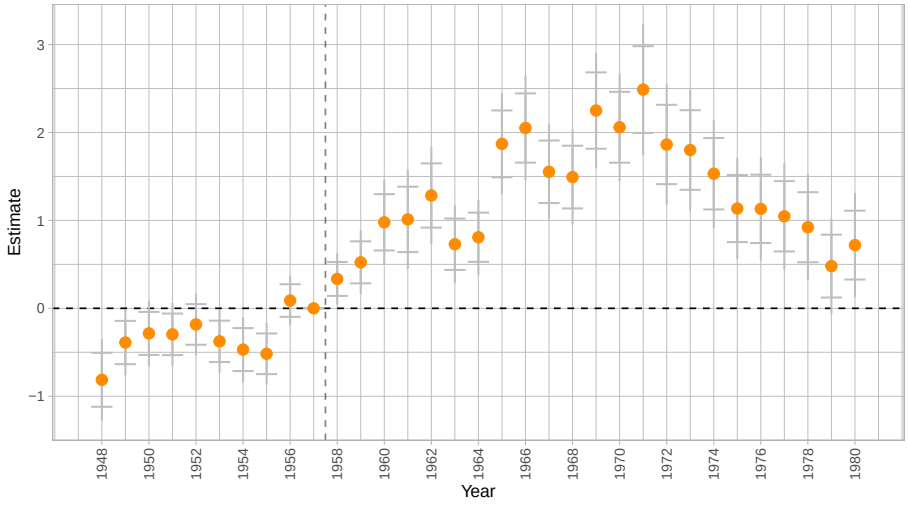
1.7.8 Alternative Control Groups

This section re-estimates the baseline specification in Equation 1.1 using two alternative control groups instead of using other government related classes.

All Classes

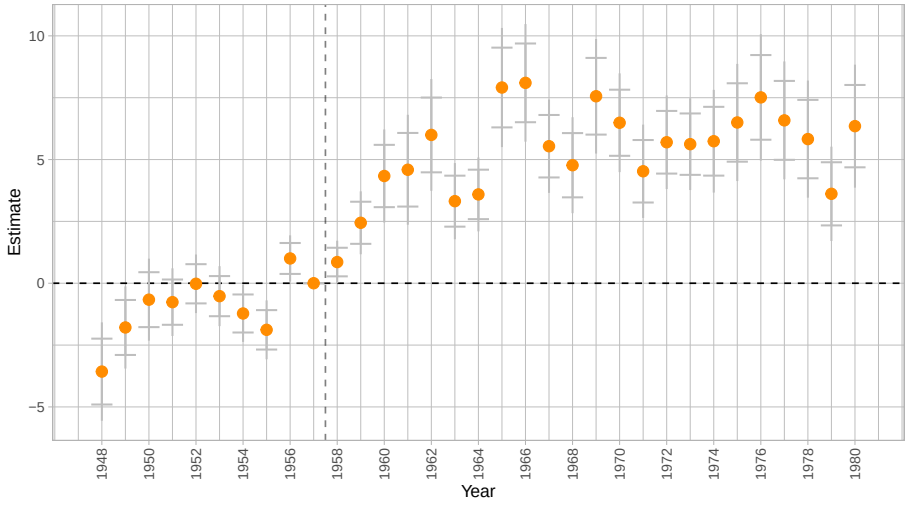
The first set of estimates uses all available technology subclasses that existed in the pre-treatment period.

Figure 1.52: Patent Issue DID Estimates, All Classes



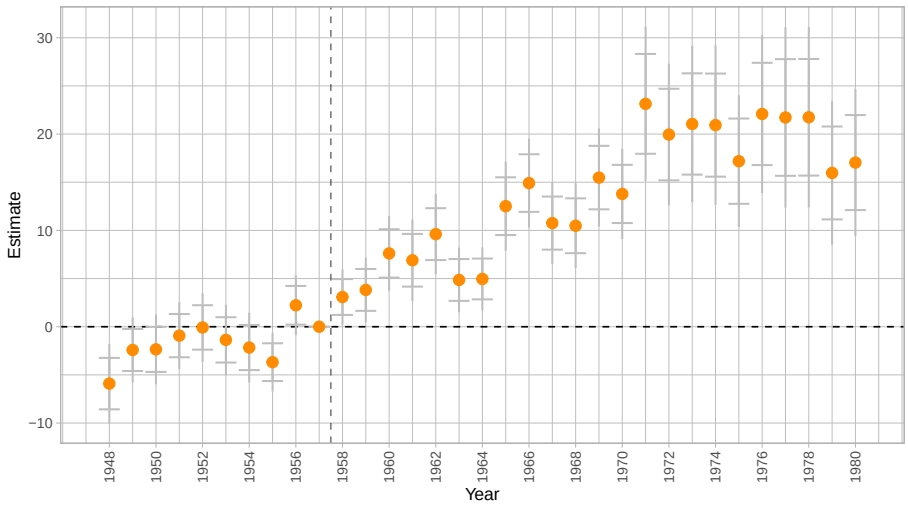
Note: S.E. clustered at subclass level, ⌶: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.53: Citations by Year DID Estimates, All Classes



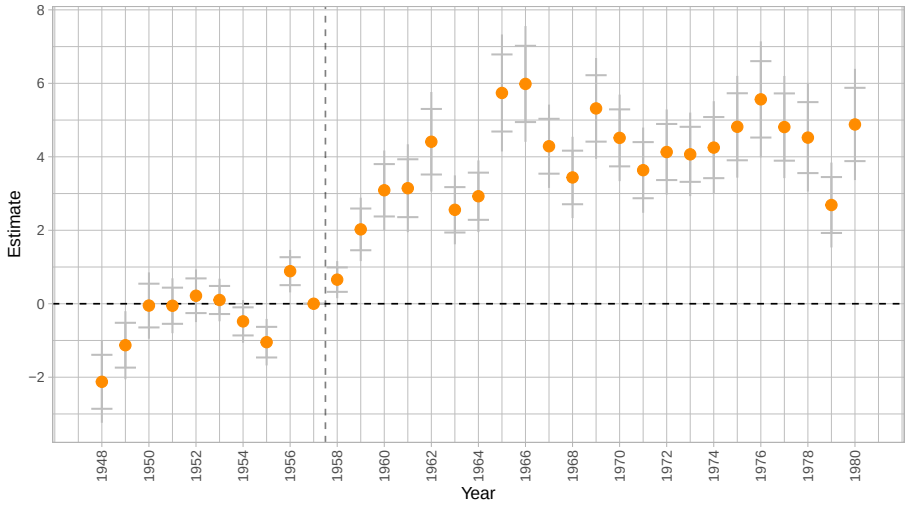
Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.54: Lifetime Citations DID Estimates, All Classes



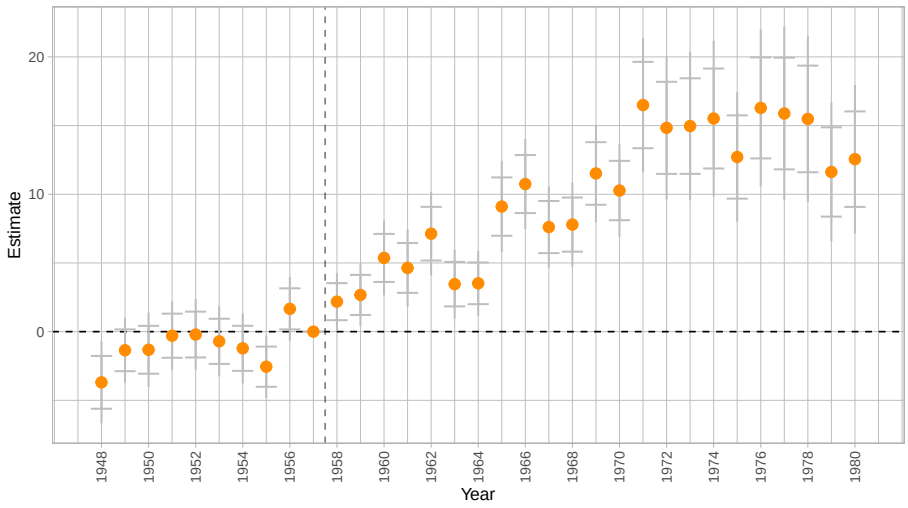
Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.55: Citations by Year (Leave-One-Out) DID Estimates, All Classes



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.56: Lifetime Citations (Leave-One-Out) DID Estimates, All Classes

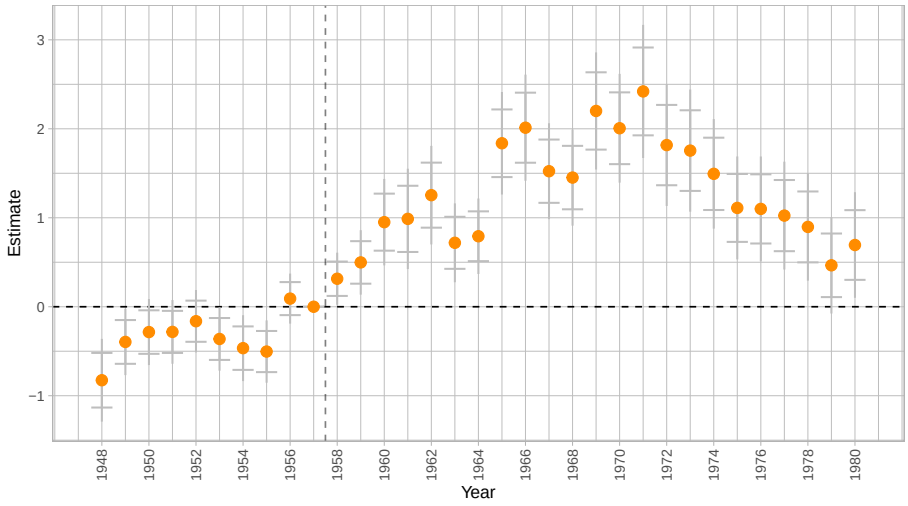


Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Treated Broad Classes

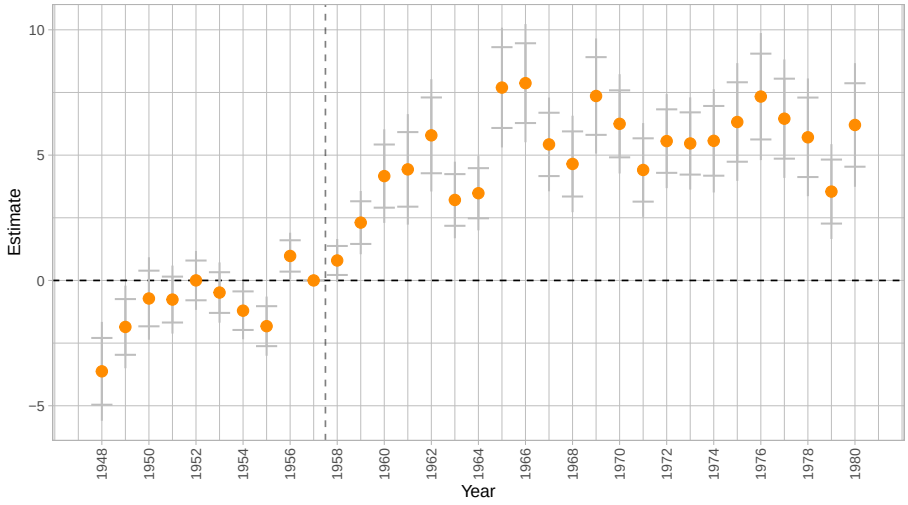
The following estimates use as controls the narrow classes that share a broad class with treated narrow classes.

Figure 1.57: Patent Issue DID Estimates, Same Broad Classes



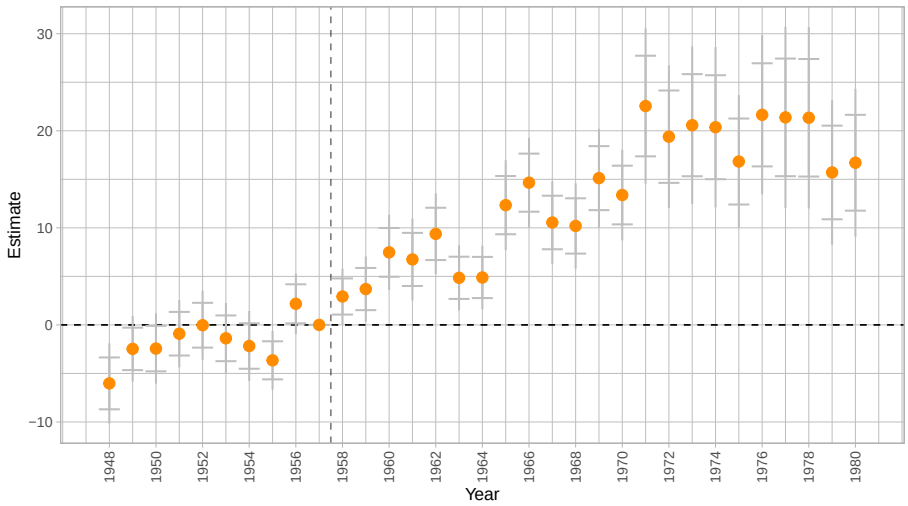
Note: S.E. clustered at subclass level, ⌶: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.58: Citations by Year DID Estimates, Same Broad Classes



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.59: Lifetime Citations DID Estimates, Same Broad Classes

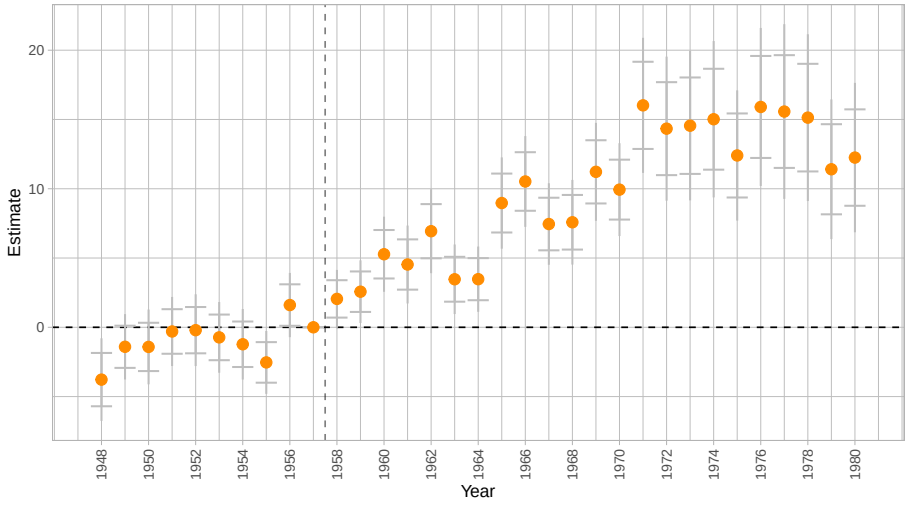


Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.60: Citations by Year (Leave-One-Out) DID Estimates, Same Broad Classes

Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

Figure 1.61: Lifetime Citations (Leave-One-Out) DID Estimates, Same Broad Classes



Note: S.E. clustered at subclass level, ⊥: point-wise 95% CI, |: sup-t 95% confidence band.

1.7.9 Linkage Procedure

This section details the procedure for matching NASA-affiliated inventors to their pre-NASA patents in Section 1.5.6. First, I clean and standardize all inventor names in the data. Then, for each NASA patent's inventor (the target inventor), I produce a set of closest candidates by taking their Jaro-Winkler string distance (Winkler, 1990) and keeping the closest 5 names in all patents or all candidates with a distance under 0.1, whichever is lesser. Because inventors can possibly migrate over time, I do not utilize the location information in the patent in this first step. This first step yields 8,315 target-candidate pairs. It is worth noting that in this step, I am also allowing for NASA to NASA matches, to account for the fact that the same inventor might have naming variations.

This approach to find near matches using names is aided by two features that are not present in other common linkage applications, such as Census of Population linkages (Feigenbaum, 2016, Abramitzky et al., 2021). Two main sources of digitization error are at the input stage, that the original document contains typos or spelling errors, and at the digitization stage, that the transcription is incorrectly carried out. Because names have to be produced in patent applications by the inventor or a third party acting on behalf of them, one could expect higher accuracy at the time the original document was generated, relative to a Census enumerator taking someone's name for the Decennial Census. Second, because documents are consistently typed and then passed through optical character recognition, even though machine recognition will produce errors, it will produce consistent transcriptions. That is to say, two humans might transcribe the exact same text differently, even though one is

correct. A program will always transcribe the same document the same way, even if it's incorrect. This second approach is more favorable for record linking.

Afterwards, I assess the match between each target inventor and their candidates. In the first pass, I only observe the name pairs to prune the list of possible candidates to those that have particularly similar names, resulting in 133 possible matches. Next, due to the lack of demographic information in the patents and the need to assess contextual information in each patent document, I manually study each scan for each inventor instead of relying on a probabilistic approach such as a supervised learning algorithm. As a non-random assessment of the optical character recognition in the CUSP dataset, I find that 66 of these potential matches where names are similar enough to look like matches but one appears to have a typo are in fact correctly transcribed as per the original patent document.

Due to the lack of identifying information, I rely on contextual clues in the patent scan such as employer and assignee information, proximity of target and candidate fields of invention, name consistency over time, and locations conditional on timing. Despite this, there is a subset of matches that have a close candidate but cannot be ruled as matches or non-matches with certainty. Specifically, there are 57 target-candidate pairs in this condition. However, none of these pairs changes the NASA or non-NASA status for all inventors' first patent in the data, thus the results are unchanged.

CHAPTER 2

LINKING HISTORIC PATENTS TO FIRMS: SUPERVISED LEARNING AND HANDLINKING APPROACHES

WITH RICHARD HORNBECK, ANDERS HUMLUM & MARTIN ROTEMBERG

This paper compares traditional rules-based data linkage approaches to a machine learning method in the context of linking patents issued between 1840 and 1900 to individual establishments in the 1870 Census of Manufactures. After applying the model to the Census at large and manually checking the set of predicted links, we find that we only agree with 57% of all positive predictions. Given the extreme class imbalance problem in this setting, where 3% of establishments are likely to hold a patent, we find that a simple rules-based approach combined with manual verification plausibly yields higher confidence links. We contribute a novel dataset for future researchers by performing this higher confidence linkage.

2.1 INTRODUCTION

A large segment of the innovation economics literature relies on patent data to measure the level of inventive activity in the economy (Williams & Bryan, 2021). Despite this, the economic value of these innovations is hard to empirically observe—patent forward citations, a commonly used measure, effectively measure scientific value, and not necessarily of economic value. To enable the study of these data issues, and to describe the relationship between innovation in the real economy and the innovation

observed in the market for patents, we propose the following novel data exercise. Using data on inventors, assignees, and location from Berkes (2018), we use a combination of handlinking and supervised learning models to link the universe of US Patent and Trademark Office (USPTO) historical patents from 1840-1900 to the recently digitized establishment-level manuscripts in the US Census of Manufactures for the years 1850-1870 (Hornbeck et al., 2023). While patents have been linked to individual inventors in the historical Census of Population to study inventors (Akcigit et al., 2017; Sarada et al., 2019), this data would allow us to study how those inventions were used and their relationship to real economic outcomes in firms.

To carry out this linkage, we pursue two distinct methods. First, we assess a machine learning procedure to generate probabilistic links. To generate training data, we handlink patents to around 2,800 unique establishments in the historical Census of Manufactures, where we consider all patents whose inventor or assignee resided in the same county as the establishment as potential matches. Then, using a supervised learning model, we train an algorithm to replicate this handlinking behavior consistently. Because full names can be ordered arbitrarily in both datasets, we compare two methods, one where names are reordered using the closest guess between two name pairs, and one where we use the historical Census of Population to order name tokens by their likelihood of being a first or last name. We find that the closest guess method provides a model that performs slightly better without having to use any additional data. We obtain a model that can detect 79.5% of all handlinked out of sample matches, and whose out of sample predictions are correct 90.6% of the time. Then, we apply this model and link the entirety of the 1870 Census

of Manufactures and all available patents between 1840-1900. The model matches 3,737 (1.99%) of establishments to some patent, and 7,882 (1.46%) of patents to some establishment. We handcheck a random sample of the predicted links and agree with 56.8% of them, which casts doubt on the model's ability to generate large-scale establishment-patent linked datasets.

Given this performance, we also apply a traditional rules-based approach relying on string distances and manual linking using the same variables to generate a final linked dataset for all 1850-1870 establishments to all 1840-1900 patents.

This paper contributes to recent work in economic history on record linkage (Feigenbaum, 2016, Abramitzky et al., 2021) by extending ideas first applied to linking individuals in the Census of Population towards the linkage of firms in the Census of Manufactures. It also relates to similar internal efforts at the U.S. Department of Commerce done on modern day administrative datasets that link the Longitudinal Business Database to USPTO patent data (Graham et al., 2018). In addition to the methodological contribution of assessing a supervised learning approach for probabilistic linkage in class imbalanced settings, this paper also contributes a novel data source that can serve as a valuable resource for future researchers.

2.2 DATA

Establishment-level data are drawn from the recently digitized manuscripts of the US Census of Manufactures (CMF) from 1850-1880 (Hornbeck et al., 2023). For linkage purposes, the main variables of interest from this data are establishment name, county, and nearest post office. Relevant production microdata such as self-

reported industry, types, quantities, and values of products and materials, and types of machinery and power used are also included. Given that for patenting purposes inventors are individuals and not firms (35 U.S.C. 100(f), 2021), a key feature of this data that makes any linkage possible is that the majority of establishment names either have a single owner's full name (77.56%) or the owner's last names (14.51%) (see Table 2.1[1]).

Table 2.1: Implied Ownership by Establishment Name, 1850-1870

Ownership Type	Percentage
Sole proprietorship	77.56%
Partnership	14.51%
Incorporated firm	7.93 %

USPTO data from 1840 to 1900 comes mainly from Berkes' (2018) Comprehensive Universe of US Patent Data (CUSP). Out of currently available historical US patent data, this is one of the most complete[2] and given the included variables, the most suitable to our purposes. CUSP includes for any given patent number the issue date, inventor and assignee's names[3], their city, and the patent's technology class[4]. To

1. That is not to say that corporate names don't have identifying information, as many will have last names just like partnerships do.

2. Andrews (2021) has an extensive review of historical patent datasets. Taking the USPTO Historical Patent Data Files (HPDF, Marco et al., 2015) as a comparison point, which are generated from internal USPTO records, CUSP covers about 99.4% of all patents from 1836 to 2016.

3. Patents always have inventors, who are individuals (35 U.S.C. 100(f), 2021), but they can also have assignees, which can be companies or other organizations, to whom ownership of the patent is transferred.

4. Patents always have a primary Original Class (OR) and can have secondary classes, known as Cross-Reference Classifications (XR). For our purposes, we are only using primary classes for now.

complement this dataset, NBER industry classifications (Hall et al., 2001) mapped from these technology classes are drawn from the USPTO Historical Patent Data Files (HPDF, Marco et al., 2015). Patents' locations are at the city level, but we will rely on counties to restrict links. Because county boundaries can change over time, each patent is assigned a year-specific county according to that year's county boundaries.[56]

Besides location, the linkage will rely the most on names, so some elaboration on the quality of these variables is worthwhile. Names in both the CMF and CUSP will contain transcription errors. However, because both datasets were constructed differently, these errors will be fairly independent of each other.

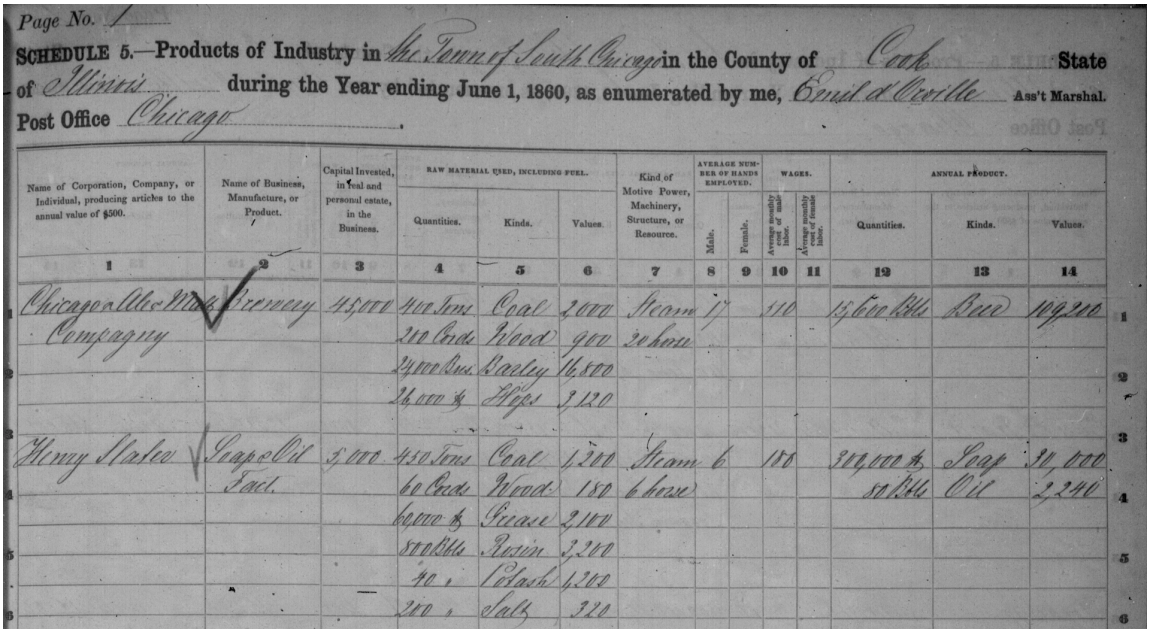
The CMF data was constructed by manually transcribing the information in the Census of Manufactures' manuscripts (see Figure 2.1). Because this information was originally handwritten in, errors might have been produced when establishments were surveyed. Additionally, transcription errors may vary between transcribers, even after double entry and discrepancy resolution. However, human transcribers are also able to discern idiosyncratic characters that would confuse a computer process such as OCR or models trained to detect handwriting. They will also be aware of whether a resulting transcription largely makes sense or not, whereas an algorithm

5. Specifically, the city centroid is placed on each Census year's county shapefile.

6. To clarify, whenever different years of CMF counties are analyzed, these changes are also considered by overlaying each year's county boundaries and calculating their overlap. Searches for candidate matches are expanded to include any counties that have an overlap larger than 1%. To illustrate, take two counties, A and B, and two years, 1 and 2. When linking establishments in county A from year 1 to 2, if county boundaries change such that by year 2 more than 1% of county A's original area lies in county B, then all establishments in county B are also considered in the candidate match pool.

can produce nonsensical transcriptions.

Figure 2.1: 1860 Census of Manufactures, Cook County



Page No. 1

SCHEDULE 5.—Products of Industry in the Town of South Chicago in the County of Cook State of Illinois during the Year ending June 1, 1860, as enumerated by me, Emil d'Orville Ass't Marshal.

Post Office Chicago.

Name of Corporation, Company, or Individual, producing articles to the annual value of $500.	Name of Business, Manufacture, or Product.	Capital Invested, in real and personal estate, in the Business.	RAW MATERIAL USED, INCLUDING FUEL.			Kind of Motive Power, Machinery, Structure, or Resource.	AVERAGE NUMBER OF HANDS EMPLOYED.		WAGES.		ANNUAL PRODUCT.		
			Quantities.	Kinds.	Values.		Male.	Female.	Average monthly cost of male labor.	Average monthly cost of female labor.	Quantities.	Kinds.	Values.
1	2	3	4	5	6	7	8	9	10	11	12	13	14
Chicago Alkali Manufacturing Company		45,000	400 Tons	Coal	2000	Steam	17		110		1500 Bbls	Beer	10,200
			300 Cords	Wood	900	30 horse							
			2400 Bus.	Barley	1800								
			26,000 lb	Hops	3120								
Henry Marleo	Soap & Oil	5,000	450 Tons	Coal	1200	Steam	6		150		300,000 lb	Soap	30,000
	Fact.		60 Cords	Wood	180	6 horse					80 Bbls	Oil	2,240
			60,000 lb	Grease	2100								
			800 Bbls	Resin	3200								
			40 "	Potash	1200								
			200 "	Salt	320								

On the other hand, the CUSP was produced using OCR and other machine-assisted tools on patent image scans which were originally typed (see Figure 2.2). In terms of input error, given that names have to be produced in patent applications by the inventor or a third party acting on behalf of them, one could expect higher accuracy at the time the original document was generated, relative to the CMF. At the transcription stage, machine scans will be more consistent than human transcribers, but will be prone to systematic errors in transcriptions, particularly when the source documents are not computer generated. For example, in the CUSP "r"s are commonly transcribed incorrectly with "e"s, "o"s are sometimes confused with "c"s, "H"s with "N"s, and vice versa. Other typos include numbers being included in names—a Henry Howard is transcribed as "Hi5Ney Howard". By searching over all

transcribed inventor and assignee names in the CUSP, we find that only 7,608 out of 517,017 (1.47%) names include numbers in them.

Figure 2.2: Patent No. 3,456

UNITED STATES PATENT OFFICE.

EZRA CORNELL, OF ITHACA, NEW YORK.

MACHINE FOR CUTTING TRENCHES AND LAYING PIPES.

Specification forming part of Letters Patent No. **3,456**, dated February 28, 1844.

To all whom it may concern:

Be it known that I, EZRA CORNELL, of Ithaca, in the county of Tompkins and State of New York, have invented a new and useful Machine or Implement for Laying Metallic Pipes in the Earth, which I denominate "Cornell's Improved Pipe-Layer;" and I do hereby declare that the following is a full and exact description of the construction and operation of the same, reference being had to the annexed drawings, making a part of this specification, in which—

Figure 1 is a perspective view, and the part thereon marked *a* is the beam, by which the implement is drawn and making a part thereof.

The part marked *b* is a cast or wrought iron furrow or trench cutter, the rear part of which is a hollow curvature, either cast with the main body of the cutter or formed separately of plate-iron and affixed to the main body by screws or bolts. Said curvature is also represented by the sectional drawings hereto annexed and made part of this specification at

arms or stops, which are marked *g* in said drawings, are turned to a position that is perpendicular, or nearly so, to the curved surface of the drum, and thereby confine the coil of pipe upon the drum.

h represents the wheels, which may be used to steady the machine and regulate the depth at which the pipe is deposited in the earth.

i represents the braces by which the wheels are attached to the beam of the machine, and which are so contrived as to raise or lower the wheels, and thereby regulate the depth at which the pipe is deposited.

j represents a section of the pipe passing from the drum through the hollow curvature in the beam and back part of the trench-cutter to the earth.

The motion of the machine, when drawn forward with the pipe once confined in the ground, draws the pipe from said drum or reel through the before-described aperture in the beam, thence through the hollow curvature above described, from the interior surface of which

Errors in the source documents and either form of transcription error will degrade the ability to produce links (manual or automated), and because these two types of transcriptions and source documents produce somewhat independent errors, a learner will be less able to relate one type of error structure to the other. An example of this is a foundry in 1870 Jefferson County, Iowa, belonging to "Demarce Antheyna" in the

CMF. Inspecting the actual Census image from which this was transcribed, the last name, "Demarce", is correctly transcribed, but the first name isn't completely clear (see Figure 2.3). In the patent data, we observe Patent No. 90,083 issued in 1869 Fairfield, Jefferson County, IA invented by an "Anthony Demab5B". By looking at the actual patent image (see Figure 2.4), we can see that the inventor's actual name is Anthony Demarce. The CMF handwritten name could be Anthony, but there are extraneous characters (the "na"). Looking at the 1870 Census of Population and the 1864 Fairfield City Military Register (Ancestry, n.d.), we can see that in 1870 Fairfield there was a 41 year-old foundryman named A. Demarce, and six years earlier in 1864 Fairfield a 35 year-old Anthony Demarce registered with the military. All of this information together points towards a match between the patent and the establishment.

Figure 2.3: 1870 CMF, Anthony Demarce's Foundry

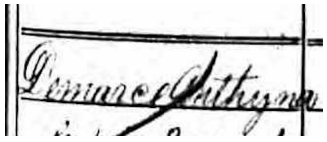


Figure 2.4: Anthony Demarce's Patent No. 90,083

ANTHONY DEMARCE, OF FAIRFIELD, IOWA.

Letters Patent No. 90,083, *dated May* 18, 1869.

IMPROVEMENT IN STEAM-GENERATORS.

Having access to both source documents and other historical records makes it fairly feasible to establish a match between the establishment and the patent, but

one can see how both types of transcription error make it much harder to conclude this exclusively from the digitized CMF and CUSP records. Because ground truth names are only available by manually consulting the source documents, the success of the linkage exercise largely relies on these errors being kept to a minimum.

2.3 LINKAGE STRATEGY

Record linkage algorithms, whether automated or manual, can all be conceptualized as solving a prediction problem. Given a pair of observations from two datasets, do the observations belong to the same establishment or not? To formalize, a pair of observations can be a match M, or a non-match U. Given a common set of characteristics in both datasets, such as name and location, define a comparison vector γ of distance measures in these common features. The goal of the algorithm is to then estimate the conditional probability of a match or non-match, given this comparison vector, $Pr(M \mid \gamma)$ (Fellegi & Sunter, 1969). Afterwards, a decision rule such as a probability cutoff can be implemented to determine the final linked dataset, or the probabilities can be used directly, as in a multiple imputation procedure.

In our application, the probabilities are estimated via supervised learning, such as in Feigenbaum (2016) and Helgertz et al. (2021). This section details the linkage strategy from start to end.

2.3.1 Handlinking

First, a subset of counties from the CMF data is selected to generate labeled data. In this case, 48 counties were randomly drawn from the subset of all counties that had

over 95% completeness in industry and name strings in the 1860 Census of Manufactures. For each of these counties, a set of candidate patents was generated by taking all patents issued up to that year and either invented or assigned to that county. Because we are not only interested in within county invention, blocking both on inventor and assignee county allows us to consider cases where the invention is generated in one county but utilized in another. Candidate patents to match are blocked on county because the number of patent to establishment comparisons quickly becomes infeasible. For example, there are 87,566 unique patents from 1840 to 1870[7], and 186,667 establishments in the 1870 Census. This would result in 16,345,682,522 comparisons to make for that single year—even under a fully automated method, this would be computationally prohibitive.

Of these 48 counties, 14 had no patents issued or assigned to them in any year up to the Census year. The remaining 34 were augmented with 3 counties that had high patent to establishment ratios in the remaining data. So the counties in this handlinked sample are not truly randomly drawn from the overall Census of Manufactures. Patents were handlinked to the establishments in these counties in 1860 and in 1870.[8]

Pooling across years, this handlinked sample consists of N = 2,755 establishments and P = 1,636 unique patents. These represent approximately 0.410% of all available establishments between 1850 and 1880[9], without considering duplicates over time,

7. The linkage does not account for patent expirations. Generating a patent indicates underlying innovative capacity, thus ever patentees (even if expired) are plausibly different from never patentees.

8. One of these additional counties, Clark County, Ohio, is missing from the 1860 data.

9. Counts for 1880 are still being determined in the microdata, but the 1880 Census reports 253,852 establishments in the entire US, which were added to 417,920 establishments from 1850 to

and 0.303% of all available unique patents from 1840 to 1900. Because the range of possible establishment-patent links is blocked on county, the total size of the available training data is not N $\times$ P. Accounting for this blocking, the resulting handlinked dataset contains 324,514 observations. The handlinks are pooled across years because for purposes of the learner, matching criteria to link patents to establishments are time invariant. However, once the model is trained, this data would also allow one to create an establishment panel with patent information.

When handlinking within county, the first criterion to decide if an establishment-patent pair is a match is name similarity between the establishment and either the inventor or the assignee. When name transcriptions were ambiguous in either dataset, the original Census manuscript images and USPTO patent document scans were consulted. The reasoning behind relying on this auxiliary data that the learner won't directly see is to have it learn underlying patterns in transcription errors whenever possible. A secondary feature used when linking was granular location. The most specific location variable in the patent data are inventor's and assignee's cities. This variable is not available in the digitized manufacturing data[10]. However, a more specific feature, nearest post office to an establishment, is recorded. In some cases, post office names and cities will match, e.g., refer to Figure 2.1 to see that the post office for Chicago is also "Chicago". To the degree that this holds more generally, we are using post office in the CMF as a proxy for city, but we do not expect this to always be true. For handlinking purposes, matching post offices and patent cities

1870 counted from the microdata for the denominator. Duplicates in the microdata are actively being fixed so these numbers are subject to change.

10. It is available in the original Census manuscript, but has not been digitized.

were considered to increase the likelihood of a match, but mismatching ones were not considered to decrease it.

Technology classes for patents and their NBER industry classifications are available along with industry classifications in the manufacturing data. However, these weren't utilized in the handlinking procedure. This is for several reasons. First, because even if establishment owners generate or buy innovations that don't directly apply to their business, they might be fundamentally different from owners that do, and their establishments might be different as well. To the degree this is true, we would like to document those matches. Second, in the process of assembling the data we observed that there is no clear mapping between patent technology classes and industries as classified in the manufacturing data. The NBER classifications were constructed with modern day industries in mind, and so do not always provide the correct industry classification. An example of this is the first photographic patent in the United States, a daguerreotype camera (Patent No. 1,582) invented by Alexander S. Wolcott and issued in 1840, a year after Louis Daguerre's original invention was made public. Under the NBER mapping, this camera is classified under the industry "Computer hardware and software". 1,277 other patents from 1840 to 1900 are classified under this same category. Finally, to the degree that technology class and industry are variables to be used in further analysis, using their comparison between both datasets to link explicitly could present endogeneity issues.

Once the entirety of this sample was handlinked, the data was then split into train, validation, and test sets. To keep results representative of unseen data, all development and testing is done using the training and validation sets, and the

test is only used to generate the final results presented in this paper. To prevent information within one establishment influencing other predictions it is a part of, the train, validation, and test split is made such that a given establishment-year appears only in train, validation, or test.

2.3.2 Supervised Learning Model

In order to choose a supervised learning algorithm, we'll evaluate the performance of logistic regression and random forests (Breiman, 2001) and choose the one with the best out of sample performance.

Comparison Features

Several cleaning steps need to be taken in order to compare names between the CMF and the CUSP. First, some names in the Census of Manufactures were entered last name first,[11] particularly in 1870, while inventor and assignee names are generally ordered first name first. More generally, both datasets contain full name strings, i.e., first and last names are not separated in the data. There are two potential ways to correct for this, and the performance of both approaches will be assessed. The first is to use the full strings and attempt to correct the name orderings in the CMF using the patent data. One can create a reordered version of an establishment's name and compare the original and reordered names to the target patent name and keep the closest. The second alternative is to rely on additional data from the Census of Population (Ruggles et al., 2021) where the names have already been split into first

11. Names that imply partnerships or incorporated firms are excluded from this cleaning step.

and last names. Using name frequencies from this data, one can predict whether any given token in an establishment is a first or last name and order accordingly. This also allows direct comparisons between specific parts of names when possible. For example, one might like to only use last names when comparing names that imply partnerships or corporations. Specifically, the algorithm is as follows:

1. Get all first and last name frequencies from the 1880 full population Census.[12]

2. For each full name string in the CMF, split it into tokens.

3. If a token appears only in the empirical distribution of first names, classify it as a first name, and vice-versa for last names.

4. If a token appears in both distributions, assign it to the name type it represents a higher share of. For example, "John" represents 3.27% of all first names and 0.01% of all last names, so it is classified as a first name.

5. If the token isn't in either name distribution, take the string distance between the token and all names in both tables. Assign it to the type it has a closest string to.

6. Once each token has been assigned a name type individually, check if the classification for the full name agrees, i.e., there is at least one first name token and one last name token.

12. Restricted use versions with full names for 1850-1940 (minus 1890, which has been lost) are only available with special access through the IPUMS or the NBER, however, 1850 and 1880 full-count Censuses with names are publicly available at IPUMS International. To refine the algorithm, we could tailor the name splitting for a given year using its specific restricted use Census, but name distributions wouldn't likely vary much in the span of a decade or two.

7. If the individual classifications disagree, i.e., they're all predicted to be first names (e.g., "James" and "Alexander") or last names (e.g., "Hyatt" and "Jackson"), find which token is likeliest to be a last name and assign the rest to the first name (resulting in "James Alexander" and "Jackson Hyatt"). [13]

Second, because patents can be linked either through inventors or through assignees, one needs to choose a target name to compare the establishment name against. A 2-to-1 comparison that transforms both distances into a scalar value is not ideal for two reasons. First, only 17.73% of the handlinked patents (23.16% in all 1840-1900 patents) have assignees. Second, using both names could also confuse a learner given that usually only one name is expected to match. For example, if the establishment name is "Terence Chau" and there is a patent invented by "Anders Humlum" and assigned to "Terence Chau", the model would observe one name that matches exactly and one that is completely different. Any transformation other than the minimum of these two distances into one scalar value dilutes the signal we'd like the learner to pick up—that there is an exact match. Therefore, to select a target name, we take the Jaro-Winkler distance between inventor and assignee names to the establishment name, and choose the closest. For consistency, we use the city that corresponds to the selected target name as well.

Unless one uses the Census of Population data in the first step, these two cleaning steps appear to be at odds with each other, because choosing a target patent name out of the two changes the distance between the potential establishment name orderings,

13. This assumes that it is likelier to have a single token last name and several first or middle names than the other way around.

which can change the target patent name, and so forth. To deal with this, we can take the four distances:

1. Establishment name, original ordering - inventor name.
2. Establishment name, inverted ordering - inventor name.
3. Establishment name, original ordering - assignee name.
4. Establishment name, inverted ordering - assignee name.

Then, take the names that correspond to the minimum distance of the four as our target establishment and patent names. Once target names are defined in both datasets, we use various distance measures to assess name similarity between them. First, following common practice in the record linkage literature in economics (Feigenbaum, 2016; Abramitzky et al., 2021; Helgertz et al., 2021), we utilize Jaro-Winkler string distance (Winkler, 1990). This distance is found by weighing the fraction of matching characters in two strings and the number of transpositions needed for these matching characters to be in the same order, with emphasis placed on the beginning of the string.

When using full name strings where first and last names haven't been identified, we potentially lose valuable information to link on due to a couple of reasons. First, longer string comparisons are penalized because the probability of two orderings of characters matching becomes lower as the strings become longer. Second, some comparisons hinge on matching last names but not matching full strings. For example, a company name such as the "Bell Telephone Company" has little string similarity with "Alexander Graham Bell" (JW = 0.506), but if one knew that Alexander's

last name is "Bell" and it matches a company name's token exactly, this should increase the probability of a match. To ameliorate these issues, we use two additional Jaro-Winkler based measures when using full name strings.

We will refer to the first measure, which deals with the fact that we are comparing long strings, as the mean Jaro-Winkler. It takes the distance between each corresponding token in both strings and then takes their average. For example, the straight Jaro-Winkler distance between "Margaret Hamilton" and "Margrt Hamiton" is 0.071, while taking the distance between "Margaret" and "Margrt", and "Hamilton" and "Hamiton" and then averaging is 0.038.[14]

The second feature, the minimum Jaro-Winkler, takes the distance between all tokens in two strings, regardless of ordering, and keeps the minimum. In the Hamilton case, this would give a minimum of 0.041, while in the Bell case this gives a minimum distance of zero. When name types have been identified using the Census of Population, we take the Jaro-Winkler between a patent name's last name and each last name identified in the establishment name, and keep the minimum.

Another set of comparison features we generate from the names in both datasets are comparisons of their Soundex (Russell, 1918; Russell, 1922) phonetic encodings. There is no straightforward way of comparing these codes—Soundex encodings are composed of the initial letter and three numbers denoting the remaining sounds (e.g., "Kodak" has a Soundex of "K320")—, so the comparison feature we generate is whether the codes match or not. We generate a phonetic code for each token, and

14. This is not a result of Jaro-Winkler distance's emphasis on string beginnings. The Jaro distances (which place no extra weight on the beginning of a string) in this case would be 0.118 and 0.063, respectively.

calculate the share of agreements between both sets of codes.

Finally, we generate other comparison features. In order to utilize the available location data, we also compute equivalent string and phonetic similarity features between an establishment's nearest post office and the city reported in the patent. Because 3% of comparison pairs are missing locations (N = 9,731), these cases are mean imputed and a missing value flag is added. Additional variables used are USPC patent class fixed effects, a dummy for whether an establishment's name implies a company-style name, i.e., anything but sole proprietorship, and the amount of years between the patent's origination and the establishment's Census year.

Linkage Performance

In order to assess the performance of the model, we will calculate various statistics based off of the test set confusion matrix such as the F-score and Cohen's κ. Given the total number of observations (O), reference positives[15] (P), reference negatives (N), along with the values generated from a confusion matrix—true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN)—, define:

- True positive and negative rates (TPR, TNR): the share of positives and negatives correctly detected, TP/P and TN/N. TPR is often also referred to as recall or sensitivity, and TNR is also known as specificity.

- Positive predictive value: the share of correctly predicted positives out of all positive predictions, $TP/(TP + FP)$. Also known as precision.

15. For our purposes, links or matches are our positive predicted class.

- F-score: The harmonic mean of the TPR and the PPV,

$$F = \frac{2 \times TPR \times PPV}{TPR + PPV}$$

- Cohen's κ: The share of agreements between the handlabel and the model, normalized by the empirically observed probability of them agreeing by chance. Let:

$$\begin{aligned} P(\text{Observed Agreement}) = P_o &= \frac{TP + TN}{O} \\ P(\text{Random Positive Agreement}) = P_p &= \frac{TP + FP}{O} \times \frac{TP + FN}{O} \\ P(\text{Random Negative Agreement}) = P_n &= \frac{TN + FN}{O} \times \frac{TN + FP}{O} \\ P(\text{Random Agreement}) = P_r &= P_p + P_n \end{aligned}$$

  Then, Cohen's κ is:

$$\kappa = \frac{P_o - P_r}{1 - P_r}$$

These measures were chosen over simple classification accuracy because we expect matches to be rare, which produces an issue of class imbalance. If the evaluation metrics don't incorporate this feature of the data, our evaluation metrics will be exceedingly optimistic. For example, if we expect 95% of comparisons to be non-matches, a model that only predicts non matches will have a classification accuracy of 95% without having detected any matches. The F-score's emphasis on detecting the

positive class with few errors, and κ's emphasis on controlling for random agreements given the empirical distribution of labels allow us to get a better evaluation of the model's predictions.

Applying the Model

Once the preferred model and specification have been chosen, the model is then retrained on all handlabeled data, candidate patents are drawn for all remaining counties, and the model is applied to the entire Census of Manufactures. Finally, instead of taking the model links as given, we handcheck the positive predicted links to confirm the final set of matches.

2.4 LINKAGE RESULTS

2.4.1 Handlink Statistics

Overall, the number of matched establishments is 69 out of 2755, or 2.505%. As for patents, 186 out of 1636, or 11.369% were matched. Grouping by year, the number of matched 1860 establishments was 13 out of 811, or 1.603%. Also, 31 out of 408 (7.598%) patents issued up to 1860 were matched to one of these establishments. The number of matched 1870 establishments was 56 out of 1944, or 2.881%. 162 out of 1635 (9.908%) of patents issued up to 1870 were matched to one of these establishments.

At the county level (Table 2.2), matches are highly concentrated. Out of 37 counties, 26 have no matches at all. Of the remaining 11, 8 have less than 10

matches, and the remaining 3 counties contain the vast majority of links: 13, 99, and 348 matches, respectively. County size, number of patents, and number of matches are highly correlated. On average, counties with no handlinks have 42.35 establishments and 10.15 patents, while counties with links on average have 150.4 establishments and 124.9 patents. For the counties with more than ten handlinks, average establishment and patent counts are 223.3 and 416.7.

Table 2.2: Number of Establishment-Patent Matches, by County

Number of Matches	Frequency
0	26
1	3
2	1
4	2
6	1
9	1
13	1
99	1
348	1

At the establishment level we can observe a similar pattern where the vast majority of establishments have no linked patents—2,687 out of 2,755 (97.5%) have zero matches, while 10 establishments have over 25 patents linked to them.

The supervised model's performance will depend on the amount of class separation that the comparison variables induce in the data. Therefore, we also present summary statistics on some of the key comparison features within the handlinked sample at the comparison level.

When using the reordered names without Census of Population data, the distri-

Table 2.3: Number of Establishment-Patent Matches, by Establishment

Number of Matches	Frequency
0	2687
1	24
2	17
3	5
4	2
5	3
7	2
8	3
9	2
25	1
26	1
27	2
29	1
32	3
45	2

butions of Jaro-Winkler distances for matches and non-matches can be seen in Figure 2.5. As expected, handlinked matches have names that are closer on average (JW = 0.285) than non-matches (JW = 0.488). The range of the string distances in the match distribution is also lower, with a max of 0.572, while non-matches are spread between zero and one. When looking at the modified string distances, the variable that manages to induce the most separation with only the available data is the minimum Jaro-Winkler (Figure 2.6)—it has a value of zero up to the 80th percentile and a mean of 0.038 within matches, while it has a median and mean of 0.482 and 0.521 among non-matches. It also increases the values at the maximum distance of 1 among non-matches. When looking at how many company-style names there are in both groups we can see the reason why the minimum JW seems to separate the

classes well: 90.25% of links have establishments with company names, while only 36.83% of non-matches have establishments with company-style names.

Figure 2.5: Jaro-Winkler Distances of Names

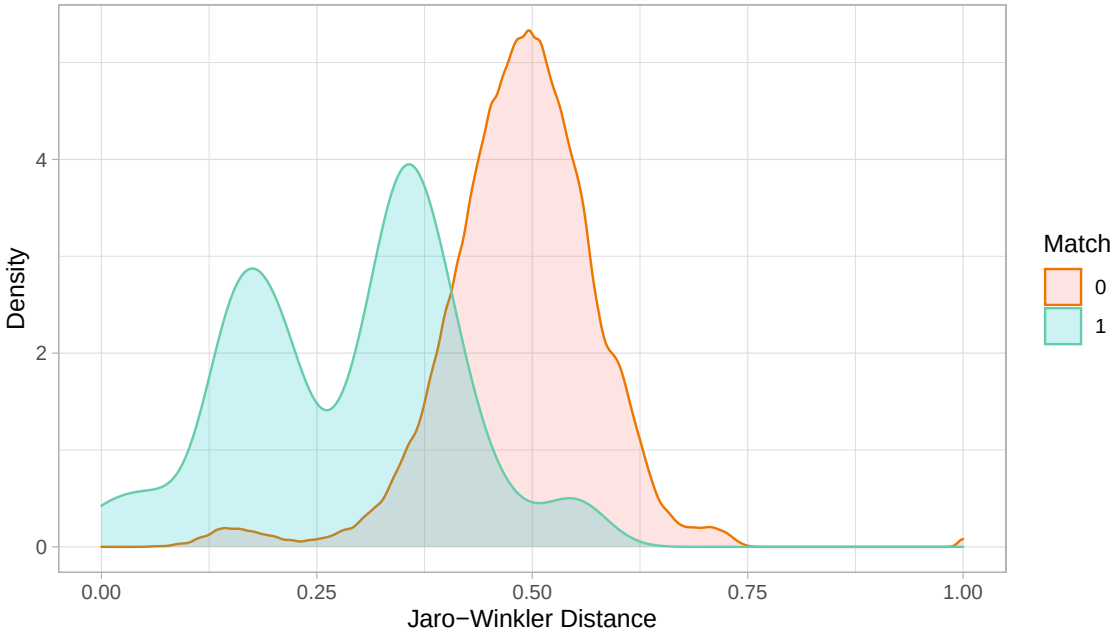
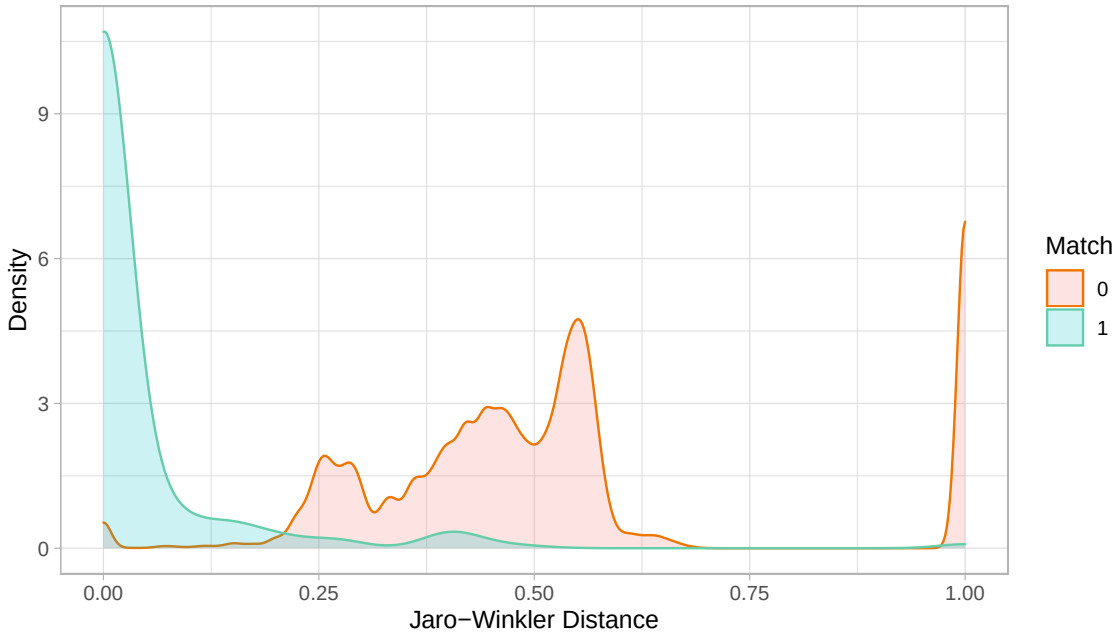
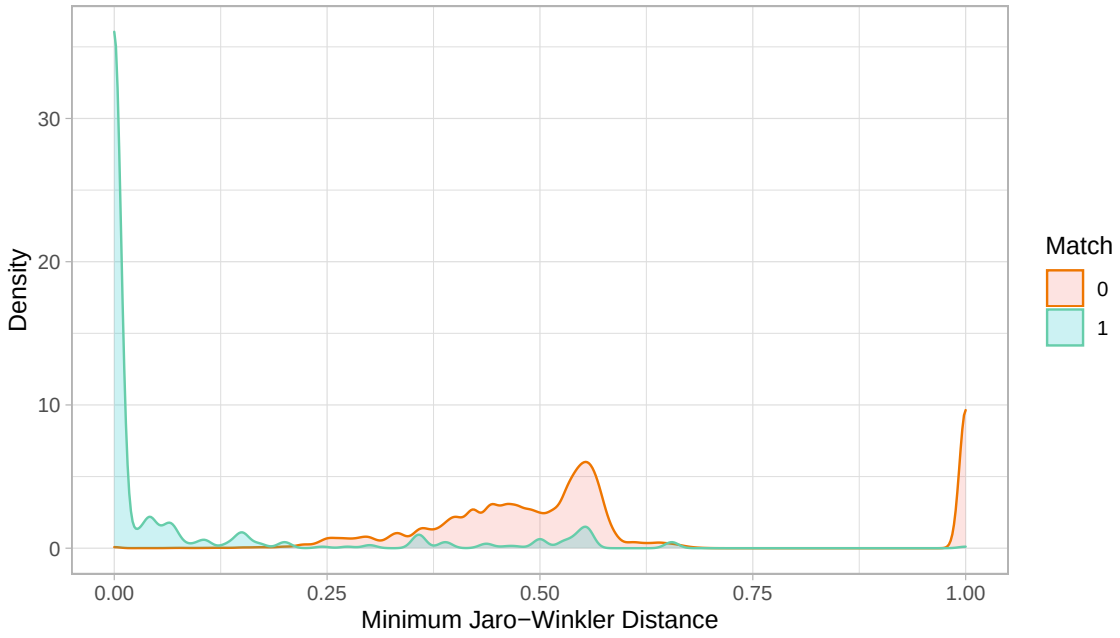


Figure 2.6: Minimum Jaro-Winkler Distances of Names



When using the Census of Population data, we find that the minimum Jaro-Winkler distance of only the inventor's last name to all establishment name tokens (as opposed to all inventor name tokens to all establishment name tokens in Figure 2.7) induces even more separation between matches and non-matches.

Figure 2.7: Minimum Jaro-Winkler Distances of Names, Using Census Names



In general, final name orderings do vary between both approaches. Comparing the original orderings of Census names to the lowest distance reordering and the population census based reordering shows that 66.33% of names maintain their original CMF ordering when using the closest patent name to select a target establishment name, while 61.26% agree when reordering names using the Census data. The final names to generate comparison features agree 65.42% of the times. Both approaches also select different target patent names, since 89.44% of them match between both approaches. These different ordering and target name choices produce different sets of comparison features which are correlated but not perfectly so (see Table 2.4).

Table 2.4: Correlation Between Features, With and Without Census of Population Data

Variable	Correlation
Name JW	0.767
Minimum JW	0.517
Mean JW	0.777
Location JW	0.979
Mean Location JW	0.985
Year Gap	1.000

String distances for locations also appear to induce separation between matches and non-matches (see Figure 2.8). We can see that the mean string distance between CMF post offices and patent cities among positive links is 0.183 and 0.460 among non-links. The average JW appears to produce slightly more distinct distributions between both groups (Figure 2.9), where the mean string distance between location names is 0.113 among matches and 0.466 among non-matches. Finally, the year gap between patent origination and the year the establishment is observed is similar between both groups (Figure 2.10).

Figure 2.8: Jaro-Winkler Distances of Post Offices and Cities

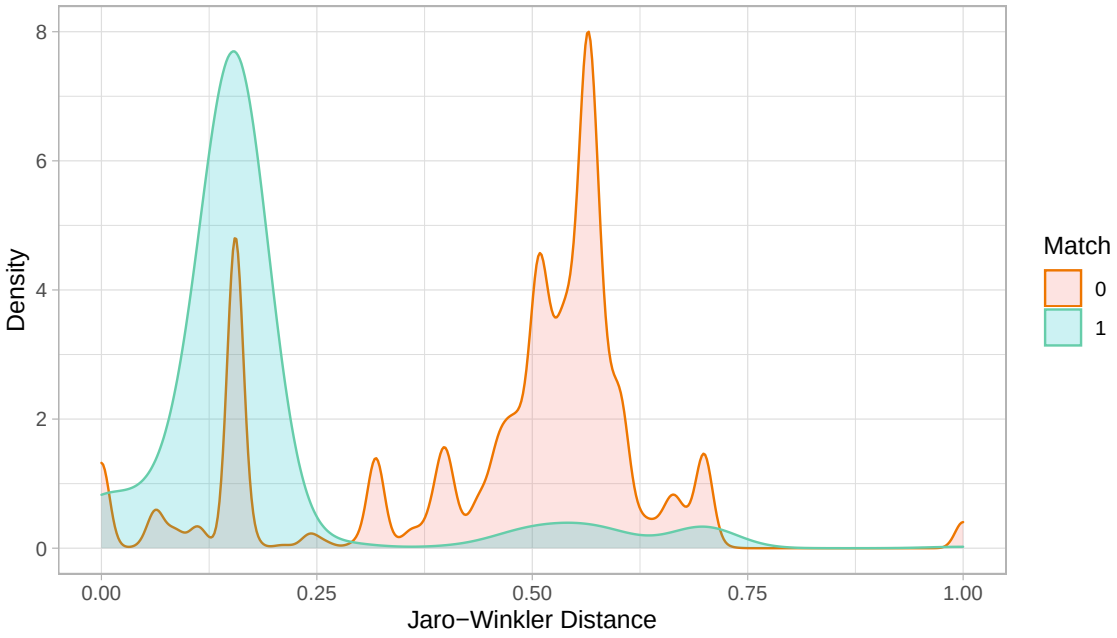


Figure 2.9: Mean Jaro-Winkler Distances Post Offices and Cities

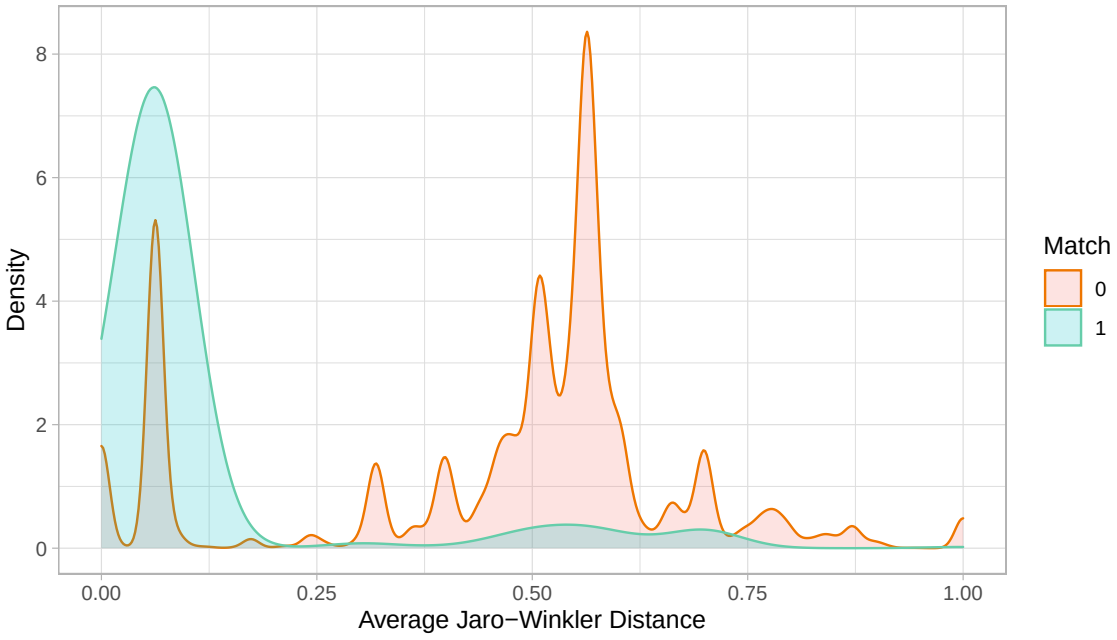
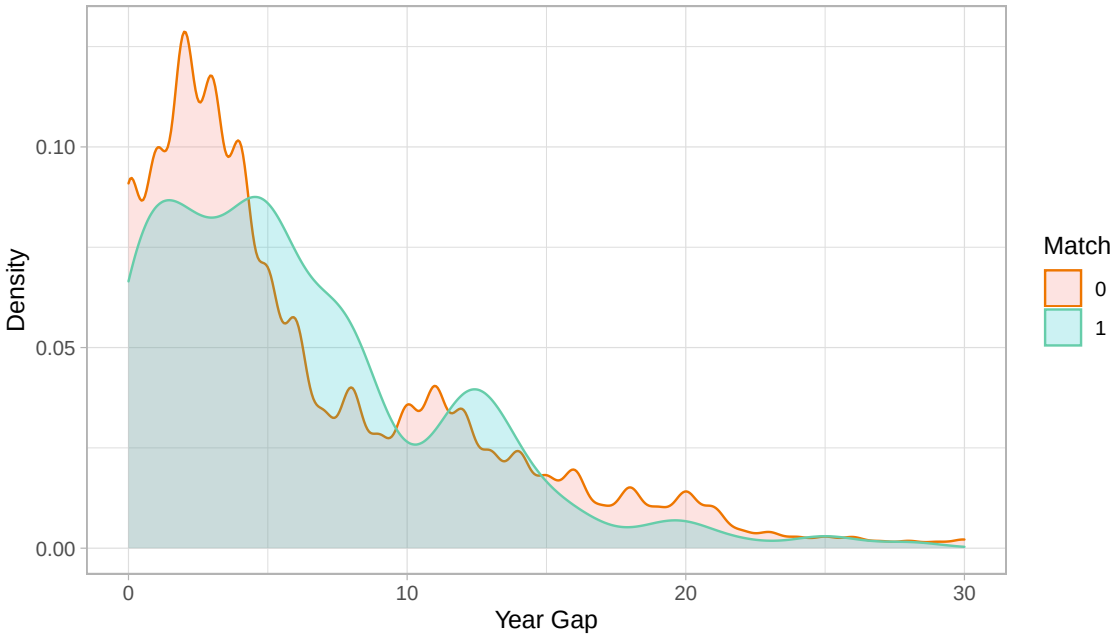
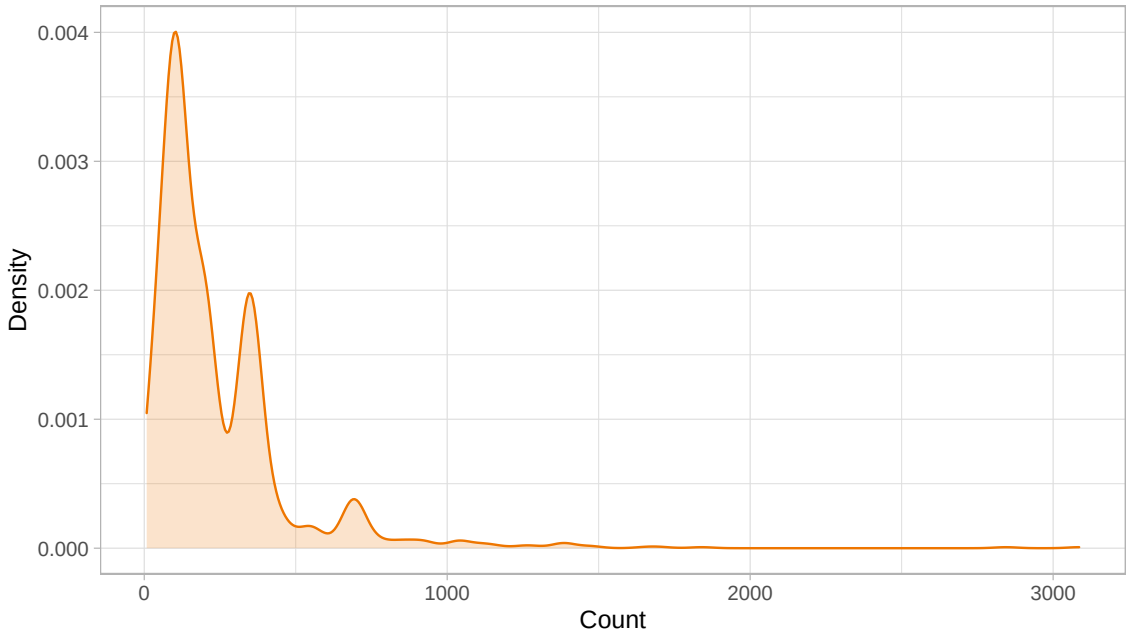


Figure 2.10: Year Gap Between Patent and Establishment Census Year



Looking at non-comparison features, we can see that principal US Patent Classification classes are unevenly distributed in the data (Figure 2.11). The median class has 184 patents and the mean class has 241.6 patents. However, some classes have up to 3,086 patents. The top ten classes are shown in Appendix Table 2.13.

Figure 2.11: Distribution of USPC Class Counts



When we subset the data to matches and non-matches, we find that only 135 classes out of 1,343 have at least one match. Out of these, 128 have between 1 and 10 matches, and 7 have more than 10, with two classes having 24 links. Appendix Table 2.14 shows the ten classes with the most handlinks. We can observe that 5 of the patent classes with most matches appear in the overall top 10 list, implying some but not perfect correlation between patent class frequency and match probability. As expected, commonly linked patent classes are all related to manufacturing in some way—machinery, agricultural tools, and weaponry.

Table 2.5 shows the count of links by broad industry in the handlinked sample, sorted by the number of matches. The majority of handlinks occur in establishments related to iron and steel products or production, which aligns with the most often

Table 2.5: Matches by Broad Industry, Handlinked Sample

Industry Broad	Match Count
Iron and Steel Products	195
Iron and Steel	65
Cooperage	29
Machinery and Fine Instruments	17
Flour and Grist Mills	6
Lumber	6
Carriages and Wagons	5
Boots and Shoes	3
Furniture	3
Liquors and Beverages	3
Yarn, Cloth, and Other Textiles	3
Construction	2
Leather Products	2
Construction Materials	1
Food Products	1

The broad industries in the sample that received no matches were: brass and other metal products, bread and bakery products, butter and cheese, chemicals, clothing, fisheries, jewelry, pottery and decorative work, leather, mining and quarrying, other consumer products, other non-manufacturing, paper, printing and publishing, ship and boat building, tin, copper, and sheet-iron ware, tobacco, and wood products.

linked patents being related to the machine tool industry.

2.4.2 Model Performance

Table 2.6 shows the results for the best logit and random forest specifications using full name strings. Additionally, the performance of the model using the 1880 Census of Population to classify name tokens is shown. Both full name models achieve near perfect performance on predicting non-links—they can detect 99.96% and 99.99% of negative labels in the test set respectively—, as would be expected from the extreme

level of class imbalance. However, when it comes to the prediction we care about, the random forest produces much better performance. Where a logit detects 60.3% of matches, the random forest has a true positive rate of 79.5%. Further, out of all positive predictions they make, the logit is correct 75.9% of the times, while the random forest has a PPV of 90.6%. This overall better performance is also captured in the F-score and κ.

When using the population census to pre-process the names (third row of Table 2.6), we find that the model is able to detect more handlinked matches (TPR = 0.836) but makes more incorrect guesses (PPV = 0.782). Since the F-score weighs both aspects equally, the baseline random forest specification using full name strings scores higher out of sample. It would appear that the minimum JW achieves much of the performance gain that using the auxiliary Census of Population data produces without producing an extra classifier, and we will use this random forest as our preferred specification.

Table 2.6: Test Set Performance, by Model

Model	TPR	TNR	PPV	F-score	κ
Logistic Regression	0.603	0.9996	0.759	0.672	0.671
Random Forest	0.795	0.9999	0.906	0.847	0.846
Random Forest + Census Names	0.836	0.9999	0.782	0.808	0.808

Tables 2.7 and 2.8 show the test set confusion matrix for both random forests.

Table 2.7: Random Forest Confusion Matrix

	Handlabel	
Prediction	Non-Match	Match
Non-Match	39987	16
Match	6	58

The main specification obtained an F-score of 0.847 and a κ value of 0.846. For additional comparison, we can derive the expected test set performance of a naive classifier that guesses at random based off of the training set's distribution of matches, i.e., a heavily biased coin flip in our setting. Given that the training and validation sets had 399 matches and 284,049 non-matches, and the test set had 73 matches and 39,993 non-matches, we could expect a TPR of 0.00140 and a PPV of 0.00182[16]. This would result in an F-score of 0.00159. Because the F-score emphasizes positive predictions, a classifier that only predicts matches leads to a slightly higher score of 0.00363. This means that by this metric the random forest does over 500 times better than an informed guess, over 200 times better than guessing that everything is a match, and 1.26 times better than logistic regression.

A useful feature from random forests is permutation variable importance. When growing a classification tree, one can permute all values in a variable and assess the delta in out-of-bag performance from the original and permuted versions. One can

16. To obtain these values, define the training set share of matches as TR and the corresponding test set share as TE. Note that over N many draws, the expected number of true positives will be $N \times TR \times TE$, the expected number of false positives will be $N \times (1 - TR) \times TE$, and so forth for each cell in the confusion matrix. Applying the definitions in section 2.3.2 implies that the expected TPR is the share of training set positive labels TR, and the expected PPV will be the share of test set positive labels TE.

Table 2.8: Random Forest (Using Census Names) Confusion Matrix

	Handlabel	
Prediction	Non-Match	Match
Non-Match	39976	12
Match	17	61

then average over all tree-level importances in the forest to proxy for the importance of a particular feature (Breiman, 2001). This method is only partially useful because it doesn't contemplate correlations between variables, however[17]. Table 2.9 shows all the variables in the final specification, along with their importances sorted in descending order. Aligning with the figures in the previous section, minimum name Jaro-Winkler and mean location Jaro-Winkler are among the most predictive variables. US patent class fixed effects are also highly predictive. As expected, the year gap is the least predictive variable.

Translating this comparison-level data to the firm and patent level, this test set contains 341 establishments and 1634 patents. Out of these, 9 establishments (2.64%) were linked to some patent, and 45 patents (2.75%) were linked to some establishment, similar to the match rates in the handlinked data.

17. One could in theory permute all possible sets of variables, but much like best subset regression, this becomes computationally infeasible quickly. Interesting recent advances in the field include applications of game theory tools to variable importance, such as using Shapley values from cooperative games to calculate variable importances (see Lundberg and Lee, 2017).

Table 2.9: Random Forest Variable Importance

Variable	Importance
Minimum name JW	0.0013181
USPC class	0.0010244
Mean location JW	0.0007532
Name JW	0.0006335
Location JW	0.0005887
Mean name JW	0.0004538
Company-style name	0.0003847
Year gap	0.0001058
Mean location JW missing flag	0.0000089
Location JW missing flag	0.0000082

2.4.3 Linked Dataset

Applying our preferred specification to the entirety of the 1870 Census of Manufactures (N = 187,404) and all patents between 1840-1890[18] (P = 539,079), we find that the model links 3,737 (1.99%) 1870 establishments to a patent, and 7,882 (1.46%) 1840-1900 patents to an 1870 establishment. In total, the model made 14,322 positive predictions at the establishment-patent level.

Like the handlinked data, we find that predicted matches on the remaining 1870 counties are highly geographically concentrated. Only 369 (15.93%) counties receive a predicted match. On the other hand, eleven counties receive more than two hundred predicted establishment-patent links (see Tables 2.10 and 2.11). At the establishment level, we find similar results, where very innovative firms can be linked to up to 144

18. The model was applied both backwards and forwards in time to capture potential matches where the patent is generated after the firm is seen the Census as well, which might allow us to observe the timing of innovations more precisely.

patents, while the vast majority of firms have none (see Table 2.12). As before, we find that the correlation between the number of predicted links and the number of establishments in a county is high (0.60, conditional on having at least one link).

Table 2.10: Counties With Over 200 Predicted Links

County	Links	Linked Establishments	Linked Patents
New Haven, CT	2617	239	975
Hamilton, OH	2134	411	812
Cook, IL	1099	136	624
Hartford, CT	816	168	464
Fairfield, CT	725	84	339
St. Louis, MO	667	142	427
Marion, IN	426	95	180
San Francisco, CA	414	153	267
New Castle, DE	326	56	178
Rensselaer, NY	261	76	157
Onondaga, NY	220	83	156

The second column shows the total amount of predicted links, while columns 3 and 4 show the number of unique establishments and patents that received at least one predicted match.

Table 2.11: Match Count Distribution, Counties With At Least One Link

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	2.00	5.00	38.81	15.00	2617.00

In order to verify the quality of the model on this unseen data, we take a random 1,000-sized subset of the predicted matches for manual inspection, and we find that we agree with only 56.8% of the model's predicted matches. This casts doubt on the model's out of sample performance in the Census at large by itself without manual

Table 2.12: Match Count Distribution, Establishments With At Least One Link

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	1.00	1.00	3.83	3.00	144.00

inspection of matches. In expectation, this agreement rate would imply that in actuality, only about 1% of establishments would receive a machine suggested and manually confirmed link.

The large discrepancy between the test set positive predictive value and the PPV in this newer data could hint at systematic differences between both sets of counties, (e.g., transcription quality). Handlinking counties were randomly drawn from the set of counties that had complete strings in the manufacturing census, along with three other particularly innovative counties. If these more complete or innovative counties are particularly different from the remainder of the Census, the model would have overfit this data and produced overly optimistic performance metrics.

2.5 A TRADITIONAL LINKING APPROACH

Given the model performance in the previous section, we provide an alternative set of links created using traditional linkage techniques.

We pursue a hybrid approach, where we prune potential links similarly to the ABE method (Abramitzky et al., 2021), but then confirm a large portion of the remaining links manually. The following section explains our method in detail.

In order to generate a set of higher confidence links while reducing the amount of manual inspection, we apply the following algorithm:

1. Separate establishments into company and people-style names, as before. Also separate names in patents the same way.

2. For a Census of Manufactures year in 1850, 1860, and 1870:

    (a) Take all patents and assign them the year-correct county.

    (b) For all company-style establishments:

        i. For each county in the year:

            A. Collect all possible pairs of company-style establishments in the county and all patents assigned to a company-style entity in the county.

            B. Keep pairs where the Jaro-Winkler distance in names is smaller than 0.2.[19]

        ii. Collect all potential matches. This results in a dataset with 35,482 potential pairs for 4,312 establishments and 10,772 patents.

        iii. Manually inspect all potential matches. This results in 10,550 matches for 1,176 unique establishments and 5,216 unique patents.

    (c) For all person-style establishments:

        i. For each county in the year:

            A. Collect all pairs of people-named establishments in the county

19. Consider that as a pruning step, this distance is liberal relative to ABE links, which keep matches within a JW of 0.1. In their approach, these pairs are considered to be matches conditional on age proximity and other constraints. In our approach, we use a more liberal cutoff because it reduces the false negative rate for potential matches, while not lowering the false positive rate, given that we follow this pruning step with manual verification.

and all patents issued or assigned to a person-named entity in the county.

B. Using the Census of Population data, apply the algorithm in Section 2.3.2 to separate all names into first and last names.

C. Keep pairs where the Jaro-Winkler distance in first and last names is both 0.2 or lower.

ii. Collect all potential matches. This results in a dataset with 178,042 potential matches for 35,691 establishments and 75,054 patents.

iii. Manually inspect a random sample of 8,000 matches. Set a JW cutoff determined empirically from the confirmed matches in this sample. Within the handlinked sample, 93.8% have a distance under 0.05, and 93.5% of links have a first name distance of 0. For last names, 76% of handlinks have a distance under 0.05, and 73.4% have a distance of zero. Given these results, we provide a flag both for matches that are exact, and for matches within a JW of 0.05.

(d) Collect all confirmed matches.

2.6 CONCLUSION

In this paper we have detailed an automated method to scale up a relatively small set of handlinks to create large linked manufacturing and patenting datasets. We find that the method detects around 79.5% of handlinks in unseen data and is correct about its predictions about 90.6% of the time, despite handlinks being extremely

rare—about 2.5% of handlinked establishments have patents associated with them. At the comparison level, which is what the model sees, this represents 0.18% of comparisons. However, as extra validation, we find that we only agree with 56.8% of the model's predicted matches in data that has not been previously linked by hand, which casts doubt on the feasibility of this algorithm to generate large-scale linked establishment-patent datasets without extensive manual confirmation of links after machine predictions are generated.

2.7 APPENDIX

Table 2.13: Top 10 Patent Classes by Count, Handlinked Training Sample

Patent Class	Count
42/61: Firearms/Revolvers/Cylinder loaded from muzzle end	3086
56/164: Harvesters/Cutting and conveying/ Reciprocating-cutter type/Self-raking mechanism	2844
460/119: Crop threshing or separating/Machine component arrangement and structure	1840
33/39.1: Geometrical instruments/Scriber/Straight line/Ink/Blank space	1702
37/467: Excavating/Scoop or excavating and transporting container Mounted rearwardly of vehicle/Handled or hand operated	1656
111/67: Planting/Drilling/Frame and planting-element arrangement/ Main and auxiliary frame/Floating auxiliary/Tool-bar type/ With lift and ungear	1472
56/269: Harvesters/Cutting/Reciprocating side cutter/Rear cut/ Co-axial wheels	1472
144/150: Woodworking/Shaping machine/Rotary disk cutter, end thrust	1384
248/268: Supports/Brackets/Rod type/Shade roller type/ Independent bracket/Mounted on opposing walls	1384
42/67: Firearms/Revolvers/Firing Mechanism/Cylinder stops	1384

Table 2.14: Top 10 Patent Classes by Matches, Handlinked Training Sample

USPC Class	Match Count
111/67: Planting/Drilling/Frame and planting-element arrangement/ Main and auxiliary frame/Floating auxiliary/Tool-bar type/ With lift and ungear	24
56/164: Harvesters/Cutting and conveying/ Reciprocating-cutter type/Self-raking mechanism	24
460/119: Crop threshing or separating/Machine component arrangement and structure	19
56/268: Harvesters/Cutting/Reciprocating side cutter/Rear cut	17
56/269: Harvesters/Cutting/Reciprocating side cutter/Rear cut/ Co-axial wheels	13
172/271: Earth working/Overload shifting/Frangible lock	12
172/343: Earth working/Guided by walking attendant; supported propelled, or held in position by attendant/ Tool manipulated with respect to mounting frame/ Arched wheel frame/Foot operated	12
56/218: Harvesters/Tongue adjustments and supports	9
417/524: Pumps/Expansible chamber type/Plural pumping chambers/ Including valved piston/Unitary or interconnected elements form inlet or discharge distributors for plural chambers	8
42/61: Firearms/Revolvers/Cylinder loaded from muzzle end	8

CHAPTER 3

DOCUMENTATION FOR ESSAYS IN THE ECONOMICS OF INNOVATION AND ECONOMIC HISTORY

3.1 INTRODUCTION

This chapter documents the various datasets used in this dissertation. Section 3.2 begins with a broad overview of the usage of patent data to study innovation in economics, institutional details about the United States Patent and Trademark Office (USPTO), and a broad outline of patent data collection efforts. Section 3.3 delves into details about the specific datasets used throughout the dissertation, including but not limited to dataset construction, reported coverage, and included variables. Summary statistics for each dataset are provided, including patent count and citation summary statistics for each dataset that contains them, and more broadly, the overlap and inconsistencies between datasets. The results of the dataset comparison are summarized in Section 3.4. Finally, a description of the replication package for this dissertation is provided in Section 3.5.[1]

1. This final section also serves as the readme file for the repository hosting the package, located at: https://github.com/terencechau/space-race-spillovers and https://github.com/terencechau/firm-patent-links.

3.2 PATENT USAGE IN ECONOMICS & DATA COLLECTION EFFORTS

3.2.1 Patents in Economics

Patent data has been utilized in economics since the early twentieth century as a proxy measure for the stock of knowledge in an economy (Griliches, 1990). An early selection of works using aggregate patent statistics evaluated patent data's utility to measure the level of inventive activity (Schmookler, 1954), leveraged it to study which sectors of the population invent (Schmookler, 1957), and studied how patent holding by firms correlates with their profits, sales, market power, and product line diversification (Scherer, 1965a, 1965b). Many of the questions and hypotheses seeded in this early work are still being tested empirically today, using more comprehensive and detailed data, e.g., Sarada et al.'s (2019) study of inventor demographics using inventor records linked to full-count Census of Population data. Formalizations of these relationships, such as Aghion and Howitt's (1992) model of the relationship between market power and technological growth have also allowed for more nuanced hypothesis tests in empirical settings.

For more details on earlier uses of patent statistics in economics, please see Griliches (1990). For a more recent survey with an emphasis on modern empirical methods, such as causal inference, and recent questions, see Williams and Bryan (2021).

3.2.2 USPTO Institutional Details & Data Collection Efforts

The USPTO has kept records of all patents issued since the first patent issued in July of 1790, a patent for making potash by Samuel Hopkins. However, a fire in 1836 destroyed the estimated 9,957 patents issued up to that date.[2] Since then, new patents have been assigned unique, consecutive numbers up to today, with the first numbered patent issued to John Ruggles for a locomotive wheel, and the latest patents issued in May 2023 in the 11,647,000 range (US Patent and Trademark Office, 2023, n.d.).[3]

In 1975, the USPTO started to digitally store patent records, which has fostered the proliferation of empirical work in modern settings using patent-level data. One of the earliest and most comprehensive efforts to wrangle these data into usable formats for economists is the National Bureau of Economic Research group, consisting of Hall et al. (2001), who collected complete patent and citation data for patents starting in 1975[4] and linked it to Compustat records to facilitate firm-level studies.

In order to obtain a more complete view of the patenting landscape before 1975, modern day researchers across various fields have leveraged advances in optical character recognition and large scale efforts to collect patent document scans to build

2. Patents issued up to this point are now referred to as "X" patents, because of the X added to their identifier numbers. These patents initially did not have unique patent numbers, but received new identifiers after the post-1836 numbering system was implemented.

3. Listings of new granted patents are issued every Tuesday by the USPTO in the Official Gazette, which has been published weekly since 1872 and can be found at: https://www.uspto.gov/learning-and-resources/official-gazette/official-gazette-patents. Prior to its creation, The Scientific American fulfilled a similar role between 1845 and 1869, by listing detailed schematics and information about recently granted and upcoming patents, available at: https://www.jstor.org/journal/scieamer.

4. They also include incomplete data for patents issued between 1963 and 1975.

historically comprehensive datasets. Andrews (2021) provides a thorough examination of some of these datasets, including two of the three patent datasets I rely on for this dissertation. To complement this effort, I follow a similar structure to his article to detail the data he does not examine, in hopes this allows future researchers to decide which dataset best suits their scholarly endeavors.

3.3 OVERVIEW OF DATASETS

3.3.1 Dataset Overview & Construction

As detailed in Chapter 1, the main patent datasets used are the USPTO Historical Patent Data Files (HPDF) (Marco et al., 2015), the Comprehensive Universe of U.S. Patents (CUSP) (Berkes, 2018), and finally, Fleming et al.'s (2019) patent and federal reliance data, which Andrews (2021) does not examine.[5]

The HPDF files are constructed using internal, administrative USPTO records, which makes them the most complete patent dataset available, with the known universe of patents from 1836 to 2014.[6] However, as it will be discussed in Section 3.3.3, they also contain little additional information on each patent. For more recent years, they also contain information on published or publicly-available, non-published applications.

5. The HPDF can be found at: https://developer.uspto.gov/product/historical-masterfile and the Fleming et al. data can be found at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DKESRC. To obtain access to the CUSP data, please reach out directly to Enrico Berkes.

6. The HPDF also contains all 2,633 X patents that have been recovered post-fire. However, the remaining 7,324 pre-fire patents have been effectively lost.

The CUSP data contains a host of additional variables parsed from each patent document's text, while containing the majority of patents in the HPDF (Andrews, 2021). To assemble this data, Enrico Berkes collected scans of every single US patent document from 1836 to 2015 stored in the Google Patents website, then applied a separate OCR algorithm to the text. This allowed him to then parse different sections of the text to extract usable variables for each patent document.

Finally, the Fleming et al. dataset, while containing less variables than CUSP, contains an essential variable for the analysis in Chapter 1—whether the patent relied on federal funding or not, and which federal agency it relied on. Similar to the CUSP data, the authors applied their own OCR to pre-1976 patent scans and parsed relevant fields from this text to generate a dataset spanning 1926-2017.

3.3.2 Coverage

Given the HPDF contains the known universe of patents, it serves as the most useful benchmark of coverage for other datasets. Table 3.1 shows the overlap in unique patents in each dataset relative to this benchmark. For consistency, comparisons are made using the sample years in Chapter 1, 1940 to 1980.[7] The CUSP data contains the closest approximation to the administrative files, mirroring the findings in Andrews (2021). While the Fleming et al. data misses a larger share of unique patent numbers, the overall missing rate is fairly low for both.

7. Given the private nature of the CUSP dataset, note that I only have access to the data for the specific sample periods in each chapter.

Table 3.1: Overlap in Unique Patent Numbers Relative to HPDF, 1940-1980

Dataset	# HPDF IDs Missing	% HPDF IDs Missing	Total HPDF IDs
CUSP	47	0.002%	2,052,239
Fleming et al.	899	0.04%	2,052,239

3.3.3 Variables Included

The HPDF files, while containing the most complete set of patent numbers, contain little additional information. The dataset includes patent numbers for issued patents, along with application numbers for published applications[8], and relevant dates.

Essential for the analysis in Chapter 1, the files contain current USPC classification codes, and not the USPC codes at time of issue. Patents are assigned to a technology class and subclass when the application is undergoing examination, however, the state of each art changes over time—new technologies are created all the time, and when reasonable, patents in previously existing subclasses must be folded into or separated into new subclasses. This reassignment process ensures that classifications are consistent at a given point in time, and that they reflect the correct assignments to a particular art unit and set of examiners at the USPTO (Marco et al., 2015). If all patents are queried using their current USPC codes, as is the case when using administrative data, this implies that all patents will be correctly classified regardless of when they were issued, and a panel dataset can be constructed without issue. However, this also implies that if one generates a patent dataset from the original document scans and parses the USPC classifications in the

8. These are a subset of total applications, see Marco et al., 2015 for details.

original scan, the patent will be assigned to its original classification and not the current, consistent classification. Due to this, the USPC codes in the HPDF files are the most useful when attempting to do class or subclass-level analyses, such as those in Chapter 1. In addition to USPC codes, the HPDF files also include mappings to the NBER industrial classification codes from Hall et al. (2001), allowing for industry-level analyses.

The CUSP dataset contains a much broader set of variables given they exploit the text data of each patent document as opposed to the scarce administrative fields in the HPDF data. They include the information in the HPDF, plus inventor and assignee names, their locations, and patent to patent citations. Names and locations allow for linkages, for example, to the Census of Population or other datasets, or as used in Chapter 1, to link inventors to themselves over time. Locations allow for the calculation of aggregate location-level statistics to study the geographic distribution of invention at a cross-section or longitudinally, and citation data allows for the study of knowledge spillovers, as in Chapter 1. In terms of technology classifications, CUSP provides codes beyond USPC, such as IPC and CPC, but suffers from the reassignment issue described above. Despite this, Table 3.2 shows the overlap in the current USPC classification for all 1940-1980 patents with the original classification as parsed from the CUSP scans, and the level of USPC mismatch is around 200 mismatches per 100,000 at the subclass level, and approximately 4.8 per 100,000 at the broad class level. Depending on the technologies of interest, this measurement error might be relevant, but broadly, the original classifications track fairly well with current ones.

Table 3.2: USPC Overlap, Current vs. Original (CUSP), 1940-1980

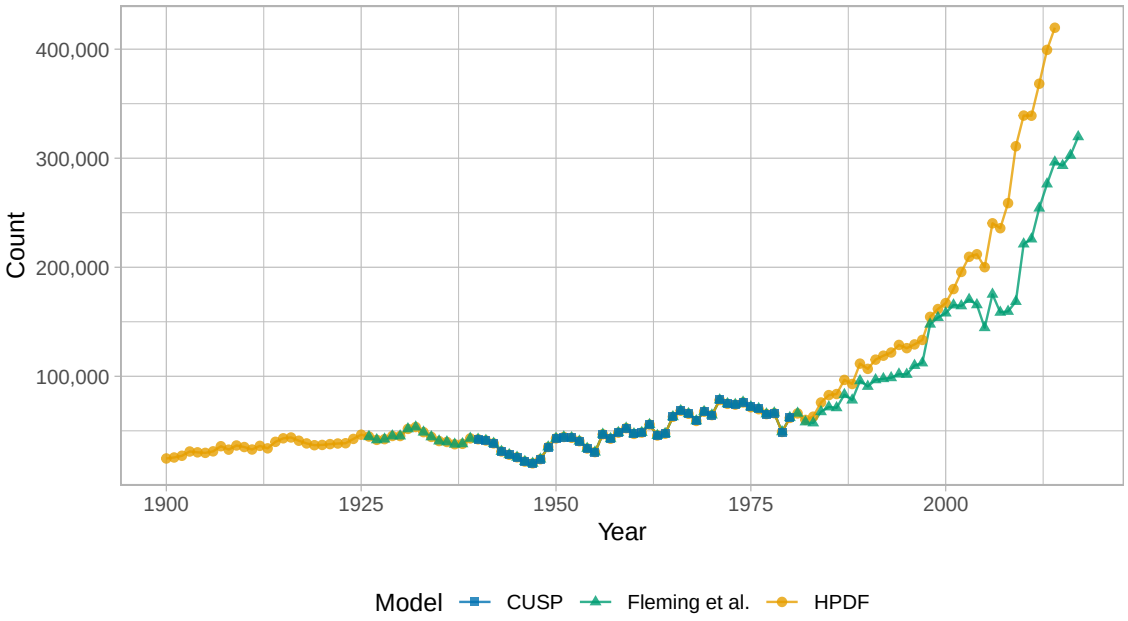
Classification Level	# Mismatched USPCs	% Mismatched USPCs	Total
Subclass	4,330	0.2%	2,052,237
Class	99	0.004%	2,052,237

The Fleming et al. data contains grant year, government reliance, citations, non-patent references, and country of origin. While it contains less information than the CUSP data, it exclusively contains the government acknowledgements used to determine NASA patents in Chapter 1 (see Chapter 1, Figure 1.6).

3.3.4 Patent Count Time Series

Figure 3.1 shows patent counts for each dataset over their available span. As implied by Table 3.1, for the 1940-1980 period, all datasets track closely. For 1926-1940, where Fleming et al. and the HPDF overlap in coverage, they track similarly well. However, for the post-1980 period, it appears that the Fleming et al. data under-counts patents at a much higher rate, relative to the HPDF.

Figure 3.1: Patent Counts by Dataset, 1900-2017



3.3.5 Patent Citation Overlap

The CUSP and Fleming data contain citations in long format. To assess the overlap between both datasets in the 1940-1980 years, I check whether the Fleming et al. data contains every single backward citation the CUSP data contains, and vice-versa.[9] For clarity, patents making the citation will be referred to as "citing" patents, and patents receiving citations will be referred to as "cited" patents.

Tables 3.3 and 3.4 show the number of unique citing patents in the 1940-1980

9. Chapter 1 relies on forward citations, i.e., the citations a given patent receives, not the citations a patent makes. However, the way they are constructed is through recording all backward citations in all patent documents, then inverting them. Therefore, the simplest way to compare two sets of citations from scans is to verify the overlap in backwards citations, particularly when the two datasets span different time periods.

period for the Fleming et al. and CUSP data, and the percentage overlap between each.[10] To maintain consistency in sample periods, all comparisons in citation coverage only use citing and cited patents issued between 1940 and 1980. The CUSP data contains almost all citing and cited patents in the Fleming et al. data, while the converse is true for 98.52% to 99.09% of patents in Fleming et al. Conditional on the cited patent being in both datasets, Table 3.5 shows the distribution of differences in citation counts at the cited patent level. 84.91% of cited patents that appear in both datasets have the same amount of citing patents in both datasets, and 99.81% have at most a five citation difference.

Table 3.3: Citing Patent ID Overlap, 1940-1980

Dataset	# Unique Citing	% Overlap With Other Dataset
CUSP	1,646,738	99.99%
Fleming et al.	1,541,343	99.09%

Table 3.4: Cited Patent ID Overlap, 1940-1980

Dataset	# Unique Cited	% Overlap With Other Dataset
CUSP	1,479,323	99.99%
Fleming et al.	1,457,622	98.52%

10. Unlike previous sections, where the HPDF could serve as as ground truth benchmark, neither of the datasets containing citations is guaranteed to contain all citations. Because the Fleming data contains citations originating as far ahead as 2017, Chapter 1 uses this dataset as opposed to the CUSP, which I only have access to up to 1980.

Table 3.5: Citation Count Differences for Cited Patents, CUSP & Fleming et al.

Citing Difference	Count	% Cited Patents
0	1,237,478	84.91%
(0, 5]	217,167	14.90%
(5, 10]	2,452	0.17%
(10, 20]	298	0.02%
(20, 43]	48	0.003%

3.4 DATASET COMPARISON SUMMARY

For the 1940-1980 period, both the Fleming et al. and CUSP data paint extremely similar pictures of patenting in the United States. In terms of patent counts, they track well with the benchmark administrative data from the HPDF, with the Fleming et al. data undercounting patents in the post-1980 period. In terms of citation counts, they both paint a similar picture for available years, with differences smaller than 2.48% in either direction on coverage, and 84.91% of patents that have been cited receiving the same amount of citations in both datasets. Finally, despite the concern that patents' USPCs were reassigned over time, I find a small 0.2% mismatch rate between the original USPCs assigned to patents from the CUSP scans and the current administrative file's CUSP assignments.

3.5 REPLICATION PACKAGE

Each chapter in this dissertation has a replication package hosted in a Github repository. The first, hosted at https://github.com/terencechau/space-race-spillovers, contains R scripts that carry out all the analyses in Chapter 1.[11] Following is a description of every script included.

- `functions.R`: Is a script loaded at the start of every subsequent file. Defines aesthetic parameters such as plot sizes, creates a uniform formatting for `kable` tables, and sets a uniform theme for `ggplot2`, which is used to generate all figures in the dissertation. Defines a function for estimating sup-t confidence error bands, `calculate_sup_t` for event study estimates, which relies on the `suptCriticalValue` package by Ryan Kessler[12]. To plot all event study estimates in Chapter 1, I define multiple functions, `plot_event_study_prep`, which takes an `lfe::felm` regression object that includes sup-t upper and lower bounds and prepares it for a standard event study plot with an omitted event-time period of -1. `plot_event_study` takes the output of this function and plots the dynamic two-way fixed effects estimates using sensible aesthetics. `plot_event_study_overlaid` does the same, but takes multiple event study objects and overlays them for easier comparison (e.g., for the leave-one-out estimates in Chapter 1).

11. Partial samples of this code in other languages, like Python and SQL queries that replicate the data preparation steps are provided but are only for illustration and were not used for any of the results in the dissertation.

12. https://github.com/ryanedmundkessler/suptCriticalValue

- `1_prepare_data.R`: Takes the Fleming et al. and HPDF datasets to construct technology subclass by year panels for the analyses in Chapter 1. The resulting datasets contain all outcomes in the paper for each relevant sample: `event_study_df.csv` for the main sample, `event_study_all_class_df.csv` for the sample including all possible control subclasses, and `event_study_within_treated_class_df.csv` for the sample that only uses subclasses within broad classes that contain at least one treated subclass. Also produces other datasets, like `diff_in_means_df.csv` for the naive comparison of NASA and non-NASA patent citations in in the post-1958 period, and takes random samples of patents for further handchecking.

- `2_summary_statistics.R`: Creates summary statistics tables and other miscellaneous calculations used throughout the paper, including, the citation difference in means of NASA and non-NASA patents, the difference in differences baseline year balance table, tabulates the largest treatment and control subclasses, and calculates how many classes NASA seeded.

- `3_baseline_event_study.R`: Takes the output from `1_prepare_data.R` and estimates the main static and dynamic two-way fixed effects regressions in the paper. Each static regression is saved as a LaTeX table, while every dynamic regression is saved as a `ggplot2` figure.

- `4_space_essential_classes.R`: Defines the space essential classes, then re-estimates the regressions from the previous file. Creates the mission deviation plot for space-essential classes, which relates essential classes to actually treated

classes.

- `5_excluding_weaponry.R`: Re-estimates the analyses in `3_baseline_event_study.R` omitting military-related classes.

- `6_alternative_control_groups.R`: Re-estimates the baseline analyses using the two sets of alternative control groups.

- `7_callaway_santanna.R`: Uses the Callaway and Sant'Anna (2021) estimator to estimate all the main regressions in the paper.

- `8_figures_other.R`: Produces all figures in the paper that don't rely on the subclass panel directly. Makes all figures related to NASA and federal R&D outlays, computer science enrollments, and top NASA USPC classes and subclasses at the patent level.

- `9_documentation.R`: Calculates all summary statistics and makes all figures in the documentation chapter.

The second set of files, hosted at https://github.com/terencechau/firm-patent-links, contains scripts used in Chapter 2, both to train the supervised learning models in the paper, and to create an alternative, handlinked dataset.

- `prep_patent_data.R`: Prepares CUSP data for linking tasks.

- `border_fixes.R`: Takes the CUSP geolocation data, which pins each patent to a city in the year 2000, and remaps it back to historical, year appropriate counties.

- `handlinks/company_linking.R`: Takes all 1850-1870 CMF establishments with company-style names and matches them to all 1840-1900 patents with company-style assignees by string distance, where company-style refers to establishments whose names do not refer to sole ownership. The resulting dataset is filtered to potential matches whose Jaro-Winkler string distance is 0.2 or smaller. Potential handmatches in this set are then determined to be matches or non-matches manually.

- `handlinks/people_linking.R`: Takes all 1850-1870 CMF establishments with people's names and matches them to all 1840-1900 patents with inventors or assignees with people's names by string distance, where potential matches are determined by having a first and last name Jaro-Winkler of 0.2. A sample of potential handmatches in this set are then determined to be matches or non-matches manually.

- `handlinks/people_linking_2.R`: After manual linking of the output from `people_linking.R`, collects and summarizes the manual links.

- `machine_links/prep_matching_sheet.R`: Prepares sheets for handlinking of training data.

- `machine_links/summarize_matches.R`: Collects the sheets from the previous script to create a training dataset.

- `machine_links/linkage_model.Rmd`: Trains a random forest and other models to predict linkage probability.

- `machine_links/linkage_model_split_names.Rmd`: Trains a random forest and other models to predict linkage probability using additional data from the Census of Population.
- `machine_links/performance_metrics.R`: Defines helper functions to measure model performance.
- `machine_links/prep_sheets_apply_model.R`: Prepares candidate matches for all other counties to apply pre-trained model.
- `machine_links/apply_model.R`: Applies pre-trained model to all other counties and creates predictions.
- `machine_links/split_large_files.R`: Splits large counties into multiple files to reduce computational burden.
- `machine_links/collect_split_files.R`: Collects split counties.
- `machine_links/collect_links.R`: Collects all links after applying the model at scale.

REFERENCES

Aaronson, D. E. (1966). Appendix A: Legislative History of the Property Rights in Inventions Provisions of the National Aeronautics and Space Act of 1958. In D. S. Watson & M. A. Holman (Eds.), *An Evaluation of the Patent Policies of the National Aeronautics and Space Administration: Report of the Committee on Science and Astronautics, U.S. House of Representatives, Eighty-ninth Congress, Second Session* (pp. 95–140). U.S. Government Printing Office.

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., & Pérez, S. (2021). Automated Linking of Historical Data. *Journal of Economic Literature*, *59*(9), 865–918.

Aghion, P., & Howitt, P. (1992). A Model of Growth Through Creative Destruction. *Econometrica*, *60*(2), 323–351.

Akcigit, U., Grigsby, J., & Nicholas, T. (2017). *The Rise of American Ingenuity: Innovation and Inventors of the Golden Age* (NBER Working Paper 23047).

Allen, B. (2017). *Maxime A. Faget*. https://www.nasa.gov/langley/hall-of-honor/maxime-a-faget

Ancestry. (N.d.). *Ancestry.com*. Retrieved May 10, 2021, from https://www.ancestry.com/

Andrews, M. J. (2021). Historical Patent Data: A Practitioner's Guide. *Journal of Economics & Management Strategy*, *30*(2), 368–397.

Arrighi, R. S. (2019). *George Low Spurred Moon Landings*. https://www.nasa.gov/feature/glenn/2019/george-low-spurred-moon-landings

Arrow, K. (1962). "Economic Welfare and the Allocation of Resources for Invention". In National Bureau of Economic Research (Ed.), *The rate and direction of inventive activity: Economic and social factors* (pp. 609–626). Princeton University Press.

Azoulay, P., Zivin, J. S. G., Li, D., & Sampat, B. N. (2018). Public R&D Investments and Private-sector Patenting: Evidence from NIH Funding Rules. *The Review of Economic Studies*, *86*(1), 117–152. https://doi.org/10.1093/restud/rdy034

Berkes, E. (2018). *Comprehensive Universe of U.S. Patents (CUSP): Data and Facts* (Working Paper).

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Brooks, C. G., & Ertel, I. D. (Eds.). (1973). *The Apollo Spacecraft: A Chronology, Volume III, October 1, 1964-January 20, 1966*. National Aeronautics and Space Administration, Scientific and Technical Information Branch.

Brooks, C. G., Grimwood, J. M., & Swenson, L. S. (1979). *Chariots for Apollo: A History of Manned Lunar Spacecraft*. National Aeronautics and Space Administration, Scientific and Technical Information Branch.

Bureau of Labor Statistics. (2023). *CPI for All Urban Consumers (CPI-U)*.

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods [Themed Issue: Treatment Effect 1]. *Journal of Econometrics*, *225*(2), 200–230. https://doi.org/https://doi.org/10.1016/j.jeconom.2020.12.001

David, P. A., Hall, B. H., & Toole, A. A. (2000). Is Public R&D a Complement or Substitute for Private R&D? A Review of the Econometric Evidence. *Research Policy*, *29*, 497–529.

de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, *110*(9), 2964–96.

Donaldson, D., & Hornbeck, R. (2016). Railroads and American Economic Growth: A Market Access Approach. *Quarterly Journal of Economics*, *131*(2), 799–858.

Dunbar, B. (2017). *History of John F. Kennedy Space Center*. https://www.nasa.gov/offices/history/center_history/kennedy_space_center

Ertel, I. D., & Morse, M. L. (Eds.). (1969). *The Apollo Spacecraft: A Chronology, Volume I, Through November 7, 1962*. National Aeronautics and Space Administration, Scientific and Technical Information Branch.

Ertel, I. D., Newkirk, R. W., & Brooks, C. G. (Eds.). (1978). *The Apollo Spacecraft: A Chronology, Volume III, January 21, 1966-July 13-1974*. National Aeronautics and Space Administration, Scientific and Technical Information Branch.

European Space Agency. (2011). *The Flight of Vostok 1*. https://www.esa.int/About_Us/ESA_history/50_years_of_humans_in_space/The_flight_of_Vostok_1

Feigenbaum, J. J. (2016). *A Machine Learning Approach to Census Record Linking* (tech. rep.). https://scholar.harvard.edu/jfeigenbaum/publications/automated-census-record-linking

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210.

Fleming, L., Greene, H., Li, G.-C., Marx, M., & Yao, D. A. (2019). Government-Funded Research Increasingly Fuels Innovation. *Science*, *364*(6646), 1139–1141. https://doi.org/10.1126/science.aaw2373

Freyaldenhoven, S., Hansen, C., Pérez, J. P., & Shapiro, J. (forthcoming). Visualization, Identification, and Estimation in the Linear Panel Event Study Design. *Advances in Economics and Econometrics: Twelfth World Congress*.

Ginzberg, E., Kuhn, J. W., Schnee, J., & Yavitz, B. (1976). *Economic Impact of Large Public Programs: The NASA Experience*. Olympus Publishing Company.

Glennan, T. K. (1993). *The Birth of NASA: The Diary of T. Keith Glennan* (J. Hunley, Ed.). NASA History Office.

Goodman-Bacon, A. (2021). Difference-in-differences with Variation in Treatment Timing [Themed Issue: Treatment Effect 1]. *Journal of Econometrics*, *225*(2), 254–277.

Graham, S. J., Grim, C., Islam, T., Marco, A. C., & Miranda, J. (2018). Business Dynamics of Innovating firms: Linking U.S. Patents with Administrative Data on Workers and Firms. *Journal of Economics & Management Strategy*, *27*(3), 372–402. https://doi.org/https://doi.org/10.1111/jems.12260

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

Grimwood, J. M. (Ed.). (1963). *Project Mercury: A Chronology*. National Aeronautics and Space Administration, Scientific and Technical Information Branch. https://history.nasa.gov/SP-4001/app9.htm

Grimwood, J. M., Hacker, B. C., & Vorzimmer, P. J. (Eds.). (1968). *Project Gemini Technology and Operations: A Chronology*. National Aeronautics and Space Administration, Scientific and Technical Information Branch.

Gross, D. P., & Sampat, B. N. (2022). *America, Jump-started: World War II R&D and the Takeoff of the U.S. Innovation System* (Working Paper).

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). *The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools* (NBER Working Paper 8498).

Hansen, J. R. (Ed.). (1995). *Spaceflight Revolution: NASA Langley Research Center, From Sputnik to Apollo*. National Aeronautics and Space Administration, Scientific and Technical Information Branch.

Helgertz, J., Price, J. R., Wellington, J., Thompson, K., Ruggles, S., & Fitch, C. R. (2021). *A New Strategy for Linking Historical Censuses: A Case Study for the IPUMS Multigenerational Longitudinal Panel* (IPUMS Working Paper).

Hornbeck, R., Hsu, S. H.-M., Humlum, A., & Rotemberg, M. (2023). *The transition from water to steam power* (Working Paper).

Howell, S. T. (2017). Financing Innovation: Evidence from R&D Grants. *American Economic Review*, *107*(4), 1136–64. https://doi.org/10.1257/aer.20150808

Jacob, B. A., & Lefgren, L. (2011). The Impact of Research Grant Funding on Scientific Productivity. *Journal of Public Economics*, *95*(9), 1168–1177. https://doi.org/https://doi.org/10.1016/j.jpubeco.2011.05.005

Jaffe, A. B., Fogarty, M. S., & Banks, B. A. (1998). Evidence from Patents and Patent Citations on the Impact of NASA and other Federal Labs on Commercial

Innovation. *The Journal of Industrial Economics*, *46*(2), 183–205. https://www.jstor.org/stable/117548

Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *The Quarterly Journal of Economics*, *108*(3), 577–598. https://doi.org/https://doi.org/10.2307/2118401

Jovanovic, B., & Rousseau, P. L. (2005). General Purpose Technologies. In P. Aghion & S. N. Durlauf (Eds.). Elsevier. https://doi.org/https://doi.org/10.1016/S1574-0684(05)01018-X

Kantor, S., & Whalley, A. (2022). *Moonshot: Public R&D and Economic Growth* (Working Paper).

Keeter, B. (2017). *History of John H. Glenn Research Center at Lewis Field*. https://www.nasa.gov/offices/history/center_history/glenn_research_center

Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring Technological Innovation over the Long Run. *American Economic Review*, *3*(3), 303–20.

Kennedy, J. F. (1961). *Address to Joint Session of Congress May 25, 1961*. John F. Kennedy Presidential Library and Museum. https://www.jfklibrary.org/learn/about-jfk/historic-speeches/address-to-joint-session-of-congress-may-25-1961

Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017). Technological Innovation, Resource Allocation, and Growth. *The Quarterly Journal of Economics*, *132*(2), 665–712.

Kraemer, S. K. (1999). NASA, Monopolies, and the Cold War: The Origins and Consequences of NASA Patent Policy, 1958-1996. *Annual Meetings of the Society for the History of Technology*.

Low, G. M. (1961). *A Plan for Manned Lunar Landing* (tech. rep.). National Aeronautics and Space Administration. Washington, DC. %7Bhttps://www1.grc.nasa.gov/wp-content/uploads/Plan-for-Manned-Lunar-Landing-1961.pdf%7D

Low, G. M. (1999). *"Before This Decade is Out....": Personal Reflections on the Apollo Program* (G. E. Swanson, Ed.). NASA History Office.

Lundberg, S., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. *CoRR, abs/1705.07874*. http://arxiv.org/abs/1705.07874

Marco, A. C., Carley, M., Jackson, S., & Myers, A. F. (2015). *The USPTO Historical Patent Data Files: Two Centuries of Invention* (USPTO Working Paper).

M.I.T. Libraries. (2005). *Jerome Bert Wiesner, 1915-1994*. https://libraries.mit.edu/mithistory/institute/offices/office-of-the-mit-president/jerome-bert-wiesner-1915-1994/

Mohon, L. (2008). *NASA's Michoud Assembly Facility*. https://www.nasa.gov/centers/marshall/michoud/maf_history.html

Moretti, E., Steinwender, C., & Reenen, J. V. (Forthcoming). The Intellectual Spoils of War? Defense R&D, Productivity and International Spillovers. *Review of Economics and Statistics*.

Morse, M. L., & Bays, J. K. (Eds.). (1973). *The Apollo Spacecraft: A Chronology, Volume II, November 8, 1962-September 30, 1964*. National Aeronautics and Space Administration, Scientific and Technical Information Branch.

Moser, P., & Nicholas, T. (2004). Was Electricity a General Purpose Technology? Evidence from Historical Patent Citations. *American Economic Review*, *94*(2).

Murray, C. A., & Cox, C. B. (2004). *Apollo*. South Mountain Books.

Myers, K. R., & Lanahan, L. (2022). Estimating Spillovers from Publicly Funded R&D: Evidence from the US Department of Energy. *American Economic Review*, *112*(7), 2293–2423.

NASA. (1962). *Map for John Glenn's Friendship 7 Space Flight [Image]*.

Nicholas, T. (2010). The Role of Independent Invention in U.S. Technological Development, 1880-1930. *Journal of Economic History*, *70*(1).

O'Brien, F. (2010). *The Apollo Guidance Computer: Architecture and Operation*. Springer Praxis. https://doi.org/https://doi.org/10.1007/978-1-4419-0877-3

Office of Management and Budget. (2021). *Budget FY 2022 - Historical Tables, Budget of the United States Government, Fiscal Year 2022*.

Rosenberg, N., & Trajtenberg, M. (2004). A General-Purpose Technology at Work: The Corliss Steam Engine in the Late-Nineteenth-Century United States. *Journal of Economic History*, *64*(1).

Rosholt, R. L. (Ed.). (1966). *An Administrative History of NASA, 1958-1963*. National Aeronautics and Space Administration, Scientific and Technical Information Branch.

Russell, R. C. (1918). *Index* (U.S. Patent 1,261,167).

Russell, R. C. (1922). *Index* (U.S. Patent 1,435,663).

Ruzic, N. P. (1976). *Spinoff 1976: A Bicentennial Report*. National Aeronautics & Space Administration Technology Utilization Office.

Sarada, S., Andrews, M. J., & Ziebarth, N. L. (2019). Changes in the Demographics of American Inventors, 1870–1940. *Explorations in Economic History*, *74*, 101275. https://doi.org/https://doi.org/10.1016/j.eeh.2019.05.003

Scherer, F. M. (1965a). Corporate Inventive Output, Profits, and Growth. *Journal of Political Economy*, *73*(3), 290–297.

Scherer, F. M. (1965b). Firm Size, Market Structure, Opportunity, and the Output of Patented Inventions. *The American Economic Review*, *55*(5), 1097–1125.

Schmookler, J. (1954). The Level of Inventive Activity. *The Review of Economics and Statistics*, *36*(2), 183–190.

Schmookler, J. (1957). Inventors Past and Present. *The Review of Economics and Statistics*, *39*(3), 321–333.

Sidey, H. (1994). *Why We Went to the Moon*. https://content.time.com/time/subscriber/article/0,33009,981167-1,00.html

Slavtchev, V., & Wiederhold, S. (2016). Does the Technological Content of Government Demand Matter for Private R&D? Evidence from US States. *American Economic Journal: Macroeconomics*, *8*(2), 45–84.

Snyder, T. D. (1993). *120 Years of American Education: A Statistical Portrait*. U.S. Dept. of Education, Office of Educational Research; Improvement, National Center for Education Statistics.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects [Themed Issue: Treatment Effect 1]. *Journal of Econometrics*, *225*(2), 175–199. https://doi.org/https://doi.org/10.1016/j.jeconom.2020.09.006

Thomas, J. R. (2016). *March-In Rights Under the Bayh-Dole Act* (tech. rep. R44597). Congressional Research Service.

United States Code Title 35 - Patents (2021). https://www.uspto.gov/web/offices/pac/mpep/consolidated_laws.pdf

Uri, J. (2021). *60 Years Ago: The Manned Spacecraft Center Makes Houston its Home*. https://www.nasa.gov/feature/60-years-ago-the-manned-spacecraft-center-makes-houston-its-home

Uri, J. (2022). *60 Years Ago: John Glenn, the First American to Orbit the Earth aboard Friendship 7*. https://www.nasa.gov/feature/60-years-ago-john-glenn-the-first-american-to-orbit-the-earth-aboard-friendship-7

U.S. House of Representatives. (1958). H.R.12575-An Act to provide for research into problems of flight within and outside the earths atmosphere, and for other purposes. *United States Statutes at Large*, *72*(426). https://www.govinfo.gov/app/details/STATUTE-72/STATUTE-72-Pg426-2

U.S. House of Representatives. (1959). *The Next Ten Years in Space, 1959-1969 : Staff report of the Select Committee on Aeronautics and Space Exploration*. U.S. Government Printing Office.

U.S. House of Representatives. Committee on Appropriations. (1960). Estimates of Appropriations, Fiscal Year 1961, Volume II: Research & Development. In

Amendments to the Budget, Fiscal Year 1961, for the National Aeronautics and Space Administration. 86th Congress, 2nd Session, Document No.329.

US Patent and Trademark Office. (2023). *Official Gazette of the United States Patent and Trademark Office, May 9th, 2023* (Vol. 1510). United States Department of Commerce, U.S. Patent and Trademark Office, Electronic Information Products Division. https://patentsgazette.uspto.gov/week19/OG/Cpc-h.html

US Patent and Trademark Office. (n.d.). *Milestones in U.S. Patenting*. https://www.uspto.gov/patents/milestones

Watson, D. S., & Holman, M. A. (1966). *An Evaluation of the Patent Policies of the National Aeronautics and Space Administration: Report of the Committee on Science and Astronautics, U.S. House of Representatives, Eighty-ninth Congress, Second Session*. U.S. Government Printing Office.

Wiesner Committee. (1961). *Report to the President-Elect of the Ad Hoc Committee on Space*. NASA History Office. https://www.hq.nasa.gov/office/pao/History/report61.html

Williams, H., & Bryan, K. (2021). Innovation: Market Failures and Public Policies. In K. Ho, A. Hortacsu, & A. Lizzeri (Eds.), *Handbook of Industrial Organization* (281–388).

Winkler, W. E. (1990). *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage* (U.S. Bureau of the Census Working Paper). https://files.eric.ed.gov/fulltext/ED325505.pdf