Ecole doctorale 465 – Economie Panthéon Sorbonne

Ecole d'Economie de Paris

Doctorat

Discipline: Economie

TROIS ESSAIS SUR LA DIFFUSION DE L'INNOVATION

Essays on the Diffusion of Innovation

Thèse dirigée par: Philippe Aghion

Date de soutenance: Septembre 2022

Rapporteurs:

1. Bronwyn Hall, University of California Berkeley
2. Adam Jaffe, Brandeis University

Jury:

1. Lorenzo Cassi, Ecole d'Economie de Paris
2. Bronwyn Hall, University of California Berkeley
3. Adam Jaffe, Brandeis University
4. John van Reenen, London School of Economics

Résumé et mots-clés

Résumé: Cette thèse de doctorat s'intéresse à la diffusion de l'innovation.

Nous commençons par nous interroger sur la mesure de la diffusion de l'innovation. Nous montrons que les citations inscrites *dans le texte* des brevets constituent des candidats de premier plan pour mesurer cette diffusion. Nous constituons une base de données fiable et exhaustive de ces citations dans le corpus des brevets américains.

Par la suite, nous nous penchons sur l'histoire de l'innovation telle que décrite par le corpus des brevets en Europe et aux Etat-Unis. Nous publions une base de données nouvelle nous permettant de retracer les caractéristiques des inventeurs (en particulier leur adresse, profession et titre académique) depuis le XIXe siècle en Allemagne, en France, au Royaume-Uni et aux Etats-Unis. Cela nous permet notamment de faire apparaître des divergences majeures entre ces différents pays, notamment dans les domaines de la concentration géographique des l'innovation, du rôle de l'immigration dans l'innovation et des caractéristiques des inventeurs.

Pour finir, nous nous intéressons à la contribution des principaux pays développés et émergents aux technologies “frontières”. Il apparaît notamment que si la Chine présente toujours un héritage caractéristique d'un pays en rattrapage technologique, cet héritage est en passe de s'effacer pour laisser place à une puissance technologique de premier plan, comparable aux Etat-Unis.

Mots-clés: Diffusion de l'innovation, Traitement automatisé du langage, Citations de brevets, Histoire de l'innovation, Technologie frontière

Foreword

In this PhD thesis, I developed a research programme on innovation diffusion. This research programme resulted in three papers constituting this PhD thesis.

I started by questioning the standard measurement device of virtually any work related to innovation diffusion using patents, that is the use of front-page patent citations as a proxy of knowledge flows. In Verluise et al (2022)[1], we show that the data generating process of in-text patent citations makes them immune to many of the sources of noise affecting the traditional front-page patent citations. This makes in-text patent citations arguably a better proxy of knowledge flows. We provide a novel and accurate dataset of in-text patent citations covering more than 10 million US patents (https://zenodo.org/record/4391095). The data generating process has been open-sourced (https://github.com/cverluise/PatCit). We hope that our results and dataset will help improve empirical measurements of knowledge flows.

Next, I realized that most of our understanding of the historical stylized facts of innovation diffusion are based on the US only, while the rest of the world remained almost terra incognita. In Bergeaud and Verluise (2022)[2] we tried to rectify this situation. We used modern Natural Language Processing to extract key patentees' data (location, occupation, citizenship) from German (including East German), French, British and US patents since the late 19th century. While such data were inexistent (except for the US) before 1980, we leverage this novel database to deep dive into the long run dynamics of innovation. We find a large degree of heterogeneity between countries regarding the geographic concentration of innovative activities, the globalisation of innovation and the role of immigrants in importing knowledge. The data-generating codebase has been open-sourced (https://github.com/cverluise/patentcity) and the dataset is publicly available (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PG6THV). We hope that this will stimulate research on the long run heterogeneity of innovation diffusion and its determinants.

Eventually, my time spent as a PhD student has been marked by rising China-US tensions around technological leadership, giving way to many discussions, often missing empirical foundations. I had the feeling that patenting could help bring a pragmatic approach to the question. In Verluise and Bergeaud (2022)[3], we draw on recent development in Artificial

[1] Verluise, Cyril and Cristelli, Gabriele and Higham, Kyle and de Rassenfosse, Gaétan, "The Missing 15 Percent of Patent Citations" (2022).

[2] Bergeaud, Antonin and Verluise, Cyril. "What Do we Learn from One Century of Innovation in Europe and the US?" (2022).

[3] Verluise, Cyril and Bergeaud, Antonin, "The Rise of China's Technological Power: the Perspective from Frontier Technologies" (2022).

Intelligence in the field of "automated patent landscaping" to accurately delineate frontier technologies from the worldwide corpus of patents. Tracking half a dozen technologies which are deemed to be the key technologies of the next decade, we observe that China's contribution to the patenting of frontier technologies has been rapidly rising over the last decade, sometimes competing with the US in terms of the quantity of patents published in these frontier technologies. However, China still exhibits the stigma of a catching-up economy characterized by a stock of lower quality patents and a lower share of patents generated by domestic inventors. The underlying code and dataset are to be open sourced and made publicly available.

Overall, to me, the contribution I might be the most excited about is the release of new datasets and their generating processes. I hope that these assets will be leveraged and extended by the community to open new research avenues.

Acknowledgements

This PhD thesis has benefited from the help of many people, funders and institutions.

I would like first to thank Philippe Aghion without whom this PhD thesis would not have been possible. Philippe has been much more than a supervisor. He has been a life inspiration in many respects.

I am also deeply indebted to my co authors, Antonin Bergeaud, Gabriele Cristelli, Kyle Higham and Gaétan de Rassenfosse and the research assistants who helped with our various projects Juliette Coly, Francesco Gerotto, Aymann M'Hammedi and Lucas Violon. They were all a constant source of learning and excitement.

I would like to thank the members of my Jury who have accepted to be part of the defense and provide generous remarks and recommendations: Lorenzo Cassi, Bronwyn Hall, Adam Jaffe and John van Reenen. I am especially humbled by this Jury as all of them have had a clear influence on my PhD thesis.

Obviously, a PhD venture is also a matter of comradeship and I am thankful to all the members of the Economics of Innovation Lab at College de France who have provided this key ingredient. This includes (but is not restricted to) Emma Bursztejn, Olympia Cutinelli-Rendina, Maxime Gravoueille, Simon Margolin, Antoine Mayerowitz and Thomas Zuber. I am also thankful to the Innovation Information Initiative community members for their support, in particular Agnes Cameron and Samuel Klein.

I am also thankful to my family and close friends for their constant support across my PhD time and before. This includes in particular my parents, Pierre and Sylvie Verluise, my brothers Erwan and Tanguy Verluise and my partner Heloïse Barbat.

Eventually, I would like to thank the European Research Council, Collège de France and Google for their financial support.

Contents

1 The Missing 15 Percent of Patent Citations 9

1.1 Introduction . . . 10

1.2 The epistemology of in-text citations . . . 12

1.2.1 A legal perspective on in-text patent citations . . . 12

1.2.2 In-text patent citations as valuable paper trails of knowledge flows . . . 13

1.2.3 In-text patent citations as valuable signals of patent importance . . . 14

1.3 Methods . . . 16

1.3.1 Data . . . 16

1.3.2 Extraction task . . . 16

1.3.3 Parsing task . . . 18

1.3.4 Consolidation task . . . 19

1.3.5 Pipeline . . . 20

1.4 Technical validation . . . 20

1.4.1 Data consistency . . . 21

1.4.2 Extraction task . . . 22

1.4.3 Parsing task . . . 23

1.4.4 Matching task . . . 25

1.5 A first look into in-text citation data . . . 26

1.5.1 Order of magnitudes . . . 27

1.5.2 Overlap between in-text and front-page patent citations . . . 27

1.5.3 Textual similarity between citing and cited patents . . . 28

1.5.4 Forward citations . . . 30

1.5.5 The internationalization of patent citations . . . 31

1.5.6 Self-reliance . . . 32

1.5.7 Geographic distribution . . . 32

1.5.8 Going further . . . 33

1.6 Concluding remarks . . . 34

1.7 Tables . . . 36

1.8 Figures . . . 41

1.9 Appendix . . . 52

2 What Do we Learn from One Century of Innovation in Europe and the US? 56

2.1 Introduction . . . 57

2.2 Data . . . 61

2.2.1 Data collection and coverage . . . 62

2.2.2 Information extraction . . . 63

2.2.3 Data enrichment . . . 67

2.2.4 Interoperability . . . 70

2.2.5 Other data . . . 70

2.3 Concentration of innovation . . . 71

2.3.1 Measure of concentration . . . 72

2.3.2 Innovation hubs . . . 72

2.3.3 Change in the level of concentration . . . 75

2.4 International flows . . . 77

2.4.1 Foreign patentees . . . 78

2.4.2 Gravity . . . 80

2.5 Who are the inventors? . . . 83

2.5.1 Occupation of inventors . . . 84

2.5.2 Innovation and poverty . . . 86

2.6 Migration and innovation . . . 89

2.6.1 Share of immigrants in US and British patents . . . 90

2.6.2 Immigration and novelty . . . 91

2.7 Conclusion . . . 94

2.8 Appendix . . . 95

3 The Rise of China's Technological Power: the Perspective from Frontier Technologies 134

3.1 Introduction . . . 135

3.2 Technology definition and selection . . . 138

3.2.1 Definition . . . 138

3.2.2 Selection . . . 139

3.3 Automated patent landscaping with humans in the loop . . . 140

3.3.1 The traditional approach . . . 140

3.3.2 Automated patent landscaping . . . 140

3.3.3 Extension . . . 142

3.4 Algorithm deployment and validation . . . 142

3.4.1 Algorithm deployment . . . 142

3.4.2 Performance and consistency . . . 144

3.4.3 External validation . . . 147

3.5 The rise of China . . . 150

3.5.1 The bi-polarization of frontier innovation by the US and China . . . 150

3.5.2 The Chinese catch up in quality . . . 151

3.5.3 China's frontier innovation domestic capacities build up . . . 152

3.6 Conclusion . . . 154

3.7 Appendix . . . 155

Bibliography 163

Chapter 1

The Missing 15 Percent of Patent Citations

Cyril Verluise. Collège de France & Paris School of Economics
Gabriele Cristelli. École Polytechnique Fédérale de Lausanne
Kyle Higham. Hitotsubashi University.
Gaétan de Rassenfosse. École Polytechnique Fédérale de Lausanne

This version: January 2022

Abstract

Patent citations are one of the most commonly-used metrics in the innovation literature. Leading uses of patent-to-patent citations are associated with the quantification of inventions' quality and the measurement of knowledge flows. Due to their widespread availability, scholars have exploited citations listed on the front-page of patent documents. Citations appearing in the full-text of patent documents have been neglected. We apply modern machine learning methods to extract these citations from the text of USPTO patent documents. Overall, we are able to recover an additional 15 percent of patent citations that could not be found using only front-page data. We show that "in-text" citations bring a different type of information compared to front-page citations. They exhibit higher text-similarity to the citing patents and alter the ranking of patent importance. The dataset is available at patcit.io (CC-BY-4).

JEL classification: C81, O30
Keywords: Citation, Patent, Open data

1.1 Introduction

Patent documents represent an invaluable source of information about technological progress. They provide a detailed account of inventive activities, sometimes as early as the mid-nineteenth century (Sokoloff, 1988; Moser and Nicholas, 2004; Akcigit et al., 2017a; Andrews, 2020a). Researchers across all fields of sciences and engineering exploit them as a knowledge repository as well as for technology foresight and competitive intelligence analysis, among other applications (Porter et al., 2008; Benson and Magee, 2015; Candia et al., 2019). Researchers in the social sciences exploit them to study various facets of the innovation process (Jaffe and de Rassenfosse, 2017).

Early work exploiting patent documents focused on easily accessible metadata, including citations and technology classes. Citation data are a particularly popular object of study; a Google Scholar search with the keyword "patent citation" returns about 15,000 results. Use cases are too numerous to list but cover the measurement of invention 'quality,' the placement of inventions in the broader invention network, and the tracking of knowledge flows. More recently, the field has been moving towards exploiting the full text of patent documents. Applications cover, e.g., keyword extraction, topic identification, and invention similarity (Kaplan and Vakili, 2015; Younge and Kuhn, 2016; Arts et al., 2018; Righi and Simcoe, 2019)

In this work, we focus on one aspect of full-text data that has eluded the attention of scholars, namely *in-text citations to patent documents*. Patent offices—and, therefore, the major patent datasets—provide structured data on so-called front-page citations. These citations are made for procedural reasons; they list prior art that is relevant for assessing the patentability of the claimed invention. They originate from applicants (or their attorneys and inventors), examiners, and third parties.[1] They may originate directly at the time of filing, during the substantive examination before grant as well as after grant in case of opposition, re-examination, revocation, etc. By their nature, front-page citations are thus conceptually different from citations typically found in scientific papers (Meyer, 2000).

By contrast, in-text patent citations appear in the patent text itself. They are made to fulfil enablement requirements; to make arguments for novelty and non-obviousness; and to make arguments for usefulness. As these justifications for adding in-text citations do not perfectly overlap with those that drive the generation of front-page citations, in-text citations contain truly novel information over and above that reflected in front-page citations.

Scholars have recently extracted in-text citations to the scientific literature, that is, patent-to-article citations (Bryan et al., 2020; Marx and Fuegi, 2020; Verluise and de Rassenfosse, 2020). Given the importance of citation data, the lack of treatment of in-text patent-to-patent citations is an obvious gap. Such data are likely to be particularly important for specific applications,

[1] An example of citations by third parties is Section 801 of the Patent Cooperation Treaty (Administrative Instructions), which allows third parties to make observations referring to relevant prior art.

such as for the measurement of knowledge flows. Indeed, inventors often contribute to the drafting of the text, and the references they mention are likely to be a better way of capturing knowledge flows than front-page references. Despite our strong suspicion that these data might be relevant for some applications, little research exists to confirm it—precisely because data were not readily available until now. It is thus critical to process these data and make them widely accessible.

We have extracted patent citations from the full-text of 16,781,144 publications filed in the U.S. Patent and Trademark Office (USPTO) from 1790 to 2018. About 95 percent of these publications are granted patents or patent applications.[2] For the sake of simplicity, unless specified, we use the term 'patent' to designate all publications in the dataset in the rest of the paper. We relied on "Grobid", an open-source machine learning library leveraging Natural Language Processing (NLP) to extract and parse citations.[3] We performed an extensive validation exercises, revealing high performance: our extraction task in particular achieves a satisfying 97 percent precision and 82 percent recall (f1-score nearing 90 percent). Overall, we extracted 63,854,733 in-text patent citations, suggesting that in-text patent citations are by no means a marginal phenomenon. A total of 49,409,629 (77.5 %) of them were matched to a standard publication number ensuring interoperability with other patent datasets. The data collection effort is part of PatCit, an open source project that aims at building a comprehensive patent citation dataset.

We have performed an in-depth quantitative analysis of the difference between in-text and front-page citations. We discovered three noteworthy elements. First, by-and-large, in-text citations do not overlap with front-page citations. Overall, we are able to identify an additional 15 percent more citations than one would get using front-page data alone (that is, these citations are not listed on the front page). This figure jumps to 100 percent before 1947, meaning that our data will be an invaluable help to researchers interested in the pre-WWII period. Second, the data generation process of in-text citations intrinsically differs from that of front-page citations and, we believe, is particularly suited to capture knowledge flows. This intuition is reinforced by measures of textual similarity; we find that in-text citations are more similar to the focal citing patent than front-page citations. Third, we find surprisingly low correlation between the front-page forward citation count and the in-text forward citation count. Scholars have used such counts to measure invention importance (Trajtenberg, 1990a; Lanjouw and Schankerman, 2004; Hall et al., 2005). The low correlation suggests that in-text citations provide valuable information to assess invention importance.

The dataset is publicly available on Google Cloud Big Query and Zenodo. Additional technical documentation and usage guides are available on the project repository and the documentation website.[4] In addition to the final output, we also release the validation data and the code with

[2] The remaining 5 percent is composed of design patents, plant patents, reissued patents and statutory invention registration (SIR).

[3] Grobid (2008-2020) https://github.com/kermitt2/grobid

[4] See http://patcit.io for the project documentation.

a view of ensuring replicability and follow-on improvements by the community.[5]

The remainder of the document is organized as follows. Section 1.2 discusses the nature of in-text citations. Section 1.3 sets forth the processing pipeline and provides technical details about the methods. Section 1.4 describes our validation procedure and reports performance measures for various critical steps of the data pipeline. Section 1.5 offers a quantitative overview of in-text citation data and compares them with front-page citations. Section 1.6 concludes.

1.2 The epistemology of in-text citations

This section describes the characteristics of in-text patent citations, with a particular focus on how they differ from 'traditional' patent citations reported on the front page of patent documents.

There are three patentability requirements enshrined in U.S. patent law that give rise to in-text citations to all types of prior art: to fulfil *enablement* requirements; to make arguments for *novelty and non-obviousness*; and to make arguments for *usefulness*. As these justifications for adding in-text citations do *not* perfectly overlap with those that generate front page citations, in-text citations contain truly novel information over and above that reflected in front-page citations. Further, we suggest that this novel information is likely to be associated with inventor input into the drafting process and, therefore, knowledge flows (Bryan et al., 2020). For a similar reason, we argue that in-text patent citations provide a valuable signal of patent importance.

1.2.1 A legal perspective on in-text patent citations

The justifications above relate to specific legal obligations that an applicant must fulfil in order for their application to be deemed patentable. While *novelty and non-obviousness* are usually judged by the examiner using direct comparison to the prior art, *enablement* and *usefulness* are also necessary for patentability and are primarily argued by the applicant in the detailed description of the patent application. Appendix 1.9 gives real examples of citations in each of these contexts.

Enablement is necessary due to 35 U.S Code § 112, which explicitly states:

> *"The specification shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same, and shall set forth the best mode contemplated by the inventor or joint inventor of carrying out the invention."*

[5] https://github.com/cverluise/PatCit/tree/0.3.0

The enablement requirement is core to the modern conception of a government-issued patent. It ensures that when a patent falls into the public domain, others can (in theory) replicate and use the invention after reading the information in the patent description. Prior art citations may be incorporated by reference where appropriate and can make this description much more succinct; if the construction or use of an invention relies on previously patented or published information, the applicant may reference this in the text of the patent specification.[6] These kinds of citations are not necessarily material to the invention's patentability and, when this is the case, not required to be disclosed by the applicant via an information disclosure statement. As such, these 'enablement' citations are not necessarily duplicated on the front page of the patent document. This is particularly true of citations accompanying specific examples that describe how the invention may be used in practice ('best modes'), which may be complementary (and not necessarily similar) to the invention described and may even be hypothetical (Freilich, 2019).

The *novelty and non-obviousness* requirements depend crucially on prior art.[7] For the most part, they are argued for implicitly through Information Disclosure Statements submitted by the applicant throughout the application and patent prosecution processes—these are the citations that appear on the front page of a patent.[8] However, the applicant can also make these arguments explicitly in the patent text by pointing out shortcomings of, or distinctions from, the most pertinent prior art, accompanied by citations to this art. As such, one may expect that citations intended to bolster an argument for novelty or non-obviousness would be duplicated on the front page.

Usefulness, perhaps the most subjective requirement, is described in 35 U.S. Code § 101. It requires the described invention to be 'new and useful' to be patentable. The first part of this clause is covered by the novelty and non-obviousness requirements described above. However, the second (usually referred to as the 'utility' requirement) requires the invention to be useful to the public as described and, as such, may overlap with *enablement* requirements. The word 'useful' is particularly open to interpretation, but generally requires the patented invention to work, and is something that people may want or need (Machin, 1999). In the former case, while there is no burden on the applicant to prove that the invention works (Cotropia, 2009), citations may be added to allay doubts that, for example, a claimed function of the invention is physically possible. The latter is unlikely to be questioned by an examiner (Machin, 1999).

1.2.2 In-text patent citations as valuable paper trails of knowledge flows

Applicants add in-text citations (to both patents and other bibliographic sources) on their patents for several reasons, necessitated by patentability requirements laid out in U.S. law, as

[6] 37 CFR 1.57
[7] 35 USC 102; 35 USC 103
[8] 37 CFR 1.56

discussed above. Some of these reasons overlap with those that require applicants to submit Information Disclosure Statements, the prior art listed on which often reach the front page of a granted patent. However, some prior art, and particularly those items deemed necessary to meet enablement or usefulness requirements, do not need to be submitted to the patent office in the form of an Information Disclosure Statement because they do not directly limit the scope of the claims in the patent application. Further, examiners do not need specific pieces of the prior art to justify a rejection under the enablement or usefulness requirements.[9] Therefore, the front-page will not contain in-text citations added for these purposes (unless, of course, they are also relevant for the assessment of novelty and non-obviousness).[10]

Due to their resemblance to citations in academic articles, it is tempting to assume that in-text citations are more likely than front-page citations to have been added by the people directly involved in the discovery process, namely the inventors. We suggest that this is probably true, for two reasons. First, the in-text citations that are duplicated on the front page, as prior art material to patentability, are likely the most relevant pieces of prior art against which the invention needs to be judged as novel and non-obvious. The fact that these citations are also in the patent description would imply that they either fulfilled multiple requirements, or were so technologically close to the citing patent that applicants need to make explicit arguments for novelty in the description with reference to specific items in the prior art (see Appendix 1.9). In either case, the inventor was likely aware of this prior art during the invention process.

Second, those citations that are *not* duplicated on the front page are most likely included to address the enablement or usefulness requirements. While utility is often assumed, and rejections based on lack of utility are rare for most technology types (providing little incentive to add citations; Chien and Wu, 2018), the enablement requirement states that a 'person skilled in the art' should be able to make and use the invention, and applicants add in-text citations to assist these hypothetical persons.[11] As such, this information was almost certainly necessary during the invention process, and the inventors were, therefore, aware of it. Believing otherwise would come with the implication that it is the *attorneys* who are writing instructions for those 'skilled in the art' and, hence, are at least as skilled as these readers.

Both of the arguments above point towards inventors having more input into selecting in-text citations than they do for front-page citations. For these reasons, we suggest that in-text citations provide a promising measure of knowledge flow.

1.2.3 In-text patent citations as valuable signals of patent importance

In addition to their utility for capturing noisy signals of knowledge flows, researchers have also used front-page forward citations for decades as indicators of technological impact (Carpenter et

[9] Manual of Patent Examining Procedure, Section 2107.02; Manual of Patent Examining Procedure, Section 2164
[10] Manual of Patent Examining Procedure, Section 2120
[11] 35 USC 112

al., 1981; Albert et al., 1991). Even if a particular cited patent was not a real knowledge input, the fact that it appears on the front page means that it is likely to be in the same technological space as the citing patent. As such, a patent receiving many front-page citations is either: useful and frequently reused information for the production of new inventions; in a dense technological space against which many new technologies happen to abut, or; a combination of these. This interpretation of front-page forward citation counts is a consequence of the legal purpose of front-page citations; namely, to delineate the prior art material to the patentability of the citing patent. However, this is not the sole purpose of in-text citations.

In-text citation counts, as described above, also serve to fulfill enablement and utility requirements. Applicants sometimes do so by referring to their own patents; for example, firms producing consumer goods may have patents on multiple complementary inventions that, while not necessarily technologically similar, come together in the final product and are cited to demonstrate how the invention is used in practice. In-text citations are also more likely to come from inventors themselves, perhaps independently from the motives for citing. For these reasons, the interpretation of a patent accumulating a large number of in-text forward citations is more complicated than for front-page citations.

On the one hand, the technologically similar inventions cited in-text are those from which the applicant of the citing patent or application has had to provide additional distinction, and therefore are likely to be those most likely to be justification for rejection. On the other hand, the technologically complementary inventions cited in-text are likely to be more generalizable technologies, as they are not technologically close enough to the citing patent to be considered material to patentability. Sometimes this relationship is made explicit, as indicated in U.S. patent 8,524,730 (emphasis added):

> *"More concretely, examples of the other active ingredients that can be combined with a compound of the invention as different or the same pharmaceutical compositions are shown below, which, however, do not restrict the invention."*

Patents cited in this fashion are not in the same technological space as the citing patent and are cited for their compatibility with other inventions. A large number of these kinds of citations may, therefore, indicate generality outside of the technical domain of the cited invention.

These reasons for making in-text citations color our understanding of how exactly a large number of forward in-text citations relate to the intrinsic properties of the cited patent or invention. However, we know that these citations are more likely to originate with the inventors themselves, rather than the attorneys or examiners. This scenario is an interesting one from the point of view of interpretation. The number of reasons for citing a patent in-text are more numerous than those made on the front page, but the resulting citations (often accompanied by context) are more thought-out and meaningful. As an analogy, if front-page citations were a single radio station plagued by significant and persistent static, in-text citations result from numerous stations broadcasting loud and clear the same frequency, to the point where it is difficult to make out what any individual station is saying. However, some may prefer this to static. The disentangling of these frequencies is undoubtedly possible; with both data and code publicly available, future research can build on this work to add the context to in-text citations and, ultimately, better understand what a highly-cited patent represents in this setting.

1.3 Methods

In this section, we describe the data sources and the different steps of the processing pipeline. We want to provide extensive insights into our technical choices in order to stimulate and enable future extensions or improvements.[12]

1.3.1 Data

The processing pipeline starts with the full-text of 16,781,144 patent documents filed at the U.S. Patent and Trademark Office (USPTO) since 1790.[13] We extracted the full-text data from the IFI CLAIMS dataset, made available by Google Patents as part of its public datasets.[14]

The text we are considering is the specification of the patent. The specification is a written description of the invention and of the manner and process of making and using the invention. It also includes information about related applications and government interest statements (de Rassenfosse et al., 2019a). It does *not* include the patent's claims or the information on the front-page.

The starting point is a long chain of characters without any structure and indication about which characters might refer to a patent citation.

1.3.2 Extraction task

The first step involves identifying the relevant strings of characters referring to a patent citation in the full text. An early attempt to do so dates back to Galibert et al. (2010), who combined a set of regular expressions to identify the cited patent number itself (e.g., country codes followed by a series of digits) based on the neighbouring text (e.g. "herein described by"). A similar approach was implemented by Berkes (2018a) for U.S. patents published before 1947. Although intuitive, these approaches lead to moderately satisfying results. Galibert et al. (2010) report a precision of 64.4 percent, a recall of 61 percent and a f1-score of 62.9 percent while Berkes (2018a) does not report performance metrics. The fundamental reason behind these low scores

[12] Readers who are not specifically versed into technical considerations can skip this section without much harm to their understanding of the nature of the data.

[13] The first extracted citation is in 1846.

[14] https://console.cloud.google.com/marketplace/partners/patents-public-data

is that language is highly variational and there are many ways of citing a patent. On this point, Adams (2010) warned the community about the complexity of the extraction task. Using a random sample of USPTO patents, he found an "alarming" (p. 26) degree of variation in the form of in-text patent citations. In this context, any attempt to use a list of predefined rules is likely to have mixed results and, above all, to lack generalisation.

In order to overcome this limitation, NLP researchers have developed statistical models that can learn to find and tag entities, such as cited patents, using a training set of annotated documents, where a researcher has labeled the presence (or not) of the entities of interest. Although an in-depth presentation of the related Named Entity Recognition (NER) literature is out of the scope of this paper, we summarize the general working principles of these models below.[15]

The key is to see a text as two sequences: a sequence of tokens and a corresponding sequence of latent labels (e.g. "PATCIT" for patent citations versus "O" for other). The task is to predict the sequence of labels. The algorithm is trained on an annotated set of documents, that is, a set of documents for which we know both the sequence of tokens and the sequence of labels. The probability of each token to belong to a given label is a recursive function of the token itself and its features (digits, capital letters, etc), the neighbouring tokens (its context) and the *neighbouring labels*. The overall goal of the algorithm is to predict rightly the full sequence of latent labels for a given sequence of tokens. If a token (or a sequence of tokens) is unknown or deviates from the learning examples, the algorithm can still leverage the other attributes to decide which sequence of labels is the most probable for the whole sentence, leading to a considerable generalization improvement.

For example, let us assume that the algorithm has been trained on a corpus of texts where citations come in the following form (with *d* denoting any digit): "described by patent *d,ddd,ddd*" and where the corresponding sequence of labels is [O, O, O, PATCIT]. Let us further assume that the algorithm is supplied a new text with a slightly different form of citation such as "described by Pat 9,535,657". Although the algorithm has never seen the token "Pat", it has learnt from the training data that the sequence of token "described by" frequently precedes a PATCIT label by two tokens. Combined with the fact that the token "9,535,657" exhibits the features frequently associated with a PATCIT (digits and commas), then the algorithm is expected to override the absence of the "patent" token and still to predict the right sequence of labels, [O, O, O, PATCIT].

The aforementioned limitations and improvement opportunities have been well identified by the machine learning community in the second half of the 2000s. In particular, Lopez (2010) developed the Grobid library in 2008 (and has been continuously improving it since then) with the goal of overcoming the limitations of a "rule-based" approach using a statistical approach. Grobid has now become an open source project leveraging modern NLP to efficiently struc-

[15] See Li et al. (2020) for a recent survey of this literature.

ture scientific documents in general, but retains a specific focus on patents. It includes models trained at extracting and structuring bibliographical references (scientific articles, books, proceedings, etc.) and patents from full-text documents. The algorithmic backbone of Grobid is the Conditional Random Fields (CRF) model. This model belongs to the family of sequence labeling models described above and was first introduced in 2001 (Lafferty et al., 2001). The CRF model has been widely used in various fields and applications.[16]

Grobid's patent citations' extraction model was originally trained on 200 annotated full-text patents.[17] This training set included 62 percent of EPO patents, 19 percent of WIPO patents and the remaining 19 percent of USPTO patents. As for the rest of Grobid models, the patent extraction model is a CRF model. The specific features entering the CRF model to support patent citation detection include the relative position of the current token in the document, the matching of a common country code (e.g., US, EP, WO, etc) and the matching of a common kind code (e.g., A1, A2, B1, B2, etc).

The output of the extraction tasks is a set of text spans that were tagged as patent citations (e.g., "United States Patent 9,535,657"). The information extracted at this stage is not structured and, therefore, improper for researchers.

1.3.3 Parsing task

The next step involves parsing the extracted patent citation strings. We take the raw span of the extracted citation as an input, with the goal of obtaining the following normalized attributes: the country code of the patent authority, the patent number and the type of the patent. This task is challenging due to the many forms in which patent citations occur in the text. Typically, the patent authority can appear as a code or a name (e.g "US Patent 9,535,657" or "United States Patent 9,535,657") either immediately next to the patent number or relatively far from it (e.g., "US Patent number 9,535,657" or "US Patents 9,911,050, 9,607,328, 9,535,657").

Lopez (2010) proposes an efficient solution for tackling this task. The fundamental idea is that both the sets of possible inputs and outputs for each patent attribute are finite (e.g., the list of patent organisation names and the list of their codes respectively). In addition, each element of the input vocabulary should be mapped with a unique element of the output vocabulary (e.g. "United States" with "US" or "European Patent Office" with "EP"). In the end, for any given patent attribute, the parsing operation can be thought of as a translation operation between two languages with a finite vocabulary. If this still seems a bit abstract, the reader can simply consider that the aforementioned task consists in regular expression matching followed by string rewriting.[18] This task perfectly fits the usage of Finite State Transducers (FST) which appeared

[16] See Sutton and McCallum (2006) for a survey.

[17] The training set was enriched since that time and now includes 270 patents, including 51 percent of EPO patents, 33 percent of WIPO patents and the remaining 26 percent of USPTO patents.

[18] Let us assume that we are interested in the organisation attribute and that we have extracted the following span "United States Patent 9,535,657". This span would trigger a match for "United States" which would then

early in the history of automated translation.[19] Importantly, FSTs have been developed with computational efficiency in mind in the early ages of computer science, making them highly efficient in todays' context.

The output of this task is a well-structured set of attributes describing the cited patent.

1.3.4 Consolidation task

The final task consists in matching each extracted patent citation to a unique and consolidated identifier, in order to connect each cited patent document to commonly used patent dataset. For patents, the identifier common to most (if not all) patent datasets is the DOCDB publication number.[20] On this point, note that we depart from Grobid which relies on the European Patent Office (EPO) search API[21] to perform the matching process and uses the EPO document number as its target and consolidation device.

Unfortunately, in a large majority of cases, in-text patent citations do not report the kind code of a patent, or report the original patent number rather than the version used in the DOCDB publication number, making it impossible to assemble the DOCDB publication number using the parsed attributes only. In order to overcome this limitation, we have relied on the Google Patents Linking Application Programming Interface (API).[22] Taking various kinds of inputs, such as the patent office code, the patent number and kind code, the API returns the associated DOCDB publication number. At a high level, the internal mechanism of this service is the following.[23] First, a large number of variations of each publication number was generated. For each variation, the original patent office and DOCDB formatted versions were indexed. Variations include adding and removing 0 padding, two and four digit year dates inside of patent number, Japanese emperor year variants and different combinations of country code, patent number and kind code. Altogether, these variations constitute a large lookup table linking many variations of a publication number to its DOCDB formatted version. Then, at the time of lookup, punctuation is stripped and the country code, number and kind code are searched for before being used to look-up for matches in the large variation table. Note that there are two distinct services, one for applications and one for patents.[24] We decide which one to call based on the status attribute parsed by Grobid which can take four values: "application", "provisional", "patent" and "reissued". The first two trigger the application service, while the last two trigger the patent service.

be rewritten as "US".

[19] See Roche and Schabes (1997) for an in-depth review of Finite State Transducers.

[20] For the sake of simplicity, we use the "publication number" terminology for both the publication number (for published patents) and the application number (for patent applications).

[21] http://v3.espacenet.com/publicationDetails/biblio

[22] https://patents.google.com/api/match

[23] We thank Ian Wetherbee from Google Patents for this explanation.

[24] Applications: https://patents.google.com/api/match?appnum
Patents: https://patents.google.com/api/match?pubnum

Using the unique publication number returned by the Google Patents Linking API we were able to connect each cited document with richer information from patent datasets generally used by researchers (e.g., PATSTAT, PatentsView, IFI CLAIMS, etc.). We enriched each cited patent with the following attributes: publication date, application identifier, patent publication identifier, INPADOC and DOCDB family identifiers.

1.3.5 Pipeline

Let us illustrate the process using an example. Consider the following excerpt from the description of US-9606907-B2, which cites two U.S. patents:

> "Examples of circuits which can serve as the control circuit ... are described in more detail by U.S. Pat. Nos. 7,289,386 and 7,532,537, each of which is incorporated in its entirety by reference herein."

After the Grobid processing, we know that the patent US-9606907-B2 cites two patents from the U.S. patent office ("US" patent authority code) and that their original numbers are 7,289,386 and 7,532,537. Using the Google Patents Linking API, we find that the two patent citations embedded in the text can be uniquely identified by their publication numbers, namely US-7532537-B2 and US-7289386-B2.

The above pipeline was deployed remotely on a large-size compute engine from Amazon Web Services.[25] In order to increase speed, we used multi-processing, a technique consisting in running multiple processes in parallel at the same time. This technique is especially useful for 'cpu bound' rather than 'io bound' operations, that is when computation is the main limiting factor, not internal communication. Processing documents at an average pace of 400,000 to 500,000 per day, this operation took us approximately one month for a total cost of about 120 USD.[26] Overall, from the 16,781,144 patent documents that we processed, we were able to extract 63,854,733 in-text patent citations. These citations point to 13,611,323 unique patent documents. We matched 49,409,629 of the extracted in-text cited documents with a publication number.

1.4 Technical validation

In order to assess the quality of the citation dataset, we undertook a thorough validation exercise of the data and the extraction, parsing and matching tasks. To do so, we relied on Prodigy,

[25] We used a t2.xlarge (4 cores and 16Gb of Ram) located in the "USA East Ohio" computing zone.

[26] Note that we simultaneously extracted in-text Non Patent Literature citations (scientific articles, books, proceedings, etc.) and tried to match them with Crossref at runtime. To do so, we used biblio-glutton, a high performance bibliographic reference matching service, and an ElasticSearch index hosted on a separate engine. It appeared that the major processing speed limitation came from ElasticSearch queries. Processing only in-text patent citations would certainly take significantly less time and resources.

a scriptable annotation tool.[27] Lopez (2010) reports performance metrics for all these tasks, however the set of documents we are considering partly differs from the corpus he used. In particular, a significant part of the patents in our corpus is much older than any document considered for Grobid training and evaluation. We also carried out detailed error analyses as a way to support future improvement efforts.

1.4.1 Data consistency

USPTO patent documents' format and the quality of the scanned document (for older patents) has changed throughout the years. Before 1971, patents were largely unstructured with no clear delimitation between the metadata and the specification text itself (see Figure 1.1). The modern patent format was introduced in 1971 and progressively replaced the old format before becoming the unique format after 1976. This format is semi-structured and clearly distinguishes between the metadata sections and the specification section *inter alia* (see Figure 1.2). These specificities of the source data have some notable implications on our output data.

First, the text of patents published in the old format includes the header of the patent. The header summarizes the main attributes of the patent, including its technological classes, title and most importantly its number. In this case, the extraction algorithm is likely to extract a patent citation which does not correspond to the kind of object we are looking for. Fortunately, this specific pitfall is relatively easy to spot as the citation appears very early in the text. Figure 1.3 reports the distribution of the rank of the first character of the extracted citations before and after 1976. We observe a clear excess mass between 0 and 50 characters before 1976. Building on this observation, we focused on the corpus of patents published before 1971 and randomly drew 50 citations starting before character 50. Confirming our doubts, we found that 88 percent were self references, 8 percent were technological classes and 4 percent were dates. In this context, we chose to flag all citations detected in a patent published before 1976 and starting before character 50 to make it easy to exclude them from analysis.

Second, in the old format, what we now call 'front-page citations' were printed *after* the patent specification, and these are also sometimes mistakenly included in our source data as part of the full-text of the patent. Since all patents have a different number of characters, looking at the distribution of citations by starting character does not make sense. However, we can still look at the relative place of the detected citations. Figure 1.4 shows their distribution as a function of their relative place (expressed in percentile) in the full text. Comparing the distribution before and after 1976 reveals a sizable excess mass for the pre-1976 distribution in the last four percent of the full-text characters. Additionally, looking at a random sample of 100 citations extracted from patents published before 1976 and occurring in the last four percent of the characters, we find that 99 percent belong to the 'front page' citations section. Hence, for patents published

[27] Prodigy (2018-2020) https://prodi.gy/.

before 1976, we chose to flag all citations detected in the last 4 percent of the full-text and encourage the user to exclude them from their analysis.

Third, during the transition period between the old and modern formats (approximately throughout 1971–1975) there were two patent formats in use, complicating the delineation of the specification text section during this time period. As a result, we observed that 'full-texts' from this time, mistakenly include the front-page of patents that are in the modern format. This can lead to the incidental extraction of 'in-text' citations that are actually front-matter, including front-page citations and references to the patent itself (including priority filings). Unfortunately, there is no straightforward solution to this problem. We encourage data users to systematically ignore patents that are both in text and front page citations during this time span.

All figures reported above and below exclude flagged patent citations as they are most likely not to correspond to real in-text patent citations (unless explicitly specified).

1.4.2 Extraction task

Lopez (2010) reports high performance metrics for the extraction task. Using cross-validation, a technique consisting in training the model ten times using 80 percent of the sample and testing it on the remaining 20 percent, the author reports the following average performance metrics: 94.66 percent of precision, 96.16 percent of recall and a f1-score of 95.4 percent. As far as we know, these are the best performances reported in the literature to date. Although this motivated our choice to use Grobid, we are fully aware that our dataset partly differs from the Grobid training set and performance could thus be affected.

In order to evaluate the quality of the extraction in our specific case, we randomly sampled 160 U.S. patents and annotated them by hand. As previously discussed, the citation of a patent can come in various ways. For instance, the country of the patent office can be reported as a code preceding the patent number, as a name anywhere in the surrounding of the patent number, etc. In this context, the only stable element of a patent citation is the patent number itself. That is why Grobid returns the first and the last character of the patent number of detected patent citations. Hence, our validation exercise consisted in comparing the spans detected by Grobid as a patent number and the spans labelled by humans as a patent number. Each patent was annotated by a single human annotator using the platform featured by Figure 1.5a.[28] The body of the text is displayed together with annotations from Grobid predictions and the annotator goes through the text to correct missing and wrong annotations. The tagged spans are saved upon exit.

As depicted by Figure 1.6, the validation sample and the universe of citing patents display very similar distributions by publication year.

[28] "Human annotators" are not undergraduate RAs but coauthors of this paper.

From the 160 random U.S. patents in the validation set, human annotators found that 103 (64.4 percent) patents cited at least one patent for a total of 470 in-text patent citations. Table 1.3 reports the extraction performance that we obtained together with the Galibert et al. (2010) and Galibert et al. (2010) benchmarks. Comparing 'gold' annotations from human annotators with the predictions obtained from Grobid, we find that Grobid exhibits a satisfying 97 percent precision and 82 percent recall (f1-score nearing 90 percent). Importantly, these results largely outperform Galibert et al. (2010), who used regular expressions for the same patent citations extraction. They reported a precision of 64.4 percent and a recall of 61 percent. This result clearly confirms that a statistical approach to in-text citation extraction is much more relevant than a regular expression approach. Interestingly, the performance obtained by Grobid on our extended corpus is very similar to the benchmark reported by Lopez (2010) regarding precision (97.44% vs 97%) but lower in terms of recall (97.74% vs. 82%). This difference means that, applied to our extended corpus, Grobid is as reliable as reported in Lopez (2010) when it has detected a patent citation. However, it misses patent citations more often in our extended corpus due to older forms of citations appearing in early-twentieth century patents.

The error analysis suggests that both false positives and false negatives exhibit patterns that could be specifically addressed by future improvements of the Grobid training set. Table 1.4 provides examples for each category of errors that we were able to identify. Starting with false negatives, that is patent citations that were not detected by Grobid, we find three categories of context generating this type of errors: 1) the context does not clearly mention "patent" or "application" but rather implicitly suggests a patent citation; 2) the patent is cited in the form "inventor (date) <PATCIT>" and 3) the patent is cited as "Serial Number <PATCIT>". While category 1) could have been expected and would certainly be hard to correct without generating a large number of false positives, categories 2) and 3) might certainly be partly addressed by augmenting the training dataset with older patents that tend to adopt this form of citations more often. Now, looking at false positives, that is text spans that were wrongly identified by Grobid as patent citations, we can find three categories of errors as well: 1) technological classes reported as "dd/ddd", 2) date and 3) docket number. Note that the categories 2 and 3 have only one occurrence each.

1.4.3 Parsing task

Grobid FST was built manually based on 1,500 patent citation examples. It was then evaluated on 250 references which were unseen before. Lopez (2010) reports a 97.2 percent accuracy for the full parsing task (patent organisation code, number and kind code). Once again, we thought that it was important to confront those results with our specific dataset.

In order to validate the quality of the parsing, we randomly sampled 300 extracted citations with their parsed attributes. As already discussed, the attributes can be relatively far from the patent number that serves as the citation anchor. Hence, it was necessary to provide the human annotators with a contextualized citation. In practice, using the patent number reported by

Grobid as an anchor, we extracted a chunk of text containing a window of ten tokens on the right and left of the detected patent. This text and the tagged patent were then displayed to the annotator together with the Grobid parsed attribute as illustrated by Figure 1.5b. The annotator would then accept or reject the attribute depending on what he actually found in the text. Each example was validated by a single annotator whose decisions were saved upon exit.

Lopez (2010) reports an *overall* 97.2 percent accuracy. Since the attributes can be used independently, we believe that a detailed understanding of the performance and errors for each attribute is also valuable for the community. Hence, we performed three distinct validation exercises, one for each attribute. Our results are summarized in Table 1.6.

Considering the parsing of the patent organisation, we first checked for sample representativity. Table 1.5 reports the distribution of the patent organisations in the validation sample. It appears that two-thirds of the citations in the sample were mapped to the U.S. patent office. This result is very much in line with the results that we report on the full dataset (see Section 1.5). Similarly, the patent organisations in the remaining third of the validation sample are also the most represented organisations at scale, including the Japanese Patent Office, the World Intellectual Property Organisation, the European Patent Office and the German Patent Office *inter alia*. On the 300 examples that we validated, we found only five errors, leading to a 98.3 percent accuracy score. Errors spread over five distinct patent offices and we do not observe any systematic confusion between patent offices, which suggest that errors generate noise rather than a systematic bias.[29]

When it comes to the parsing of the patent number, there is no specific way of checking sample representativeness. Overall, on the 300 examples that we validated, we found thirteen errors, leading to a 95.7 percent accuracy score. Among the errors, we find two recurring cases. First, patent citations in their Paris Cooperation Treaty (PCT) form (e.g., PCT/EP2005/008238) generate patent numbers mixing part of the letters in the prefix and the patent number itself (e.g., PTEP2005008238). Second, as already reported in Lopez (2010), we found that Grobid removes the first letter of the patent number of Japanese applications with date prior to 2000 (e.g., H08-193210 where H stands for the Heisei era that spanned from 1989 to 2019). However, this indication is key to uniquely identify the application. This letter refers to the era and acts as the time marker. Note that this specific issue is partly fixed by the Google Patent matching API as explained in Section 1.3.

Lastly, we validated the parsing of the so-called kind code, that is the code indicating the specific kind of document the citation refers to (granted patent, application, reissue, design, etc.). Over the 502 random examples, we obtain an accuracy of 97.6 percent. Note, however, that this measure includes a large proportion of null results as the kind code is in fact rarely reported in the text. In order to further characterize the quality of the parsing, we drew a

[29] The five offices are: SA (Saudi Arabia), AL (Albania), CH (Switzerland), DE (Germany) and BE (Belgium).

sample of 50 citations where the parsed kind code was not null. We found 7 mistakes, meaning a ‘conditional’ accuracy of 86 percent. Specifically, we found three groups of parsing errors: errors due to unconventional formatting, OCR issues and Grobid mistakenly interpreting ‘Cl’ (class abbreviation) for the ‘C’ kind code. Importantly, every instance in standard form was correctly parsed.

1.4.4 Matching task

The matching task involves associating the extracted attributes with a unique identifier, which is the DOCDB publication number in our case. In order to validate this step of the process, we randomly sampled 200 citations from our final dataset and we compared the concatenation of the parsed attributes with the publication number provided by the Google Patent's Linking API. The annotator's task was to answer the following questions: i) if there is a matched publication number, is it the right one? ii) if there is no match, would it be possible to find one for a human reasonably well trained in the task? A single human annotator fulfilled this validation exercise. Based on that, we can assign each annotated example to a standard classification outcome category and derive the associated performance metrics. Table 1.7 summarizes these categories, their contents and the results from the validation exercise.

On the 200 examples in the validation sample, we find that 147 were matched and 53 remained unmatched. Among the 147 matches, 137 were correct (True positives) and 10 were incorrect (False positives) including six patents that could have been matched and four non-patent items that should not have been matched. Among the 53 unmatched examples, we found that 17 could have been matched (False negatives) while no match could be found for the remaining 36 (True negatives). Overall, we find that the matching procedure achieves a 93.2 percent precision and a 88.96 percent recall, leading to a 91.06 percent f1-score.

Next, we delved into the nature of the errors and non-matches. Tables 1.8 and 1.9 respectively detail errors occurring during matching and cases classified as unmatchable by the human annotator. We find that errors arising at this final step of the processing pipeline are partly inherited from upstream steps. Among the ten incorrect matches, half are due to either a parsing error or an extraction error. In the same way, among the thirty-six unmatched citations that were judged unmatchable, 56 percent were directly related to either a parsing error or an extraction error. Another group of errors seems to arise from the specificities of in-text citations and their intrinsic ambiguities. This group includes citations of provisional patent applications (which might well never appear in standard patent datasets) and partial citations that even a human cannot match.[30] This family of errors represent 41 percent of the thirty-six unmatchable detected citations in our validation sample. Eventually, focusing on the unmatched citations

[30] A provisional application is a legal document filed at the patent office that establishes an early filing date, but does not mature into an issued patent unless the applicant files a regular non-provisional patent application within one year.

that a human can match reveals some blind spots of the Linking API. Over the seventeen cases in this category, 52 percent are caused by missing zeros after the country code/year or a Japanese publication number reporting the year after the serial number rather than before it as is usually expected.

While the previous step can characterize the performance of the matching procedure with high precision, due to the small size of the validation sample it cannot uncover rare irregularities that might still be of sizable magnitude at large scale.

Considering the full dataset, Figure 1.7 show, the yearly number (1.7a) and share (1.7b) of citing patents according to the matching status of the extracted in-text citations.[31] Patents with all in-text citations matched to a publication number represent 42.7 percent of the total, whereas those with only some in-text citations matched represent 32.7 percent. Patents with no in-text citations matched account for the remaining 24.6 percent. From 1947 to 1964, patents with all in-text citations matched report an increasing yearly share, from around 40 percent to almost 70 percent. For patents published between 1965 to 1975, the performance of our matching procedure worsens, as the proportion of patents with only some citations matched or no citation matched grows. From 1976 onwards, the share of patents with all citations matched returns to be the largest (around 77 percent in 1976), although it progressively decreases for patents published during the following years in our dataset.

These aggregate figures mask high variation depending on the patent office of the cited patent documents. Table 1.10 reports the number of extracted in-text citations and the number and relative share of matched citations for the top five patent offices in our dataset. More than half of in-text citations are made to patents filed at the USPTO (about 58% of the total). We are able to match 89 percent of them to their correct publication number. Patents filed at the World Intellectual Patent organisation (WIPO) and the Japan Patent Office (JPO), with respectively around 6.5 millions (10% of the total) and 5.7 millions (9% of the total) citations are the second and third largest groups. We match almost 82 percent of the citations to WIPO patent filings and around 77 percent to JPO patent filings. We obtain a similar match rate (i.e., 73%) for citations to patents filed at the German Patent and Trade Mark Office (DPMA), around 1.4 million of extracted citations. We obtain less satisfactory match rates for citations to EPO patent filings. They are 2.2 millions and we match only 51 percent of them.

1.5 A first look into in-text citation data

Front-page patent citations have been extensively used over the past decades and multiple studies have assessed their validity as indicators and discussed their pitfalls. As far as we can ascertain, we are the first to introduce a consistent and validated dataset of in-text patent

[31] We consider only citing patents with at least one extracted in-text citation.

citations covering all U.S. patents. The purpose of this section is to provide an overview of the characteristics of in-text citations as compared to 'traditional' front-page citations.

We find that in-text and front-page patent citations are two largely distinct sets. We also find that in-text citations are semantically and technologically more similar to the citing patents than their front-page counterparts. This result suggests that in-text patent citations might be a better proxy for knowledge flows than front-page citations, as argued in Section 1.2. We report that the forward citations counts obtained from the front page and the in-text citations are only weakly correlated. Additionally, we find that in-text citations are more internationalized and reveal a higher degree of self reliance. Table 1.11 summarizes the key figures of the section. We use our dataset for in-text citations and the IFI CLAIMS dataset for front-page citations. Unless specified, we consider all U.S. patents published from 1790 to 2018.

1.5.1 Order of magnitudes

From the 16,781,144 U.S. patents in our dataset, we find that 9,453,181 U.S. patents cite at least one patent in the body of their description, corresponding to 56.3 percent of all U.S. patents. Looking at the same set of patents, we observe that 11,923,551 patents (71.3% of the total) exhibit at least one front-page patent citation. In-text citations exhibit high variability over time. The share of U.S. patents citing at least one patent has increased from less than five percent in the second half of the nineteenth century to 70 percent in the 2010s.

When we consider the total number of citations, we find that the number of in-text citations reaches one-third of the front-page citations. We extracted 63,854,733 in-text patent citations while the total number of front-page citations listed by U.S. patents during the same period amounts to 203,557,2015. On average, the body of a patent contains 3.8 patent citations, 6.7 patent citations conditional on citing at least one patent. Once again, there is high variability over time, from less than one in-text patent citation until the early 1960s to more than five since the beginning of the twenty-first century (unconditional on having at least one in-text patent citation).

1.5.2 Overlap between in-text and front-page patent citations

A natural question is how large is the overlap between in-text and front-page patent citations. To answer it, we list all unique pairs of citing-cited patents, called 'citations' thereafter, for both in-text and front-page citations. Comparing the two lists of citations yields three exclusive and exhaustive sets: citations appearing in the text only, citations observed on the front page only, and citations recorded in both.

We find that citing-cited patent pairs resulting from in-text and front page citations are largely exclusive from one another. Figure 1.8 depicts the number of patent citations appearing in the text only, on the front page only, and in both. There are 11,868,037 patent citations appearing both in the text and on the front page, which represents only 5.79 percent of all front-page citations and 24.2 percent of all in-text citations.[32] Note also that, before 1947, front-page patent citations did not exist: before that date, all patent-to-patent citations were available only in-text. In the end, over the whole period, considering in-text patent citations adds 37,541,592 citations that are not found among front-page citations. Focusing only on front-page patent citations leads to missing 15.34 percent of all patent citations. That is what we call *the missing 15 percent of patent citations*.

A potential explanation of the missing 15 percent of patent citations could be patents cited as a "Translation of Patent . . . " or as "Patent Abstracts . . . ". These references are listed in the front page as part of the non-patent literature (NPL). A legitimate question, therefore, is whether these missing 15 percent are (at least partly) available as front-page NPL. If true, extracting in-text patent citations would not bring more information than parsing 'patent' citations reported in the front page NPL section, a much simpler task. To delve deeper into this question, we trained a text classifier to determine whether a front-page NPL citation contains information on a patent. This classifier achieves a sufficient 78.31 percent precision and 89.04 percent recall on the test set. We then applied it to the universe of front-page NPL citations recorded in the DOCDB database. In the end, we estimate that there are 1,714,260 such 'patent' citations reported in the front page NPL sections of U.S. patents since 1947. Making the bold assumption that all these citations appear in the text as well, these patent citations would reduce the missing patents to 14.63 percent (-0.71 percentage points).

It is now clear that in-text and front-page patent citations exhibit very little overlap. Thus, quantitatively, considering in-text patent citations does bring new information. Next, we try to understand whether and how their qualitative characteristics differ.

1.5.3 Textual similarity between citing and cited patents

Figure 1.9 shows distributions of semantic similarity between citing and cited documents for in-text and front-page citations. Semantic similarity is calculated as the dot product of Google Patent's document embedding vectors, which were recently made available to researchers.[33] The embeddings are trained to predict CPC categories from each patent's full-text with a WSABIE algorithm (Weston et al., 2011). Figure 1.9 also shows two reference semantic similarity distributions. The first one ('Within art unit') is based on the similarity between randomly chosen pairs of patents examined by the same art unit. The second one ('Random') is based on the similarity between cited in-text patents matched to a random citing patent.

To produce Figure 1.9 we considered only citing and cited patents that were granted by the

[32] These figures include only in-text citations which were matched with a standard publication number.

[33] https://cloud.google.com/blog/products/gcp/google-patents-public-datasets-connecting-public-paid-and-private-patent-data. Note that a myriad of similarity measures exist, including Younge and Kuhn (2016); Arts et al. (2018, 2020).

USPTO in the years 2000–2009 (for ease of interpretation). In Figure 1.9a we removed all within-INPADOC-family citations occurring for in-text and front-page citations (N=325,247). Pairs of patents used for the 'Within art unit' and the 'Random' distributions have been randomly omitted to match this sample size. INPADOC families, also known as 'extended patent families', include all patents that can be linked through their priorities (but not necessarily to a single common priority filing).[34] Citations between patents belonging to the same INPADOC family are much more common in the patent text than on the front page, and removing them improves the comparability of the similarity distributions. In Figure 1.9b we report the same similarity distributions, excluding citations between patents belonging to the same DOCDB family. These families, also known as 'simple patent families,' consist of sets of patents linked to a common priority filing. They are smaller and more selective than INPADOC families. The in-text citation similarity distribution shown in 1.9b clearly includes many near-identical patents, owing to the complexity of priority filing strategies. For this reason, we will focus on the distributions excluding within-INPADOC-family citations (Figure 1.9a), as they are more comparable to front-page citations.

One can make a number of observations from this graphical comparison of similarities. First, in agreement with our validation measures, there are unlikely to be a large portion of in-text citations that are incorrectly matched, as these would be drawn from the random distribution. Indeed, because we cannot see a conspicuous lump in the in-text similarity distribution in the region where the random distribution peaks and because the shape is similar to that of the front-page citations, we may conclude that the error rates in these two sets of citations are roughly similar. Second, the in-text citation distribution is shifted to slightly higher levels of similarity when compared to the distribution for front-page citations. This shift indicates that patents cited in-text are, on average, more technologically similar to the citing patent than patents cited on the front page. Lastly, the in-text citation distribution displays a fatter tail at lower similarity levels, particularly around the similarity level expected from patents examined by the same art unit. This pattern is expected. Because patents cited in the patent text do not necessarily impact on patentability and do not have to be technologically similar to serve their purpose, they are drawn from a wider (but still related) set of prior art.

This evidence reinforces our view of in-text citations as a promising indicator of knowledge flows, potentially less "noisy" than front-page ones (Jaffe et al., 1998; Corsino et al., 2019; Kuhn et al., 2020) and more closely related to the focal inventors' prior knowledge, less affected by the complex patent examination procedure (Choudhury et al., 2020) through examiners' (Alcacer and Gittelman, 2006) or patent attorneys' practices (Jaffe et al., 2000).[35]

[34] https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/patent-families/inpadoc.html

[35] Interestingly, this result also suggests that the tracking of long distance technological relationships by in-text patent citations (hypothesis formulated in section 1.2.1) is only of second order (at most).

1.5.4 Forward citations

The count of forward citations, that is, the number of times a patent is cited by another patent, has been widely used in various contexts as a way to measure the quality of a patent, but also as an output measure in settings where innovation or knowledge flows are susceptible to be affected by another economic variable (see Jaffe et al., 1993; Almeida, 1996; Kerr, 2008; Agarwal et al., 2009 *inter alia*). Due to the large interest of the community for this forward citations count and its central role in innovation research, it seems natural to use our dataset to compute the forward citations count based on in-text citations rather than the usual front page citations. Further, we compare the forward citations counts obtained using in-text citations and front page citations.

To do so, we consider U.S. patents and their in-text and front-page citations. We slightly depart from raw forward citations counts in two ways. First, in order to make our results immune to potential variations between in-text and front-page citing patterns, we compute the forward citations count at the invention level, as defined by the DOCDB family, rather than at the publication level.[36] Second, we exclude citations from patents belonging to the same extended invention family as defined by the INPADOC family. This is a conservative choice aiming at excluding self-references in a broad sense.

The first observation is that in-text and front page citations are directed to two partly disjoint sets of DOCDB patent families. In-text citations point to 5,506,374 distinct families, front page citations point to 13,817,609 distinct families and 4,262,548 of these families are in both sets. This result further confirms the fact that in-text citations bring additional and distinct information from their front page counterparts.

The second observation is certainly the most important and puzzling one: the count of forward citations based on in-text citations is only weakly related to the same metric obtained from front-page citations. Restricting to the set of DOCDB patent families with a positive count of forward citations both on the front page and in the text, the correlation between the two measures is 0.23. Figure 1.10 shows these two forward citations count for a random sample of 10 percent of the patent families cited both in the text and on the front page. The regression line corresponds to a univariate model where the dependent variable is the front page forward citations count and the in-text count is the independent variable. The associated R-squared is close to zero (0.03), highlighting the poor predictive power of the dependent variable over the independent variable. Roughly speaking, the two forward citations counts are almost orthogonal.[37] The above result is puzzling and raises a host of questions about the use of in-text forward citations count.

The third set of observations relates to the distribution of forward citations counts. Figure 1.11

[36] The kind of pitfall that we want to avoid is, for example, a higher tendency to cite applications in the text instead of granted patents on the front page and vice versa, including for the exact same invention.

[37] This result is robust to considering only patents with more than five in-text and front-page citations.

compares the empirical probability (panel 1.11a) and cumulative (panel 1.11b) distribution function of forward citations counts. It reveals two notable properties. First, the front page distribution stochastically dominates the in-text distribution. Second, the tail of the front page distribution is larger. These observations can be partly explained by a fundamental difference in the citation generating process. Whereas in-text citations are mostly in the hand of the inventors, hence decentralized among many agents, front page citations are determined by a finite number of examiners who, by nature, are likely to be aware of a limited number of patents on each subject. This leads to the emergence of highly cited patents, the so-called 'focal patent,' which participate in the larger tail observed in the distribution of front page forward citations count. It is interesting to note that 'focal patents' might well be so partly independently on their intrinsic social or private value but because of examiners' biases.

Among our results, the orthogonality puzzle is certainly the most challenging to grasp for the community.

1.5.5 The internationalization of patent citations

International patent citations, that is, citations to and from patents granted at a foreign patent office, have been used as a way to measure countries' contribution to the creation and diffusion of innovation and ultimately productivity growth. For example, Eaton and Kortum (1996a,b, 1997, 1999) have used international patent citations to infer the direction and magnitude of the international diffusion of technology. Such studies have a profound impact on our representation of who are the main contributors of technological progress and worldwide productivity growth. Here we look at the distribution of U.S. citations by country of patent office obtained using the 'traditional' front-page citations and the newly available in-text citations.

We find that in-text patent citations are almost three times as much internationalized than front-page citations. Figure 1.12 represents the number of citations from U.S. patents by country of the cited patent for front-page (panel 1.12a) and in-text (panel 1.12b) citations. Since 1947, the share of in-text patent citations to non-U.S. patents has reached 28.82[38] percent while it is 11.0 percent for front-page citations. Thus, considering only front-page citations leads to a more U.S.-centric view of knowledge flows. Going further, we find that some countries are dramatically under-represented in front-page citations as compared to in-text citations. For example, the share of Japanese patents in-text citations is almost three times as large as their share in front-page citations. We believe that our representation of the direction and magnitude of international knowledge flows might well improve in light of our newly available data.

[38] This result is immune to the exclusion of "self-citations". Excluding within-INPADOC-family citations, citations to non-U.S. patents represent 31.38 percent of all citations.

1.5.6 Self-reliance

In the context of the present study, we call 'self-reliance' the citation of one or more patents belonging to the same family or originating by the same patentees as the citing patent itself. There are two main reasons to be interested in the role of self-reliance in patent citations. First, the diffusion of a piece of knowledge is likely to be conveyed primarily by the persons and organisations who created it. Second, one might be worried that in-text citations are mostly self-reference, that is citations of patents belonging to the same family of invention. Consequently, they would not bring much information compared to already available patent family information.

Starting with same-family citations, we map each citing and cited patent to its patent family and compute the share of citations citing a patent belonging to its own family. We consider both the DOCDB families and the INPADOC families. As previously explained, INPADOC families are more permissive as they include in the same group all the documents sharing directly or indirectly (e.g., via a third document) at least *one* priority. We find that the share of in-text citations belonging to the same DOCDB (INPADOC) family is 6.42 (10.65) percent. This is higher than front-page citations' self-references figures, which are 0.69 (1.63) percent. That being said, even considering the most permissive definition of invention families, 90 percent of in-text citations are not self-references, bringing useful information of patented inventions' knowledge background outside their respective patent family.

Turning to same-patentee citations, we look at the share of citations having at least one common inventor or at least one common assignee. We rely on the harmonized names reported in the IFI CLAIMS dataset, labeling as same-patentee citations those where the name of at least one inventor (assignee) is the same for the citing and cited patent. We find that 17.43 (22.46) percent of in-text patent citations have at least one inventor (assignee) in common with their citing patent, against 5.98 (9.26) percent for front-page citations, that is almost three (two) times as much. This result confirms the relative importance of self-reliance in knowledge creation which appears to be even more visible through the lens of in-text citations.

1.5.7 Geographic distribution

A large literature has documented how geography restricts knowledge flows' breadth (Jaffe et al., 1993; Audretsch, 1998; Peri, 2005; Belenzon and Schankerman, 2013). Scholars have pointed to labor mobility within regional labor markets (Almeida and Kogut, 1999) and localized co-invention networks (Breschi and Lissoni, 2009) as leading mechanisms of knowledge flows' geographic concentration. Patent (front-page) citations have been a crucial data source for these studies, proxying the elusive "paper trail" of knowledge (Krugman, 1991) connecting patented inventions.

In this section, we compare in-text and front-page citations in the geographic space. Specifically, we take citing-cited inventor dyads in the two citation groups and calculate the distance between the two inventors' geocoded addresses (de Rassenfosse et al., 2019b), comparing their geographic distribution. Despite being a mere descriptive exercise, this analysis can provide useful insights about differences between in-text and front-page citations along the geographic dimension.

Figure 1.13 shows the probability distribution function and cumulative distribution function of in-text and front-page citing-cited inventor dyads. The x-axis quantifies distance in kilometers. All graphs using all kind of citations portray in-text citations as more localized than front-page ones. Panel 1.13e in particular, shows a higher share of citations within 25km of distance from the cited inventor's location for in-text citations, relative to front-page ones. We also report the same distributions excluding all self-citations between patents appearing in the same INPADOC family and all self-citations at the assignee-level.[39] While in-text citations seem to still be slightly more localized, the difference with front-page ones is minimal and substantially less sharp than suggested by unconditional figures, mostly the result of a higher share of in-text citations occurring at "zero" distance (see panel 1.13f). The higher geographic localization of in-text citations portrayed in Figure 1.13 when considering all citations seems to be explained by a larger occurrence of self-citations for in-text relative to front-page citations.

At the descriptive level, in-text and front-page citations do not display particular differences in terms of their geographic distributions. Nevertheless, we believe that an econometric investigation will be needed to probe this question properly (e.g., following the approach pioneered by Jaffe et al., 1993).

1.5.8 Going further

At this point, it is clear that more work needs to be done to precisely determine the methodological implications of in-text patent citations and how they should be articulated with the existing research toolkit. That being said, the previous results might be a bit embarrassing in the meantime. Front-page patent citations have been used for many different things, including the measurement of patents' importance and knowledge flows. Assuming that in-text patent citations are not just noise, the tiny overlap with front-page citations (see 1.5.2), the orthogonality puzzle (see 1.5.4) and the similarity results (see 1.5.3) are legitimately worrying. These results raise a simple question: which of the in-text and front-page citations provide the best proxy of patent importance and knowledge flows? This question goes far beyond the scope of this paper. The difficulty of this task is notably due to the absence of ground-truth data for knowledge flows and patent importance. This makes the task of comparing the respective ability of front-page and in-text citations to proxy these concepts even harder. Still, in the following lines, we discuss the strategies that seem the most promising to address this challenging question. First, starting with *patent importance*, a standard approach would be to extend Hall et al. (2005) to compare the predictive power of in-text *versus* front-page citations with

[39] We identify self-citations using the same procedure employed in section 1.5.6.

respect to patents' market value. Such analysis could benefit from the newly available data from Kogan et al. (2017). Second, regarding *knowledge flows*, we see two promising approaches. One would be to resort to inventors' survey as in Jaffe et al. (2000). This is our favourite option and would certainly be the closest to a ground truth benchmark. Unfortunately, it would also require a large amount of resources. Another less resource-intensive approach would be to focus on patents filed at the European Patent office which patent citations are systematically accompanied by a code indicating their degree of relevance. Restricting to in-text citations which are also reported on the front page could provide insights on their relative degree of relevance compared to front-page citations. Although much more actionable, this approach might suffer from various selection biases.

1.6 Concluding remarks

This paper introduces a novel dataset on patent citations. It provides 63,854,733 million citations identified in the full-text of 16,781,144 million U.S. patent documents from 1790 to 2018. To the best of our knowledge, it is the first openly-released and extensively validated dataset of the sort. Given the importance of citation data in various fields of the social sciences, we expect these data to be of considerable interest to the scientific community.

Three main messages are particularly noteworthy. First, we found little overlap between the 'traditional' front-page citations and the novel in-text citations. We estimate that the inclusion of in-text citations adds a net 15 percent more patent citations compared to using front-page citations alone.

Second, in addition to adding *more citations*, the inclusion of in-text citations also adds information of *a different nature* due to a different data generation process compared with front-page citations. In particular, we have argued and provided tentative evidence that in-text citations offer a particularly relevant trace of knowledge flow compared to front-page citations. We have also explained why in-text citations represent valuable signals about patent importance. Capturing knowledge flow and measuring patent importance are two of the most popular uses of patent citations and, therefore, we encourage researchers to explore the present data.

Finally, we have relied on best-in-class techniques from NLP and have performed in-depth validation exercises to ensure the quality of the data, achieving highly satisfactory results. We see these results as a proof of the considerable potential offered by the open source community and more particularly applications of modern NLP to information extraction in applied economics and management. In this context, we have made the codebase and the replication material (including code and validation data) natively open source and the data open access.[40] We encourage the community to contribute to the continuous improvement of the dataset. Of

[40] The code is licensed under the MIT license https://opensource.org/licenses/MIT. The data are licensed under the CC-BY-4 license https://creativecommons.org/licenses/by/4.0/legalcode.

particular interest will be the deployment of our pipeline to other jurisdictions.

In conclusion, we hope that the public release of the dataset will enable the community to shed new light on studies exploiting citation data to track knowledge flows and measure patent importance. Furthermore, the data may open new research questions related, e.g., to strategic knowledge disclosure (*à la* Lampe, 2012) or knowledge sourcing (*à la* Wagner et al., 2014). On a more technical level, we see value in leveraging the context of patent citations to determine citation intent (enablement, usefulness, non obviousness, improvement, etc). Such contextual information could lead to a more accurate usage of patent citation data.

1.7 Tables

Table 1.1: Composition of the dataset

Kind code	Kind of document		Number	Share
	Pre 2001	*Post 2001*		
A	Patent	Patent application	11,909,035	0.71
B	Reexamination certificate	Patent	4,188,597	0.25
S	-	Design patent	613,050	0.04
P	Plant patent	Plant patent & Plant patent application	34,852	2.00E-3
E	-	Reissued patent	32,226	2.00E-3
H	-	Statutory invention registration (SIR)	2,255	1.00E-4
I	-	-	1,129	6.00E-5

Table 1.2: Composition of the dataset: focus on patents and applications

Kind code	Kind of document		Number	Share
	Pre 2001	*Post 2001*		
A	Patent	-	6,145,197	0.37
A1	-	Patent application publication	5,753,613	0.34
A2	-	Patent application publication (republication)	1,742	1.00E-4
A9	-	Patent application publication (corrected publication)	8,483	5.00E-4
B1	-	Patent (no pre-grant publication)	776,074	0.04
B2	-	Patent	3,412,523	0.2

Notes: Share of full dataset.

Table 1.3: In-text patent citations extraction performance

	Number of patents in the test set	Avg number of patent tags per patent	Precision	Recall	F1-score
Galibert et al. (2010)	760	12.75	64.4%	61.0%	62.6%
Lopez (2010)	20	9.96	97.44%	97.74%	97.68%
Verluise et al (2020)	160	2.93	97%	82%	89.2%

Table 1.4: In-text patent citations extraction error analysis

Error type	Category	Example
False negative	1	"introduced into a mold (as in Example 1 of 2,154,639) wherein it is polymerized to form a A"
	2	"Aug. 20, 1935 2,255,030 Tholstrup Sept. 2, 1941 2,394,733 Wittenrnyer Feb. 12, 1946 2,433,349 Drewell Dec. 30"
	3	"Filed May 25, 1973, Ser. No. 364,196 Int. Cl. Blk 1/00, 3/06; C01b"
False positive	1	"US. Cl ..29/492, 29/497, 29/498, 29/502, 29/589, 29/628 [51] lnt.Cl."
	2	"Aug. 12, 1941. ALKAN&#39; emoumnmrc COMPASS I iled July 15, 1936 3"
	3	"No. 09/808,790, (Attorney Docket No. 20468-000110), previously incorporated herein by reference. FIG"

Notes: The underlined span of text triggered the error. In the false negative case, it was not detected by Grobid as a patent citation while it should have been the case. In the false positive case, it was detected by Grobid as a patent citation while it is not.

Table 1.5: Distribution of U.S. patent citations by patent office

Patent office	Number of occurrences in validation sample	Share in validation sample	Share in universe of U.S. patents
US	203	0.67	0.61
JP	52	0.17	0.09
WO	18	0.06	0.10
DE	9	0.03	0.02
EP	5	0.02	0.03
KR	4	0.01	7.00E-3
FR	4	0.01	6.00E-3
BE	2	7.00E-03	3.00E-3
SA	1	3.00E-03	3.00E-3
CH	1	3.00E-03	3.00E-3
AL	1	3.00E-03	0.02

Table 1.6: In-text patent citations parsing accuracy

	Number of examples in the test set	Organisation name	Original number	Kind code	All
Lopez (2010	250	-	-	-	97.2%
Verluise et al (2020)	300	98.4%	95.7%	97.6%	-

Notes: Lopez (2010) does not distinguish between the accuracy on the three attributes and reports the overall accuracy of the Finite State Transducers to translate the natural language citation into a fully structured citation represented by its three attributes.

Table 1.7: In-text patent citations matching performance

	True		**False**	
	Content	*Number*	*Content*	*Number*
Positive	A publication number was correctly matched	137	A publication number was incorrectly matched	10
Negative	No matched publication number and no match found by the annotator	36	No matched publication number but a match was found by the annotator	17

Table 1.8: In-text patent citations matching error analysis

Error type	**Category**	**Sub-category**	**Example**	**Number of occurrences**
False match	Incorrect patent	Badly formatted pre-2000 Japanese patent	JP5064281 instead of JPS5064281	5
		Incorrect extraction of pre-1970 U.S. patent due to bad OCR	CA-8465T-T (from 2,936,846 5/60 Tyler et al, in reference list)	1
	Non patent	Garbled table	-	2
		Technology class	US-32537 extracted from "... U.S. Cl. 325/392, 325/37..."	1
		Date	US-312012 extracted from "...filed Aug. 31, 2012, . . . "	1
False no-match	Formatting	Missing leading zeros after country code or date	EP592106 instead of EP0592106	6
		Year reported after instead of before patent number	JP3518222000 instead of JP2000351822	3
		Incorrect extraction of country code	SU-14553625 extracted from "U.S. Utility application Ser. No. 14/553,625"	1
	Wrong service call	-	-	7

Notes: Error analysis based on 200 random examples.

Table 1.9: Extracted citations judged unmatchable by the annotator

Category	**Example**	**Number of occurrences**
Garbled tables	AL-1226-C extracted from "...AL C 257 75.108 67.122 6.016 1..."	11
Provisional patent applications	US-60723639 extracted from "U.S. provisional application Ser. No. 60/723,639"; provisional patent applications are not public information	8
Incorrect and ambiguous number formats	EP-87309853 extracted from "European patent specification No 87309853.7" (non-standard format of a non-searchable application number)	4
Incorrect parsed attributes	WO-PTS0767103 instead of WO-PTUS07067103	5
Non searchable	DE-19654649 (not indexed by Google Patents)	3
Non patents (technological class, dates, etc)	US-32128 extracted from "... U.S. Cl. 322/79, 310/68 D, 321/28, ..."	10

Notes: The Number of occurrences includes both matched and unmatched examples.

Table 1.10: Number and share of citations matched by patent organisation (selected)

Patent organisation	**Total number of citations**	**Share of citations matched**
USPTO	37,072,526	89.14
WIPO	6,453,099	81.89
JPO	5,659,300	77.22
EPO	2,228,096	51.27
DPMA	1,371,114	73.46

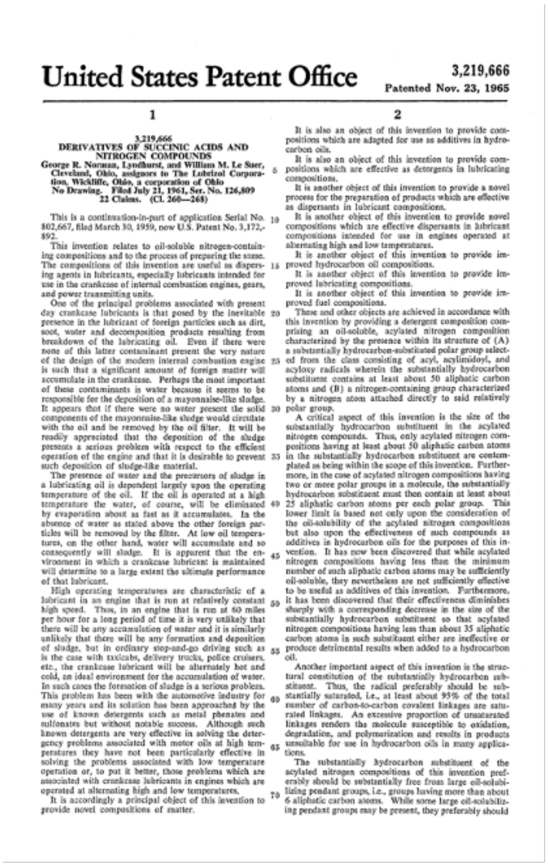
Table 1.11: In-text and front page citations at-a-glance

	Front page	**In-text**
Number of patents	16,781,144	16,781,144
Number of patents with at least one citation	11,965,720	9,453,181
Share of patents with at least one citation	71.30%	56.33%
Number of citations	203,557,205	63,854,733
Number of citations^a	203,557,011	46,115,608
Average number of citations per patent	12.13	3.81
Average number of citations per patent - conditional on citing at least one patent	17.01	6.75
Number of US patent citations^a	181,162,466	32,827,382
Share of non U.S. citations^a	11%	28.82%
Median pairwise similarity (dot product) between citing and cited patent [lower quartile, upper quartile]^{a,c}	0.71 [0.62, 0.78]	0.80 [0.68, 0.88]
Share of citations in the same DOCDB family^b	0.69%	6.27%
Share of cited patents in the same INPADOC family^b	1.63%	10.51%
Share of cited patents with at least one shared inventor^b	5.98%	17.43%
Share of cited patents with at least one shared assignee^b	9.26%	22.46%

Notes: ^{a}: After 1947 only. ^{b}: Matched in-text only. ^{c}: After removing within-DOCDB family citations.

1.8 Figures

Figure 1.1: Example of the USPTO “old” patent format (US-3219666-A)



United States Patent Office

3,219,666
Patented Nov. 23, 1965

1

3,219,666
DERIVATIVES OF SUCCINIC ACIDS AND NITROGEN COMPOUNDS
George R. Norman, Lyndhurst, and William M. Le Suer, Cleveland, Ohio, assignors to The Lubrizol Corporation, Wickliffe, Ohio, a corporation of Ohio
No Drawing. Filed July 21, 1961, Ser. No. 126,809
22 Claims. (Cl. 260—268)

This is a continuation-in-part of application Serial No. 802,667, filed March 30, 1959, now U.S. Patent No. 3,172,892.

This invention relates to oil-soluble nitrogen-containing compositions and to the process of preparing the same. The compositions of this invention are useful as dispersing agents in lubricants, especially lubricants intended for use in the crankcase of internal combustion engines, gears, and power transmitting units.

One of the principal problems associated with present day crankcase lubricants is that posed by the inevitable presence in the lubricant of foreign particles such as dirt, soot, water and decomposition products resulting from breakdown of the lubricating oil. Even if there were none of this latter contaminant present the very nature of the design of the modern internal combustion engine is such that a significant amount of foreign matter will accumulate in the crankcase. Perhaps the most important of these contaminants is water because it seems to be responsible for the deposition of a mayonnaise-like sludge. It appears that if there were no water present the solid components of the mayonnaise-like sludge would circulate with the oil and be removed by the oil filter. It will be readily appreciated that the deposition of the sludge presents a serious problem with respect to the efficient operation of the engine and that it is desirable to prevent such deposition of sludge-like material.

The presence of water and the precursors of sludge in a lubricating oil is dependent largely upon the operating temperature of the oil. If the oil is operated at a high temperature the water, of course, will be eliminated by evaporation about as fast as it accumulates. In the absence of water as stated above the other foreign particles will be removed by the filter. At low oil temperatures, on the other hand, water will accumulate and so consequently will sludge. It is apparent that the environment in which a crankcase lubricant is maintained will determine to a large extent the ultimate performance of that lubricant.

High operating temperatures are characteristic of a lubricant in an engine that is run at relatively constant high speed. Thus, in an engine that is run at 60 miles per hour for a long period of time it is very unlikely that there will be any accumulation of water and it is similarly unlikely that there will be any formation and deposition of sludge, but in ordinary stop-and-go driving such as is the case with taxicabs, delivery trucks, police cruisers, etc., the crankcase lubricant will be alternately hot and cold, an ideal environment for the accumulation of water. In such cases the formation of sludge is a serious problem. This problem has been with the automotive industry for many years and its solution has been approached by the use of known detergents such as metal phenates and sulfonates but without notable success. Although such known detergents are very effective in solving the detergency problems associated with motor oils at high temperatures they have not been particularly effective in solving the problems associated with low temperature operation or, to put it better, those problems which are associated with crankcase lubricants in engines which are operated at alternating high and low temperatures.

It is accordingly a principal object of this invention to provide novel compositions of matter.

2

It is also an object of this invention to provide compositions which are adapted for use as additives in hydrocarbon oils.

It is also an object of this invention to provide compositions which are effective as detergents in lubricating compositions.

It is another object of this invention to provide a novel process for the preparation of products which are effective as dispersants in lubricant compositions.

It is another object of this invention to provide novel compositions which are effective dispersants in lubricant compositions intended for use in engines operated at alternating high and low temperatures.

It is another object of this invention to provide improved hydrocarbon oil compositions.

It is another object of this invention to provide improved lubricating compositions.

It is another object of this invention to provide improved fuel compositions.

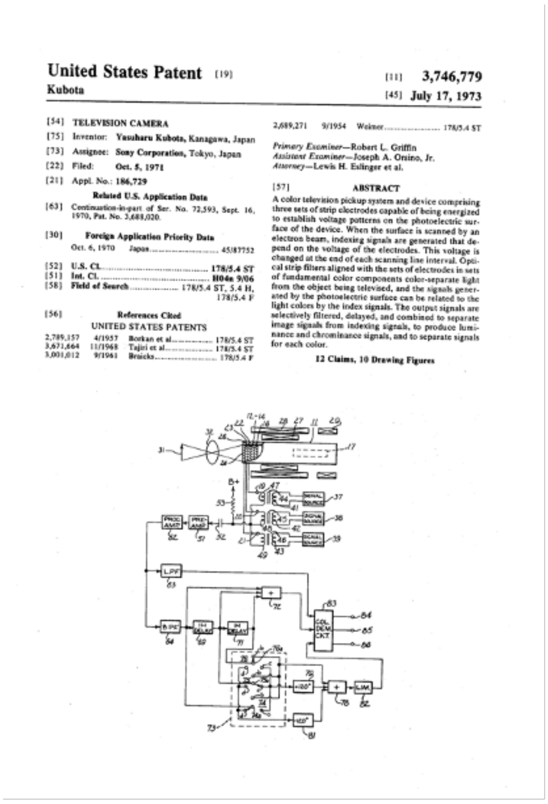
These and other objects are achieved in accordance with this invention by providing a detergent composition comprising an oil-soluble, acylated nitrogen composition characterized by the presence within its structure of (A) a substantially hydrocarbon-substituted polar group selected from the class consisting of acyl, acylimidoyl, and acyloxy radicals wherein the substantially hydrocarbon substituent contains at least about 50 aliphatic carbon atoms and (B) a nitrogen-containing group characterized by a nitrogen atom attached directly to said relatively polar group.

A critical aspect of this invention is the size of the substantially hydrocarbon substituent in the acylated nitrogen compounds. Thus, only acylated nitrogen compositions having at least about 50 aliphatic carbon atoms in the substantially hydrocarbon substituent are contemplated as being within the scope of this invention. Furthermore, in the case of acylated nitrogen compositions having two or more polar groups in a molecule, the substantially hydrocarbon substituent must then contain at least about 25 aliphatic carbon atoms per each polar group. This lower limit is based not only upon the consideration of the oil-solubility of the acylated nitrogen compositions but also upon the effectiveness of such compounds as additives in hydrocarbon oils for the purposes of this invention. It has now been discovered that while acylated nitrogen compositions having less than the minimum number of such aliphatic carbon atoms may be sufficiently oil-soluble, they nevertheless are not sufficiently effective to be useful as additives of this invention. Furthermore, it has been discovered that their effectiveness diminishes sharply with a corresponding decrease in the size of the substantially hydrocarbon substituent so that acylated nitrogen compositions having less than about 35 aliphatic carbon atoms in such substituent either are ineffective or produce detrimental results when added to a hydrocarbon oil.

Another important aspect of this invention is the structural constitution of the substantially hydrocarbon substituent. Thus, the radical preferably should be substantially saturated, i.e., at least about 95% of the total number of carbon-to-carbon covalent linkages are saturated linkages. An excessive proportion of unsaturated linkages renders the molecule susceptible to oxidation, degradation, and polymerization and results in products unsuitable for use in hydrocarbon oils in many applications.

The substantially hydrocarbon substituent of the acylated nitrogen compositions of this invention preferably should be substantially free from large oil-solubilizing pendant groups, i.e., groups having more than about 6 aliphatic carbon atoms. While some large oil-solubilizing pendant groups may be present, they preferably should

Figure 1.2: Example of the USPTO “new” patent format (US-3746779-A)



United States Patent [19] — [11] **3,746,779**

Kubota — [45] **July 17, 1973**

[54] **TELEVISION CAMERA**

[75] Inventor: **Yasuharu Kubota,** Kanagawa, Japan

[73] Assignee: **Sony Corporation,** Tokyo, Japan

[22] Filed: **Oct. 5, 1971**

[21] Appl. No.: **186,729**

Related U.S. Application Data

[63] Continuation-in-part of Ser. No. 72,593, Sept. 16, 1970, Pat. No. 3,688,020.

[30] **Foreign Application Priority Data**

Oct. 6, 1970 Japan........................ 45/87752

[52] **U.S. Cl.**.................................... **178/5.4 ST**
[51] **Int. Cl.** **H04n 9/06**
[58] **Field of Search**.................. 178/5.4 ST, 5.4 H, 178/5.4 F

[56] **References Cited**

UNITED STATES PATENTS

2,789,157	4/1957	Borkan et al.	178/5.4 ST
3,671,664	11/1968	Tajiri et al.	178/5.4 ST
3,001,012	9/1961	Braicks	178/5.4 F
2,689,271	9/1954	Weimer	178/5.4 ST

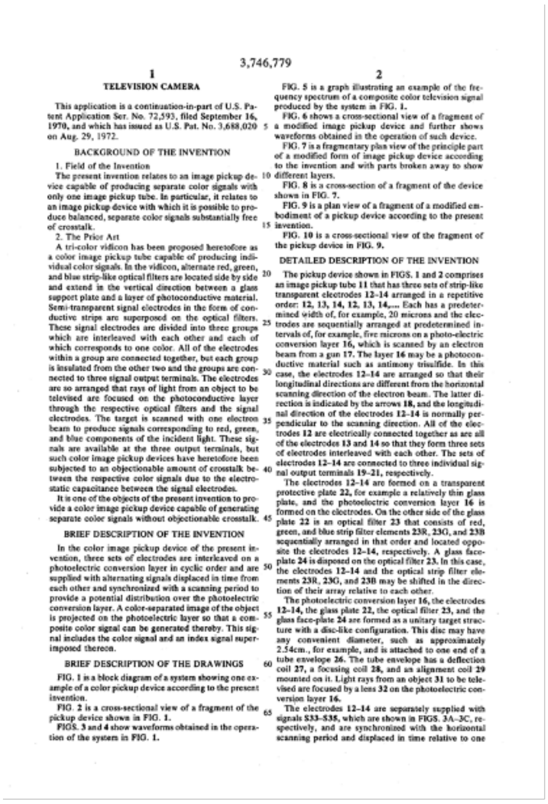
Primary Examiner—Robert L. Griffin
Assistant Examiner—Joseph A. Orsino, Jr.
Attorney—Lewis H. Eslinger et al.

[57] **ABSTRACT**

A color television pickup system and device comprising three sets of strip electrodes capable of being energized to establish voltage patterns on the photoelectric surface of the device. When the surface is scanned by an electron beam, indexing signals are generated that depend on the voltage of the electrodes. This voltage is changed at the end of each scanning line interval. Optical strip filters aligned with the sets of electrodes in sets of fundamental color components color-separate light from the object being televised, and the signals generated by the photoelectric surface can be related to the light colors by the index signals. The output signals are selectively filtered, delayed, and combined to separate image signals from indexing signals, to produce luminance and chrominance signals, and to separate signals for each color.

12 Claims, 10 Drawing Figures

(a) Front page



3,746,779

1

TELEVISION CAMERA

This application is a continuation-in-part of U.S. Patent Application Ser. No. 72,593, filed September 16, 1970, and which has issued as U.S. Pat. No. 3,688,020 on Aug. 29, 1972.

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to an image pickup device capable of producing separate color signals with only one image pickup tube. In particular, it relates to an image pickup device with which it is possible to produce balanced, separate color signals substantially free of crosstalk.

2. The Prior Art

A tri-color vidicon has been proposed heretofore as a color image pickup tube capable of producing individual color signals. In the vidicon, alternate red, green, and blue strip-like optical filters are located side by side and extend in the vertical direction between a glass support plate and a layer of photoconductive material. Semi-transparent signal electrodes in the form of conductive strips are superposed on the optical filters. These signal electrodes are divided into three groups which are interleaved with each other and each of which corresponds to one color. All of the electrodes within a group are connected together, but each group is insulated from the other two and the groups are connected to three signal output terminals. The electrodes are so arranged that rays of light from an object to be televised are focused on the photoconductive layer through the respective optical filters and the signal electrodes. The target is scanned with one electron beam to produce signals corresponding to red, green, and blue components of the incident light. These signals are available at the three output terminals, but such color image pickup devices have heretofore been subjected to an objectionable amount of crosstalk between the respective color signals due to the electrostatic capacitance between the signal electrodes.

It is one of the objects of the present invention to provide a color image pickup device capable of generating separate color signals without objectionable crosstalk.

BRIEF DESCRIPTION OF THE INVENTION

In the color image pickup device of the present invention, three sets of electrodes are interleaved on a photoelectric conversion layer in cyclic order and are supplied with alternating signals displaced in time from each other and synchronized with a scanning period to provide a potential distribution over the photoelectric conversion layer. A color-separated image of the object is projected on the photoelectric layer so that a composite color signal can be generated thereby. This signal includes the color signal and an index signal superimposed thereon.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram of a system showing one example of a color pickup device according to the present invention.

FIG. 2 is a cross-sectional view of a fragment of the pickup device shown in FIG. 1.

FIGS. 3 and 4 show waveforms obtained in the operation of the system in FIG. 1.

2

FIG. 5 is a graph illustrating an example of the frequency spectrum of a composite color television signal produced by the system in FIG. 1.

FIG. 6 shows a cross-sectional view of a fragment of a modified image pickup device and further shows waveforms obtained in the operation of such device.

FIG. 7 is a fragmentary plan view of the principle part of a modified form of image pickup device according to the invention and with parts broken away to show different layers.

FIG. 8 is a cross-section of a fragment of the device shown in FIG. 7.

FIG. 9 is a plan view of a fragment of a modified embodiment of a pickup device according to the present invention.

FIG. 10 is a cross-sectional view of the fragment of the pickup device in FIG. 9.

DETAILED DESCRIPTION OF THE INVENTION

The pickup device shown in FIGS. 1 and 2 comprises an image pickup tube **11** that has three sets of strip-like transparent electrodes **12–14** arranged in a repetitive order: **12, 13, 14, 12, 13, 14,...** Each has a predetermined width of, for example, 20 microns and the electrodes are sequentially arranged at predetermined intervals of, for example, five microns on a photo-electric conversion layer **16,** which is scanned by an electron beam from a gun **17.** The layer **16** may be a photoconductive material such as antimony trisulfide. In this case, the electrodes **12–14** are arranged so that their longitudinal directions are different from the horizontal scanning direction of the electron beam. The latter direction is indicated by the arrows **18,** and the longitudinal direction of the electrodes **12–14** is normally perpendicular to the scanning direction. All of the electrodes **12** are electrically connected together as are all of the electrodes **13** and **14** so that they form three sets of electrodes interleaved with each other. The sets of electrodes **12–14** are connected to three individual signal output terminals **19–21,** respectively.

The electrodes **12–14** are formed on a transparent protective plate **22,** for example a relatively thin glass plate, and the photoelectric conversion layer **16** is formed on the electrodes. On the other side of the glass plate **22** is an optical filter **23** that consists of red, green, and blue strip filter elements **23R, 23G,** and **23B** sequentially arranged in that order and located opposite the electrodes **12–14,** respectively. A glass face-plate **24** is disposed on the optical filter **23.** In this case, the electrodes **12–14** and the optical strip filter elements **23R, 23G,** and **23B** may be shifted in the direction of their array relative to each other.

The photoelectric conversion layer **16,** the electrodes **12–14,** the glass plate **22,** the optical filter **23,** and the glass face-plate **24** are formed as a unitary target structure with a disc-like configuration. This disc may have any convenient diameter, such as approximately 2.54cm., for example, and is attached to one end of a tube envelope **26.** The tube envelope has a deflection coil **27,** a focusing coil **28,** and an alignment coil **29** mounted on it. Light rays from an object **31** to be televised are focused by a lens **32** on the photoelectric conversion layer **16.**

The electrodes **12–14** are separately supplied with signals **S33–S35,** which are shown in FIGS. 3A–3C, respectively, and are synchronized with the horizontal scanning period and displaced in time relative to one

(b) Specification

Figure 1.3: Empirical probability distribution function of citation detection as a function of the starting character

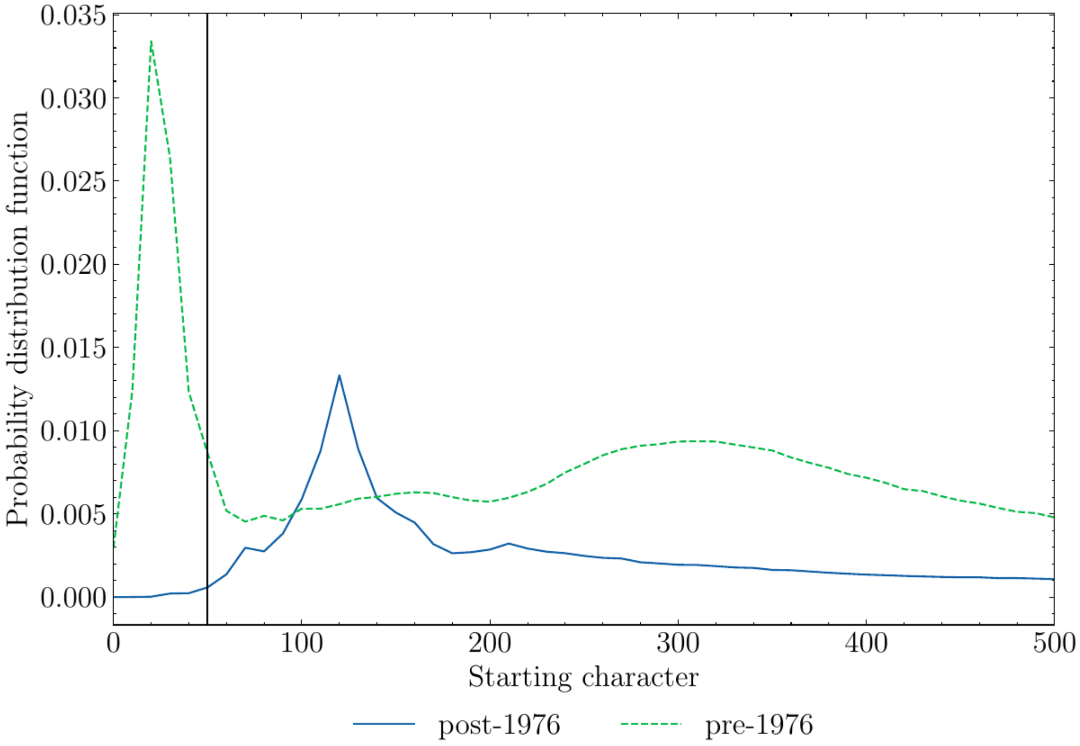


Figure 1.4: Empirical probability distribution function of citation detection as a function of the relative place of the starting character

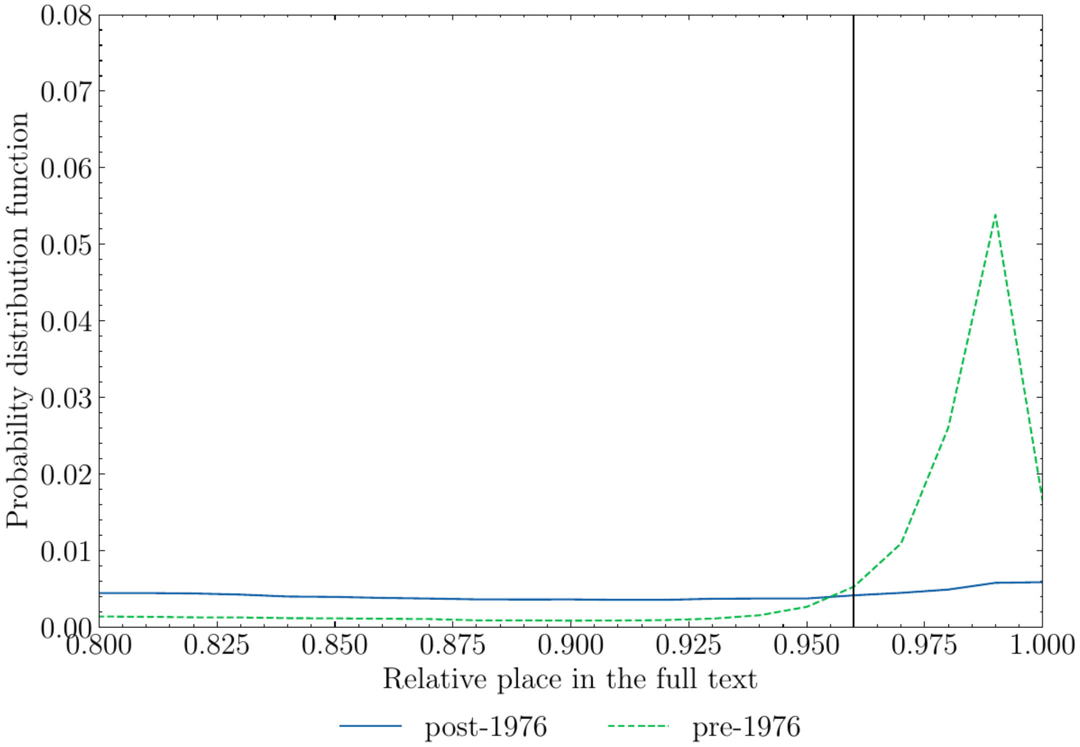
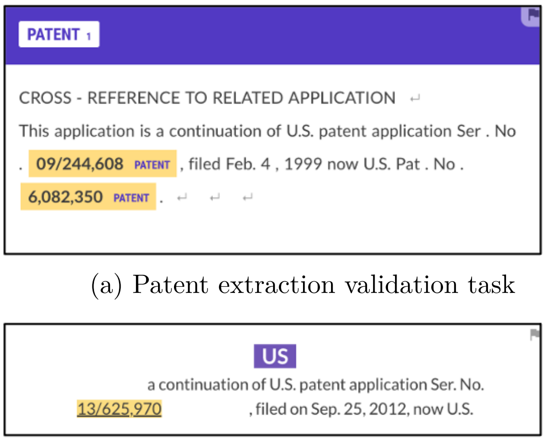


Figure 1.5: Preview of the annotation platform



PATENT 1

CROSS - REFERENCE TO RELATED APPLICATION ↵

This application is a continuation of U.S. patent application Ser . No . 09/244,608 PATENT , filed Feb. 4 , 1999 now U.S. Pat . No . 6,082,350 PATENT . ↵ ↵ ↵

(a) Patent extraction validation task

US

a continuation of U.S. patent application Ser. No. 13/625,970 , filed on Sep. 25, 2012, now U.S.

(b) Patent parsing validation task (organisation name)

Figure 1.6: Empirical cumulative distribution function of patents in the validation sample and in the universe of U.S. patents (by decade)

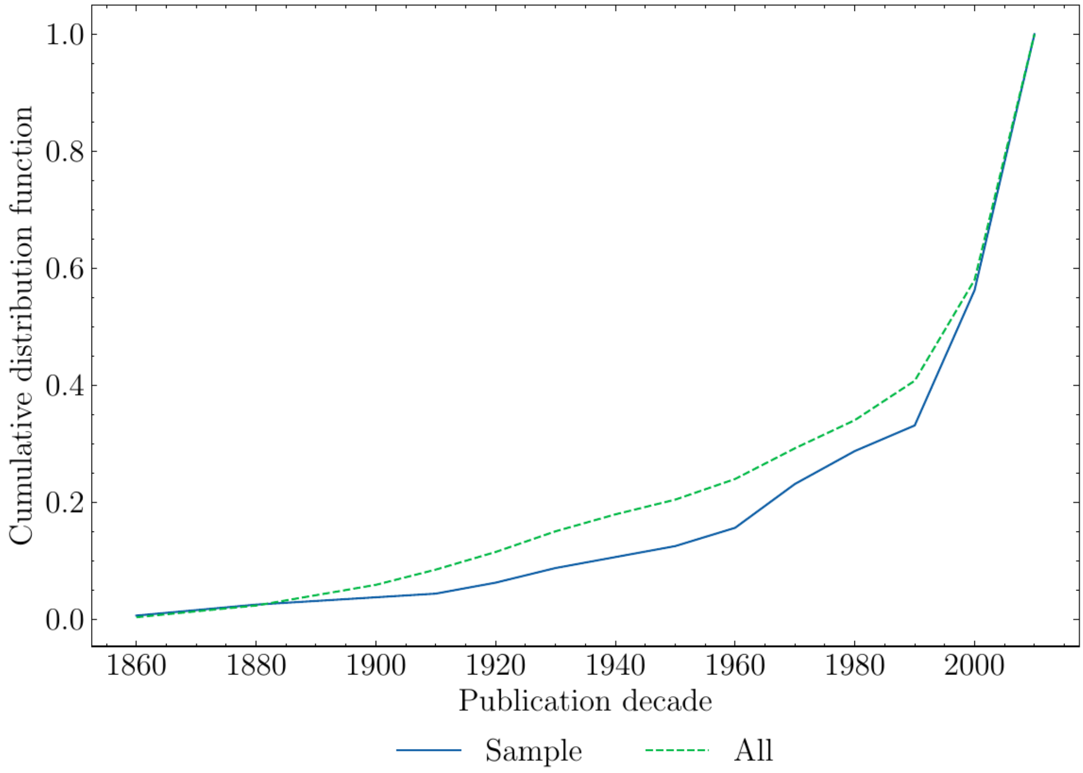
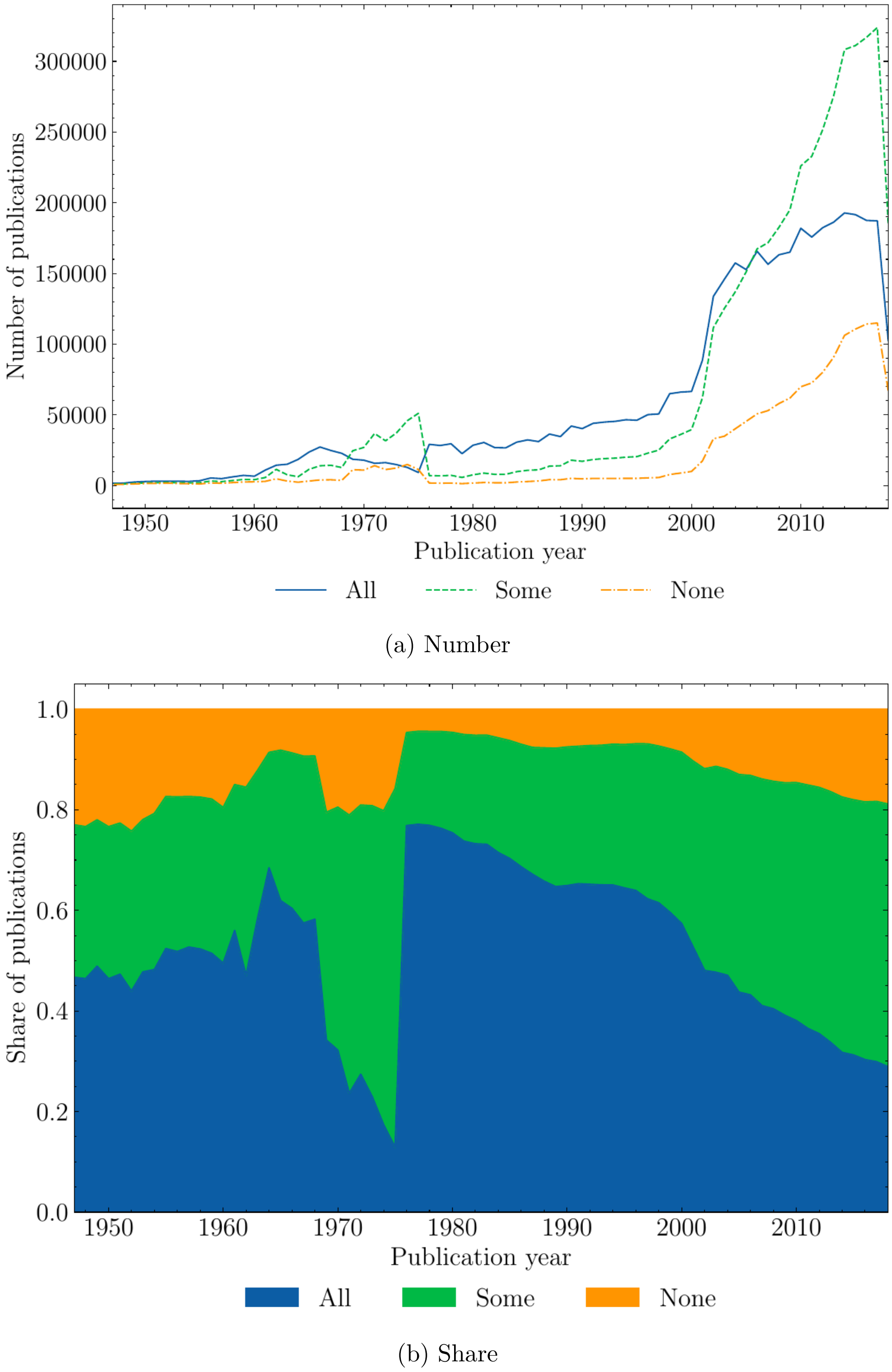


Figure 1.7: Citing patents over time by in-text citation match status



(a) Number

(b) Share

Notes: "All" (blue solid line) refers to patent publications for which it was possible to match all extracted in-text citations. "Some" (orange dashed line) refers to patent publications for which it was possible to match only some extracted in-text citations. "None" (green dash-dot line) depicts patent publications for which we could not match any extracted in-text citation.

Figure 1.8: Patent citations by origin

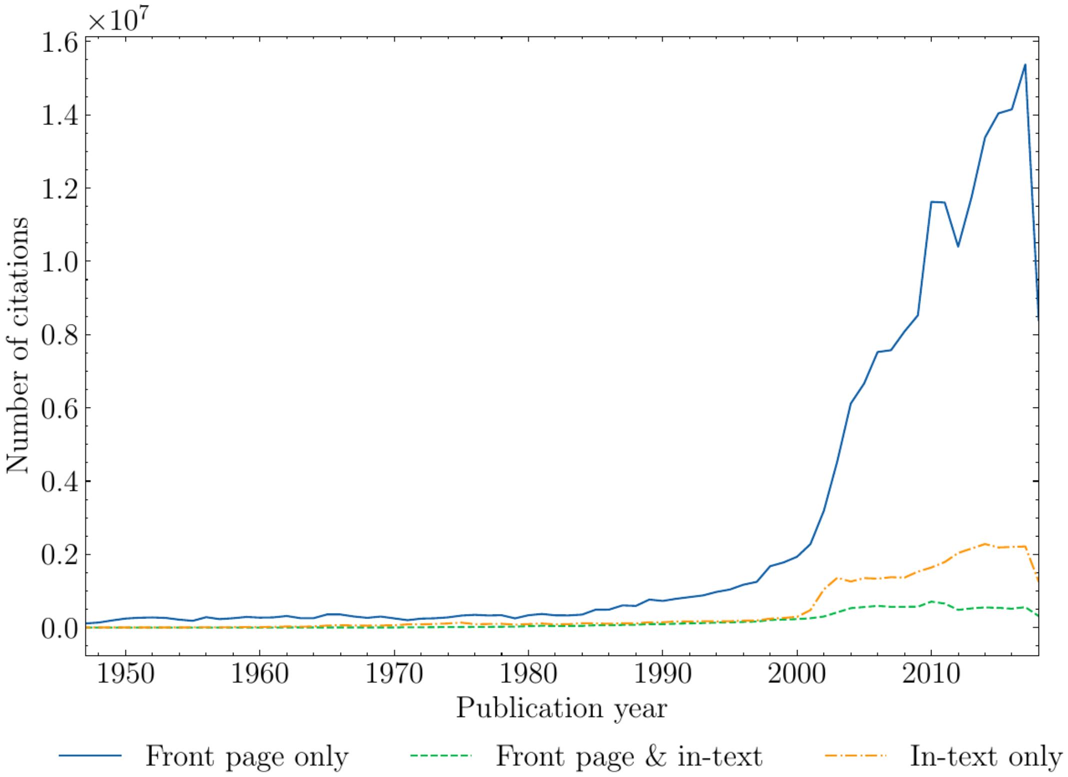
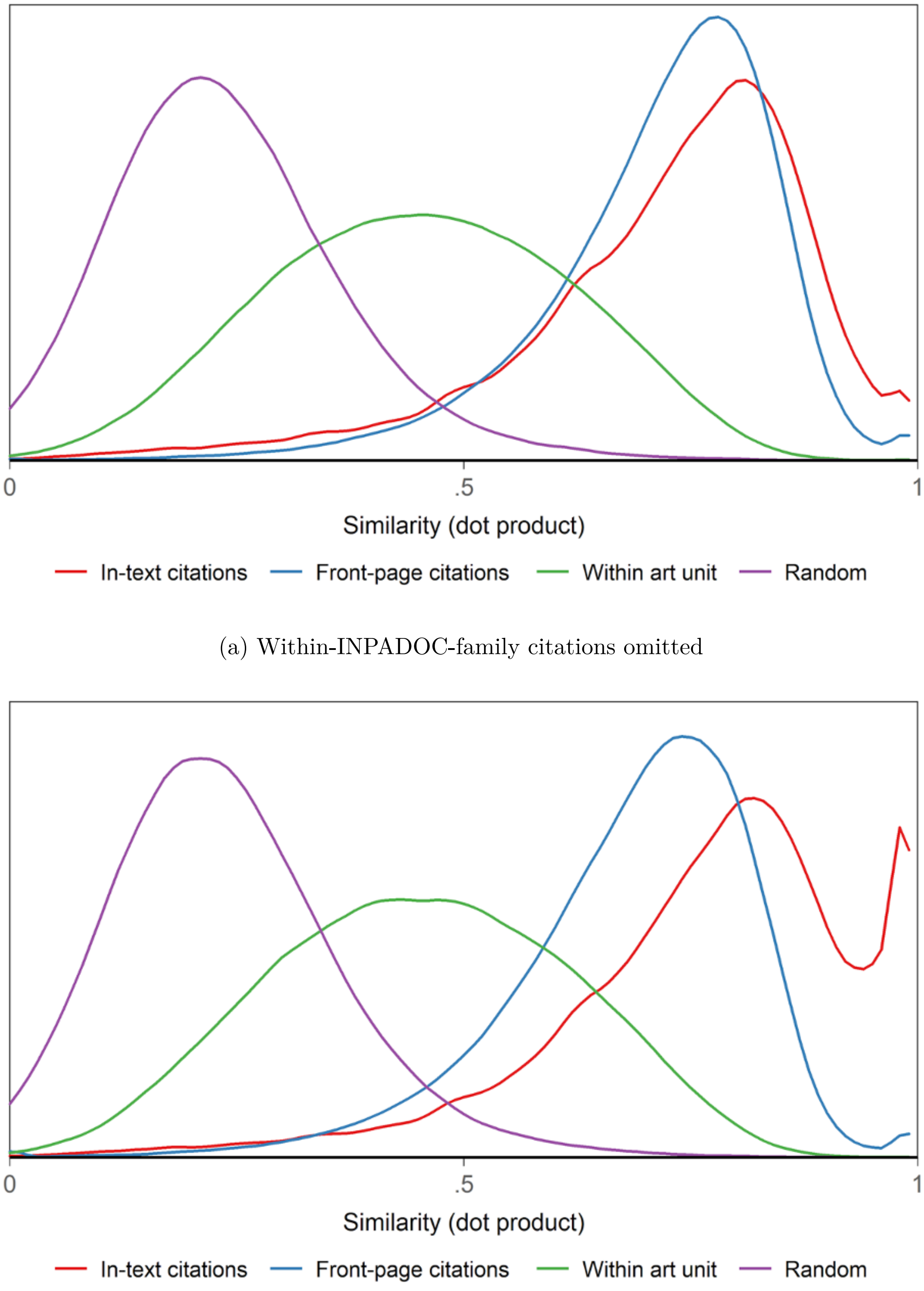


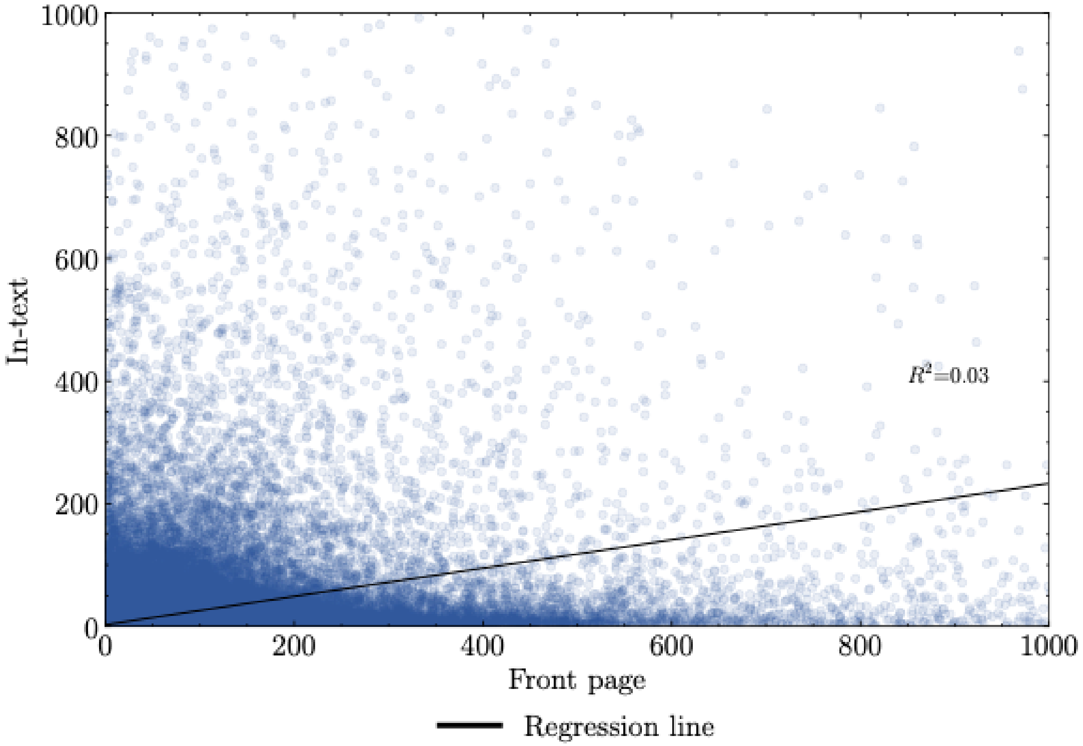
Figure 1.9: Citing-cited patent pair-wise similarity distribution



(a) Within-INPADOC-family citations omitted

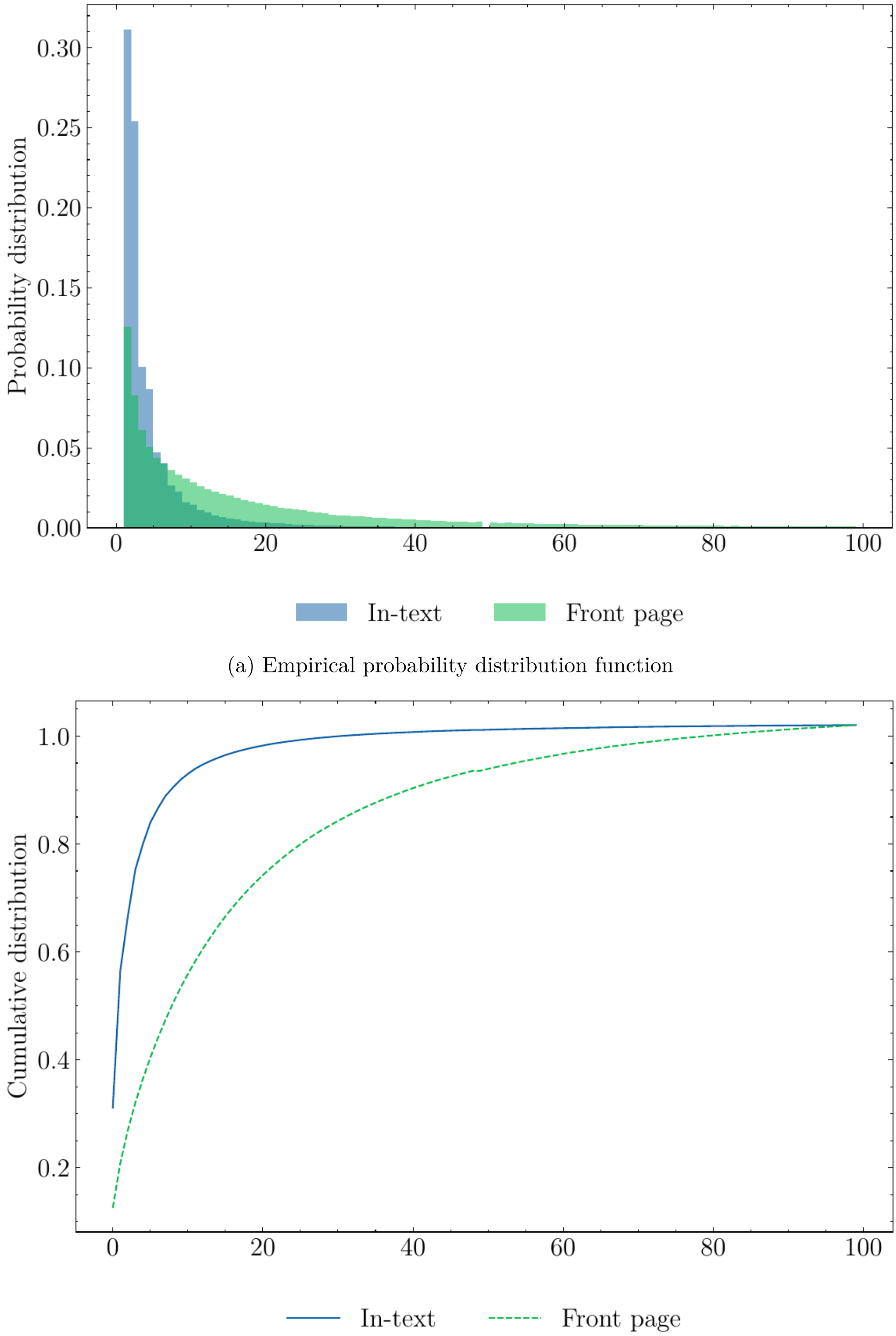
(b) Within-DOCDB-family citations omitted

Figure 1.10: Forward citations count of invention families: front page citations *versus* in-text citations



Notes: We use a 10 percent random sample of all DOCDB patent families with a positive front page and in-text forward citations count. Each data-point represents a DOCDB patent family. The regressions line corresponds to the following model: $in - text\ forward\ citations\ count = a(front\ page\ forward\ citations\ count) + b$

Figure 1.11: Empirical distribution of forward citations count

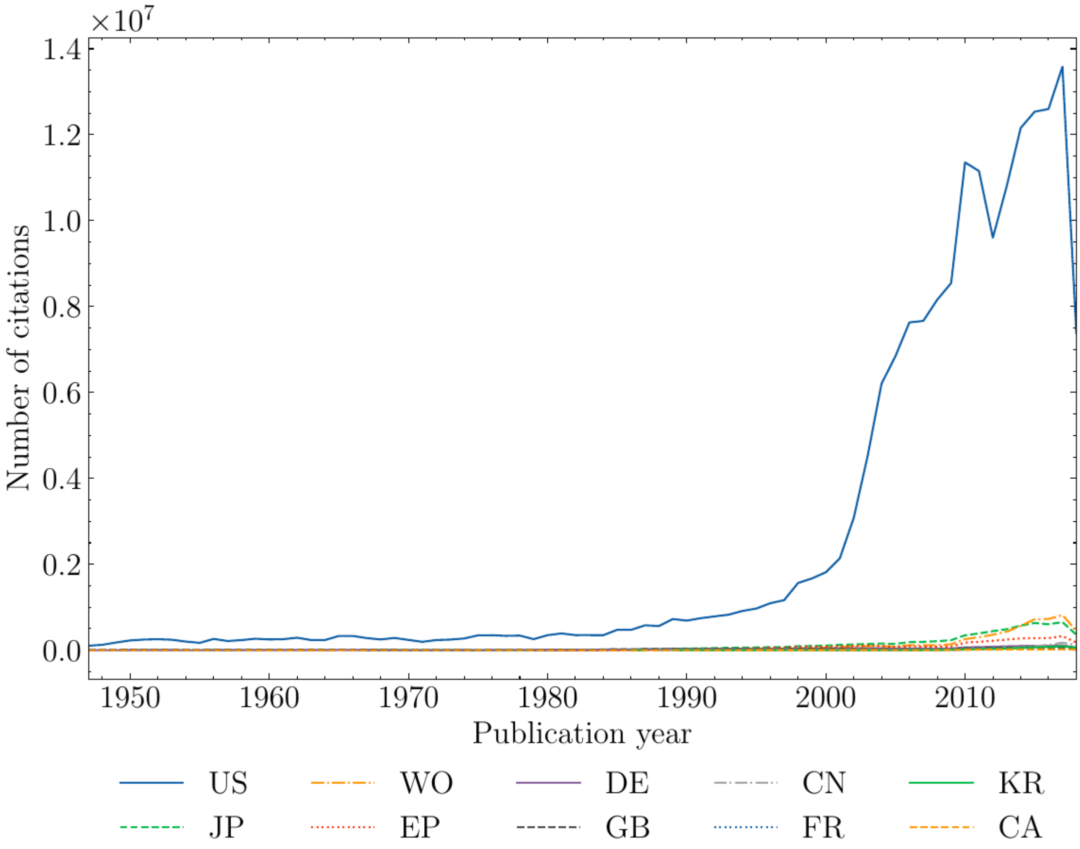


(a) Empirical probability distribution function

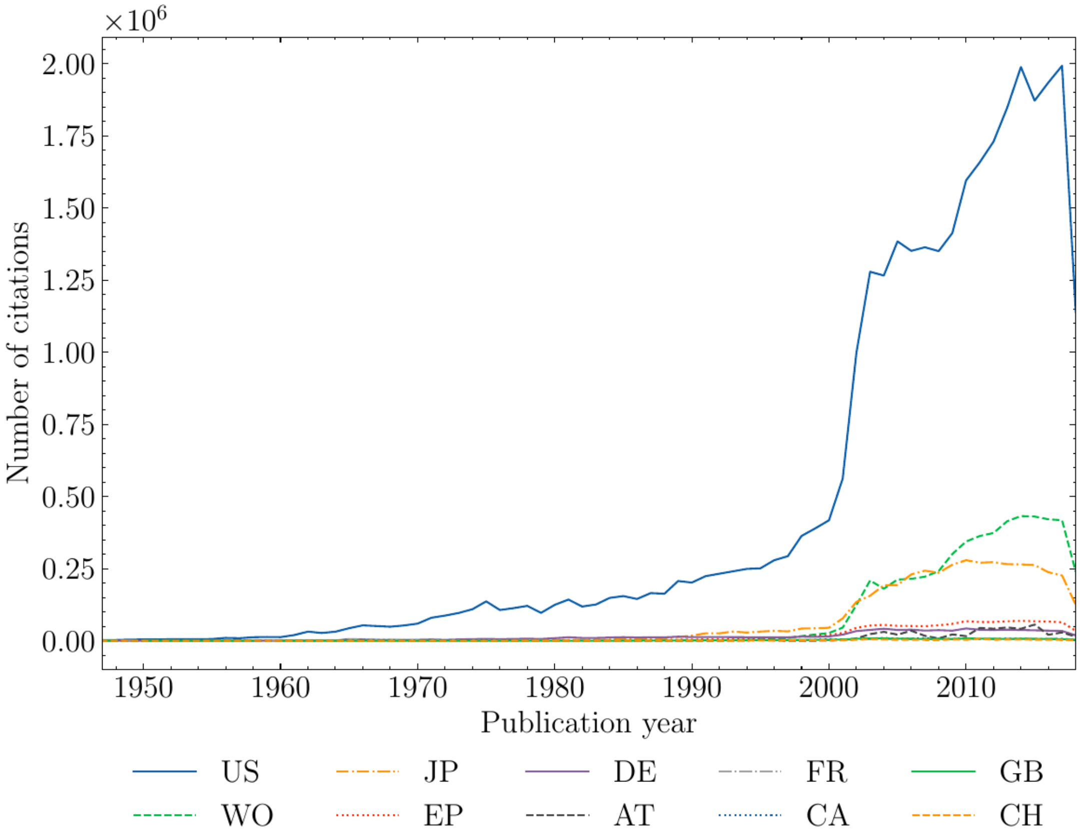
(b) Empirical cumulative distribution function

Notes: We use a 10 percent random sample of all DOCDB patent families with a positive front page and in-text forward citations count.

Figure 1.12: Patent citations by “receiving” country

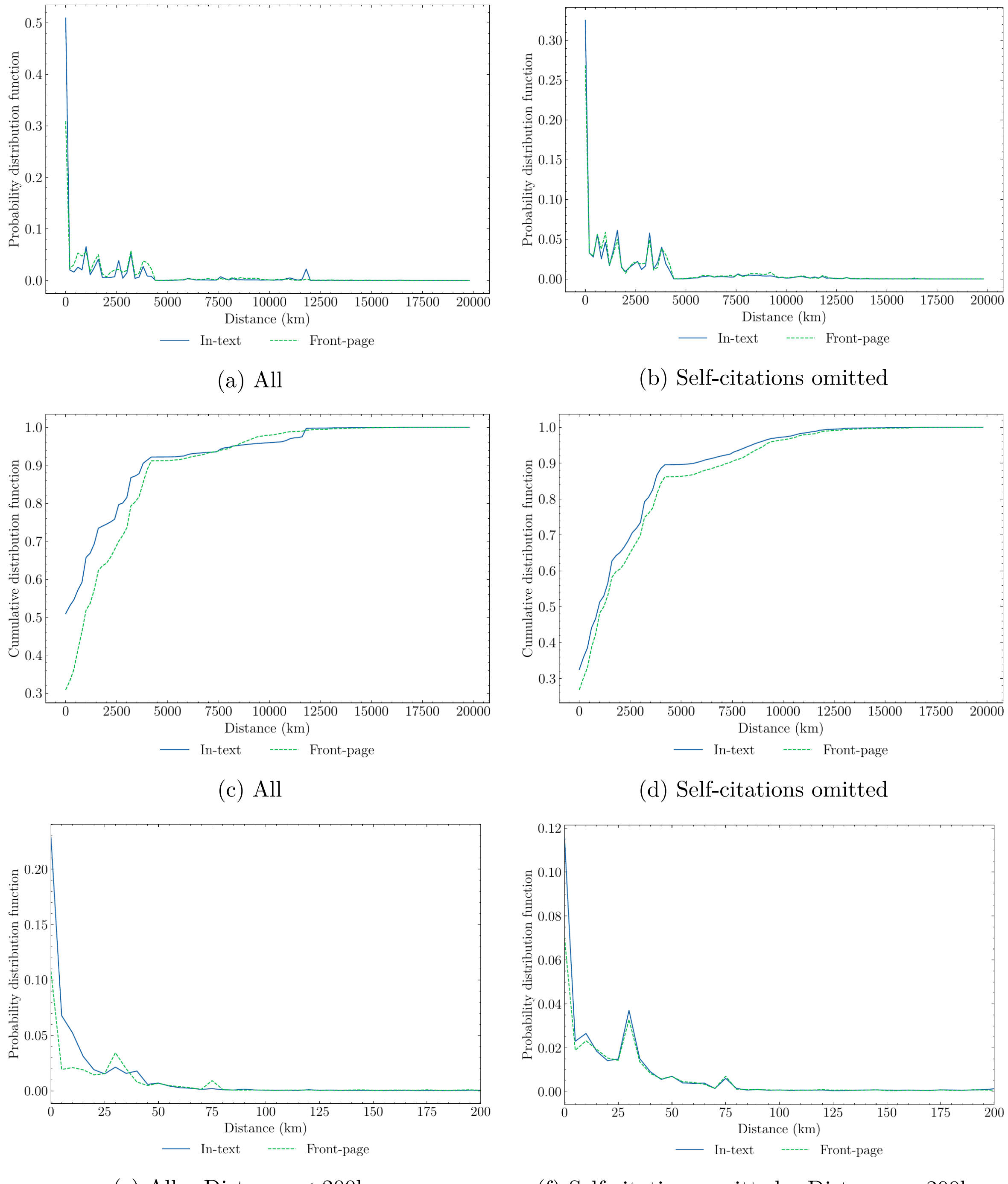


(a) Front page



(b) In-text

Figure 1.13: Distribution of citing-cited inventors distance



(a) All

(b) Self-citations omitted

(c) All

(d) Self-citations omitted

(e) All – Distance < 200km

(f) Self-citations omitted – Distance < 200km

Notes: Distance in kilometers is calculated from the latitude-longitude coordinates of the citing inventor's address to the latitude-longitude coordinates of cited inventor's address. Self citations include within-INPADOC-family citations and same assignee citations. In panel 1.13a, 1.13b, 1.13c and 1.13d we group observations by 200km bins. In panel 1.13e and 1.13f we use 5km bins.

1.9 Appendix

In-text patent citations reasons and examples

Citation Reason	**Example Patent**	**Citation and Context**
Enablement	9,607,299 (*Transactional security over a network*)	"Techniques for data encryption are disclosed in, for example, U.S. Pat. Nos. 7,257,225 and 7,251,326 (incorporated herein by reference) and the details of such processes are not provided herein to maintain focus on the disclosed embodiments."
	9,606,907 (*Memory module with distributed data buffers and method of operation*)	"Examples of circuits which can serve as the control circuit ... are described in more detail by U.S. Pat. Nos. 7,289,386 and 7,532,537, each of which is incorporated in its entirety by reference herein."
Novelty and non-obviousness	8,100,652 (*Ceiling fan complete cover*)	"U.S. Pat. No. 5,281,093, issued to Sedlak, et al., discloses a fan blade cover with a zipper. Sedlak, however, does not protect the fan's housing and motor, nor does it prevent blades from spinning."
	9,607,328 (*Electronic content distribution and exchange system*)	"One skilled in the art will readily appreciate that there is a great deal of prior art centered on methods for selecting programming for a viewer based on previous viewing history and explicit preferences, e.g., U.S. Pat. No. 5,758,257. The methods described in this application are unique and novel over these techniques as they suggest..."
Usefulness	9,607,730 (*Non-oleic triglyceride based, low viscosity, high flash point dielectric fluids*)	Applicant directly compares empirical results for the invention at hand with similar, previously granted patents.
	9,911,050 (*Driver active safety control system for vehicle*)	"For example, the interior rearview mirror assembly may comprise a prismatic mirror assembly, such as the types described in U.S. Pat. Nos. 7,249,860; 6,318,870;..., which are hereby incorporated herein by reference in their entireties."

Data record and reproducibility

Data generation and validation reproducibility is guaranteed by the codebase hosted on the project repository. Validation data are supported by Data Version Control (DVC). Since the project is open-source and continuously improving, exact replication of the data and results detailed above requires the user to choose the tag '0.3.1' of the code.[41]

The data are reported as a nested table that is structured as follows:

- Each entry corresponds to the patent document from which we extracted patent citations. Each such patent is identified by a publication number (primary key). In addition to the publication number, we also report its publication date, application identifier, and patent publication identifier. We also include DOCDB and INPADOC family codes, which identify a constellation of inter-related patents that protect the same invention across jurisdictions.
- Each entry has a citation variable in which cited patents are listed and their attributes are nested. Any detected patent is represented by the two attributes parsed by Grobid, the code of its patent office and its original number. When these two attributes can be matched with a publication number, we also report the publication date, application identifier, patent publication identifier and the DOCDB and INPADOC family identifiers. Eventually, we report a flag indicating that the extracted citation is likely to belong to the front matter or the header.

The schema of the table is detailed below.

Name	**Description**	**Type**	**Nb non null**
publication_number	Publication number.	STR	16781144
publication_date	Publication date (yyyymmdd).	INT	15862299
appln_id	PATSTAT application identification. Surrogate key: Technical unique identifier without any business meaning	INT	15862299
pat_publn_id	PATSTAT Patent publication identification. Surrogate key for patent publications.	INT	15862299
docdb_family_id	Identifier of a DOCDB simple family. Means that most probably the applications share exactly the same priorities (Paris Convention or technical relation or others).	INT	15862299

[41] github.com/cverluise/PatCit/tree/0.3.1

Name	Description	Type	Nb non null
inpadoc__family__id	Identifier of an INPADOC extended priority family. Means that the applications share a priority directly or indirectly via a third application.	INT	15862299
citation		REC	16781144
____.country__code	Country code of the cited patent. Parsed by Grobid.	STR	64185636
____.original__number	Original number of the cited patent. Parsed by Grobid.	STR	64185636
____.kind__code	Kind code of the cited patent. Parsed by Grobid.	STR	6096368
____.status	The status of the cited patent. Parsed by Grobid.	STR	64185636
____.pubnum	Concatenation of country code, original number and kind code of the cited patent. Based on attributes parsed attributes.	STR	64185636
____.publication__number	Publication number of the cited patent. Obtained from the google patent linking API.	STR	49542360
____.publication__date	Publication date (yyyymmdd) of the cited patent based on the matched publication__number.	INT	49231609
____.appln__id	PATSTAT application identification of the cited patent. Based on the matched publication__number. Surrogate key: Technical unique identifier without any business meaning.	INT	49231609
____.pat__publn__id	PATSTAT Patent publication identification of the cited patent. Based on the matched publication__number. Surrogate key for patent publications.	INT	49231609
____.docdb__family__id	Identifier of a DOCDB simple family of the cited patent. Based on the matched publication__number. Means that most probably the applications share exactly the same priorities (Paris Convention or technical relation or others).	INT	49231609

Name	**Description**	**Type**	**Nb non null**
___.inpadoc_family_id	Identifier of an INPADOC extended priority family of the cited patent. Based on the matched publication_number. Means that the applications share a priority directly or indirectly via a third application.	STR	49231609
___.flag	Flag detected citations which are likely to be in the header rather than in the specification itself. Flag is True for citations extracted from patents published in the pre-1976 format and with all occurrences detected before character 50 or in the last 4 percent of the text. It is recommended to exclude those citations from most analyses.	BOOL	71407446
___.char_start	First character of the detected cited patent. Refers to description_localized.text in patents-public-data.patents.publications.	INT	71407446
___.char_end	Last character of the detected cited patent. Refers to description_localized.text in patents-public-data.patents.publications.	INT	71407446

Notes: Nested variables are denoted by a dot. For instance, ___.country_code is the country code of a cited patent nested in the citation variable.

Chapter 2

What Do we Learn from One Century of Innovation in Europe and the US?

Antonin Bergeaud. HEC Paris & Collège de France
Cyril Verluise. Paris Scool of Economics & Collège de France

This version: January 2022

Abstract

Innovation is an important driver of potential growth but quantitative evidence on the dynamics of innovative activities in the long-run are hardly documented due to the lack of data, especially in Europe. In this paper, we introduce a novel dataset on the location and nature of patentees from the 19^{th} century using information derived from an automated extraction of relevant information from patent documents published by the German, French, British and US Intellectual Property offices. We then propose an exploration of this new dataset and describe key facts on the geography of innovation, the role of migration, and the nature of inventors over more than a century. Our results exhibit important differences across the four countries considered.

JEL classification: N90, O30, Y10
Keywords: history of innovation, patent, text as data

2.1 Introduction

Endogenous growth theories (Romer, 1990; Aghion and Howitt, 1992a) have placed innovation at the hearth of the driving forces behind long-run growth. In parallel, the availability of new quantitative data has paved the way for numerous studies analyzing the social and economic implications of innovation activities and describing the environment favorable to enhance innovation (see Hall and Harhoff, 2012 for a review). Most of these studies use patent documents to measure innovation across time and space. While patents are arguably imperfect and incomplete proxies for innovation – not all inventions are patented with heterogeneity in patenting propensity across countries, time, sectors and firm size, see Arundel and Kabla, 1998; Mansfield, 1986 – they are however widely used in the economic literature because of the rich set of information embedded in patent publications and their availability. In addition, despite their known limitations, evidence shows that the use of patents as a measure for innovation nevertheless provides a relevant signal (they are in particular well correlated with R&D activities, see Pakes and Griliches, 1980; Acs and Audretsch, 1989).

The patent system has been in place for a very long time[1], yet only limited information is available before the 1980s and publications did not systematically exist before the end of the 19^{th} century in most countries. One known exception is the United States Patent and Trademark Office (USPTO) which consistently published patents since 1836 and made them publicly available[2]. In this specific case, extracting the information of interest (e.g., inventors, assignees, locations. . .) can therefore be performed in a single step; either manually or using simple semantic rules. This has motivated early efforts to exploit and study parts of this rich corpus of documents (e.g. Lamoreaux and Sokoloff, 1997, 2000; Sokoloff, 1988). Recent improvements in large data handling and text data processing have stimulated a renewed interest in large scale use of historical patents (in particular Packalen and Bhattacharya, 2015; Petralia et al., 2016; Akcigit et al., 2017b; Berkes, 2018b; Sarada et al., 2019). Thus far, this momentum has mostly been restricted to US patents - notably due to the public availability of US patents *text* data.[3]

Consequently, our understanding of the long-term development of innovative activities is largely based on a US perspective. In contrast, we do not know much about the forces at stake in other major innovative countries, namely European technological leaders, before the dawn of the 21^{st} century. In particular, the location, occupation and citizenship of patentees (inventors or assignees), which are key to the study of innovation dynamics, are unavailable from standard patent datasets such as PATSTAT and Claims before the 1980s. However, most historical patent documents are available as scanned images. Starting from these images and using a

[1] It is commonly acknowledged that the first British patent was granted to John of Utynam in 1449, see Plasseraud and Savignon (1983).

[2] USPTO patent publication texts are publicly available for bulk download from the USPTO website and the Google Patents public dataset.

[3] With some notable exceptions that restrict to patents published before the 19^{th} century, see e.g. Hanlon (2016); Nuvolari and Tartari (2011); Nuvolari et al. (2020, 2021), but do not focus on the geography of patentees.

pipeline of data science and Natural Language Processing (NLP) steps, we extend previous work restricted to US patents, both in terms of coverage and methodology. Using raw images of patent documents as our input, we extracted and structured the embedded information and finally produced a relational database covering patents published in Germany (including East Germany), France, the United Kingdom, and the US since the 19^{th} century.

Next, from this novel database, we investigate a number of historical facts. We first show that innovation is a highly regionally concentrated activity, even when population is controlled for. Yet, the level of spatial concentration and its development over the 20^{th} century show clear differences across the four countries. The US starts from a slightly more multi-polar model with several clusters located in the North East, while at the same time, the capital regions of European countries (Paris, Berlin, and London) accounted for the very large majority of patentees. The 20^{th} was a period of progressively moving to a less concentrated model and in particular after WW2. At the end of the century, the geography of innovation in the US and Germany was characterized by multiple centers, while France continued to be highly concentrated and the United Kingdom in between. Second, we look at the internationalization of innovation, as measured by the share of foreign inventors in each patent offices. The results clearly indicate a decline in domestic bias during the period following WW2. Moreover, when looking at the country of residence of foreign inventors, we notice that flows of innovation, like trade flows, follow a gravity law: closer countries tend to exchange more. Third, we look at the inventors and exploit the information on their location and occupation to show how the typical profile of an inventor has changed over time. Specifically, we document a change in the nature of patent activities, with more and more inventors declaring to be engineers in the United Kingdom and Germany and less and less single inventor patents. We also exploit the very high details on the location of patentees in British patents to look at the correlation between the wealth of the neighborhood and the likelihood of becoming an inventor using data on London at the turn of the 19^{th} century. We find that the probability of becoming an inventor increases with wealth, except for the wealthiest districts of London. Finally, we leverage information on the citizenship of domestic inventors in the United Kingdom and the US to look at the nature of innovation done by immigrants. We show that immigrants tend to patent in different technological fields than other inventors, and typically do more radical innovation, as measured by the novelty of the publication.

Our contribution is therefore threefold. First, to the best of our knowledge, our database is the largest of its kind, both in terms of time-space coverage and scope of applications.[4] Second, we use our database to document historical facts on the nature and dynamics of innovation over the 20^{th} century. Third, we make our dataset open access and we open source tools to

[4] Key aspects of the database, such as the occupation of the inventors, are entirely absent from US patents, hence from existing databases, but still available from East-German, German and British patents. Our database therefore offers the unprecedented opportunity to shed new lights on the nature of inventors over time *inter alia* as we will develop in this paper.

help the community build on/extend our work.[5] Despite the large number of efforts in the field for US data, we are not aware of any other publicly available database to date with similar coverage. We have also made the database as interoperational as possible. Each patent and geographical information are associated with standard identifiers that should facilitate the matching of Patentcity with other data source.

We hope that this work will encourage researchers to use and extend our work to complete our knowledge on innovation in the 20^{th} century and earlier.

Related literature. Our project relates to the growing and recent literature that aims at overcoming the lack of historical data on the location of innovative activities using patent documents. We have already mentioned early effort by Lamoreaux and Sokoloff (1997, 2000); Sokoloff (1988) which are based on a small sample of patents that are manually classified and geocoded. More recently, Nicholas (2010) studied innovation activities between 1880 and 1930 in the US thanks to the construction of a new dataset that restrict to a 10% sample of USPTO patents that were not associated with a specific assignee. Since then, other datasets have extended this work by implementing automatic rules to the text of the patent publications to extract relevant information, namely Sarada et al. (2019); Packalen and Bhattacharya (2015); Berkes (2018b); Berkes and Gaetani (2019); Akcigit et al. (2017b, 2018) and Petralia et al. (2016). These datasets follow different purposes. For example Akcigit et al. (2018) use patent data to measure the impact of taxes on individual inventors and firms, Berkes and Gaetani (2019) look at the geographical concentration of innovation in history and Packalen and Bhattacharya (2015) analyze the role of physical proximity as an engine for new ideas and innovation. They also differ in the nature of the information they focus on, their time frame and the way they collect the data. The accuracy of these databases is usually high based on different criteria and despite their differences, they paint a consistent picture of the nature of inventions in the history of the US (see Andrews, 2019 for a comparison of existing datasets). However, all these datasets focus on USPTO patents only and do not include information on patents filed in other patent offices. Of course, some scholars have studied innovation in Europe and before WW2 in the past, either using alternative data (e.g., Moser, 2005) or using a subset of patents (e.g. Nuvolari and Tartari, 2011; Nuvolari and Vasta, 2017; Andersson and Tell, 2018). However, none of these projects attempted to add geographical information to a comprehensive set of patents. For the more recent period, one notable exception is de Rassenfosse et al. (2019c) who used information available from the patent office registers on the address of patentees to geocode assignees and inventors' locations all over the world since the 1980s. This of course includes the four countries we are focusing on. We view our work as completing these projects by extending these works either in time or in space thanks to substantial methodological novelties.

[5] The pipeline code base is publicly available and fully documented on the GitHub repository of the project at www.github.com/cverluise/patentcity. Non technical additional material is also available on the project website at www.patentcity.xyz.

In this paper, we also present examples of analysis that can be done from this new dataset. In particular, we derive historical facts about the geography of innovation. This relates to a rich literature that considers the spatial heterogeneity of innovative activities. Based on relatively recent data, this literature has highlighted that innovative firms, research laboratories, and universities but also venture capital funds are geographically organized around clusters (e.g., Delgado et al., 2010; Hausman, 2020). As a result, the modern geography of innovation is characterized by local specialization hubs (Egger and Loumeau, 2018; Buzard et al., 2017) and by the existence of superstar cities (Gyourko et al., 2013; Carlino et al., 2007). These clusters facilitate the diffusion of knowledge (Rosenthal and Strange, 2003; Feldman and Kogler, 2010) and maximize the positive social spillovers that innovation generates (Audretsch and Feldman, 1996; Klenow and Rodriguez-Clare, 2005). With our data, we approach this question using a long-term view which allows us to look at how the concentration of innovation has changed over time, and whether these developments are different across countries.

Our dataset also contains some information about the citizenship of inventors which allows us to derive some results about the impact of migration on innovation. This naturally speaks to the literature on the relation between immigration and innovation which consistently finds that immigration is a privileged vehicle for importing knowledge. For example, Bahar et al. (2020) uses a large set of countries and recent data and document that the probability of a country to experience an abnormal momentum in patenting activity in a technological field is positively affected by an increase in the influx of migrants coming from a country with a patenting advantage in this field. Bernstein et al. (2018) show evidence for this using data for the US since the 1990s. In addition to relying on different knowledge and being more productive than their domestic counterparts, foreign-born inventors also generate larger spillovers. This was notably the case for Jewish chemists fleeing the Nazi as studied by Moser et al. (2014) whose overall impact on innovation largely exceeded their personal contribution. On the other hand, Borjas and Doran (2012) show that immigration of scientists can have a negative business-stealing effects on the productivity of domestic scientists, but this adverse effect is more likely to materialize in very constrained labor markets (in their case, mathematicians in academia). In terms of historical trends, Akcigit et al. (2017b) and Arkolakis et al. (2020) provide large scale historical research stressing the crucial role of the 1880-1940 immigration on the dynamics of US innovation. Specifically, Arkolakis et al. (2020) find that European immigrants spurred more radical innovations compared to domestic inventors while Akcigit et al. (2017b) find that the specific expertise brought by immigrants during the 1880-1940 period resulted in more patenting in these areas in the 1940-2000 period. In the case of the United Kingdom and the US, our dataset contains direct information on the citizenship of inventors which allows us to look more directly into these questions.

More generally, looking at inventors over the 20^{th} century brings interesting information about how innovative activities have changed (see e.g., Akcigit et al., 2017b; Berkes, 2018b), in particular in time of crisis (Babina et al., 2020). This is what is done in Akcigit et al. (2017c) and Sarada et al. (2019) which have both documented that most US inventors are white males but that this pattern changes slightly over time. Sarada et al. (2019) also reports that the typical occupation of an inventor moves away from farming to engineer and scientists. In these two studies, information is obtained by matching the name reported in patent publications to different vintage of the census. In this study, we complement these findings by looking at the information directly reported in patent publications. In addition to the citizenship, we also extract and use information on the occupation of inventors whose changes in a window on the evolution of the nature of innovative activities.

From a data perspective, our work borrows a lot from modern NLP, in particular to the Named Entity Recognition (NER) field. This strand of literature seeks to develop algorithms to detect mentions of predefined semantic types, either generic (e.g., person, organization, location, etc) or domain specific (e.g., assignee, inventor, occupation, etc). Two approaches coexist in the literature. First, the rule-based and statistical methods (see Li et al., 2020 for an in-depth survey of the NER literature). Rule based approaches usually leverage large domain specific gazetteers (Etzioni et al., 2005, Sekine and Nobata, 2004) and syntactic-lexical patterns (Zhang and Elhadad, 2013). However, this approach is largely unable to handle inherent ambiguities of natural language and to generalise to new documents. To overcome these limitations, the literature has introduced statistical approaches. Starting with text data annotated by humans with entity labels, machine learning algorithms are trained to learn a model to recognise similar patterns from unseen data. The first generation of this class of algorithms, notably including Hidden Markov Models (Eddy, 1996) and Conditional Random Fields (Lafferty et al., 2001), typically rely on feature engineering. More recently, statistical approaches leveraging deep learning have repeatedly advanced the state-of-the-art performance in the field. Such models are able to exploit non linearity to uncover complex and hidden features automatically, without the need for feature engineering or built-in domain expertise (Collobert et al., 2011; Huang et al., 2015; Lample et al., 2016; Chiu and Nichols, 2016; Peters et al., 2017). The class of models we use to extract relevant data from the patent documents belongs to this latter group.

The rest of the paper is organized as follows. Section 2.2 describes our novel dataset. The following sections report major historical facts on the dynamics of innovation since the 19^{th} century in Germany, France, Great-Britain, and the United States of America. Section 2.3 chronicles the concentration of inventive activity within countries across time. Section 2.4 informs the magnitude and direction of the international flow of innovation between countries over time. Section 2.5 portrays the socioeconomic characteristics of inventors. Section 2.6 examines the role of migration in inventive activity within countries.

2.2 Data

In this section, we detail the construction of the database. The key steps are the following. We start by collecting the patent document images. We convert these document into text data using Optical Character Recognition (OCR). We then leverage modern Named Entity Recognition (NER) techniques to extract the relevant information from the patent text: the name of inventors and assignees, and, if available, their locations, occupations, and citizenship. These attributes are then tied together using a simple relationship prediction algorithm (e.g., an inventor is linked to his location). Finally, we enrich the dataset, in particular we take care to translate the extracted natural language text spans into harmonized attributes. In particular, we geocode the extracted locations and provide administrative codes to facilitate the interoperability of the database with other sources. Figure 2.46 summarizes the workflow that we describe in detail in this section.[6]

2.2.1 Data collection and coverage

Contrary to the USPTO, patent publications from the East German, German, French and British intellectual property offices are not publicly available for bulk download in text format.[7] To overcome this obstacle, we scraped the patent document images and extracted the embedded text using `Tesseract v5.0` (Kay, 2007), a popular open-source OCR software. A qualitative assessment of the results showed that the quality of the text of USPTO patents could be improved by using the latest version of `Tesseract` compared to the text provided by the USPTO itself and generated by former OCR technologies. Hence, we used the patent images made available by the USPTO and implemented in-house OCR in order to maximize the quality of the text and to make our dataset more consistent across different patent offices.

We restrict attention to utility patents. Utility patents are the class of patents which cover the creation of a new or improved –and useful– product, process, or machine.[8]. For the sake of brevity, we refer to utility patents as as patents thereafter. As previously mentioned, we focus on patents published by the East German, German, French, British and US patent offices. Data collection is subject to two conditions. First, we need patent publications to exist and to be available in a digital image format. Second, we need these documents to include at least some geographical information. These conditions have been met consistently for patents published between 1950 and 1992 for East-German patents (with the exception of the period 1973 and 1976), from 1877 for German patents, from 1903 for French patents, from 1893 for British patents and from 1836 for US patents. Starting from those publication dates, we collect all patents published until 1980. Overall, this represents around 8.9 million documents.

After 1980, we complete our data using the work of de Rassenfosse et al. (2019c) which reports the patentees location for a very large corpus of patents, including publications from the patent

[6] The codebase is open source and fully documented on the project GitHub repository.

[7] Patent search engines such as EspaceNet and Google Patents enable manual patent download on a per-document basis. Unfortunately, both of them impose quotas on the daily number of downloads.

[8] Utility patents cohabit with other types of patents. They are usually identified by a set of kind codes, that is the last letter of the DOCDB publication number. Appendix 2.8 reports the list of kind codes selected as referring to utility patents for each patent office.

offices we are interested in. When necessary, we completed their corpus with patents published after 1980 but missing from their dataset to make sure that the transition between the two datasets is smooth.[9] Our dataset comprehensively[10] spans over the following periods: 1877-1980 for German patents, 1950-1972 and 1977-1992 for East German patents,[11] 1903-1980 for French Patents, 1893-1980 for British patents and 1836-1980 for US patents. After 1980, our dataset smoothly splines over de Rassenfosse et al. (2019c)'s which provides data up until 2013 included.

Overall, we find that our dataset covers between 80% and 100% of the published patents reported by the IFI Claims and PATSTAT datasets[12] for each patent office and publication year under consideration (see Appendices 2.42 and 2.44). It should also be stressed that after the creation of the European Patent Office (EPO) in 1977, patents granted by the EPO and validated by the national offices for domestic usage are still reported by our dataset.

2.2.2 Information extraction

Our information extraction pipeline is made of two layers. First, a NER model in charge of extracting the entities of interest. Second, a relationship prediction model which role is to resolve the relations between the extracted entities. Both layers are crucial to fully exploit the potential of patent texts.

Entities

Our goal is to extract the names of the inventors, the names of the assignees but also their location, occupation and citizenship when applicable. The exact definition and actual examples by countries are reported in Table 2.1 and discussed in Appendix **??**. This is naturally subject to the actual reporting of these entities in the text of the patent. The reason why we focus on this set of information is largely influenced by the last decades of the innovation literature. The relation between geography and innovation occupies a central place in this literature. The occupation of inventors also constitutes a valuable asset to study their socio-economic characteristics. Eventually, the combination of inventors' citizenship and location provides their immigration status, which appears to be key to understand innovation dynamics. One important remark is that the very notion of inventor and assignee is mainly a US and modern times terminology. In many offices and at many points in time, there is no explicit distinction between the two. In this case, we called inventors any human being involved in the invention and assignee any

[9] In particular, we collected patents from the East German patent office until the last one in 1992

[10] Depending on the office, our coverage varies between 98% and 100% of the utility patents listed in the Google Patents Public Data, the largest publicly available bibliographic dataset of patent publications.

[11] To our knowledge, digitized copies of East German patent documents published between 1973 and 1976 are not available.

[12] The IFI Claims and PATSTAT datasets are the two standard patent datasets used in academia.

company related to the invention.[13]

Table 2.1 summarizes the entities extracted by patent office. We were able to extract the names of the inventors and assignees and their locations from all patent offices. In contrast, the occupation and citizenship are only available for some countries. Specifically, the occupation is reported in East-Germany, Germany and the United Kingdom while the citizenship is reported in the United Kingdom and the US. Importantly, even within a given patent office, the reporting of a given entity can vary over time. See Appendix 2.8 for more details on the share of patents from which we extracted at least one entity of each category by publication year and countries. Similarly, the level of precision of the location (i.e. country, state, county, ...) changes across time and countries. More details are provided in Figure 2.34.

Table 2.1: Entities extracted by countries

	DD	DE	FR	GB	US
E-Inventor	✓	✓	✓	✓	✓
E-Assignee	✓	✓	✓	✓	✓
E-Location	✓	✓	✓	✓	✓
E-Occupation	✓	✓		✓	
E-Citizenship				✓	✓
Time span	1950-1992	1877-1980	1903-1980	1893-1979	1836-1976

Notes: The prefix E refers to "Entity" and is added to make sure that they entities not confounded with relationships designated with similar names and reported with a R prefix. The actual reporting of the entities can vary over time. See Appendix **??** for more details on the share of patents from which we extracted at least one entity of each category by publication year and countries. This table only reports the entities extracted in the course of this project. Later results incorporate de Rassenfosse et al. (2019c) dataset which provides the names and locations of German, French, British and US patentees after the end of our dataset. DD stands for East Germany, DE for Germany (which only includes West Germany during the 1950-1989 period), FR for France, GB for the United Kingdom and US for the United States of America.

Named Entity Recognition

Meta-data (e.g., patentees' names and locations) on historical patents are reported in an unstructured way, most often as part of the first paragraph or in the header of the document. Table 2.2 shows typical examples for each patent office. To our knowledge, previous historical patent data projects used rule-based methods to extract such domain-specific data. Instead, we use deep-learning based statistical NER. As previously explained in the literature review, this class of models have been conceived by the NLP community specifically to improve on rule-based approaches and have repeatedly advanced the state-of-the-art since their introduction. In

[13] This is a necessary but arbitrary point which has important implication for comparability across countries. For example: French patents most of the time did not explicitly report the name of the inventor but only the name of the "déposant" (applicant). In some cases, this applicant is a firm and in other cases a physical person. In rare instances, the name of the inventors are given in addition to the name of the applicant. For this reason, we chose to define this applicant as an assignee. See Appendix 2.8 for more details.

our specific case, they also present the advantage to have considerable generalization abilities based on a relatively small amount of examples - making them robust to typos and variations in word-use which can be very frequent at some patent offices and would give rule-based models a hard time. It is also worth noting that, contrary to most previous works, we produced and release manually annotated data making which supports rigorous and transparent performance evaluation and future extensions.[14]

Table 2.2: Example of patent documents with embedded entities

Country	Example	Source
DD	*Erfinder: Wilhem Uhrig, WD. Inhaber: Dr. Plate GmbH, Bonn, WD.*	DD-79836-A
DE	*Bela Barenyi, Stuttgart-Rohr, ist als Erfinder genannt worden. DAIMLER-BENZ Aktiengesellschaft, Stuttgart-Unterturkheim*	DE-869602-C
FR	*MM. Joseph MARTINENGO et Jean-Baptiste GAUDON résidant en France (Loire)*	FR-504101-A
GB	*We William Christopher Fanner, and Henry Elfick, trading together as De Grave, Short, Fanner & Co., of Farringdon Road in the County of London, Scale and Balance Manufacturer, do hereby declare the nature of this invention...*	GB-189704983-A
US	*Be it known that I, PAUL SCHMITZ, a subject of the King of Prussia, German Emperor, residing at Cologne-Niehl, in the Kingdom of Prussia, German Empire, have invented*	US-1108402-A

Notes: Examples of patent document for each of the fifth patent offices considered. Colored text correspond at the entities that we seek to extract: red for inventors, purple for assignees, olive for locations, brown for citizenship and blue for occupations.

In practice, the NER models were trained using `spaCy v3` (Honnibal et al., 2020), a popular Python NLP library offering an efficient framework for reproducible custom domain NLP models. The manually labeled dataset was split in two subsets, the training set used for model training and the test set, used for model's performance evaluation. The goal of this approach is to avoid over-fitting, that is the tendency of the model to "learn training data by heart" which can produce very high performance on the training set while harming its ability to generalize to other data. Each office was treated independently from one another and multiple models were trained for offices to account for the large changes in the format of the patents (see Appendix 2.8). More details are provided in Appendix **??**.

In Table 2.3, we report the performance of the models on the test sets for each entity of interest. The performance metrics are respectively: the precision, that is the share of *extracted* entities which are *actual* entities; the recall, that is the share of *actual* entities which are indeed *extracted* and the F1-score, the geometric mean of the precision and the recall. In short, the higher the F1-score, the better the reliability of the model. For the sake of brevity, we average over models

[14] For the labeling tasks, we used `Prodigy v1.10` (Montani and Honnibal, 2018). Data and annotation guidelines are available on the project GitHub repository at https://github.com/cverluise/patentcity.

performance when there was more than one data format, hence models, for a given office. We report in brackets the underlying number of models. The average F1-score over all extracted entities ranges from 0.94 to 0.98 on the test set which indicates a high level of performance.

Table 2.3: Performance of the NER models

	DD (2)	DE (2)	FR (2)	GB (1)	US (4)
E-Inventor	0.95/0.95/0.96	0.98/0.97/0.98	0.99/0.99/0.98	0.95/0.96/0.96	0.99/0.99/0.99
E-Assignee	0.97/0.97/0.97	0.98/0.98/0.98	0.98/0.98/0.98	0.93/0.92/0.93	0.96/0.96/0.96
E-Location	0.98/0.97/0.97	0.99/0.99/0.99	0.99/0.99/0.99	0.92/0.92/0.92	0.98/0.98/0.98
E-Occupation	0.96/0.97/0.96	0.97/0.97/0.97	-	0.90/0.86/0.88	-
E-Citizenship	-	-	-	0.96/0.96/0.96	0.98/0.98/0.98
E-All	0.97/0.96/0.97	0.99/0.98/0.98	0.97/0.97/0.97	0.93/0.94/0.94	0.98/0.98/0.98

Notes: The prefix E refers to "Entity" and is added to make sure that they entities not confounded with relationships designated with similar names and reported with a R prefix. Reported performance metrics were computed on the test set - unseen during training. The figure in brackets indicates the number of different models used for the office. For example, for the German office, there was a major shift in the patent information display in 1881 forcing us to train two different models (see Appendix 2.8). Performance metrics are reported as follows: precision/recall/F1-score. Model by model performance for each patent offices can be found in Appendix **??**.

Relationship prediction

At this stage, we have extracted the information of interest from a patent with a high level of reliability but the output is basically a "bag" of entities. For example, assuming that we have extracted one inventor, one assignee and two locations, then we still do not know which one is located where. Such relationship can be extremely detrimental to the analysis. For instance, if we want to know whether an inventor is an immigrant, we need to link its name to a citizenship and to a location. For this reason, we go one step further and reconstruct the latent relationships between our different entities. That is what we call relationship prediction and to the best of our knowledge, we are the first to implement this kind of approach in the field.

In our case, there are three different kinds of relationships: the *location* which relates the patentee to his address, the *occupation* which relates the patentee to his occupation, or academic title and the *citizenship* which relates the patentee to its citizenship or country of origin. There are many different ways to implement such relationship prediction but we found that a simple algorithmic approach leveraging the relative position and the absolute distance of the attributes (location, occupation, citizenship) to the patentees (inventor, assignee) with a slight level of hyperparameter fine tuning performs surprisingly well. Our approach is the following: we iterate over extracted patentees, harvest all attributes positioned either at the right or left of the patentee within a distance expressed in terms of number of words (or tokens) and keep the closest element of each attribute family (if any). In this algorithm, two hyperparameters need to be chosen: the position (right, left, both) and the size of the window (expressed in tokens).

We evaluate the performance of this procedure on a set that has been manually annotated in Table 2.4. Since parameter fitting remains minor, we considered that the risk of overfitting is relatively small and did not split the labeled set in a training and test set and report performance on the training set. Same as before, we average performances over the different models for each patent offices for simplicity. The overall F1 score varies from 0.93 to 0.98 depending on the office, which guarantees a high level of confidence.

Table 2.4: Performance of the relationship prediction models

	DD (2)	DE (2)	FR (2)	GB (1)	US (4)
R-Location	0.98/0.96/0.97	0.99/0.99/0.99	0.98/0.97/0.98	0.97/0.92/0.94	0.98/0.93/0.95
R-Occupation	0.88/0.86/0.87	0.98/0.99/0.98	-	0.96/0.94/0.95	-
R-Citizenship	-	-	-	0.92/0.93/0.92	0.98/0.97/0.97
R-All	0.94/0.93/0.93	0.98/0.99/0.98	0.98/0.97/0.98	0.95/0.93/0.94	0.97/0.93/0.95

Notes: The prefix R refers to "Relationship" and is added to make sure that relationships are not confounded with entities designated with similar names and reported with a E prefix. The number in brackets indicates the number of different models used for the office (see Appendix 2.8). For example, for the German office, there was a major shift in the patent information display in 1881 forcing us to train two different models. Performance metrics are reported as follows: precision/recall/f1-score. Model by model performance for each patent offices can be found in Appendix **??**.

2.2.3 Data enrichment

At this stage, each patent is characterized by a set of extracted inventors and/or assignees who are themselves characterized by a set of attributes, as is usual in modern patent datasets. Most importantly both the extracted entities and predicted relations exhibit a high level or reliability. However, some limitations remain for research usage. Extracted attributes are reported in raw text, which requires geocoding for locations and further disambiguation for the citizenship. The publication dates from German patents published before 1919 and East German patents published before 1972 are missing from standard datasets, which calls for some additional effort ass well. In this section, we detail how we overcame these limitations and the resulting data enrichment process.

Location geocoding

Our first task is to turn natural language attributes into high quality and harmonised variables. The most challenging and crucial task was certainly the geocoding of natural language locations, that is the translation of free-text locations such as "Farringdon Road in the County of London" (from patent GB-189704983-A) into well defined geographic attributes (country, state, county, . . .) and coordinates. This "geocoding" exercise is well known as challenging and resource intensive due to the many ambiguities and typos that can be found in natural language addresses and the size of the universe of worldwide addresses. In our case, there are the additional difficulties of multiple languages and changing names and borders since the beginning of the considered time span. For all these reasons, we found that the best output quality was only achievable using a commercial geocoding supplier. Having close to 3 million unique addresses to geocode we mixed two providers (HERE and Google Maps) to maximize efficiency. Specifically, we leverage the specific features of the two services: on the one hand, HERE tends to have a low rate of errors but a relatively high rate of "unmatched" locations; on the other hand, Google Maps tends to have a very low rate of unmatched locations -notably thanks to a better understanding of locations expressed in plain – language and historical entities which have changed names (e.g., "Karl-Marx Stadt" in East Germany now known as "Chemnitz")–, sometimes at the expense of a slightly higher error rate (see Perlman et al. (2016) for a discussion of the geocoding of historical patent using modern Geographic Information System). With that in mind, we decided to get the best of both worlds. We first processed locations through HERE batch geocoding API and then restricted Google Maps geocoding to the unmatched locations.[15] The two outputs were relatively straightforward to align in a common data structure.

Table 2.5 presents the share of matched locations together with the level of quality of the geocoding (conditional on match). Tthe geocoding output was validated by hand. The human annotator was given both the extracted location and the geocoded address. He would then choose from a set of options (country, state, county, . . .) to select the finest geographic level at which the location was rightly geocoded. The share of locations matched varies from 88.3% for the British patents to 99% for French patents. Conditional on matching an address, more than 92% of the locations are rightly geocoded at the country level for all offices. This figure can even exceed 98% for French and US patents. Results at more detailed geographic levels vary depending on how detailed the location was in the patent document itself. It goes up to 95% at the city level for German and US patents versus only 33.5% for French patents.

Table 2.5: Performance of the geocoding

	DD	DE	FR	GB	US
Match	0.987	0.976	0.990	0.883	0.975
Country	0.927	0.971	0.986	0.934	0.985
State	0.576	0.957	0.483	0.924	0.982
County	0.569	0.953	0.456	0.910	0.968
City	0.569	0.950	0.335	0.887	0.951
Postal Code	0.116	0.251	0.006	0.727	0.185
District	0.109	0.226	0.006	0.690	0.085
Street	0.014	0.035	0	0.605	0.034
House number	0.007	0.010	0	0.394	0.002

Notes: The match rate is the share of locations for which either HERE or Google Maps found an address. The match rate is based on the *entire* dataset. Conditional on a match, other figures represent the share of locations which were rightly geocoded at a given geographic level based on the manually validated sample. For instance, for German patents, 97.6% of the extracted locations were matched and 95% of the matched addresses were right at the City level. These conditional figures are based on a *manually* annotated sample.

[15] Both APIs are respectively documented at the following addresses HERE API and Google Maps API.

Citizenship disambiguation

Our second task consisted in turning citizenship statements (e.g., “a citizen of the United States of America”, “a subject of the King of Great Britain”...) into harmonized and unambiguous country codes. This exercise can be seen as a translation task where we start from a finite (but large) set of possible citizenship statements which we want to map to another (smaller) finite set of country codes.[16]

A simple way to implement such mapping is to define a set of regular expressions which, when matched, trigger a pre-determined country code. We collected a list of citizenship and country names together with the corresponding country codes and authorized a small amount of edit distance between the target and the extracted text to account for typos. Confronting the output with a set of manually annotated citizenship, we find that this procedure achieves a satisfying level of accuracy defined as the share of initial citizenship statements mapped to the right country code. We achieve 98.7% and 92.9% accuracy on British and US patents respectively.

Publication date approximation

The final data enrichment exercise was especially crucial for later analysis since it has to do with the time dimension of the dataset. As previously noted, standard datasets do not report the publication date of patents German patents published between 1877 and 1919 and East German patents published between 1950 and 1972. Fortunately, in both cases the publication number can be used in some way to overcome the issue. In the case of Germany, we use Patent Gazette published by the German patent office since 1877[17], take the last publication number reported under the section “*Erteilungen*” (i.e. “Publications”) and define it as the last publication number of the year. We then iterate backward to fill the publication year until we hit the last publication number of the previous year. To our knowledge, East Germany did not generate such a Patent Gazette. Nevertheless, we were able to develop a similar approach based on publication numbers. First, we drew a random sample of undated East German patents. Second, we manually filled their publication date based on the information displayed on the patent itself. Third, we used the clear but imperfect relation between the publication number and the publication year to find thresholds similar to those found in the German Patent Gazette. Specifically, we chose the publication number thresholds so as to maximize the F1-score of the predicted publication year. Doing so, we obtain an overall 93% accuracy of the publication year.

[16] This perspective borrows from the Finite Set Transducer which was developed in early attempts to automate natural language translation.

[17] German Patent Gazette are available for download at the DPMA website.

2.2.4 Interoperability

We format the data into a ready-to-use database at the patent level with nested information. The database full schema is reported in Appendix 2.8. Importantly, every patent entry in the dataset is identified by its DOCDB publication number. A DOCDB publication number has the following form: "CC-NNNNNN-KK" where CC is a two-letter country code, NNNNNN the publication number, and KK the kind code. In addition to identification, the DOCDB publication number also serves as the natural vehicle for interoperability with external datasets including useful variables (e.g., technological class, citations, ...) that are consistently collected by usual patent datasets.

We also harmonize the geographical information that we extracted. For each address, and in addition to field presented in Table 2.5, we give the official administrative code for the corresponding regions at different level. Specifically, we report the Nomenclature of Territorial Units for Statistics (NUTS) level 1, 2 and 3 when applicable for Germany, France and Great Britain, and the county code, Commuting Zone code and state code for the US.

2.2.5 Other data

Finally, we complement our dataset with external data to help the analysis. We first construct estimates of the population at the most detail geographical level we could. That is county for the United Kingdom (in fact, NUTS2 regions), *Regierungsbezirke* in Germany, (former) *Régions* in France and county in the US. These estimates have been retrieve from backdating official recent estimates with different studies, in particular Rosés and Wolf (2018); Eckert et al. (2020) and sources (the INSEE, and Vision of Britain. See Appendix 2.8 for more details.

We also use information on trade flow, GDP per capita and standard "gravity" variables to estimate a gravity equation model for innovation diffusion. These data are taken from the CEPII geodist and Tradhist datasets (see Mayer and Zignago, 2011; Fouquin and Hugot, 2016) and the Maddison project (Bolt and van Zanden, 2020).

In the remaining parts of the paper, we use our dataset to look at some specific questions related to the long-run development of innovation. We first consider the concentration of innovation, then look at the international collaborations and exchanges, we then turn to the nature of inventors, and finally look at the role of migration. All these questions have been the subject of a large literature, but again, mostly focused on the US. We review this literature and show how our novel dataset can help to shed new light on these questions and highlight the difference between the US and Europe when applicable.[18]

[18] Due to the small number of patents retrieved for East Germany, we do not systematically present results for this patent office.

2.3 Concentration of innovation

It is well documented that the current geography of innovation is characterized by a very large degree of spatial concentration, even more than production or population (see Feldman and Kogler, 2010 for a review). This was already the case at the beginning of the 20^{th} century in the four countries that we study. Figure 2.55 in Appendix 2.8 shows that the Gini coefficient of the number of patentees per region was around 0.6 in Germany and in the United Kingdom and 0.8 in France and in the US, while the Gini coefficients of the level of population were respectively equal to 0.3, 0.35, 0.4 and 0.6.[19] This unequal distribution is partly driven by the very large importance taken by a few hubs. For example, the region of Paris (*Ile de France*) accounted for almost three-quarters of all domestic patentees during the first decade of the 20^{th} century, but only 12% of the population. Similarly, during the same decade, Berlin and (Inner) London accounted for 30% of domestic patentees and represented, respectively 7 and 12% of the total national population. In the US, Boston, Chicago and New York, the three most innovative regions during the 1900-1910 decade totaled 25% of all domestic patentees and 11% of the population.

In all cases, these hubs played a detrimental role in fostering innovation but their relative importance changed during the century. We report clear differences across the four countries. In France and in the United Kingdom, the capital regions continued to host a very large number of innovators until the 1980s, and to some extent even after this period in France. In Germany, Berlin became less dominant after WW2 and the regions of Munich and Stuttgart progressively rose as the largest innovative clusters. In the US, the 20^{th} century first saw the development of Chicago as a major innovative region and then, in a second time, the dramatic emergence of Los Angeles and San Francisco as the most innovative areas by the end of the century. These different developments lead to changes in the spatial distribution of innovation. France has remained globally a mono-polar model even though the relative share of Ile-de-France dramatically fell after World War II. Note that the dramatic fall of the relative share of Ile-de-France after WW2 seems to reflect the publication postponing of patents filed before WW2 and immediately followed by the emergence of the post WW2 Innovation landscape characterized by the rise of Lyon in the French innovation landscape. The US returned to a high level of concentration at the end of the 20^{th} century following the rise as California. Only Germany has really switched to a multipolar model with the emergence of several highly innovative regions and the gradual disappearance of Berlin. The case of Great Britain is harder to interpret. The rapid fall in the relative importance of London in the 1980s seems to reflect a change in administrative rules rather than a realistic change in the concentration of innovative activities.

[19] These Gini coefficient have been computed using observation at the regional level, see next Section 2.3.1. The average size of a region is different in the US and in European countries, which makes the comparison of the absolute values of Gini coefficients tricky. However, the difference with the Gini coefficient on the distribution of population remains comparable across countries.

2.3.1 Measure of concentration

One challenge associated with the measure of concentration over time is to find a consistent level of geographical aggregation with limited border changes and information on innovation. This is rather straightforward in the case of France where *départements* – the smallest geographical unit at which we can localize patentee – have essentially not changed since the beginning of the 19^{th} century. It is also natural in the US as county borders have undergone little modifications and can be tracked over time.[20] In the case of the United Kingdom and Germany, the problem is more complex with multiple administrative changes. For these two countries, we exploit the very high level of detail in the geographical information included in the patents to map postal codes to present day boundaries using geographical crosswalk from the European Commission.

As explained in Section 2.2.5, our dataset proposes an harmonized geolocation of patentee at three different administrative levels. The largest level corresponds to the NUTS 1 for European countries and to the states for the US. The second level corresponds respectively to NUTS 2 and to Commuting Zones (CZ) for the US and the third level to NUTS 3 and to counties. In this section, we study the spatial distribution of patentees at the level of aggregation corresponding to NUTS 2 and CZ. We can calculate standard statistics to measure the level of concentration and its development over time. While the analysis could be extended to a finer level given the granularity of the patent data, we limited to NUTS 2 and CZ as this allows us to use our estimates of the local level of population.

2.3.2 Innovation hubs

To look at the development of local innovation hubs, Figure 2.1 reports the time series of the relative share of all patents filed by domestic inventors in the top 10 regions (NUTS 2 and CZ) for each of the four countries. We also report the cumulative share of the number of patentees accounted for by the top 1, 3, 5 and 10 regions every year in Figure 2.3.

The results show some striking differences across countries. The United Kingdom and France have an extremely concentrated location of innovation in their capital region (London and Paris) up to 1980, while other hubs only account for a modest share of total innovation. The shares accounted for by these two leading regions are very high and stable until the 1980s: about 40 and 80%, respectively, but then decline at the benefit of other regions (in particular the regions of Lyon and Toulouse in France). The US exhibits a more complex pattern. First, no commuting zones account for more than 20% of total patenting activities over the whole period. Second, there is a progressive decline of the historical innovation hubs of the East Cost (New York and Boston in particular), first in favor of Chicago during the first half of the 20^{th} century, and then

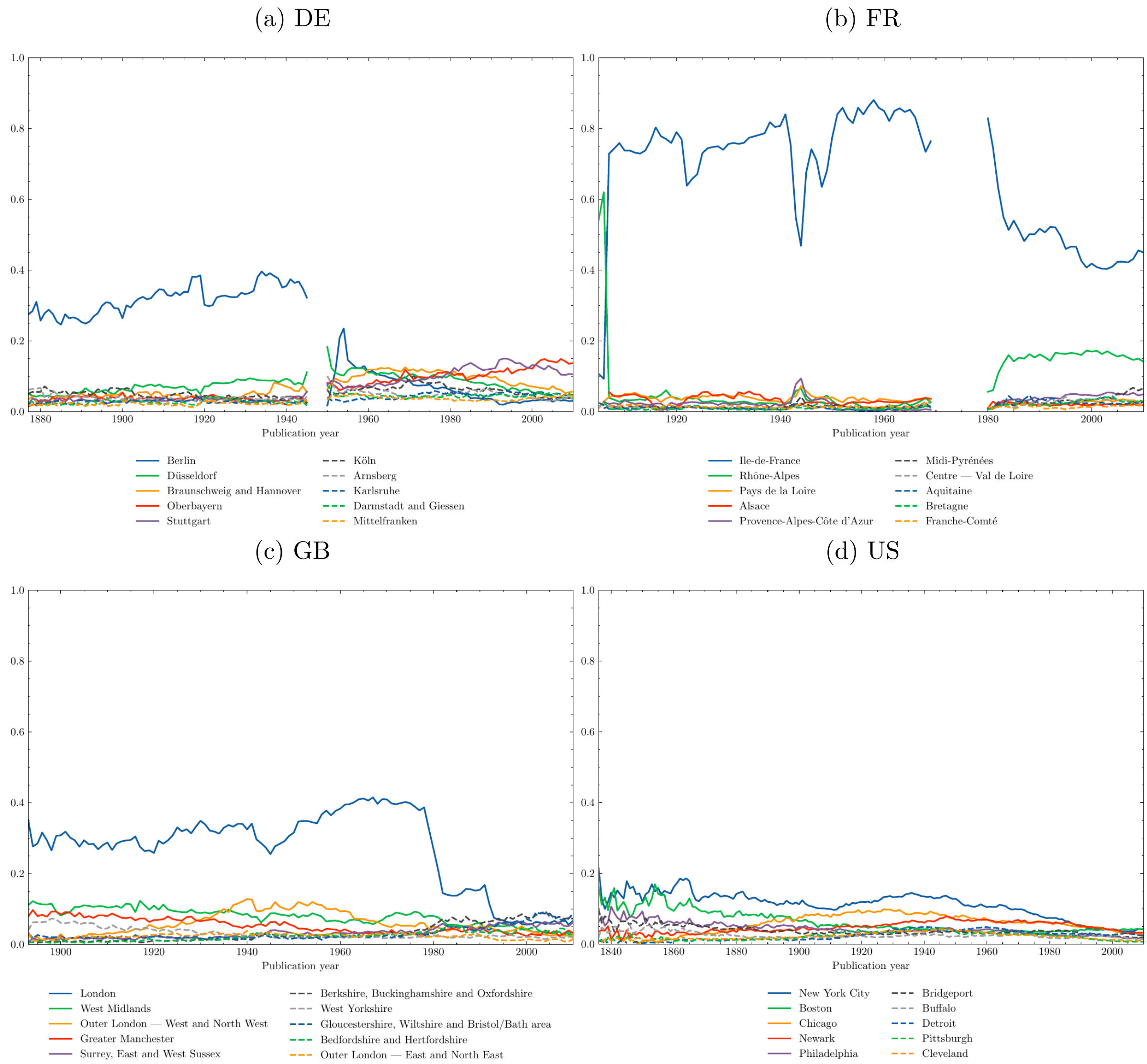
[20] In particular thank to the work of Eckert et al. (2020). See also Perlman et al. (2016) and Bazzi et al. (2020) for discussions on boarder changes in the US.

of California.[21] As a result, the share of innovators located in the top regions increases from the 1990s.

The concentration of innovation in Germany follows a different development. At the beginning of the 20^{th} century, the region of Berlin accounted for a large share of patentees – around 40%, similar to London– and this share remained before WW2. However, after WW2, the relative importance of Berlin collapsed as the regions of Munich and Stuttgart (but also North Westphalia) counted more and more patentees. This results in a clear declining trend in the development of the cumulative share of patentees located in the top regions after WW2, a pattern that only Germany experiences.

[21] The commuting zones of the San Francisco area (San Francisco-Oakland and San Jose) account for a growing share of the total number of patentees after the end of the 1990s, but they are not included in the top 10 CZ because over the whole period, their share is not one of the 10 largest. See Appendix 2.8 for a similar Figure at the state level which shows the spectacular rise of California over time.

Figure 2.1: Share of patents in top 10 regions over time



(a) DE

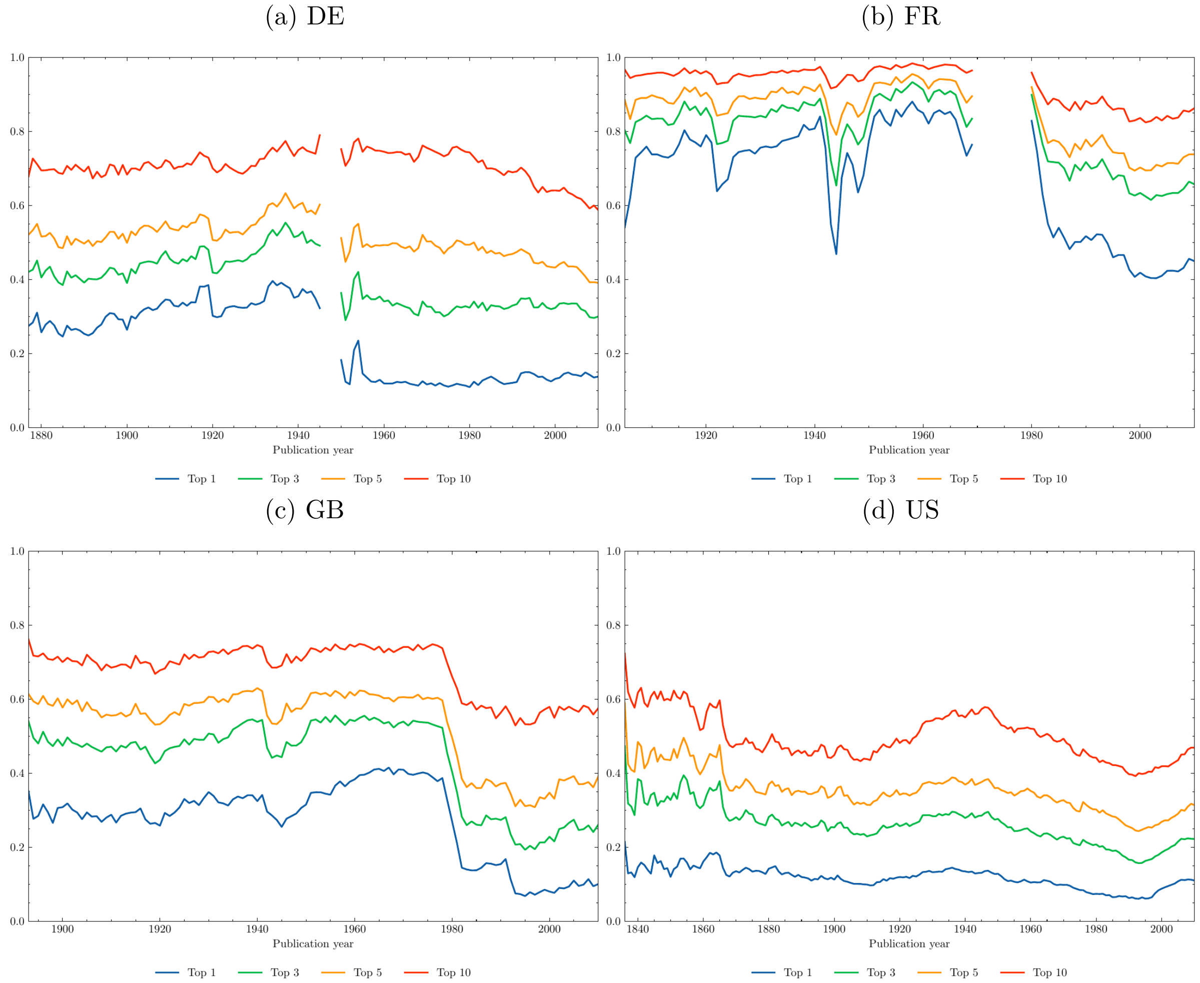
(b) FR

(c) GB

(d) US

Notes: This reports the share of total patentees by regions (NUTS2 + CZ for the US) where the location of a patent is given by the location of its patentees (whether inventors or assignees). The top 10 regions are selected based on the number of years they belong to the top 10 over the whole period of observation. Data for 1946-1950 are not available for Germany, data for 1970-1980 do not report any *département* level information for France.

Figure 2.3: Cumulative share of top regions


(a) DE

(b) FR

(c) GB

(d) US

Notes: This reports the total share of patentees (whether inventors or assignees) by regions (NUTS2 + CZ for the US) accounted for by the top 1, 3, 5 and 10 regions. Top regions are selected each year and can therefore change over time. Data for 1946-1950 are not available for Germany, data for 1970-1980 do not report any *département* level information for France.

2.3.3 Change in the level of concentration

We have so far mostly looked at the top of the spatial distribution of innovation. This does not fully inform about how unequal the overall distribution of innovating activities is, nor do they inform about how this changes over time. One natural way to summarize the full distribution is to calculate the Gini coefficient across all regions of a country every year. This is what is done in Appendix 2.8 and in Andrews and Whalley (2021) in the case of the US since 1836. However, this statistics has some limitations, especially for the US in the early period as many regions have a low level of population and consequently a very small number of patents. We thus consider an alternative metric: the Jensen-Shannon (JS) divergence between the distribution of innovative activities and the population. Intuitively, this distance measures to what extend the two distributions are different. Formally, the JS divergence between two distributions P of population and Q of innovation can be calculated as:

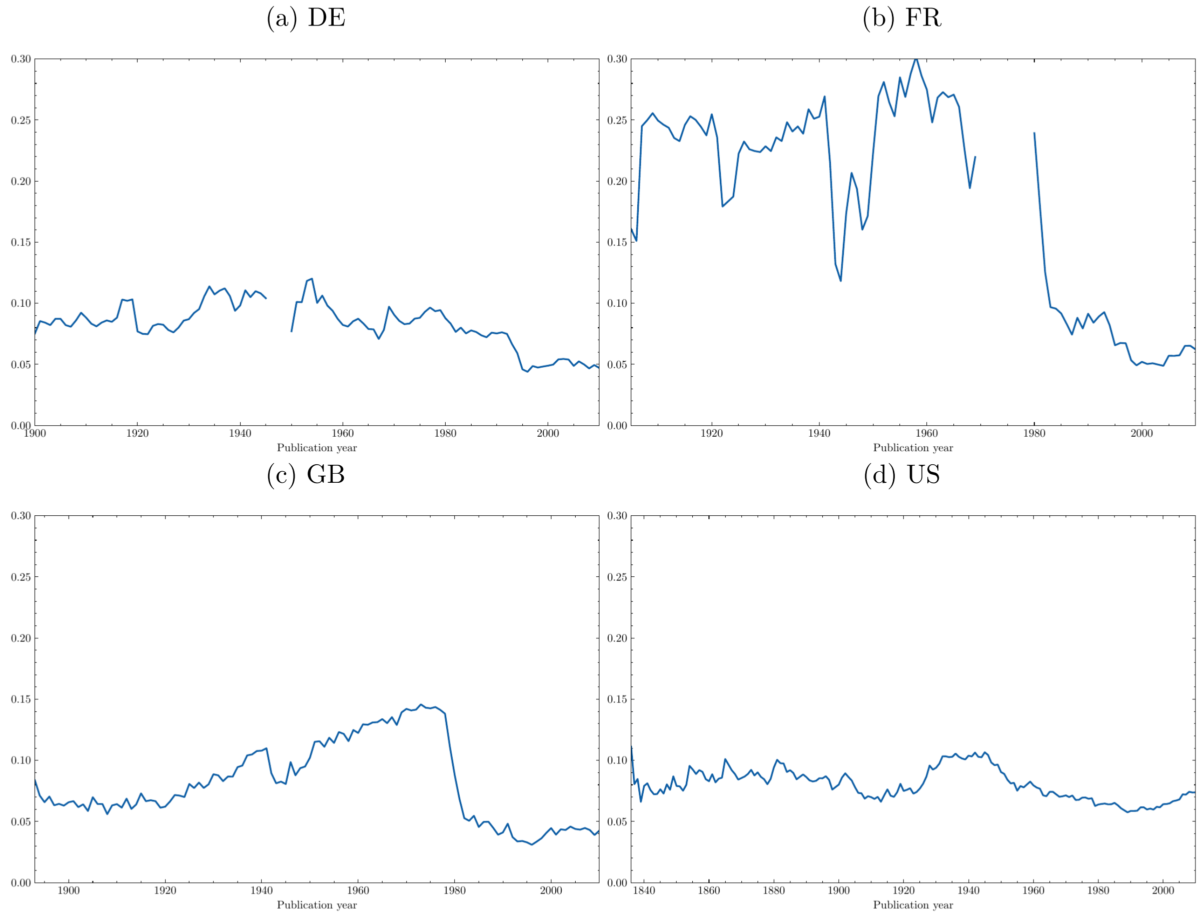
$$JS(P,Q) = \frac{1}{2}\left(KL(P,M) + KL(Q,M)\right) \text{ where } M = \frac{1}{2}(P+Q).$$

In this formula, $KL(P,Q)$ is the Kullback-Leibler divergence over the set of regions Ω:

$$KL(P,Q) = \sum_{\omega\in\Omega} P(\omega)\log\left(\frac{P(\omega)}{Q(\omega)}\right).$$

The yearly value of $JS(P,Q)$ for each country is reported in Figure 2.5. To the extent that the JS divergence measures how much the distribution of patentees differs from the distribution of the population, we interpret an increase of its value as evidence of a larger concentration of innovation activities after controlling for potential changes in the concentration of population.

Figure 2.5: Jensen-Shannon divergence between population and patenting


(a) DE

(b) FR

(c) GB

(d) US

Notes: The Jensen-Shannon divergence is computed for all patentees over the whole period of observation for each country. No data are recorded for Germany during the period 1946-1949, data for 1970-1980 do not report any *département* level information for France. Data for population are also missing for Germany before 1900.

The development of the concentration of patentees, as measured by the JS divergence, varies across countries. For the US, we find a pattern that is consistent with Andrews and Whalley (2021) and in particular a declining trend until the beginning of the 20^{th} century, followed by a period of concentration until WW2. After WW2, the level of concentration declined progressively until the end of the 1990s. In Germany, the level of concentration has reduced since WW2 following an increase during the 1930s. In the UK, the JS divergence increased between 1900 and the mid 1970s. While London continue to be the home of a large but constant share of patentees, population became less centralized during the same period. After this decade, the level of concentration of innovation rapidly collapsed. France experienced a similar trend, although the level of concentration was very high before the 1970s due to the Paris area.

Given that the regional aggregation is not completely comparable across countries, it is not possible to use the magnitude of the JS divergence to compare the level of concentration across countries. However, taken together with the previous analysis, these results paint a consistent picture of the development of the concentration of innovative activities. Patenting in France and in the United Kingdom were both heavily concentration on their capital regions until the 1980s and progressively started to converge to a more even distribution. Germany and the US have a less concentrated model with different clusters, at least since WW2, although the growing importance of California for the past 20 years tend to reverse this trend.

These different long-run changes in the spatial distribution of innovation across countries have many potential explanations. For example, in Germany, WW2 and its legacy reshaped the organization of innovation due to war destruction, the division of the country during Cold War and the brain drain of scientists to the US (see Fohlin, 2016). In the US and in the United Kingdom, universities played an important role, and continue to do so, in spurring local private innovation in the US (see e.g., Andrews, 2020b) while this is less the case in continental Europe as shown by Owen-Smith et al. (2002) in the case of life science. This has naturally impacted the location of inventors differently in the US and in the United Kingdom than it did in continental Europe. In France, the break observed during the 1980s for Paris corresponds to the political will to decentralize the country. Industrial policy is also an important driver, with the development of local specialized clusters such as the region of Toulouse for aeronautics. Demonstrating the causality of these and other factors on the geography of innovation is a critical exercise and we hope that this new dataset will spur new research in this direction.

2.4 International flows

In the previous section, we have only considered domestic innovation and have ignored foreign inventors and assignees. In this Section, we look at the international flow of innovation across countries and over time. International transfers of technology are important drivers of long-term growth and the literature has established that openness to international trade has indeed large positive impacts in terms of productivity (Coe and Helpman, 1995; Keller and Yeaple, 2009).

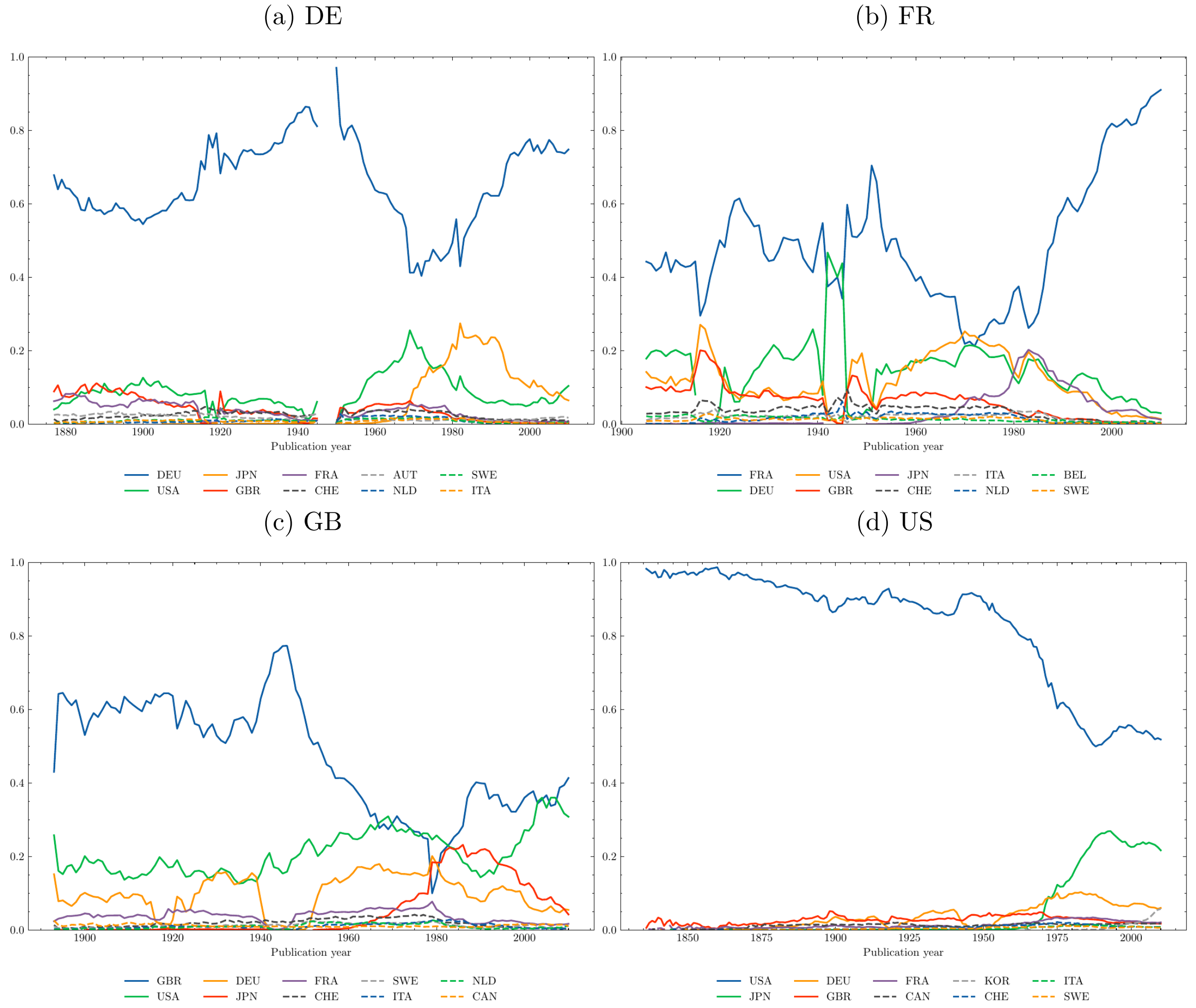
There are many dimensions through which patent data can help identify technology spillovers (see Eaton and Kortum, 1999; Guellec and de la Potterie, 2001; Keller, 2004 for a review). For example, Jaffe and Trajtenberg (1999) used the network of patent citations to measure international knowledge flows. A similar method is implemented by Aghion et al. (2021a) who show that when firms export goods, they also export new ideas as evidenced by the increasing flow of patent citations from the exporting destination.

In our data, we observe that patenting activities become increasingly international during a period of 30 years following WW2. During this period, the share of foreign inventors and assignees has increased in all countries. This dynamic is driven by large American, German, and Japanese multinational groups, but also by increasing exchanges between nearby countries.

2.4.1 Foreign patentees

To explore these technological exchanges between pairs of countries using patent data, we look at the country of residence of inventors and assignees and measure the relative importance of each foreign country in Germany, France, Great-Britain and the US. To do so, we plot the yearly number of patentees that report an address in a foreign country as a share of total patents. We restrict to the top 10 countries in each case and present the results in Figure 2.7. In this figure, we have considered all patentees regardless of whether they are defined as assignees or inventors, we report the corresponding Figures for inventors and assignees separately in Appendix 2.8, Figures 2.59 and 2.61 respectively.

Figure 2.7: Country of residence of assignees and inventors by patent offices


(a) DE

(b) FR

(c) GB

(d) US

Notes: This Figure reports the share of patentees by country of residence for each of the German, French, British and US patent offices and for the top 10 countries in terms of average share over the period of observation. Patentees are selected regardless on whether they are inventors or assignees

The shares plotted in Figures 2.7 can help identify international connections between pairs of countries. A large share of foreign patentees from a country A patenting in a country B could indicate either a technological similarity between A and B, or large knowledge exchanges between the two countries. The results highlight different facts.

First, there is a clear domestic bias: domestic patentees represent the majority of patent applications on average over the 20^{th} century and this domestic bias has experienced important changes in time. In the US, the share of domestic patentees slowly declined from about 100% to 90% until WW2. This decline then became more pronounced as more and more assignees from Germany and Japan take a growing importance (as well as Korea at the very end of the period[22]). France and the UK exhibit a similar trend: the share of domestic patentee is constant

[22] Note that China does not make the top 10 list in any patent office. This is because the number of Chinese patentees only start to rise modestly at the end of the time period, but remains at a lower level than Japan or Korea.

before WW2, declines during the 1950s-1960s and increases during the 1980s. In both case, the rise and fall of German and US patentees in the second half of the 20^{th} century explains most of the dynamics. Germany also experienced a decline in the share of domestic patentees after WW2, resulting from an increase in the number of Japanese and US patentees. Contrary to the other 3 patent offices, the share of domestic inventors and assignees increased from 60% to more than 80% during the first half of the century.

Second, the dynamics of domestic patenting changes in all countries after the 1980s. It stabilizes in the US and in the United Kingdom, at a lower level than at the beginning of the century, and increases in Germany and in France. This finding is consistent with the fact that the period following WW2 was characterized by an economic catch-up of western European countries to the world productivity leader. This catch-up dynamics was largely fueled by the adoption of new technologies from the US. After the end of the 1970s, France, Germany and the United Kingdom reached a similar productivity level than North America (see Bergeaud et al., 2016 for a quantitative illustration) and relied less on the adoption of these technologies.[23]

Third, the US, Japan, and Germany are the three countries that explain most of the dynamics in foreign patenting after WW2. In the United Kingdom, US patentees even overtook domestic patentees during the 1960s. This results from the spectacular increase of patenting from large manufacturing multinationals such as IBM, Siemens, Bayer, and General Electric during the 1960s-1970s, and Sony, Canon and Mitsubishi during the 1980s. French and British top assignees (Imperial Chemical Industries Limited, Thomson...) filed less patents in the US.

Fourth, the list of countries that patent the most in a given patent office seems to follow a type of gravity law. For example, Belgium, a French-speaking country and France's neighbor, is among the list of top foreign applicants for France. Similarly, Netherlands, Austria and Switzerland are all in the list for Germany.

2.4.2 Gravity

To check this last point more formally, we implemented simple gravity equations. Figure 2.9 shows that there is indeed a strong positive relationship between trade flows, as measured by the sum of exports and imports between two countries and the number of patents filed by one country in the other one's patent office. This relationship has motivated us to estimate a gravity equation replacing the flow of imports or export by the flow of patents. Formally, using geographical and economic data on as many countries as possible over the whole 20^{th} century, we estimate the following model:

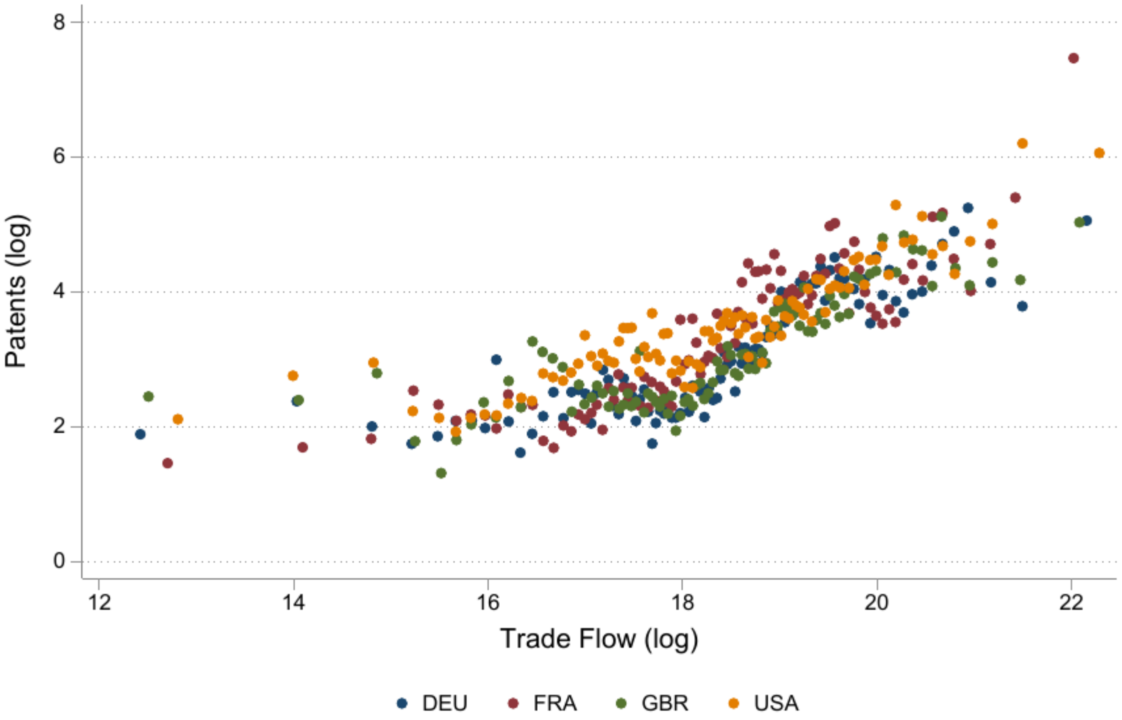
[23] The fact that the share of domestic inventors skyrocketed in France and to some extent in Germany and in the United Kingdom after the 1980s could be explained by the development of defensive patenting by domestic firms. Defensive patents are rarely filed abroad and mechanically decrease the share of foreign patentees. Interestingly, the increase is less pronounced for inventors than for assignees, see Figures 2.59 and 2.61.

$$N_{i,j,t} = \exp\left(\alpha d_{i,j} + \beta T_{i,j,t} + \gamma \mathbf{X}_{i,j} + \nu_{i,t} + \mu_{j,t} + \varepsilon_{i,j,t}\right). \tag{2.1}$$

In equation (2.1), i denotes one of the four patent office, j any other country (about 150) and t the year. $N_{i,j,t}$, the dependent variable is the number of patents filed in country i by a patentee residing in country j during year t, $d_{i,j}$ is the distance between the two countries (in km and taken in log), $T_{i,j,t}$ is the sum of export and import between the pair of countries (also taken in log). Finally, $\boldsymbol{X}_{i,j}$ is a vector of time invariant country-pair binary characteristics (same language, former colony) and $\nu_{i,t}$ and $\mu_{j,t}$ are sets of country-year fixed effects. These are meant to capture any changes within country that could explain variations in the number of patent applications (economic development, war, etc...). Results are presented in Table 2.6. We estimate the parameters of equation (2.1) first by taking the logarithm[24] in columns 1-3 and then using a Poisson regression in columns 4-6. In columns 1 and 4, we do not include ν and μ and instead control for GDP per capita. In columns 2 and 5, we add the logarithm of the total trade flow between the two countries. Finally, columns 3 and 6 estimate the full model. In all cases, the effect of the distance is negative as expected. Similarly, speaking the same language or being a former colony also increases the number of patent publications. We retain an estimate of -0.2 for α, this implies that when the distance double (say moving from 1000 to 2000 km) the average number of patents declines by about 13%. The negative coefficient on GDP per capita for country i, combined with the positive coefficient for GDP per capita in country j suggests that patenting by foreign inventors is more frequent when the two countries have similar levels of development.

[24] We accommodate for cases in which $N = 0$ by replacing $\log(N)$ by $\log(N + 1)$

Figure 2.9: Correlation between patenting and trade flows



Notes: This Figure plot the correlation between the logarithm of trade exchanges, defined as the sum of export and import between two countries i and j and the number of patents filed by a patentee residing in country j in country i's patent office. Countries i are either Germany, France, the UK or the US while countries j include 162 countries observed during the year for which we can measure their GDP. Bins average the values of the x and y variables for each percentiles of the x variable, separately for the four countries i. (i, t) and (j, t) fixed effects are additively included

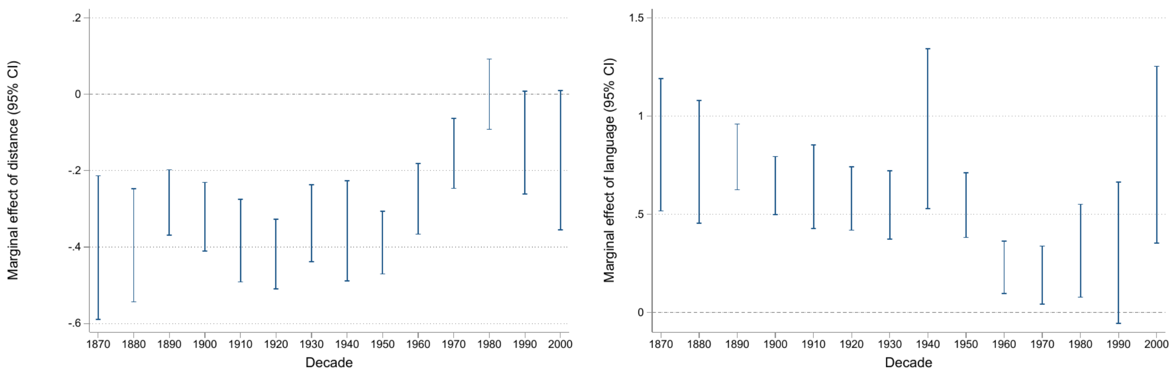
To consider the possibility that the role of distance changed over time, we then allow α to vary by decade and run our preferred specification (column 6 of Table 2.6). Results are presented graphically in Figure 2.11a. We see that the negative effect of distance has been gradually reduced following WW2 and is not distinguishable from 0 from the end of the 20^{th} century. This is consistent with the view that globalization has increased the incentives for assignees to patent in other countries to commercialize their products. Consistently, Figure 2.11b present an estimation in which the coefficient of the binary variable for speaking a common language is allowed to vary by decade. We see that the effect of sharing a language became less important after WW2 in explaining the flow of patents from one country to another.

Table 2.6: Gravity model

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
		OLS			Poisson	
Distance	-0.192**	-0.190**	-0.234**	-0.567***	-0.567***	-0.204***
	(0.090)	(0.089)	(0.092)	(0.069)	(0.072)	(0.046)
Same Language	0.312**	0.310**	0.415***	0.268**	0.268**	0.407***
	(0.125)	(0.125)	(0.138)	(0.109)	(0.110)	(0.072)
Former Colony	0.507***	0.504***	0.475***	0.263*	0.263*	0.197**
	(0.104)	(0.105)	(0.120)	(0.145)	(0.146)	(0.083)
GDP per capita country i (log)	-0.643***	-0.657***		-1.344**	-1.344**	
	(0.230)	(0.229)		(0.539)	(0.537)	
GDP per capita country j (log)	1.168***	1.178***		3.059***	3.058***	
	(0.180)	(0.181)		(0.587)	(0.577)	
Trade Flow (T)		0.005	0.023**		-0.000	0.068***
		(0.006)	(0.011)		(0.017)	(0.025)
R^2	0.795	0.795	0.935			
Observations	21,830	21,830	19,377	21,830	21,830	19,365

Notes: Estimation of parameters of equation (2.1). Distance is the number of kilometers between the centroid of a pair of country (i, j). "Same Language" and "Former colony" are two binary variables equal to 1 if the two countries share a common language (officially or *de facto*) and if country j was a former colony of country i respectively. Standard errors are clustered by pair of countries i, j. In columns 1, 2, 4 and 5, i, j and t fixed effects are added while in columns 3 and 6, (i, t) and (j, t) fixed effects are additively included . Countries i are either Germany, France, the UK or the US while countries j include 162 countries observed during the year for which we can measure their GDP. Columns 1-3 use an OLS estimation where the dependent variable is the logarithm of 1 plus the number of patents filed in country i by patentees (either assignees or inventors) located in country j. Columns 4-6 use a Poisson estimation.

Figure 2.10: Evolution of the marginal effect of distance and language


(a) Distance

(b) Language

Notes: Figure 2.11a plots the estimated value of α from equation (2.1) when α is allowed to vary by decade. The model is otherwise the same as in column 6 of Table 2.6. Figure 2.11b does the same for the coefficient on sharing a common language.

2.5 Who are the inventors?

In this section, we explore what our data can say about the inventors themselves. We exploit the information included in patent publications and document an important evolution in the typical profile of an inventor over time. In Great-Britain, we report that inventors who defined themselves as engineers increased from about 20% to more than 30% between 1895 and 1920. During the same period, the share of inventors presented as manual workers declined from more than 15% to less than 10%. A consistent pattern is found in Germany. We exploit the academic title traditionally specified just before the name of the inventor (e.g. *Dipl.-Ing. Marin Tomov*, see Appendix 2.8 for more details) and report that the share of inventors with a higher education degree increases over time.

We then combine the detailed geographical location of inventors in British patent to look at the correlation between the social and economic level of a neighborhood and the likelihood to become an inventor using the example of the late 19th century London.

2.5.1 Occupation of inventors

Patents filed in the UK patent office at the beginning of the 20th century frequently report the occupation of the inventor.[25] This represents a new source of information to document the professional activities of inventor and how this evolves over a 30 year window.

The denomination of occupation is free and as a result there is a very large number of distinct entities in the data. These can be highly precise, as for example, "Watchmaker and Jeweller", "Cemetery mason" or "Artificial limb manufacturer", or more vague like "Manufacturer" or "Engineer". The list of occupations covers a wide range of different skills. While the most frequently reported occupation is "Engineer" the list also include a large amount of low skilled occupations like "plumber", "worker" or "clerk" and more unexpected occupations like "Artist" or "professional mandolinist". At the same time, some inventors also declare to be "landowners" or "gentlemen".

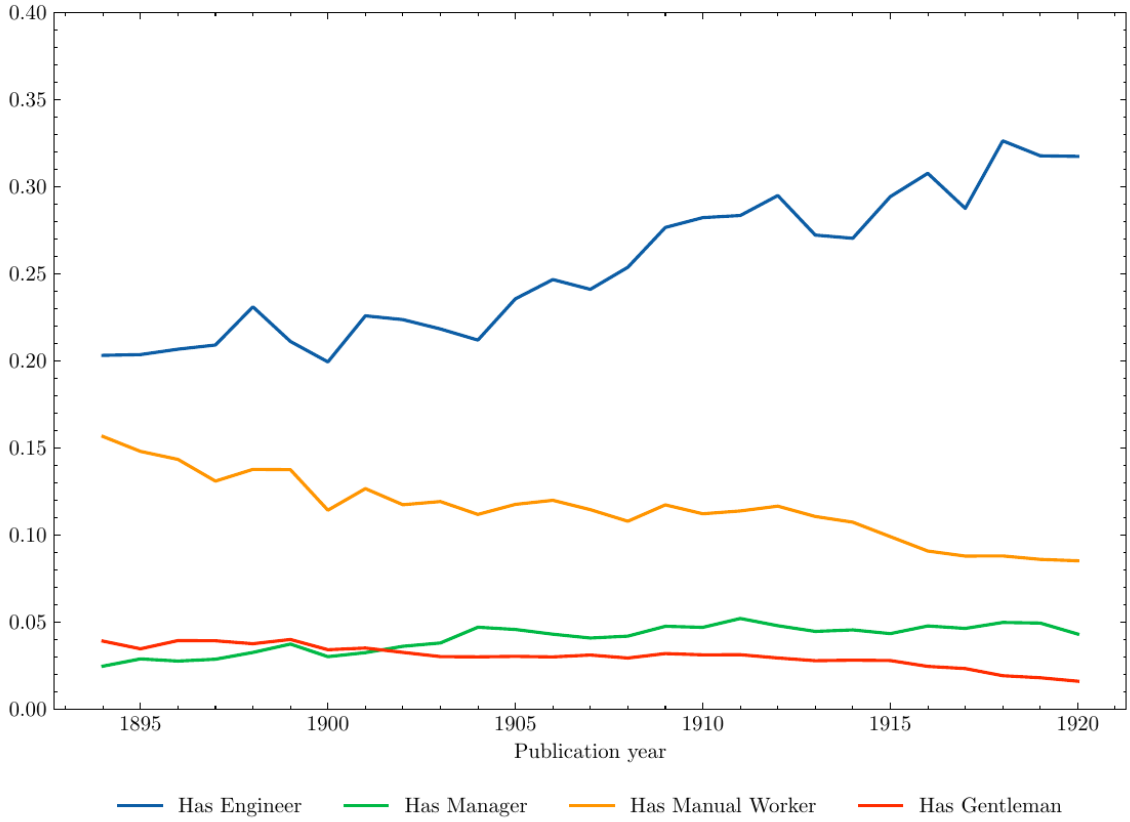
The most frequently reported occupation is engineer, or one of its derivative which is not surprising. Even as far back as during the very beginning of the industrial revolution, Squicciarini and Voigtländer (2015) have shown that human capital and upper-tail knowledge are important drivers of productivity in the most innovative sectors. Perhaps more surprising is the fact that workers in low-educated occupations can become inventor. However, this is in line with insight from historians of innovation. For example, Meisenzahl and Mokyr (2011) and Kelly et al. (2014) study the creation of new technologies in the end of the 18th century in England and emphasize the role of a skilled labor force with talented manufacturers and craftsmen who were able to adapt and adjust the equipment that they manipulate daily. Inventors in the present day are on the contrary very unlikely to work in low-educated occupations as studied by Jones (2009) and Bloom et al. (2020) who push the view that innovation becomes increasingly complicated and demanding in terms of knowledge and education.

Figure 2.12 reports the share of patents with at least one inventor declaring an occupation belonging to the following groups: engineer, manager, manual worker, and gentleman. We see that the share of patents involving an engineer increases over the period 1895-1920 from about

[25] The reporting of occupations in British patent is not systematic, but is fairly frequent over the period 1894-1920 with on average 50%-60% of inventors declaring one occupation. See Figure 2.30 in Appendix 2.8.

20% to more than 30%, while at the same time, less patents involve at least one manual worker. At the same time, although at a much lower level, the share of patents having at least an inventor reporting "gentleman" as an occupation decreases from 4% to 2% the share of patents with a manager increases from 2% to 5%.

Figure 2.12: Occupation of inventors in the United Kingdom



Notes: This Figure reports the share of patents involving at least one engineer (Has engineer), one manager (Has manager), one manual worker (Has manual worker) or one gentleman (Has gentleman) in terms of the occupation of the inventor reported in the text. Time period: 1894-1920.

The case of Germany

German patents (both East or West Germany) also offer a way to inform about the education of inventors as the names of the patentees are preceded by an academic title, when applicable. This includes the prefix "Dr.", but goes far beyond, with many different possibilities like "Dipl-Ing.", "Phy. Dr.", "Ing.", . . . We consider the presence of these elements as indications that the inventor has done some higher education. Figure 2.13 reports the share of patents where at least one inventor reports an academic title: Doctor (Has Dr), Engineer (Has Ing), Diploma (Has Dipl) and any the previous title (Has Higher Education). The time periods are restricted to 1955-1980 for West Germany and 1965-1980 in the case of East Germany due to limited reporting of inventors before those periods.

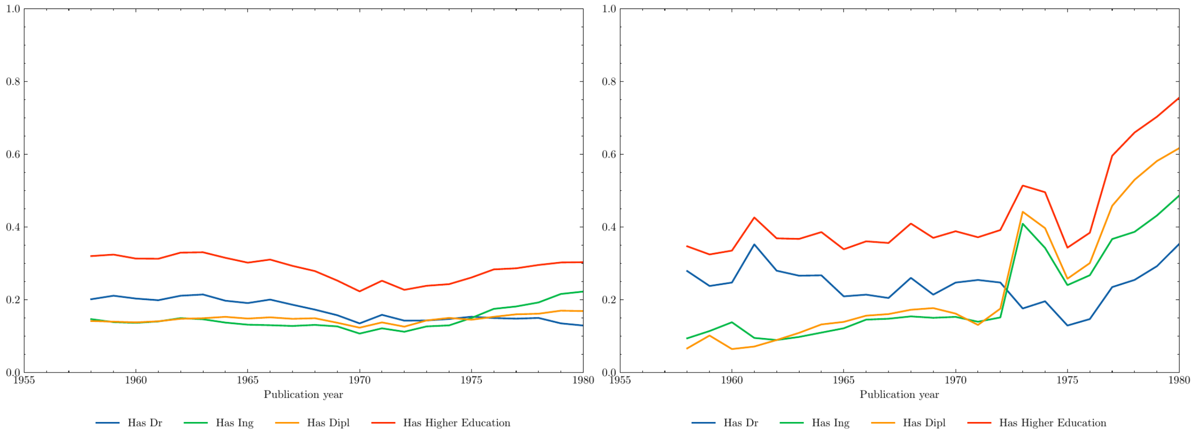
In both cases, Figure 2.13 shows that the share of patents involving an inventor reporting a title that indicates some higher education increases after the 1970s from around 25% to 35%

in West Germany and from around 40% to 70% in East Germany. In addition, this increasing share seems to be driven by inventors who report to be engineers or to have a diploma, rather than doctors or professors whose relative importance has declined in time.

Figure 2.13: Share of inventors with an academic title in Germany

(a) West Germany

(b) East Germany



Notes: This Figure reports the share of patents with at least one inventor declaring an academic title: Doctor (Has Dr), Ingenior (Has Ing), Diploma (Has Dipl). We also define "Has Higher Degree" as the union of the previous variables. Time period: 1958-1980 for West Germany and 1965-1980 for East Germany.

These results are, of course, only indicative, but they seem to testify to an evolution in the way the creation of innovation has gradually evolved. Both British and German patents report an increasing share of engineers, while at the same time, the share of invention with only one inventor has declined in the UK from about 50% in the very beginning of the 20^{th} century to 25% just before WW2.[26] In the US and in Great-Britain, the average number of inventors per patent has increased, see Figure 2.63 in Appendix 2.8. Similarly, in all countries, the share of patent with only one inventor have overall decreased over time (Figure 2.65). These results suggest that the "architecture of innovation" has evolved from the model of a single person developing her own idea to a more complex structure of R&D teams organized by large corporations or laboratories (Lerner, 2012).

2.5.2 Innovation and poverty

The economic literature has recently focused on the socio-economic characteristics of inventors and more generally scientists, with the underlying idea that if intrinsic talent is randomly distributed in the population, then any group that is underrepresented among inventors (women, blacks, families from disadvantaged backgrounds...) is evidence of a policy failure that results in misallocation of resources (Hsieh et al., 2019). For example, Bell et al. (2019) study the lives of

[26] This statistics cannot be calculated for France and Germany as information on inventors are limited for these two countries and it is difficult to distinguish the case of a single inventor innovation from the case of an invention with one assignee and no inventor.

about one million inventors in the US and link them to their tax and school records to track them from their birth. They find that exposure to innovation when young is an important predictor of the probability to become an inventor and highlight a phenomenon of "lost Einsteins" or lost "Marie Curies". Using information on Finnish patents, Aghion et al. (2017) also look at the socio-economic profile of inventors.

Looking precisely at the correlation between revenue and the likelihood of filing one patent is typically done by matching the patent data to different vintage of the census (see Akcigit et al., 2017c; Sarada et al., 2019). Unfortunately, to our knowledge, this matching approach is impossible to replicate for European countries due to the lack of available historical census data. Matching names and location to administrative data is in any case challenging because of the high frequency of typos and imperfectly recognized entities, in particular for the oldest documents.

British patents offer an alternative approach. Because the location of patentees is given with a very high degree of precision, we can map inventors to the different neighborhoods of a given city which are more or less privileged. We do this exercise for London thanks to the Maps Descriptive of London Poverty (1898-9), also known as the "Booth Poverty Maps". This set of twelve maps represent the late 19^{th} century London social cartography based on a large scale survey organized by Charles Booth, a British ship owner and social researcher. Each street (or even building in some areas) was assigned one of seven socio-economic categories ranging from "Lowest class, vicious, semi-criminal" to "Wealthy. Upper-middle and Upper classes". The category of each street resulted from Booth and fellow researchers' visits to households and the differences in lifestyle and qualitative factors (food, clothing, shelter, and relative deprivation) they could observe.

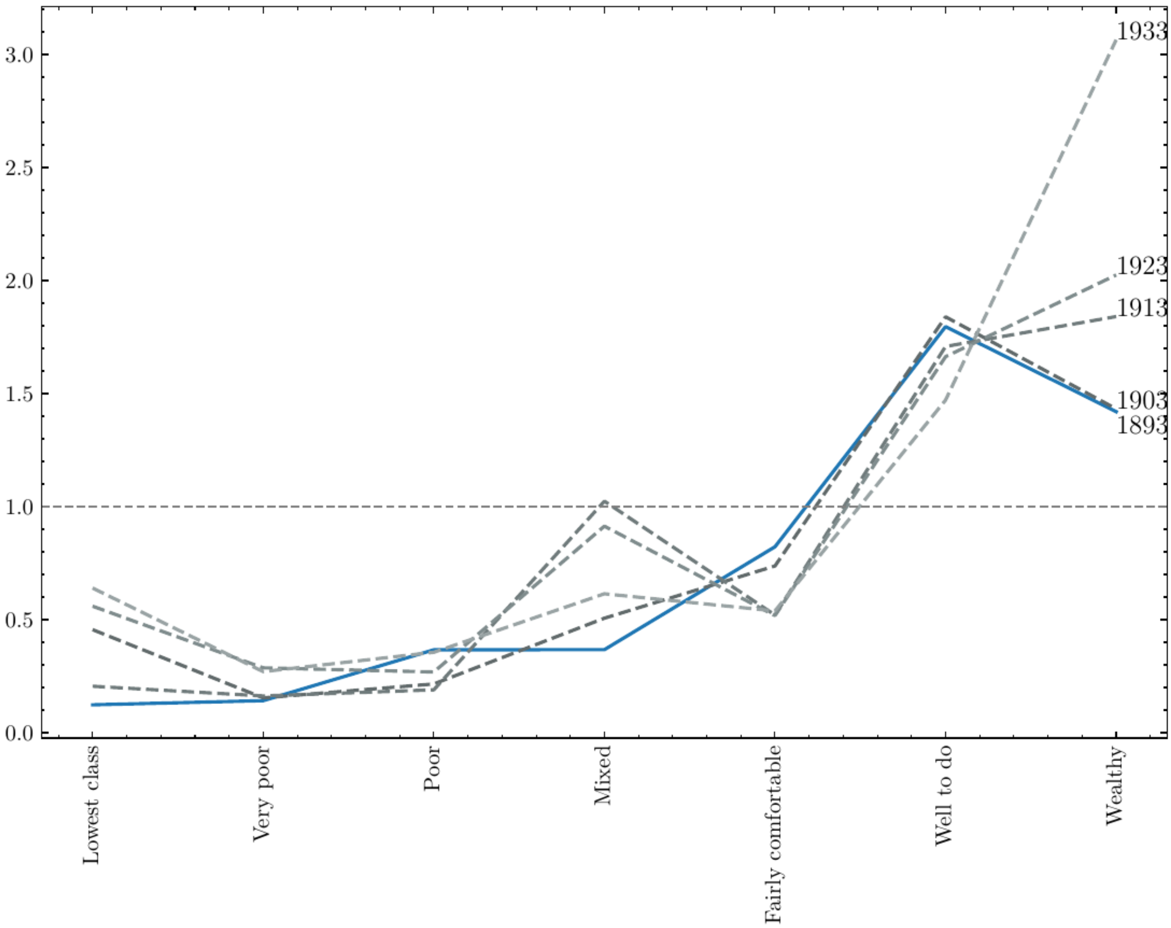
The Booth Poverty Maps have been digitized by Orford et al. (2002). The authors represent each socio-economic class area by a set of coordinate points. We use this digitized map and project the coordinates of the geocoded inventors' location and use the k-nearest neighbors (with k=3) algorithm to find the socio-economic class the inventor is most likely to belong to (based on his address). We apply this method to the 3,107 inventors extracted from British patents published between 1893 and 1903 (that is, at most 5 years away from the Booth survey) and geolocalized within the Booth Poverty Map area with a high degree of precision (either "street" or "house-number"). We then derive a measure of the socio-economic category of each inventor from the address declared on the patents. Cross referencing these results with the share of the population belonging to each of the categories at the time of the survey, we compute the odds of being an inventor by socio-economic category.

Results are presented in Figure 2.15 (blue solid line). In line with findings on historical data in the US (Akcigit et al., 2017c) or recent data in Finland (Aghion et al., 2017) who report an important correlation between parental income and the probability to become an inventor, we also clearly find that in late 19^{th} century - early 20^{th} London, innovators were under-represented in disadvantaged neighborhoods. However, one striking additional finding is that the odds ratio is larger among the category “Well to do”, the second richest than among the wealthiest category “Wealthy”. We conjecture that this result is a feature of the British society at the end of the Victorian era, with a large part of the wealthiest socio-economic group being composed of idle land-owners.

To check this hypothesis, we replicate the exercise for the next decades in addition to the baseline 1893-1902: 1903-1912, 1913-1922, 1923-1932 and 1933-1942. For all these periods, we make the assumption that the geographic distribution of revenue remains the same as in the early 20^{th} century.[27] The results are presented in Figure 2.15 (gray dashed lines) and show that indeed, the wealthy category counts more and more inventors as we move forward in time. In the decade 1933-1942, the relationship between wealth and the probability to become an inventor is clearly convex, in line with results presented by Akcigit et al. (2017c); Aghion et al. (2017); Bell et al. (2019) in other contexts.

[27] We stop in 1942 to avoid the risk of measurement error when attributing each neighborhood to one of the category of the Booth Map based on the end of the 19^{th} century. The persistence of the socio-economic characteristics of a London neighborhood is likely to be affected by the reconstruction following the destruction of WW2.

Figure 2.15: Booth Poverty Map and innovators



Notes: This Figure reports the odds ratio of innovator, defined as the share of inventor in each category divided by the share of the population in the same category. Inventor localized in London in British patents from 1893 to 1902 (blue line). The same exercise is then replicates for each following decades: 1903-1912, 1913-1922, 1923-1932 and 1933-1942.

2.6 Migration and innovation

The question on whether immigration has a positive impact on local innovation has received a lot of attention in the recent literature (see e.g. Akcigit et al., 2017b; Arkolakis et al., 2020). Most of existing studies use external data to identify immigrants, for example different vintages of the US Census or registers of inventors. We take a complementary approach using the information on citizenship included in the text of the patent publications in the US and in the United Kingdom. We focus on two distinct subperiods (respectively 1920-1950 for the United Kingdom and 1880-1925 for the US) for which we have direct information on the citizenship and the location of some inventors[28] which allows us to classify them as "immigrant". Of course

[28] Not all patentees declare a citizenship even during these subperiods. Among the set of patentee that are located in the United Kingdom, 87% report a citizenship for patents filed between 1920 and 1950. During the period 1950-1980, around 20% of inventors filing a British patent did declare their citizenship. Nevertheless, we decided not to consider this time span due to the small fraction of inventors concerned. For the US, this share is around 37% between 1880 and 1925 but is closer to 45% after 1910 (see Appendix Figure 2.32).

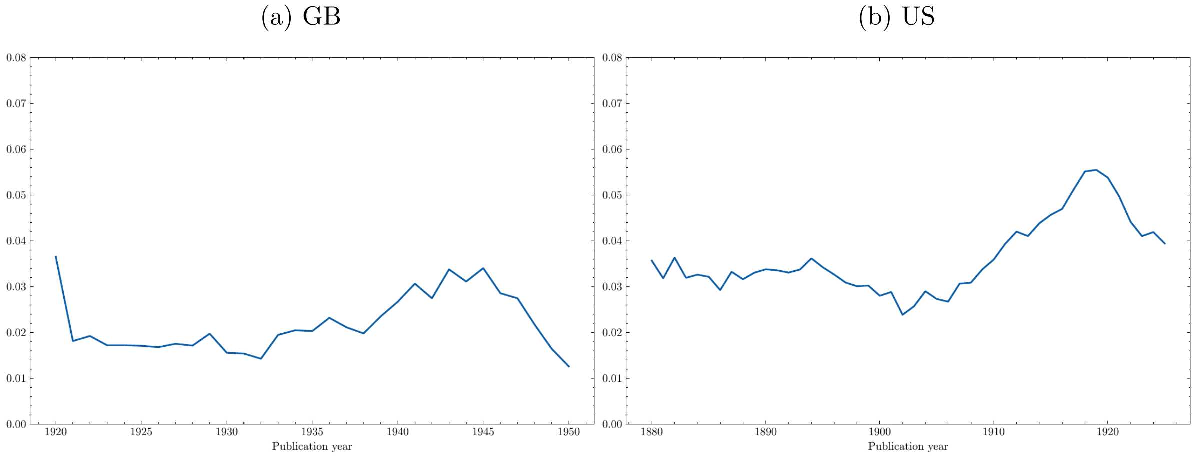
this definition is only an indirect evidence that the inventor is indeed an immigrant, it could well be that the inventor is just temporarily visiting a foreign country. However, one advantage of this method is that it does not require to implement a complex matching to external data, which is typically based on the name and location of inventors.

With these differences in mind, we looked more precisely at the evolution of migrant inventors from different nationalities over the time period considered in the United Kingdom and in the US. The composition of these nationalities obviously changes over time. In particular, a large number of German citizens living in the United Kingdom patented during the 1930s, most likely as a consequence of them fleeing the Nazi regime. We then show that these immigrant inventors make different types of innovation than other native inventors. Not only do they specialize in different technological field, but they also patent on average more novel innovation, although the differences with other inventors tend to fade over time.

2.6.1 Share of immigrants in US and British patents

We find that between 4% and 5% of inventors who report an address in the US but are *not* American.[29] In the United Kingdom, this share is lower, between 1% and 2%, at any point in time between 1920 and 1950. In Figure 2.18, we report this share every year for the two countries. We can see that the US experienced a sizeable increase in the share of immigrants during the 1910s. The United Kingdom experienced a similar upswing during the 1940 decade.

Figure 2.16: Share of immigrant inventors over time


(a) GB

(b) US

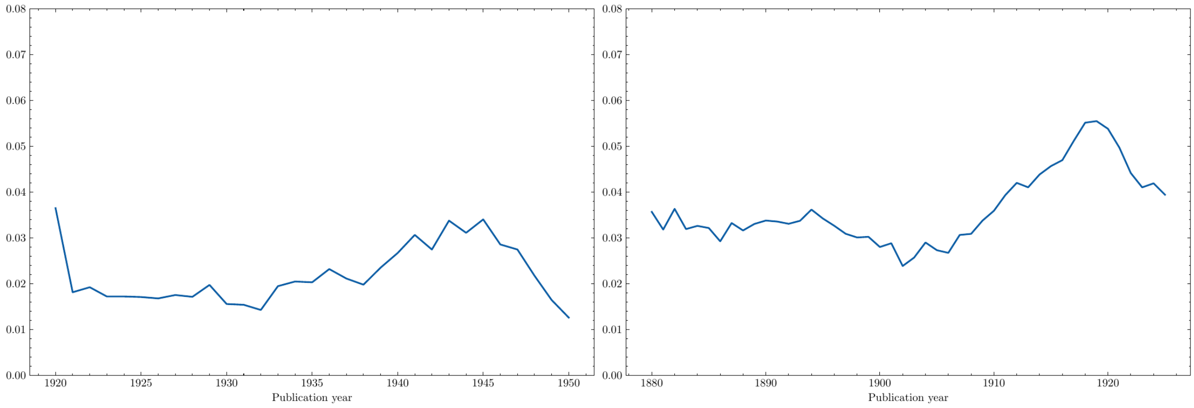
Notes: The share of immigrant is computed as the ratio of the number of inventors who report a non-domestic citizenship different over the number of inventors reporting a domestic address. Time periods: 1920-1950 (GBR) and 1880-1925 (USA).

Figure 2.59 reports the evolution of the composition of these immigrants by country of citizen-

[29] These numbers are lower than those reported by Akcigit et al. (2017b) and Arkolakis et al. (2020). This can happen for two reasons. First, it could be that immigrant inventors under-report their citizenship compared to US born inventors. Second, both Akcigit et al. (2017b) and Arkolakis et al. (2020) define an immigrant based on the country of birth, while we consider citizenship at the time of patent publication. Part of the difference might then come from inventors who acquired US citizenship but were foreign born, hence counted as non immigrants in this paper but counted as immigrants in the two aforementioned papers.

ship for the 10 most frequent nationalities respectively in the United Kingdom and the US. As expected, Europeans constituted the bulk of immigrant inventors (consistently between 70% and 90%) in the US. The share of British and German inventors alone represented close to 60% of immigrant inventors in the late 19^{th} century and gradually decreased to reach 40% in the 1920s. In the United Kingdom the 1930s were marked by the massive migration of German inventors (most likely pushed out by the Nazis) who represented up to 40% of immigrant inventors in 1940 while they were almost absent before 1930. Following the *Anschluss* and the subsequent Poland invasion, the share of Austrian and Polish inventors rose up to close to 10%. Before this decade, American and Swiss immigrants represented up to around 40% of immigrant inventors.

Figure 2.18: Share of immigrant inventors over time


(a) GB

(b) US

Notes: The share of immigrant is computed as the ratio of the number of inventors who report a non-domestic citizenship different over the number of inventors reporting a domestic address. Time periods: 1920-1950 (GBR) and 1880-1925 (USA).

2.6.2 Immigration and novelty

We now address the question of whether inventions from immigrant inventors differ from those of domestic inventors. In particular, we want to know if immigrants tend to import more novel ideas than what is produced by their domestic counterparts but also whether this difference in the degree of novelty varies over time and countries. Recent studies on the role of immigrants tend to show that migrants serve as a vehicle of new knowledge that is influenced by educational and cultural differences (see e.g. Miguélez and Morrison, 2021). This was also the case in the past.

To show this, for both countries, the US and the United Kingdom, we select the 5 most represented foreign citizenship and compute the share of patents that they file in each of the 1-letter CPC code (A, B, C, D, E, F, G and H). We then calculate the ratio of this share by the similar share for non-immigrant domestic inventors. This measure summarizes the specialization of a given group of immigrants into a given technological field. A ratio larger than 1 means that the corresponding technological class is overrepresented in the set of patents filed by the group of immigrants. Results are presented in Table 2.7. They show that some groups of immigrants are differently specialized than native inventors with some heterogeneity by citizenship. For example, German inventors tend to patent more in chemistry while Polish inventors patent more in Human Necessities (food, health, agriculture...). These results are mostly illustrative but show that immigrant inventors are not similar to natives in the technology they typically specialized in. They also indicate that migrant inventors tend to import their knowledge in some specific field. Could this mean that these inventions are more novel?

Table 2.7: Specialization by CPC

		United Kingdom					United States				
CPC code		USA	DEU	CHE	POL	FRA	GBR	DEU	AUT	RUS	SWE
A	Human Necessities	0.9	0.9	0.5	2.1	1.3	0.7	0.8	1.1	1.4	0.9
B	Performing Operations	1.2	0.9	0.4	0.6	1.1	0.8	0.8	0.9	0.8	0.9
C	Chemistry	1.3	1.5	3.1	0.8	0.9	1.9	2.6	1.3	0.6	1.1
D	Textiles	0.3	0.6	4.1	0.6	0.9	1.5	1.2	0.7	1.1	0.9
E	Fixed Constructions	0.6	0.4	0.4	0.9	0.6	0.6	0.6	1.0	1.0	0.8
F	Mechanical Engineering	0.7	0.8	1.0	0.8	0.8	1.0	0.9	0.9	0.8	1.1
G	Physics	1.0	1.1	0.4	0.9	1.0	1.1	1.1	0.9	1.1	1.0
H	Electricity	1.0	1.5	0.9	0.8	0.5	1.7	1.5	1.0	0.9	1.2

Notes: This table presents the ratio of the share of patents filed in each of the 8 CPC code by inventors of different citizenships, as described in the patent publication filed in the UK patent office and the US patent office, by the similar share calculated by native inventors. The ratio is calculated over the set of domestic inventors that declare a citizenship.

Measuring novelty

Measuring the novelty of an invention is non-trivial. Empirical economics defines novelty as the dissimilarity (or inverse of similarity) of a given invention with the existing corpus of inventions at the moment of its publication. In practice, recent research has favored text similarity measures to assess the degree of novelty of a patent. Immediately related to our work, Arkolakis et al. (2020) use the "Term Frequency Backward Inverse Document Frequency" (TFBIDF) introduced by Kelly et al. (2018) to study the novelty of immigrants' patents. This measure captures the extent to which the distribution of words used by a given patent differ from the words used by the corpus of patents published before - typically in a 5-year single-sided window. Our approach is very similar, except that we do not use the TFBIDF. There are two main reasons for that. First, TFBIDF relies on very large vectors. Each patent is represented by a vector of the size of the word dictionary of the entire corpus of patents, which makes the computation very demanding. Second and more fundamentally, in the context of TFBIDF, two patents are considered similar, if and only if they use the exact same words - leaving no room for synonymy or sentences close in meaning but using different wording. Since language is known to be highly variational[30] this is a major pitfall of this kind of textual similarity measures. To circumvent this, we use the Google Patents embedding which is made publicly available by the Google

[30] e.g., "The President is at the Office" and "Joe Biden is at the White House" are two largely distinct ways to say the exact same thing but these two sentences would have very low similarity scores using TFBIDF. The intuition is that they have only very frequent words ("is", "at", "the") in common and that common words receive low weight in this similarity measure. Common words are even not considered in the case of the so-called stop-words (here "at" and "the").

Patents research team. This embedding builds on recent advances in NLP, specifically on the WSABIE algorithm described in Weston et al. (2011). It addresses the two main limitations of TFBIDF. First, each patent is represented by a dense vector of size 128 which makes them handy to manipulate. Second, the text embedding is "learnt" with the objective of maximizing the similarity of two patents having the same technological classes.

In this setting, the similarity between two patents is given by their dot products, which returns a value between 0 and 1 – the larger the value the more similar the patents. Using this methodology, we can measure the novelty of any patent as the average dissimilarity (1 - similarity) of this patent with respect to the corpus of patents published up to 5 years before its publication. Furthermore, we consider three groups of patents: a group of random patents which can be authored by either immigrants or non-immigrants – this is the group of patents that is used to define the "average corpus"; a group of patents authored by immigrants only and a group of patents authored by non-immigrants only. Once we have computed the novelty of all patents in the last two groups, we use a simple OLS regression to determine the effect of being an immigrant on patent novelty. Similarly to Arkolakis et al. (2020), we also introduce time, space and industry fixed effects.[31] We find that being an immigrant has indeed a positive and highly significant effect on patent novelty in both the US and Great Britain (see Table 2.8).

Table 2.8: Patent novelty and migration

	United Kingdom	United States
Dummy Immigrant	0.0075***	0.0058***
R^2	0.07	0.28
N	10,065	84,649

Notes: OLS estimation results of a linear model in which the dependent variable is the novelty of a patent and the regressor is a binary variable set to 1 if the inventor reports a citizenship that is different from its location. Column 1 uses British patents and column 2 uses US patents. Time span is respectively 1920-1950 and 1880-1925. Regressions include additive decade, county (Great Britain) or state (US) and technological class (CPC) fixed effects. Both coefficients are significantly different from 0 at the 1% threshold.

The nature of immigration has changed in time, is this also the case for this "novelty premium"? To check this, we apply the same approach but we run distinct regressions for each country and decade. In both countries, the results are less unequivocal. We report the estimated coefficients and the associated confidence interval at the 5% level in Figure 2.20. In Great Britain, the effect of being an immigrant is no longer significant at the 5% level in the 1940s. In the US, the additional novelty brought by immigrant is decreasing and its effect is indistinguishable from 0 during the 1920s. We also observe variations in the magnitude of the estimated effect

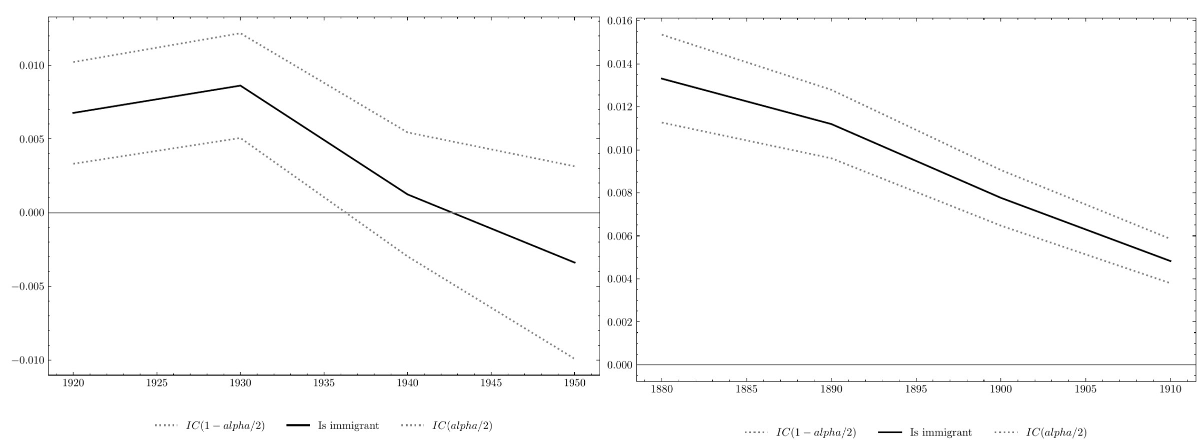
[31] We use respectively the publication decade, the county (Great Britain) and state (US) and the 1 letter-CPC class.

of immigrant on the measured novelty of inventions. As suggested by Arkolakis et al. (2020), part of the explanation might come from structural changes in the average population of inventors in terms of time spent since immigration, pre-immigration occupation or origins. Other explanations could include institutions such as migration policy or the strength of pre-existing international interactions between inventors[32]. Solving this question is undoubtedly a promising research avenue on the way toward a deeper understanding of the relation between immigration and innovation. We believe that exploring the difference between the US and other countries could bring important insight in this direction.

Figure 2.20: Regression result by decade

(a) GBR

(b) USA



Notes: Coefficients and 95% confidence intervals from an estimation of the same model as in Table 2.8 for each decade with the OLS.

2.7 Conclusion

In this paper, we have presented a novel dataset constructed from an automated text analysis of patent documents published in the German (including East German), French, British and US patent offices. The data cover as many years as possible and include most of the 20^{th} century, and part of the 19^{th} century. The information extracted from these publications offer a novel opportunity to acquire a better understanding of the long-term determinants of innovation. Our findings confirm the intuition that the long-run development of innovation activities, as revealed by patents, in Europe and in the US are different, and vindicate the view to extend historical datasets to embrace the actual heterogeneity characterizing the long-term dynamics of innovation.

Our work could be prolonged in different directions. One natural improvement would be to include more countries in the dataset. Patents have existed since the end of the 19^{th} century in

[32] Note that the aforementioned results cannot be solely attributed to the fact that foreign-born inventors' writing-style is different from their US-born counterparts since embeddings are based on both text *and* technological classes which are set externally.

many places that are important R&D actors: Japan, Sweden, Switzerland... The methodology presented in this paper has been designed with the goal of reducing the required efforts to apply it to new patent corpuses. We also hope that making the codebase open source will support a collective data design and continuous improvement momentum. Combining our database and future extensions with historical datasets also has the potential to foster our understanding of the role of institutions on innovation. We see two particularly promising research directions: the role of universities for local innovative activities in Europe vs in the US and the role of the dense urban network connecting Western Europe largest cities in innovation diffusion.

2.8 Appendix

Selection of utility patents

Table 2.9: Granted utility patents

Patent office	Time span (publication year)	Kind code(s)
DD	1950-1992	A, A1, A3, B
DE	1877-2013	A1, B, B3, C, C1, D1
FR	1902-2013	A, A1, A5*, B1*
GB	1893-2013	A, B*
US	1836-2013	A, B1*, B2*

Notes: The selected kind codes try to emulate the USPTO concept of “Granted Utility Patent”. We restrict to the first publication or second publication without first publication kind codes in order to avoid double counting issues. We exclude patent *applications* and *revised* publications for the same reason. In the case of DD, we are limited by the availability of raw patent images and therefore include all types of publications. * indicates that the kind-code is considered only after 1980. This can be due to changes in the meaning of the kind-code or to its creation date.

Formats

Entities by country

In this Section, we detail the different types of entities matched for each country and what they usually means.

United States In the case of the US, the inventors and assignees are clearly separated entities. The inventor is the name of the person who conceived the invention while the assignee is the entity (either a person, a firm, the government, a university...) who own the right of the patent. US patents also give information on the citizenship of patentees. In the case of inventors, this is the country of citizenship (e.g., “a citizen of the kingdom of Italy”) and in the case of assignee the legal origin of the firm when applicable (“a company duly organized under the laws of New Jersey”). Finally, the entity location gives the address of the inventor and assignee, usually at the city level. For more details, see the Annotation guidelines for the US

Table 2.10: Publication number and patent format

Patent office	Publication number (range)	Format number
DD	DD1 - DD123499	1
DD	DD123500 -	2
DE	DE1C - DE977922C	1
DE	DE1000001B -	2
FR	FR317502A - FR1569050A	1
FR	FR1605567A -	2
GB	GB189317126A - GB2000001A	1
GB	GB2000001A -	2
US	US1A - US1583766A	1
US	US1583767A - US1920166A	2
US	US1920167A - US3554066A	3
US	US3554067A -	4

Notes: The structure of a patent document can change over time. We track these changes and adapt the statistical model to each cases. The table shows the different formats for each patent offices and the associated first and last patents of each format.

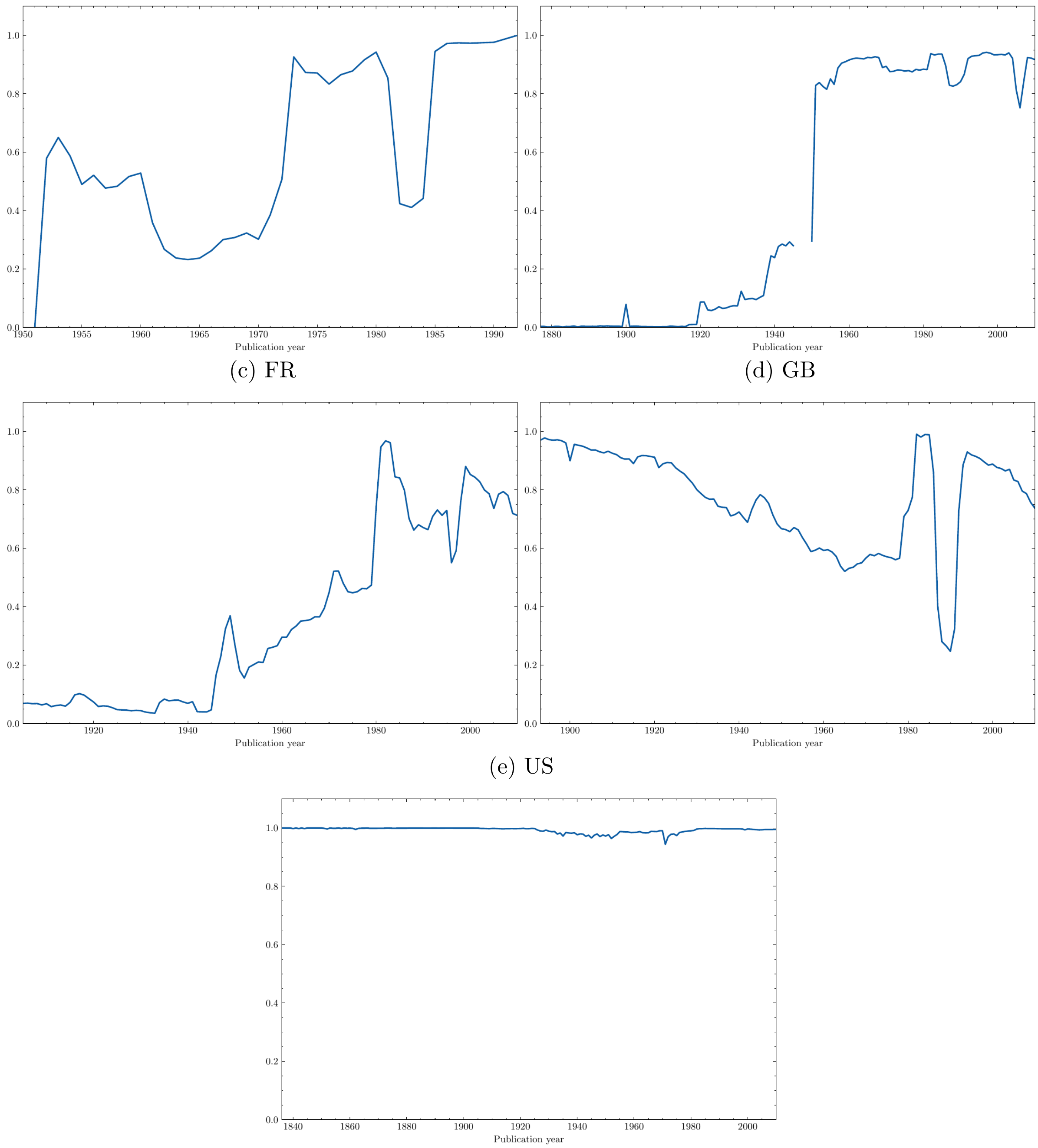
Germany In the case of Germany, inventors are referred to as *"Erfinder"* and assignees as *"Anmelder"*. Both entities can represent physical people while assignees can also be companies. Most of the patents filed before the 1950s do not include any inventor. Although it is likely that in that case, the inventor and the assignee can be the same person, we only label the entity as inventor when the term *"Erfinder"* is explicitly mentioned. German patents also give some information on the occupation of inventors or assignees from the denomination of their academic title (e.g., *"Dr."*, *"Ing."* or *"Pr."*). Finally, the location is usually given by the city of the inventor or assignee. For more details, see the Annotation guidelines for Germany and the specific guidelines for East-Germany

France The case of France is similar to the case of Germany regarding inventors and assignees. Most of the patents have a *"déposant"* which we label as assignee while some patents also have an *"inventeur"* which we label as inventor. French patents do not give information on occupation or citizenship, except if extremely rare instances. The location is given at the county (*"département"* level in the case of a patentee located in France and at the country level for foreign inventors. For more details, see the Annotation guidelines for France

United Kingdom In the British case, the inventor and the assignees are not explicitly distinguishable. By convention, we denote each firm by an assignee and each person as an inventor. The British patents also include information on the occupation of the inventor, and in some case on the occupation of the assignee (e.g., "a clock manufacturing company"). Information on the citizenship of inventor and assignee are also provided like in the US. Finally, the location of the assignee and of the inventor is given as a full postal address. For more details, see the Annotation guidelines for British patents.

Data coverage

Figure 2.22: Share of patents with at least one inventor



(a) DD

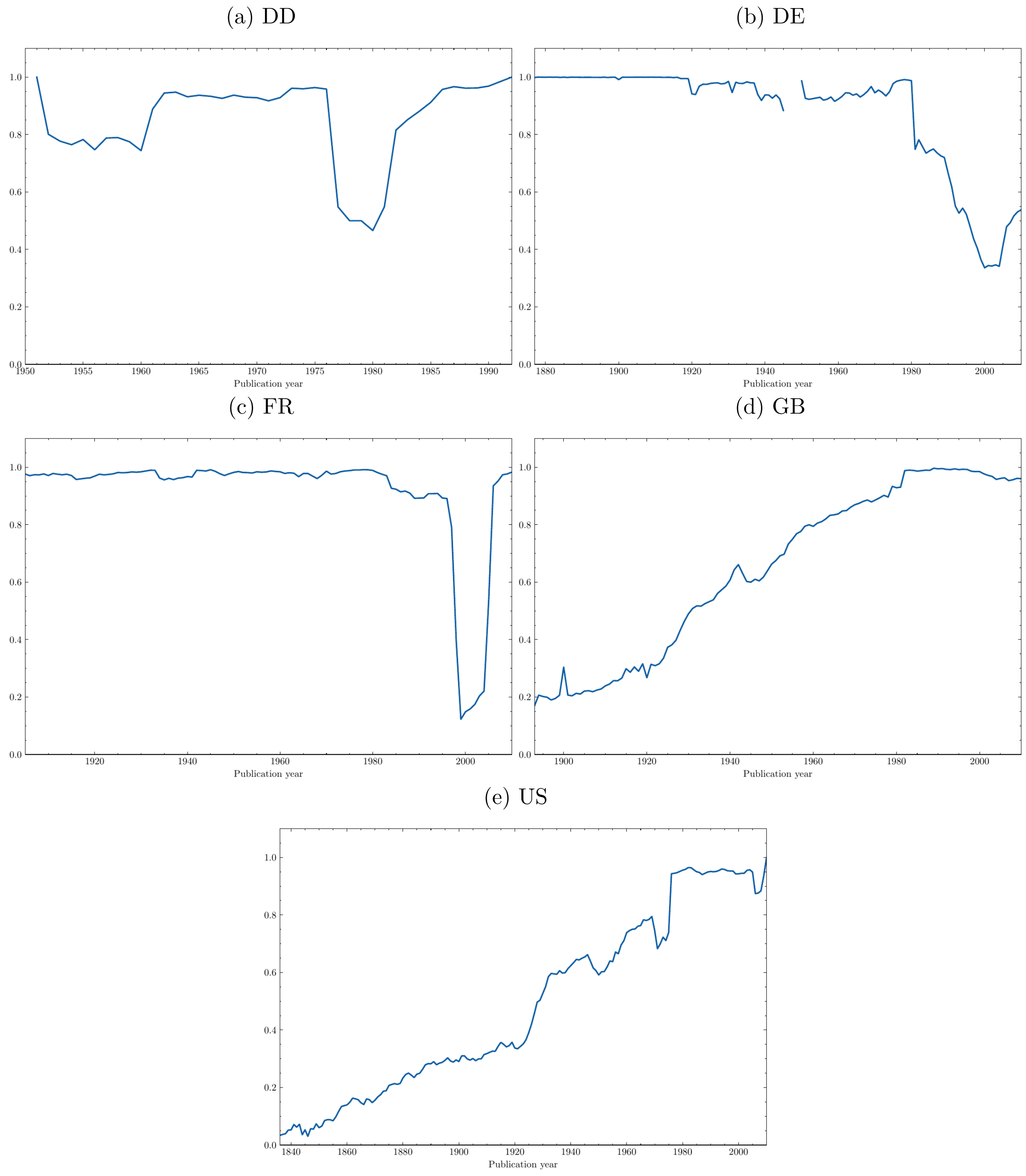
(b) DE

(c) FR

(d) GB

(e) US

Figure 2.24: Share of patents with at least one assignee



(a) DD

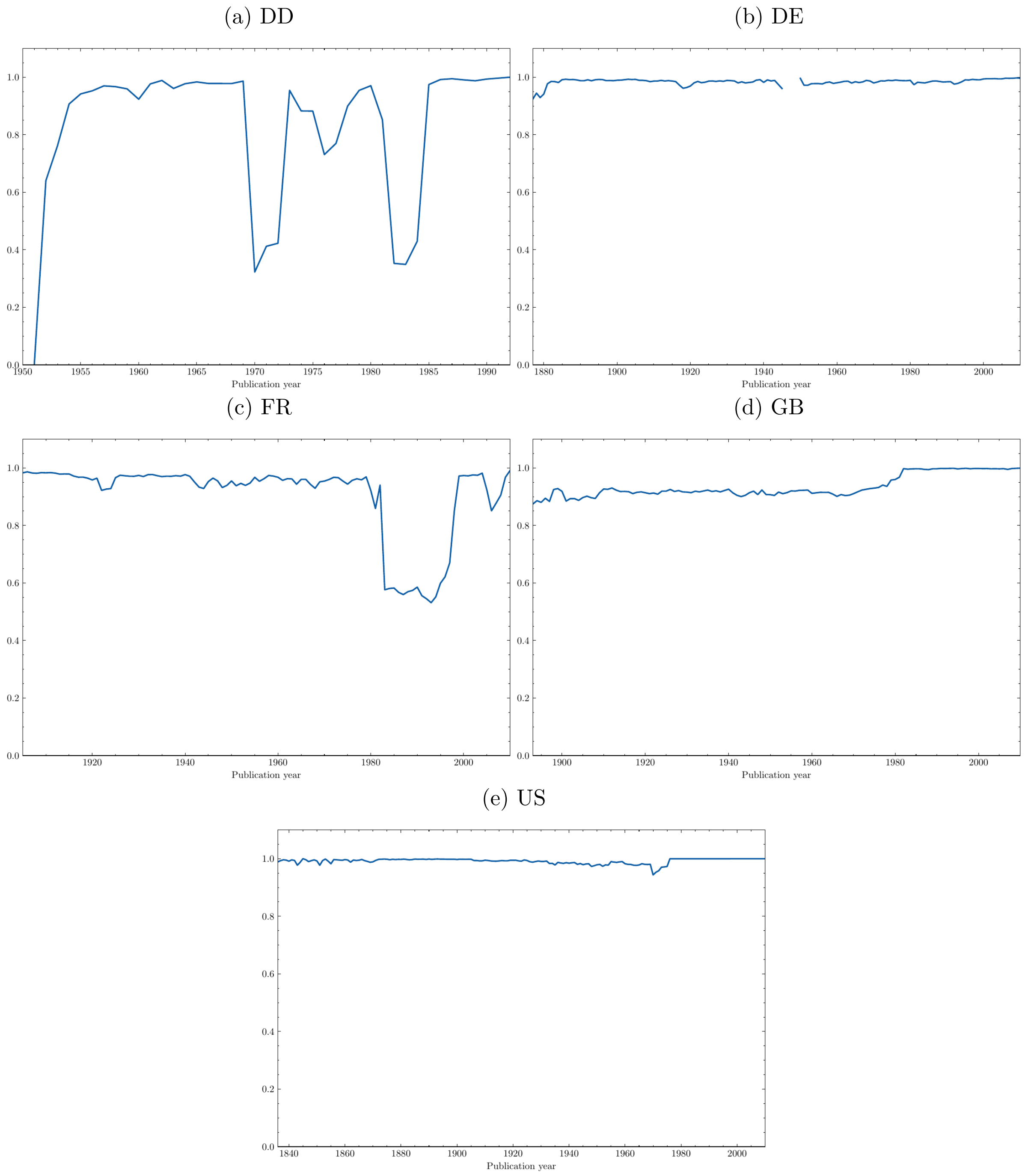
(b) DE

(c) FR

(d) GB

(e) US

Figure 2.26: Share of patents with at least one location



(a) DD

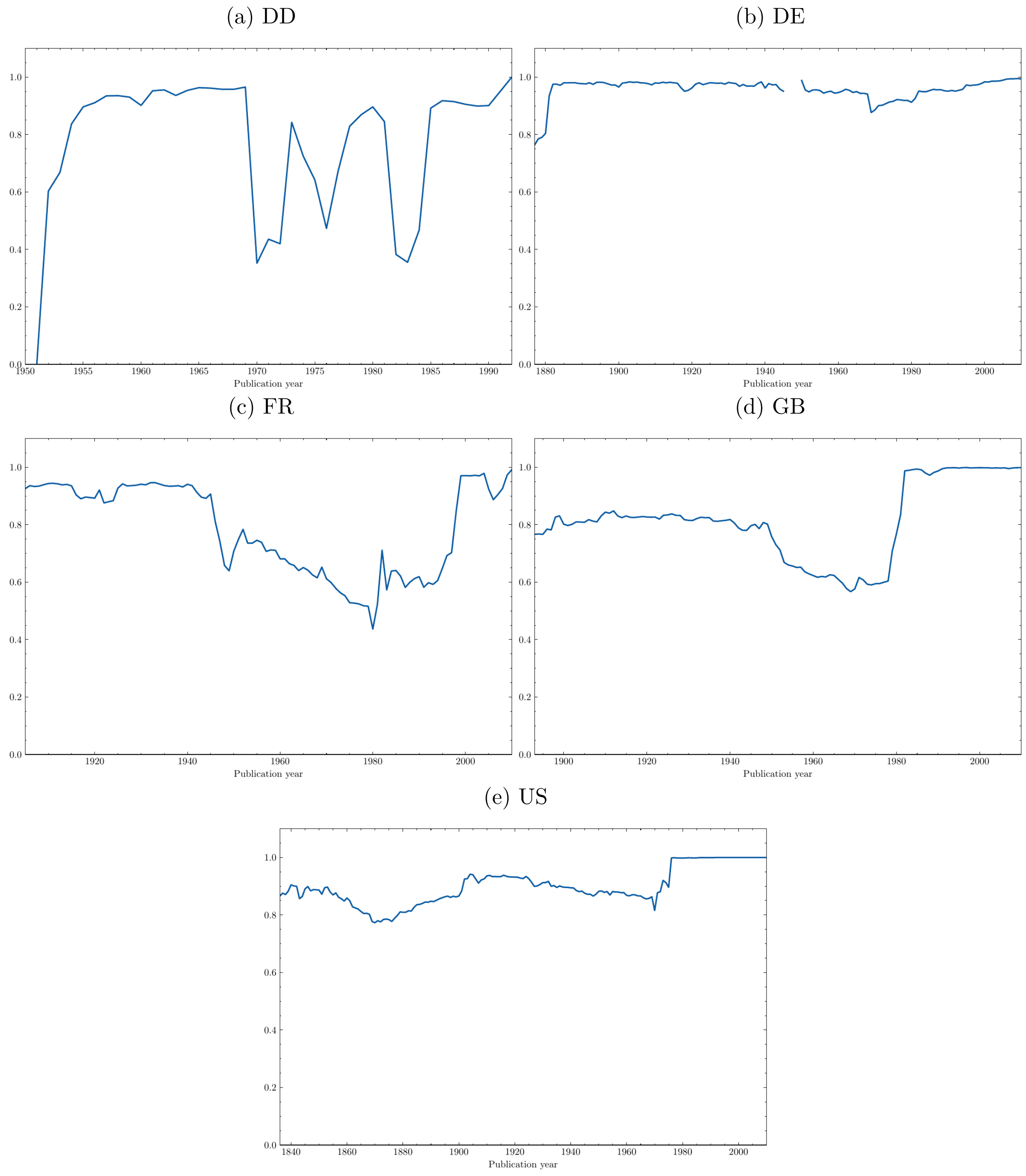
(b) DE

(c) FR

(d) GB

(e) US

Figure 2.28: Share of patentees with a detected location



(a) DD

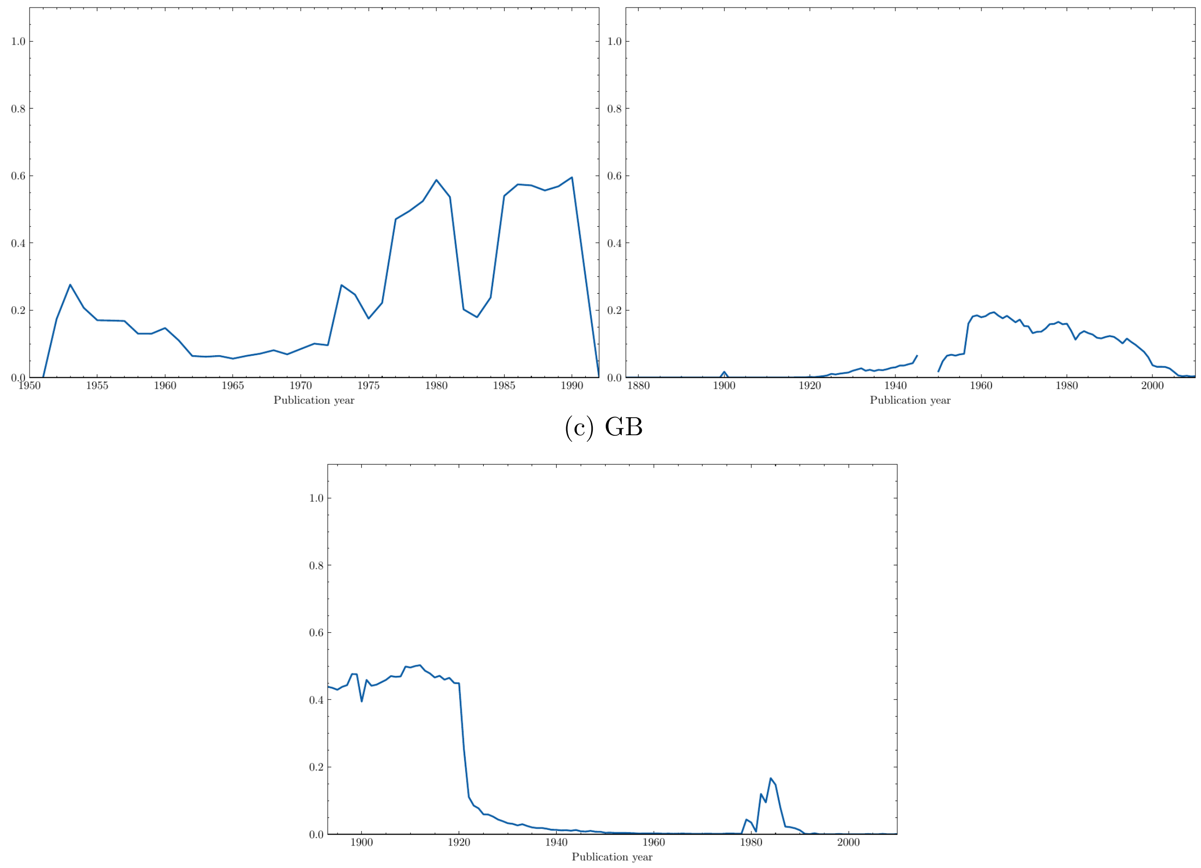
(b) DE

(c) FR

(d) GB

(e) US

Figure 2.30: Share of inventors with a detected occupation

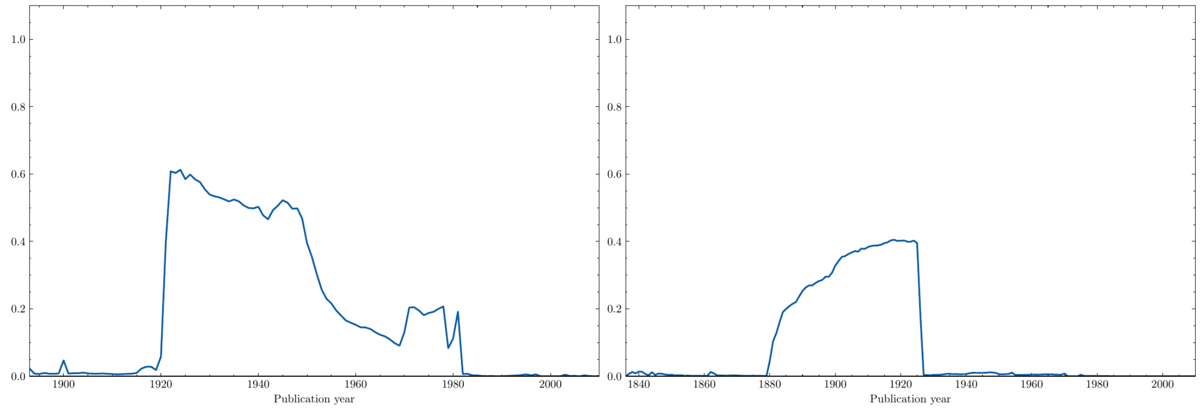


(a) DD

(b) DE

(c) GB

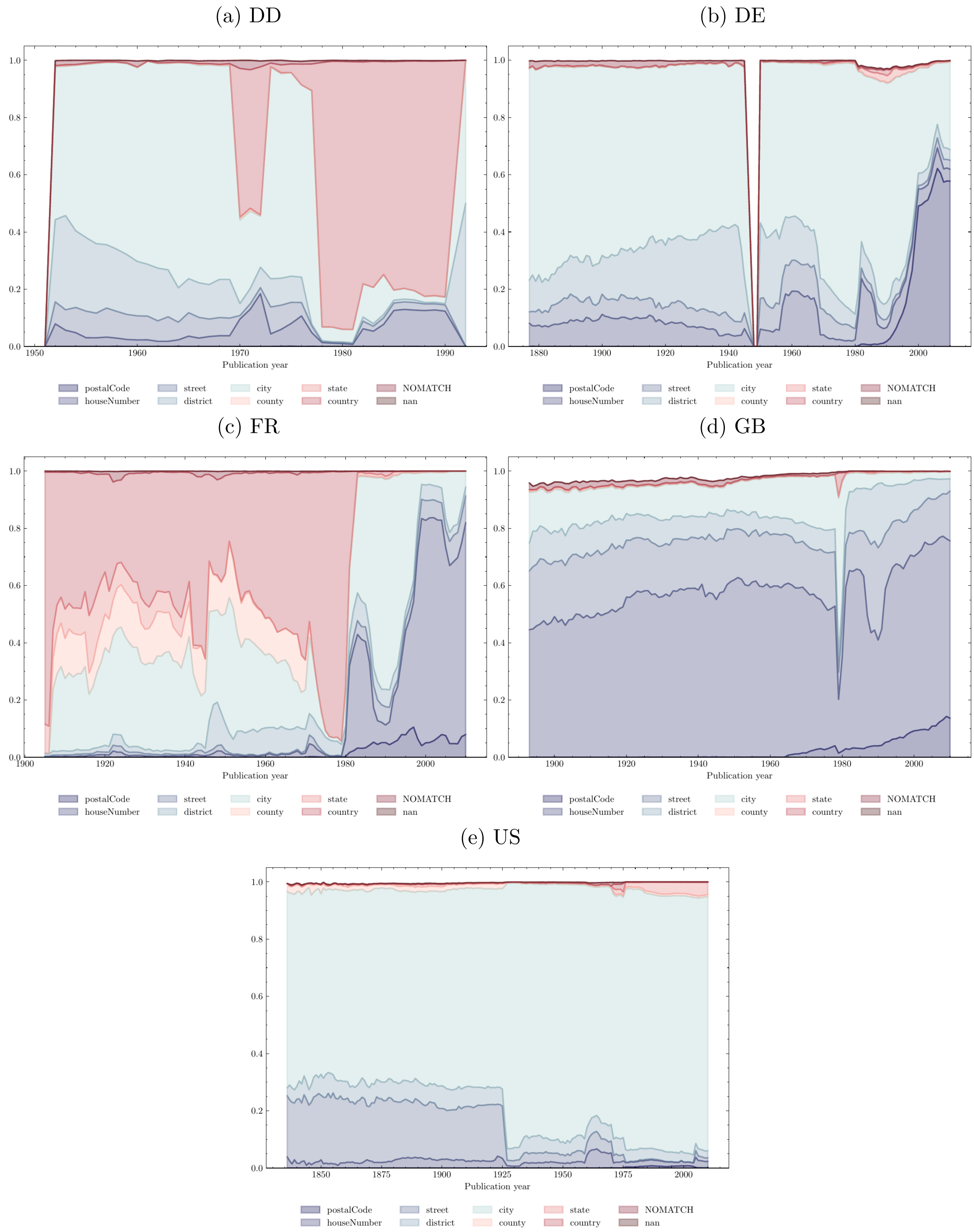
Figure 2.32: Share of inventors with a detected citizenship



(a) GB

(b) US

Figure 2.34: Composition of the most detailed level of geocoding



(a) DD

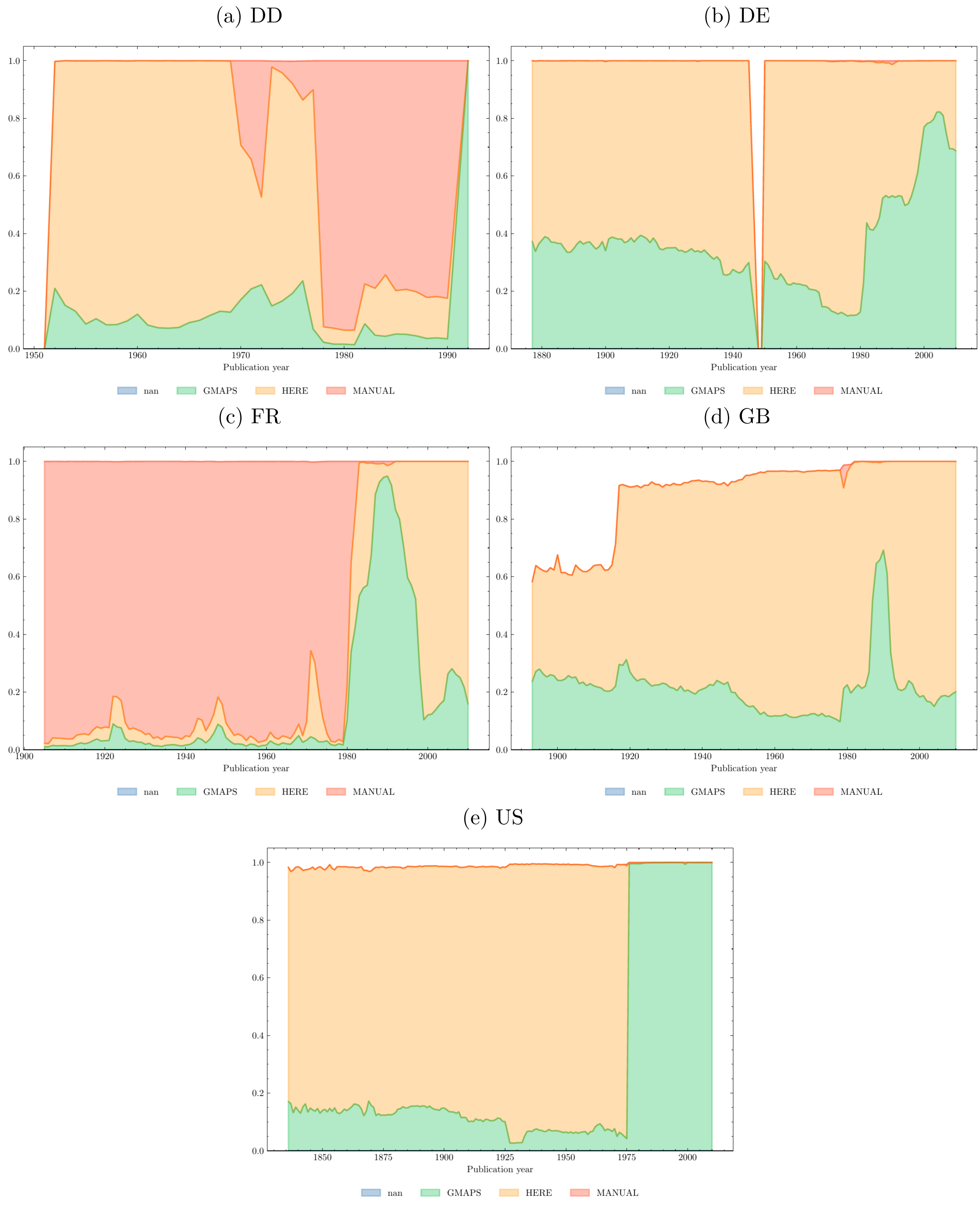
(b) DE

(c) FR

(d) GB

(e) US

Figure 2.36: Composition of geocoding by geocoding source



(a) DD

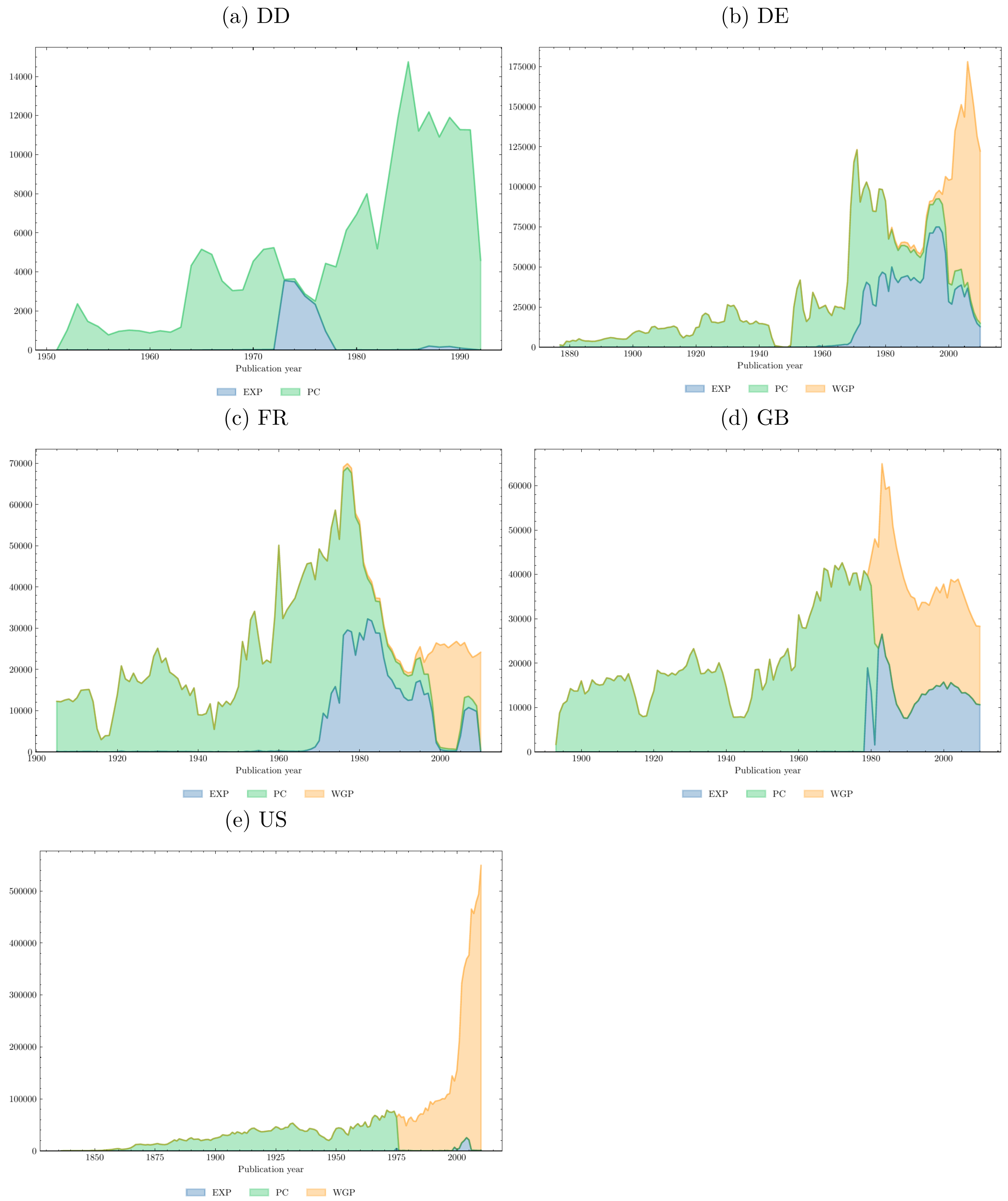
(b) DE

(c) FR

(d) GB

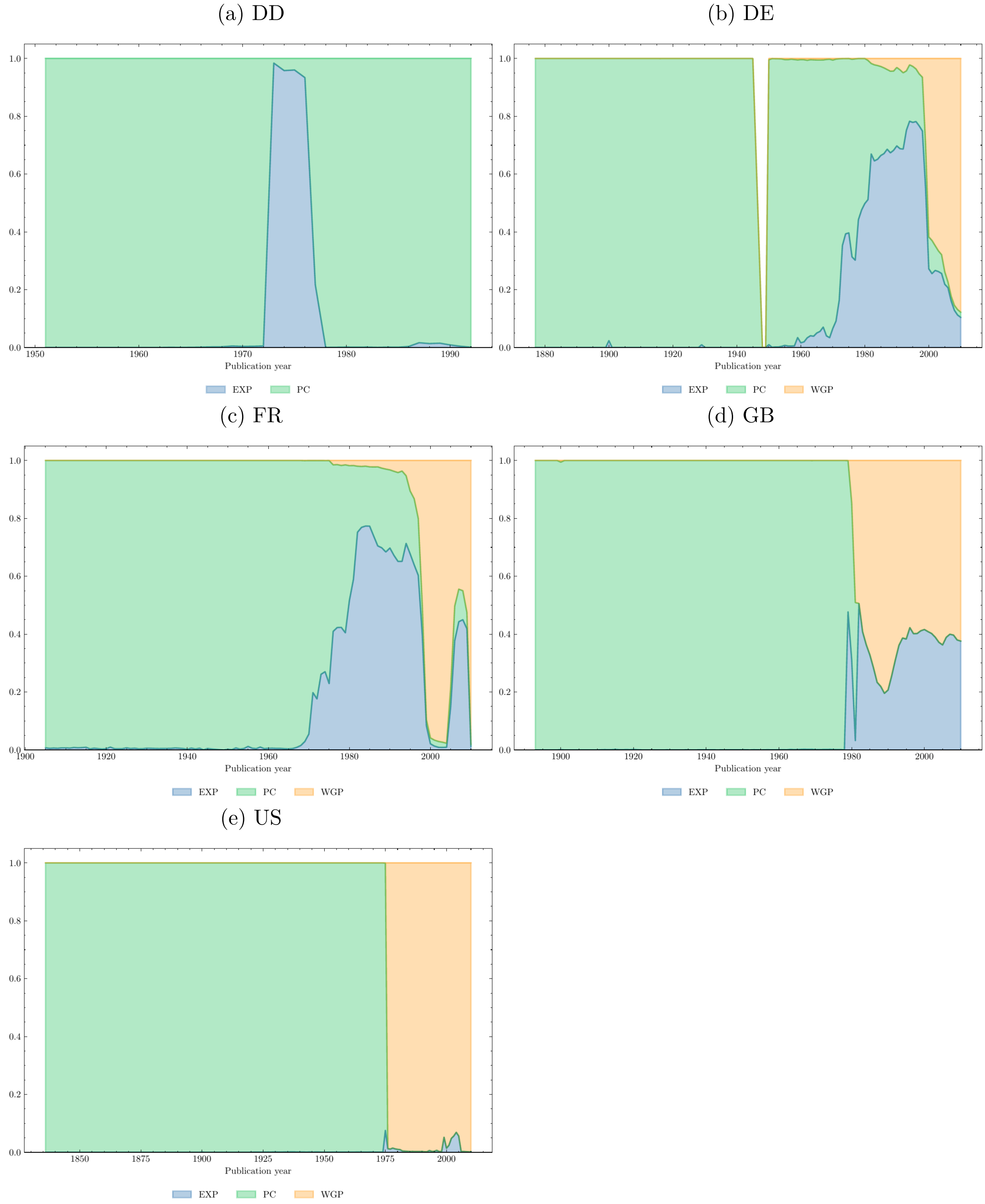
(e) US

Figure 2.38: Database composition by source (number of patents)



Notes: PC refers to PatentCity data, WGP refers to de Rassenfosse et al. (2019c) data and EXP refers to data collected from family expansion from patents included in either PC or WGP.

Figure 2.40: Database composition by source (in share)



(a) DD

(b) DE

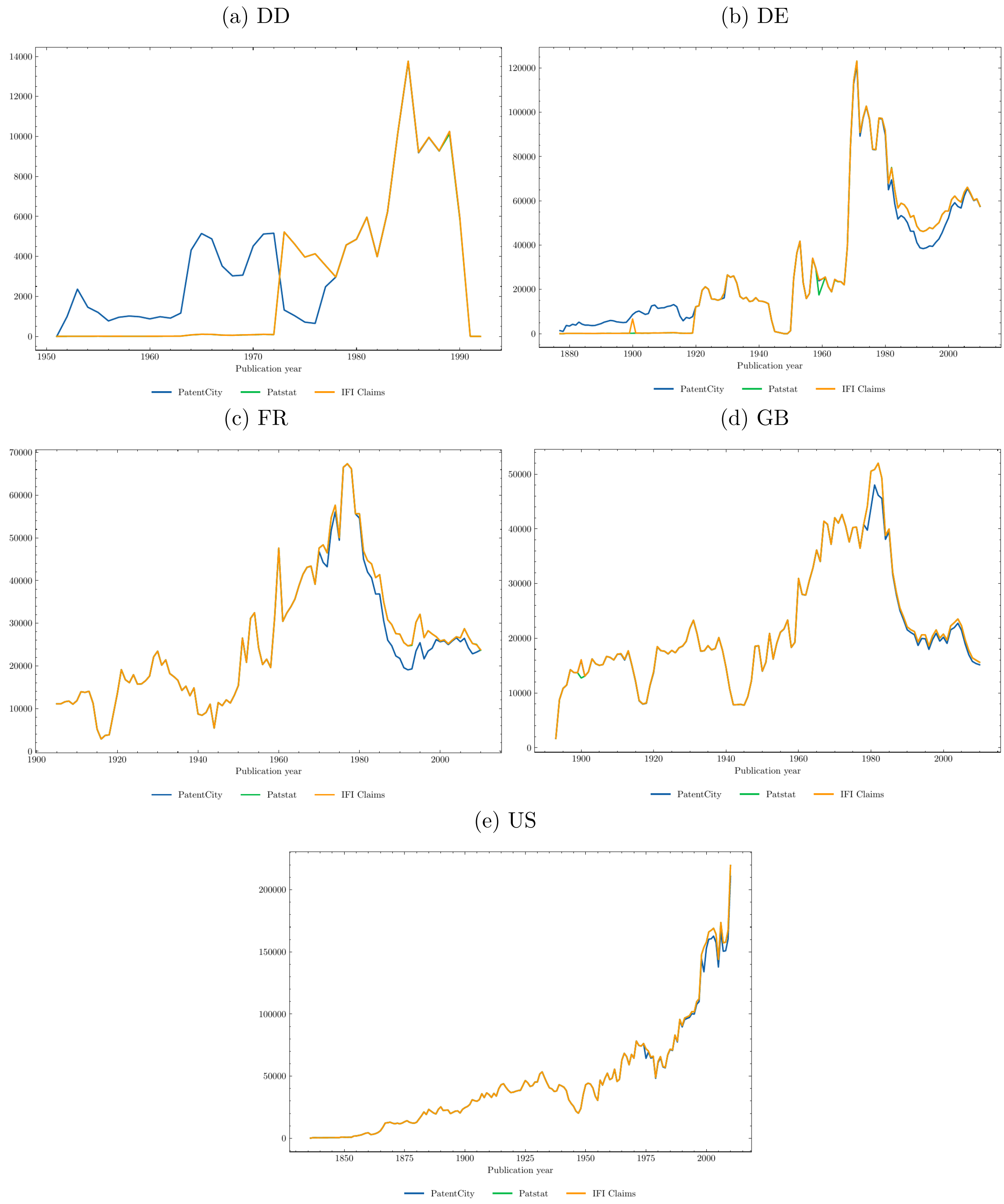
(c) FR

(d) GB

(e) US

Notes: PC refers to PatentCity data, WGP refers to de Rassenfosse et al. (2019c) data and EXP refers to data collected from family expansion from patents included in either PC or WGP.

Figure 2.42: Database coverage by office and publication year (in absolute values)


(a) DD

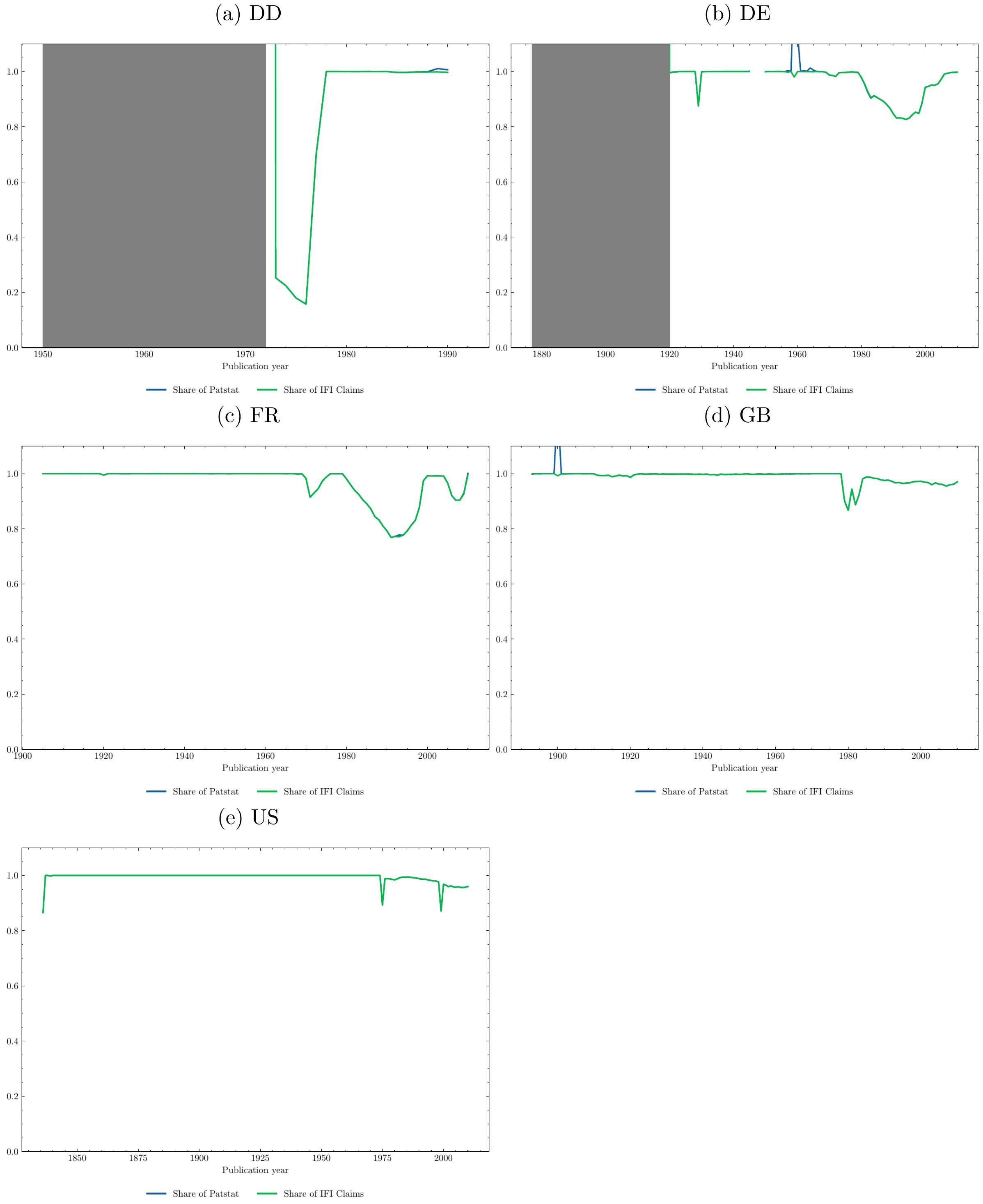
(b) DE

(c) FR

(d) GB

(e) US

Figure 2.44: Database coverage by office and publication year (in share of the Claims database coverage


(a) DD

(b) DE

(c) FR

(d) GB

(e) US

Notes: We report the share of patents which are reported in our database at the office-publication year level as compared to the coverage of the IFI Claims database (publicly available as part of the Google patents public dataset). Shaded areas correspond to office and publication years where patents reported in the IFI Claims database miss dates, meaning that we miss a proper denominator.

Additional annotation guidelines

Table 2.11: Entity annotation guidelines

Patent office	Entity	Content	Example
DD	ASG	Assignee full name	Inhaber: Rhône Poulenc S.A , Paris (Frankreich).
	INV	Inventor full name (*Erfinder*)	Erfinder: Dr. Karl Jellinek , WD
	LOC	Location of the assignee/inventor	Erfinder: Jean Auguste Phelisse, Lyon (Frankreich).
	OCC	Occupation of the assignee/inventor (academic title)	Dr. Elisabeth Kob, WD.
DE	ASG	Assignee full name	ANTON KLEBER in SAAR-BRUCKEN
	INV	Inventor full name (*Erfinder*)	Frutz Doring , Berlin-Frohnau ist als Erfinder genannt worden
	LOC	Location of the assignee/inventor	Demag Akt-Ges. in Duisburg.
	OCC	Occupation of the assignee/inventor (academic title)	Dipl-Ing Georg Werner Gaze, Ingolstadt
	CLAS	Technological class (German system)	KLASSE 49h GRUPPE 27 D 16736VI/49h
FR	ASG	Assignee full name	M. Robert John Jocelyn SWAN résidant en Angleterre
	INV	Inventor full name	(Demande de brevet déposée aux Etats-Unis d'Amérique au nom de M. Ladislas Charles MATSCH)
	LOC	Location of the assignee/inventor	M. Louis LEGRAND résidant en France.
	CLAS	Technological class (French system)	XII Instruments de précision 3 POIDS ET MESURES, INSTRUMENTS DE MATHEMMATIQUES
GB	PERS	Person full name	Maxim Hanson Hersey , Lighting Engineer
	ORG	Firm full name	We, The Convex Incandescent Mantle Company Limited , Manufacturers
	CIT	The origin of the firm or citizenship of the person	a subject of the king of Great Britain and Ireland ,
	LOC	Location of the person/firm	Maxim Hanson Hersey, Lighting Engineer, of 145, Bethune Road, Amhurst Park, London N..
	OCC	Occupation of the person	Maxim Hanson Hersey, Lighting Engineer .
US	INV	Inventor full name	Be it known that I, JAMES M. GARDINER , ...
	ASG	Assignee full name	ASSIGNOR OF ONE-HALF TO SMITH FULMER
	LOC	Location of the assignee/inventor	residing at Mikkalo, in the county of Gilliam and State of Oregon
	CIT	Citizenship of inventor	JOHN SCHLATTER, a citizen of United States

Notes: Colored text corresponds to the entities that we seek to extract: red for inventors, purple for assignees, olive for locations, brown for citizenship and blue for occupations. Full annotation guidelines available at https://cverluise.github.io/patentcity/ (section Guides).

Table 2.12: Relation annotation guidelines

Patent office	Relation	Content	E.g
DD	LOCATION	Links an ASG/INV to a LOC	Rhône Poulenc S.A⟶LOCATION⟶Paris (Frankreich)
	OCCUPATION	Links an ASG/INV to an OCC	Dr⟵OCCUPATION⟵Elisabeth Kob
DE	LOCATION	Links an ASG/INV to a LOC	MARIUS ALBERT de DION⟶LOCATION⟶PUTEAUX (Seine, Frankr.)
	OCCUPATION	Links an ASG/INV to an OCC	Dr.⟵OCCUPATION⟵KARL HENKEL
FR	LOCATION	Links an ASG/INV to a LOC	M.Frederic PERDRIZET⟶LOCATION⟶ France (Gironde)
	CITIZENSHIP	Links an ORG/PERS to its CIT	Maxim Hanson Hersey⟶CITIZENSHIP⟶subject of the king of Great Britain and Ireland
GB	LOCATION	Links an ASG/INV to a LOC	Maxim Hanson Hersey⟶LOCATION⟶145, Bethune Road, Amhurst Park, London N.
	OCCUPATION	Links an ASG/INV to an OCC	Maxim Hanson Hersey⟶OCCUPATION⟶Lighting Engineer
	CITIZENSHIP	Links an INV/ASG to its CIT	WILLIAM H. BAKER⟶CITIZENSHIP⟶citizen of the United States
US	LOCATION	Links an ASG/INV to a LOC	SEDWARD WILLIAM YOUNG⟶LOCATION⟶Tytherley, Wimborne, Dorset, England

Notes: Examples of relations between extracted entities for each patent office. Colored text corresponds to the entities extracted: red for personal inventors, purple for assignees, olive for locations, brown for citizenship and blue for occupations.

Overview of the dataset

Table 2.13: Database schema

Name	Description	Type	Nb non null
publication_number	Publication number.	STR	18,626,068
publication_date	Publication date (yyyymmdd).	INT	18,625,367
family_id	Family ID (DOCDB).	STR	18,625,353
country_code	Country code of the patent office.	STR	18,626,068
pubnum	Publication number.	STR	18,626,068
kind_code	Kind code.	STR	18,626,068
origin	Indicates the origin of the patentee data (PC: patentcity, WGP25: Worldwide Geocoding of Patent - slot 25, WGP45: Worldwide Geocoding of Patent - slot 45, EXP: expansion).	STR	18,626,068
patentee	Patentee	REC	18,626,068
___.is_inv	True if the patentee is an inventor, else False.	BOOL	45,537,241
___.is_asg	True if the patentee is an assignee, else False.	BOOL	45,537,241
___.name_text	Name.	STR	43,402,865
___.person_id	Person ID (PATSTAT).	INT	23,763,520
___.name_start	Name start.	INT	19,639,345
___.name_end	Name end.	INT	19,639,345
___.occ_text	Occupation text.	STR	1,354,930
___.occ_start	Occupation start.	INT	1,354,930
___.occ_end	Occupation end.	INT	1,354,930
___.cit_text	Citizenship text.	STR	3,996,958
___.cit_code	Citizenship code.	STR	3,861,775
___.cit_start	Citizenship start.	INT	3,996,958
___.cit_end	Citizenship end.	INT	3,996,958
___.loc_text	Location text.	STR	42,232,737
___.loc_start	Location start.	INT	16,334,841
___.loc_end	Location end.	INT	16,334,841
___.loc_addressLines	Formatted address lines built out of the parsed address components.	STR	16,003,816
___.loc_locationLabel	Assembled address value for displaying purposes.	STR	41,901,699
___.loc_country	ISO 3166-alpha-3 country code.	STR	41,898,330

Continued on next page

___.loc_state	First subdivision level(s) below the country. Where commonly used, this is a state code (for instance, CA for California).	STR	41,428,298
___.loc_county	Second subdivision level(s) below the country. Use of this field is optional if a second subdivision level is not available.	STR	34,200,971
___.loc_city	Locality of the address.	STR	40,391,684
___.loc_district	Subdivision level below the city. Use of this field is optional if a second subdivision level is not available.	STR	18,276,320
___.loc_subdistrict	Subdivision level below the district. Used only for India.	STR	16,003,816
___.loc_postalCode	Postal code.	STR	23,837,493
___.loc_street	Street name.	STR	18,145,660
___.loc_building	Building name.	STR	16,130,485
___.loc_houseNumber	House number.	STR	17,710,245
___.loc_longitude	Longitude.	FLOA	41,517,796
___.loc_latitude	Latitude.	FLOA	41,517,796
___.loc_relevance	Indicates the relevance of the results found; the higher the score the more relevant the alternative. The score is a normalized value between 0 and 1.	FLOA	12,203,353
___.loc_matchType	Quality of the location match. pointAddress: Location matches exactly as point address. interpolated: Location was interpolated.	STR	41,268,017
___.loc_matchCode	Code indicating how well the result matches the request. Enumeration [exact, ambiguous, upHierarchy, ambiguousUpHierarchy].	STR	16,003,816
___.loc_matchLevel	The most detailed address field that matched the input record.	STR	41,643,215

Continued on next page

____.loc__matchQualityCountry	MatchQuality provides detailed information about the match quality of a result at attribute level. Match quality is a value between 0.0 and 1.0. 1.0 represents a 100% match. Here, matchQuality is defined at country level.	FLOA	2,658,311
____.loc__matchQualityState	Same at state level.	FLOA	6,553,671
____.loc__matchQualityCounty	Same at county level.	FLOA	1,547,347
____.loc__matchQualityCity	Same at city level.	FLOA	11,331,772
____.loc__matchQualityDistrict	Same at district level.	FLOA	1,361,402
____.loc__matchQualityPostalCode	Same at postalCode level.	FLOA	147,862
____.loc__matchQualityStreet	Same at street level.	FLOA	2,452,802
____.loc__matchQualityHouseNumber	Same at houseNumber level.	FLOA	1,034,844
____.loc__matchQualityBuilding	Same at building level.	FLOA	410
____.loc__key	Key used for statistical area mapping (internal use).	STR	31,137,221
____.loc__statisticalArea1	Name of the high level Statistical Area.	STR	31,061,188
____.loc__statisticalArea1Code	Code of the high level Statistical Area.	STR	31,061,188
____.loc__statisticalArea2	Name of the mid level Statistical Area.	STR	31,061,165
____.loc__statisticalArea2Code	Code of the mid level Statistical Area.	STR	19,738,673
____.loc__statisticalArea3	Name of the low level Statistical Area.	STR	31,055,300
____.loc__statisticalArea3Code	Code of the low level Statistical Area.	STR	31,067,057
____.loc__recId	Identifier of the input address in the response.	STR	42,232,737
____.loc__seqLength	Number of results for the corresponding input record.	INT	12,244,380
____.loc__seqNumber	Consecutively numbers the different results for the corresponding input record starting with 1.	INT	29,657,332
____.loc__source	Geocoding source (in [HERE, GMAPS, MANUAL]).	STR	41,901,712

Continued on next page

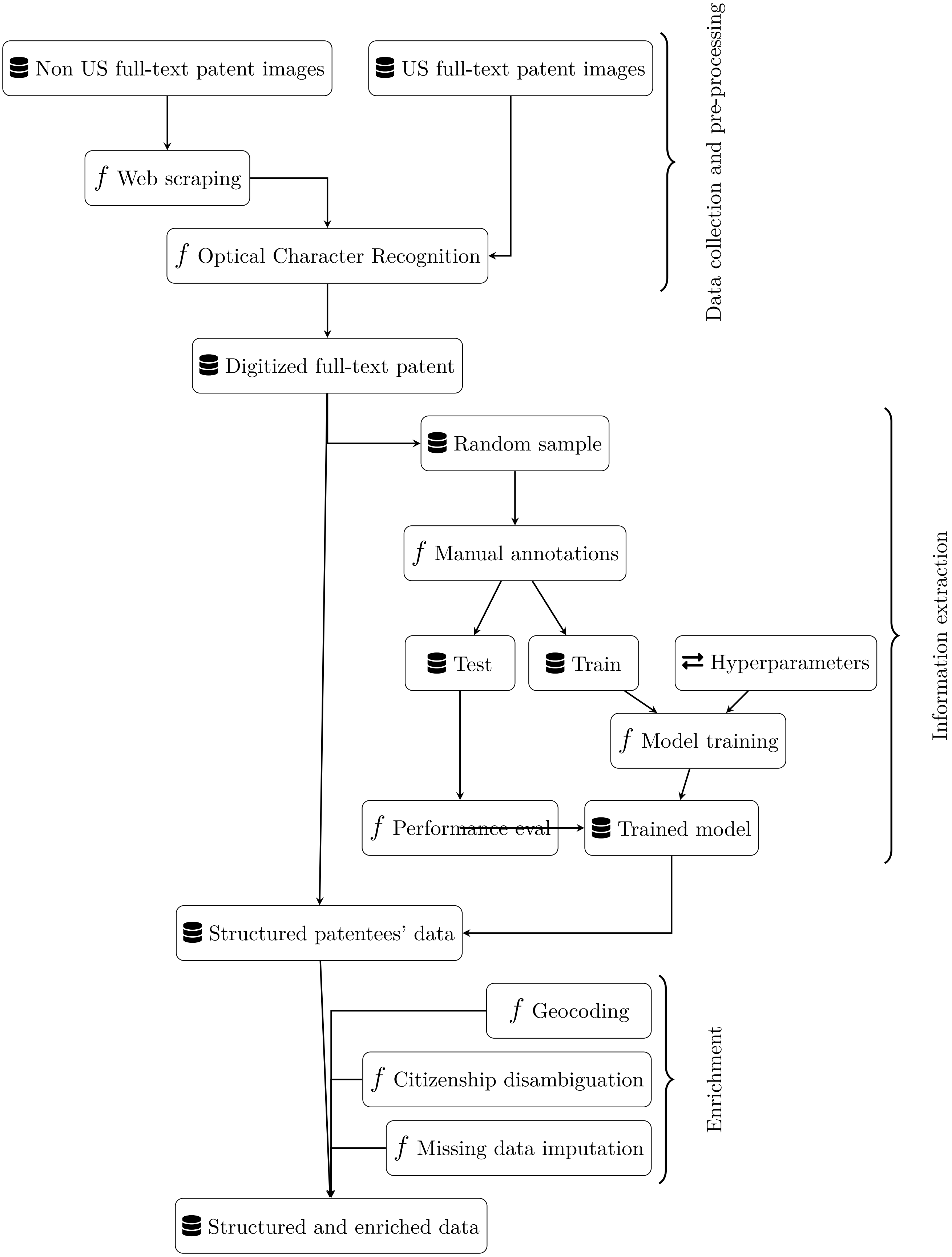
___.is_duplicate	True if a patentee with the 'same' name has been detected in the same patent. Only one of the two is marked as duplicate.	BOOL	3,985,815

Notes: Variable names prefixed by a «___.» are nested variables. For example, «___.is_inv» is nested in the «patentee» variable.

Pipeline

Figure 2.46: Workflow pipeline

2.46.


Non US full-text patent images
US full-text patent images
Web scraping
Optical Character Recognition
Data collection and pre-processing
Digitized full-text patent
Random sample
Manual annotations
Test
Train
Hyperparameters
Model training
Performance eval
Trained model
Information extraction
Structured patentees' data
Geocoding
Citizenship disambiguation
Missing data imputation
Enrichment
Structured and enriched data

Other data

Population by region

We construct population data at regional level using official sources for the most recent period from as long as possible. The data are constructed at different levels and then aggregated at the NUTS2 (or Commuting Zone for the US) level when necessary. These levels are respectively: Government regions (*Regierungsbezirke*) for Germany (NUTS 2), *Départements* for France (NUTS 3), county or equivalent for the United Kingdom (NUTS 2) and county for the US.

These official sources used as a benchmark are respectively Eurostat for Germany, the INSEE for France, the ONS for the United Kingdom and the Census Bureau's Population Estimates for the US. They are then completed by historical sources namely from Rosés and Wolf (2018) and Eckert et al. (2020). Details on the assumptions made to construct these estimates are reported in the project's website.

Gravity variables

The analysis presented in Section 2.4 use country specific variables such as GDP per capita and country to country bilateral variables like trade flow (imports and exports), distance and other standard gravity variables.

We use the following sources:

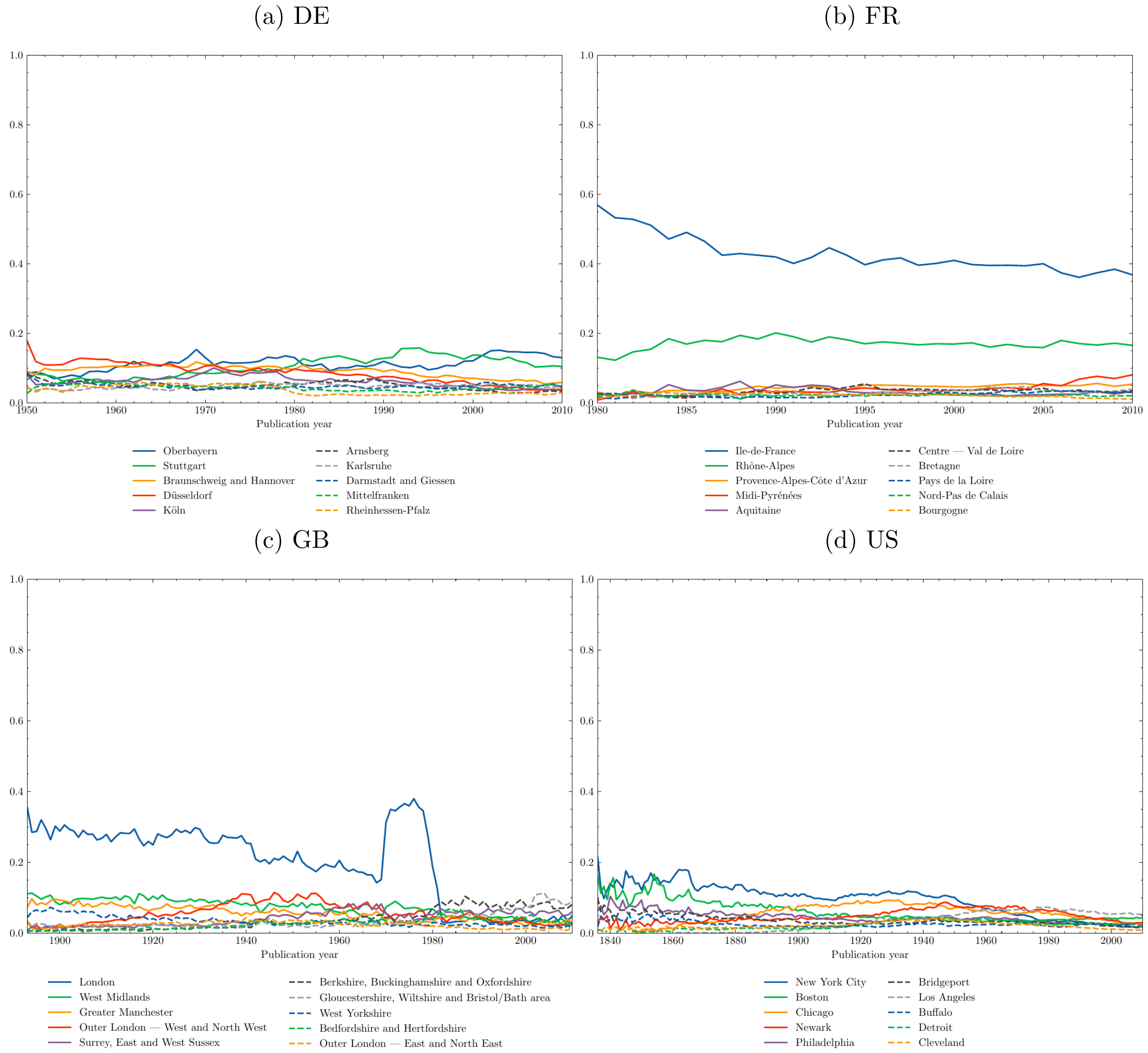
- GDP per capita comes from Bolt and van Zanden (2020)
- Distance between pairs of countries, dummy variable for being a former colony and sharing a same language comes from Mayer and Zignago (2011)
- Trade flows comes from Fouquin and Hugot (2016)

We have restricted the sample to countries and year for which we can measure GDP per capita, which represent 165 countries. The country of origin is one of the four countries considered in the analysis (France, Germany, the United Kingdom and the US). More details can be found in project's website.

Additional results for Section 2.3

Top regions for inventors

Figure 2.47: Share of patents in top 10 regions over time - inventors only



(a) DE

(b) FR

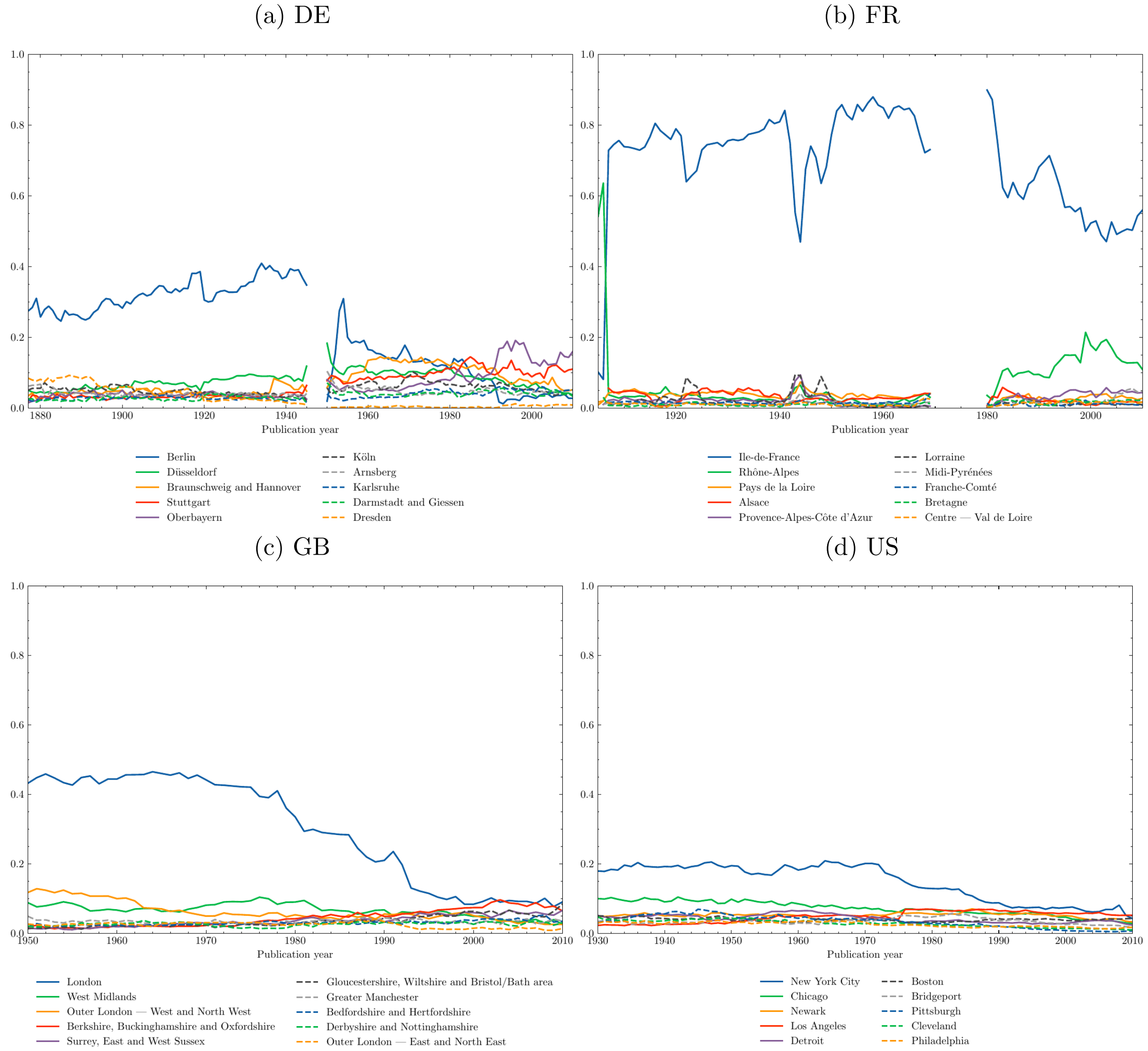
(c) GB

(d) US

Notes: This reports the share of total patentees by regions (NUTS2 + CZ for the US) where the location of a patent is given by the location of its inventors. The top 10 regions are selected based on the number of years they belong to the top 10 over the whole period of observation. Data for 1946-1950 are not available for Germany, data for 1970-1980 do not report any *département* level information for France.

Top regions for assignees

Figure 2.49: Share of patents in top 10 regions over time - assignees only


(a) DE

(b) FR

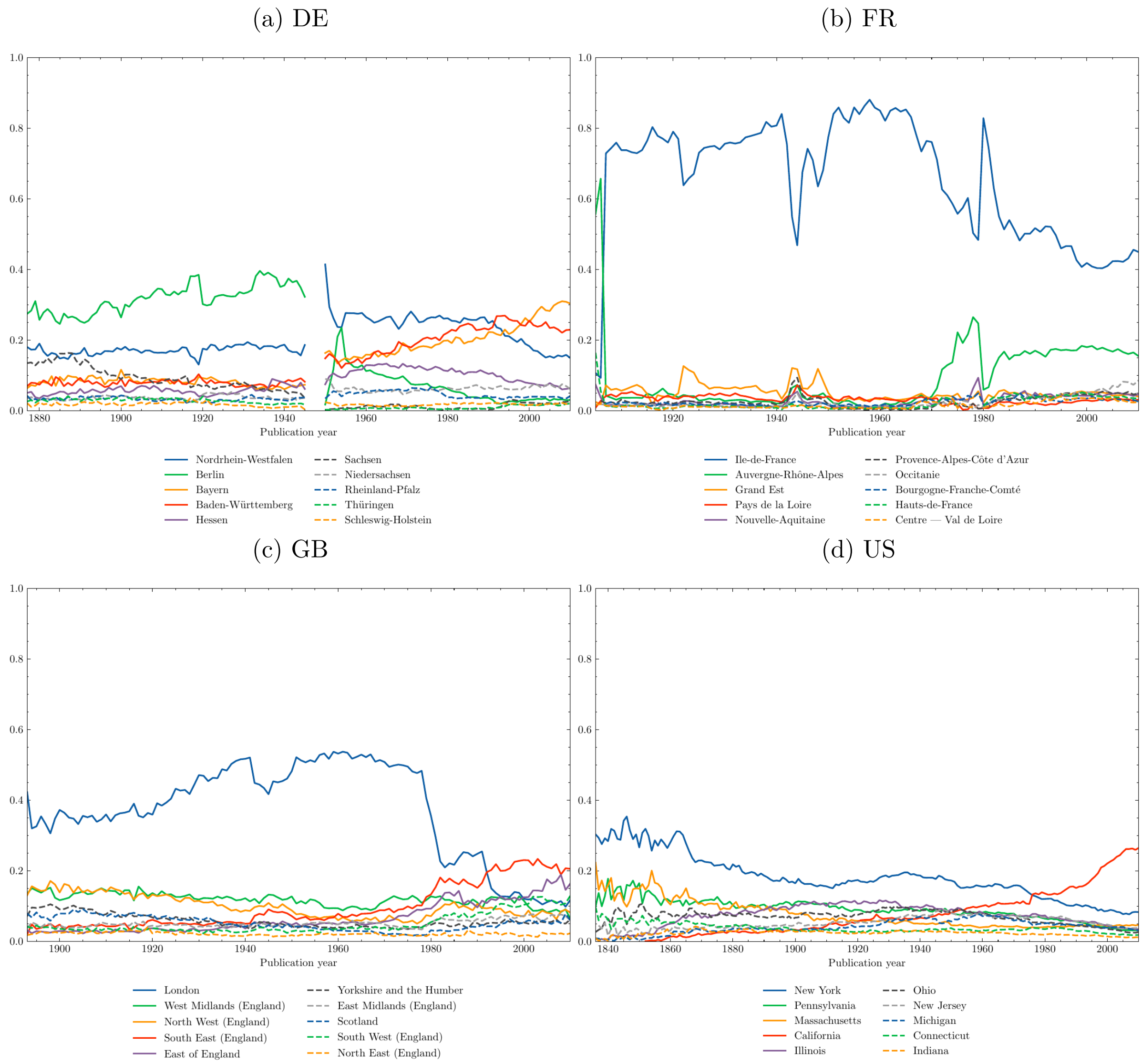
(c) GB

(d) US

Notes: This reports the share of total patentees by regions (NUTS2 + CZ for the US) where the location of a patent is given by the location of its assignees. The top 10 regions are selected based on the number of years they belong to the top 10 over the whole period of observation. Data for 1946-1950 are not available for Germany, data for 1970-1980 do not report any *département* level information for France.

Top regions at a more aggregated level

Figure 2.51: Share of patents in top 10 regions over time - larger regional level



(a) DE

(b) FR

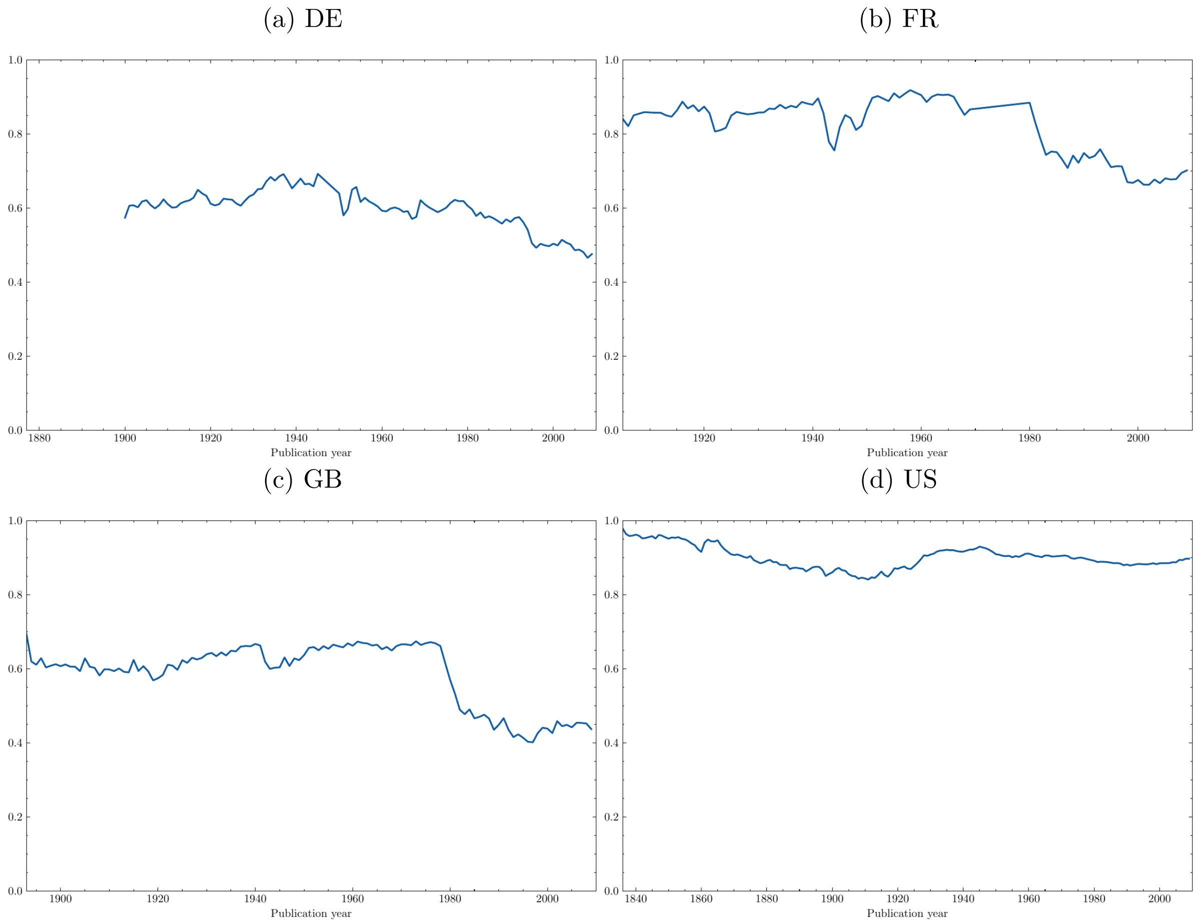
(c) GB

(d) US

Notes: This reports the share of total patentees by larger regions (NUTS1 + state for the US) where the location of a patent is given by the location of its patentees. The top 10 regions are selected based on the number of years they belong to the top 10 over the whole period of observation. Data for 1946-1950 are not available for Germany, data for 1970-1980 do not report any *département* level information for France.

Gini coefficient

Figure 2.53: Gini coefficient of the distribution of patents



(a) DE

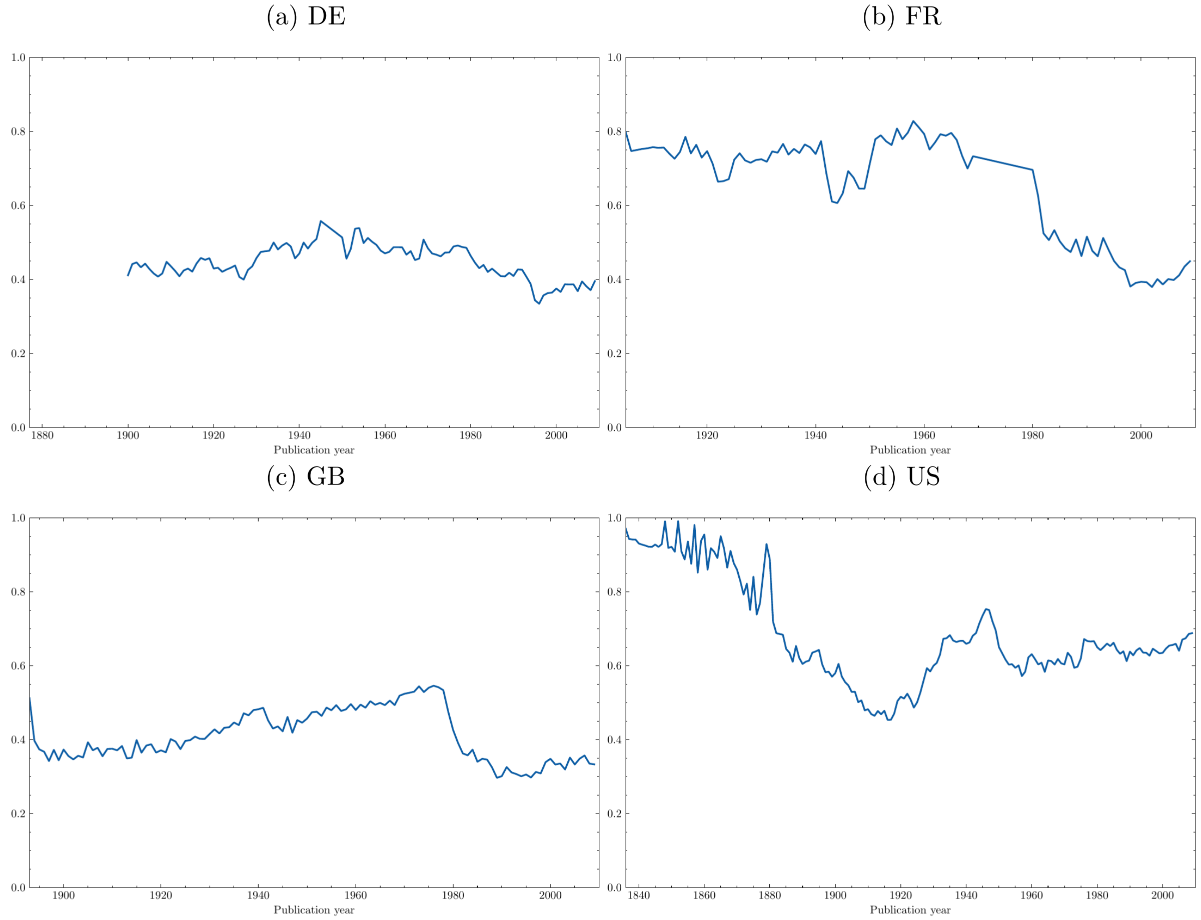
(b) FR

(c) GB

(d) US

Notes: This reports the Gini coefficient on the distribution of number of patents per regions (NUTS2 + CZ for the US) where the location of a patent is given by the location of its patentees. Data for 1946-1950 are not available for Germany, data for 1970-1980 do not report any *département* level information for France.

Figure 2.55: Gini coefficient of the distribution of patents per capita


(a) DE

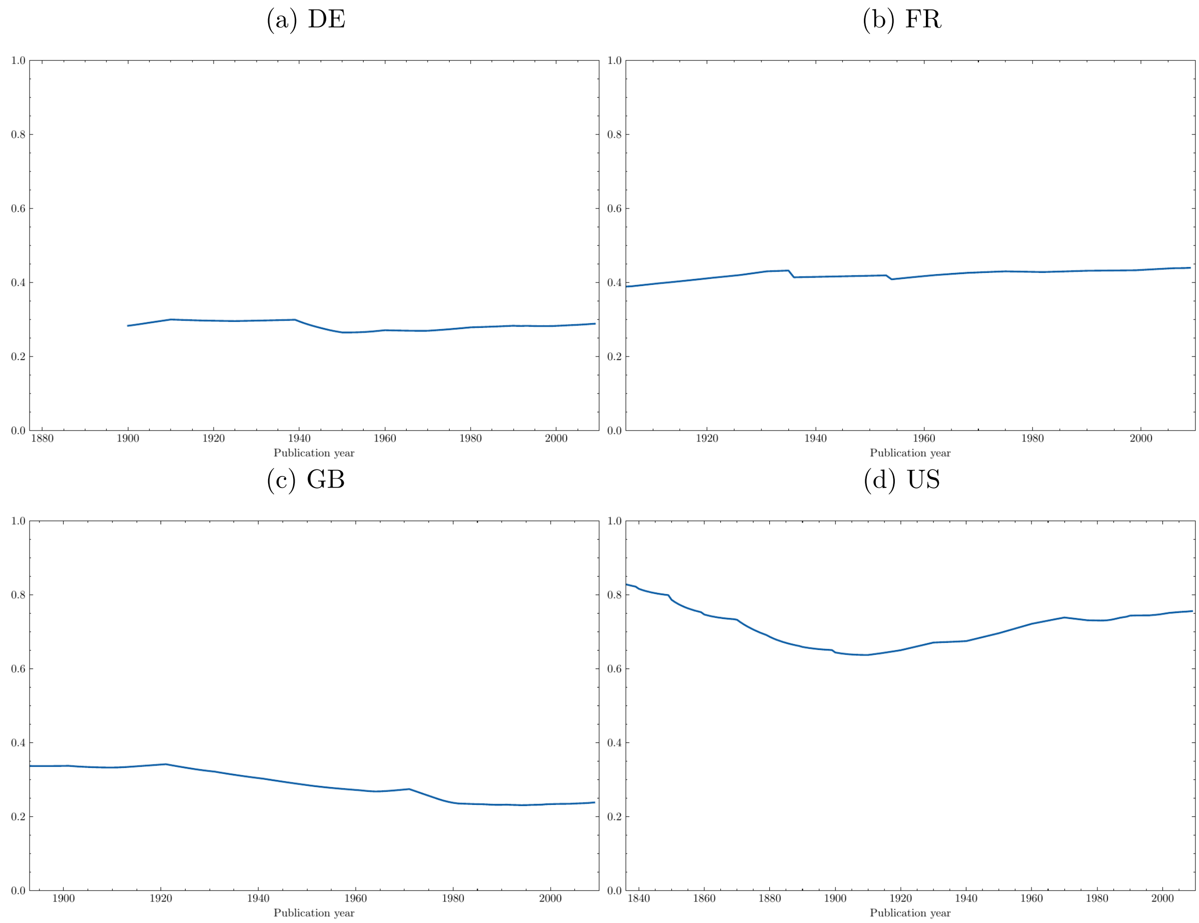
(b) FR

(c) GB

(d) US

Notes: This reports the Gini coefficient on the distribution of number of patents per capita per regions (NUTS2 + CZ for the US) where the location of a patent is given by the location of its patentees. Data for 1946-1950 are not available for Germany, data for 1970-1980 do not report any *département* level information for France.

Figure 2.57: Gini coefficient of the distribution of population


(a) DE

(b) FR

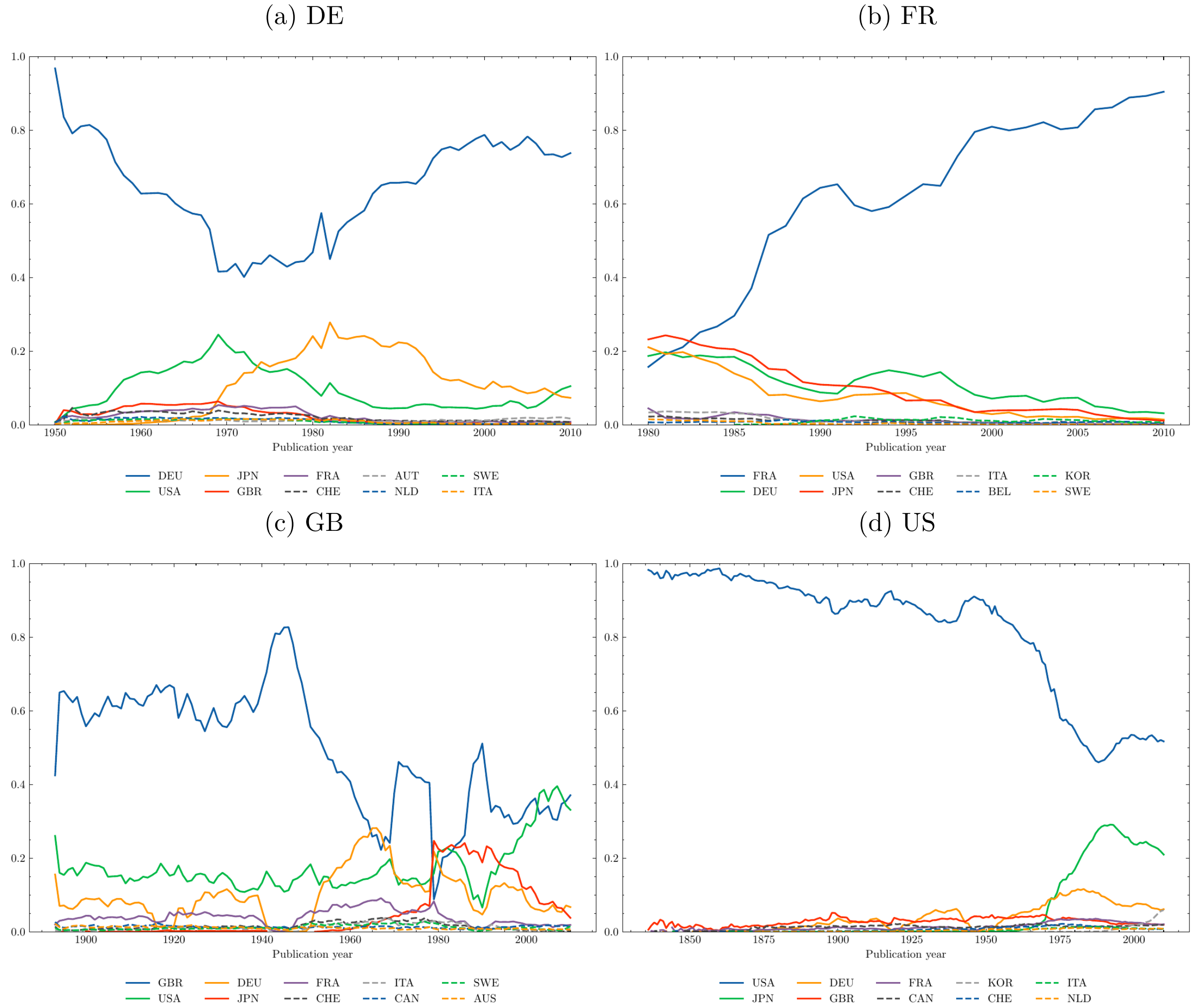
(c) GB

(d) US

Notes: This reports the Gini coefficient on the distribution of population per regions (NUTS2 + CZ for the US).

Robustness Figure for Section 2.4

Figure 2.59: Country of residence of inventors by patent offices


(a) DE

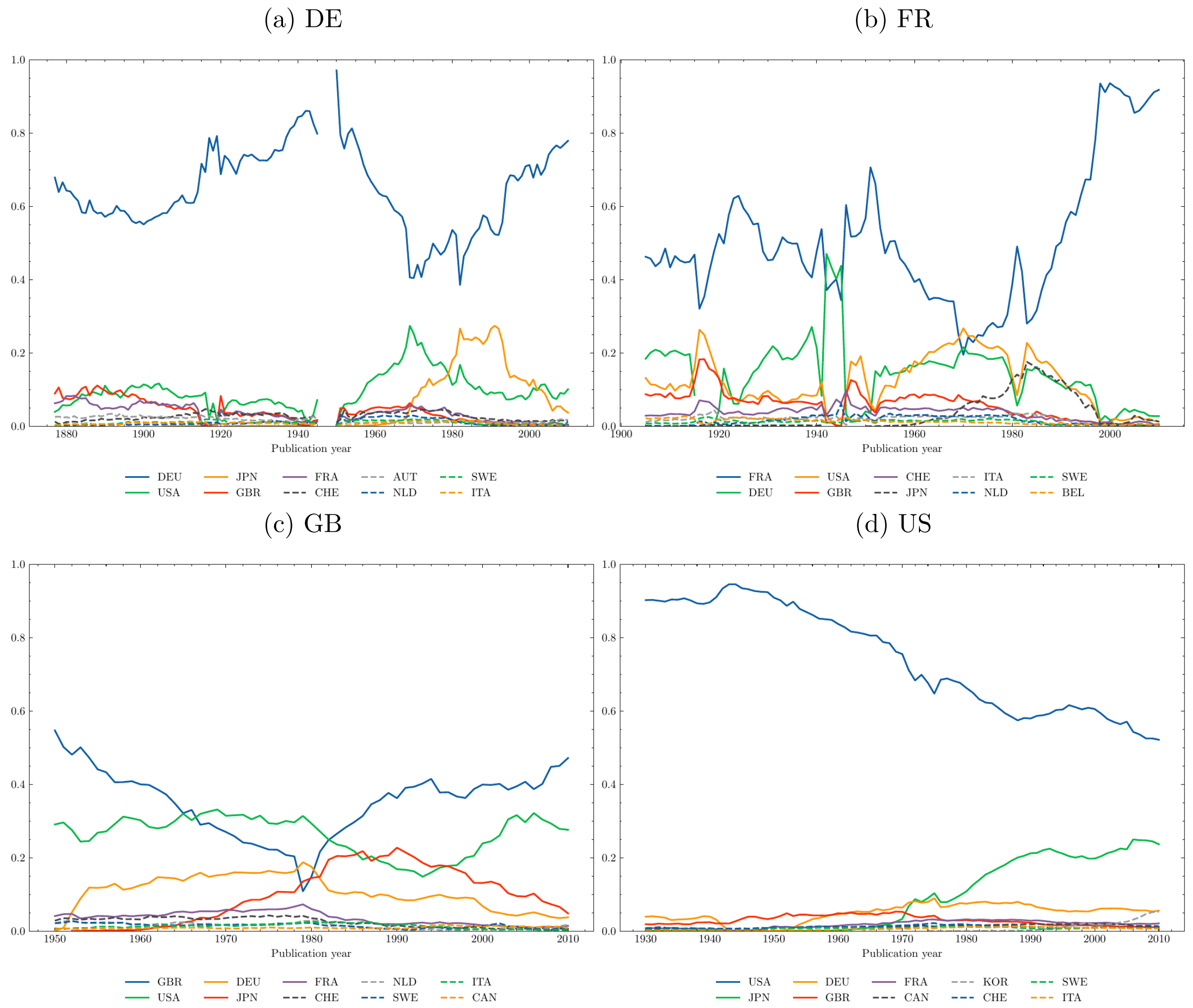
(b) FR

(c) GB

(d) US

Notes: This Figure reports the share of inventors by country of residence for each of the German, French, British and US patent offices and for the top 10 countries in terms of average share over the period of observation.

Figure 2.61: Country of residence of assignees by patent offices



(a) DE

(b) FR

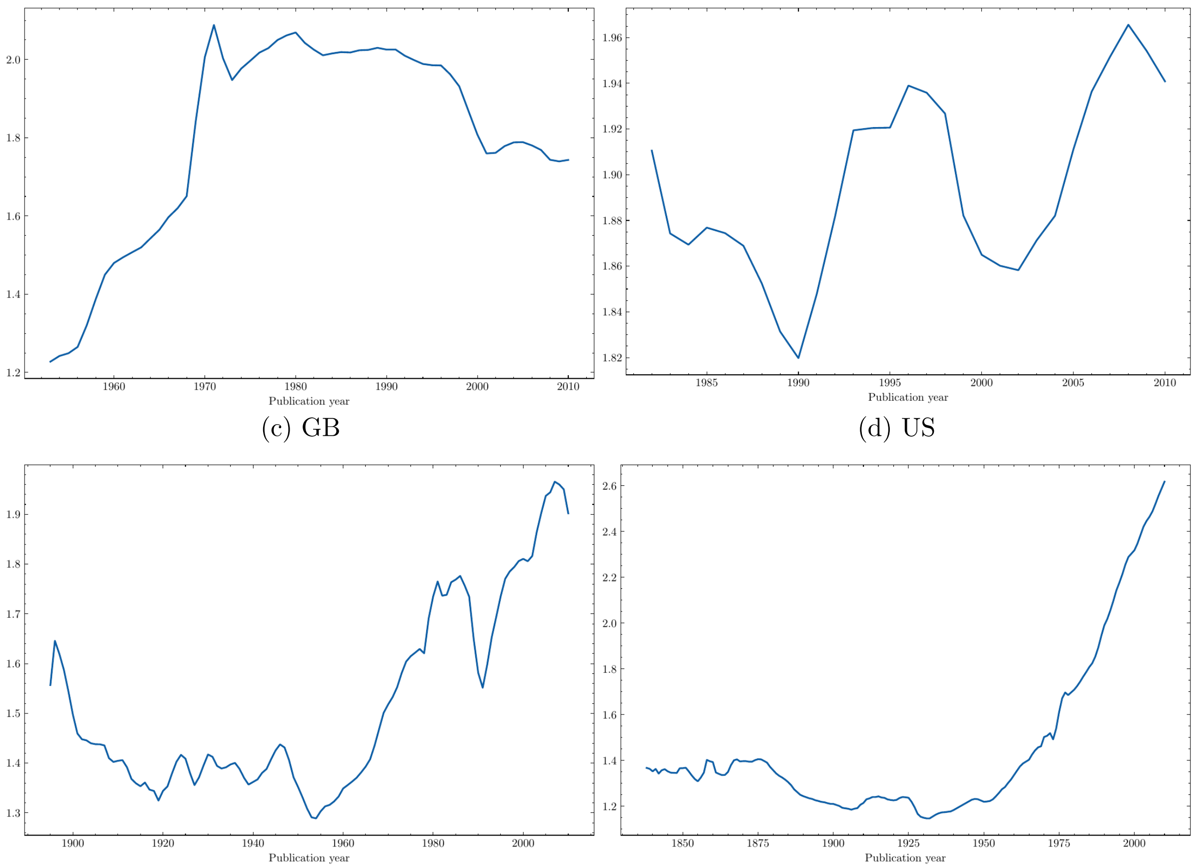
(c) GB

(d) US

Notes: This Figure reports the share of assignees by country of residence for each of the German, French, British and US patent offices and for the top 10 countries in terms of average share over the period of observation.

Robustness Figure for Section 2.5

Figure 2.63: Average number of inventors per patent



(a) DE

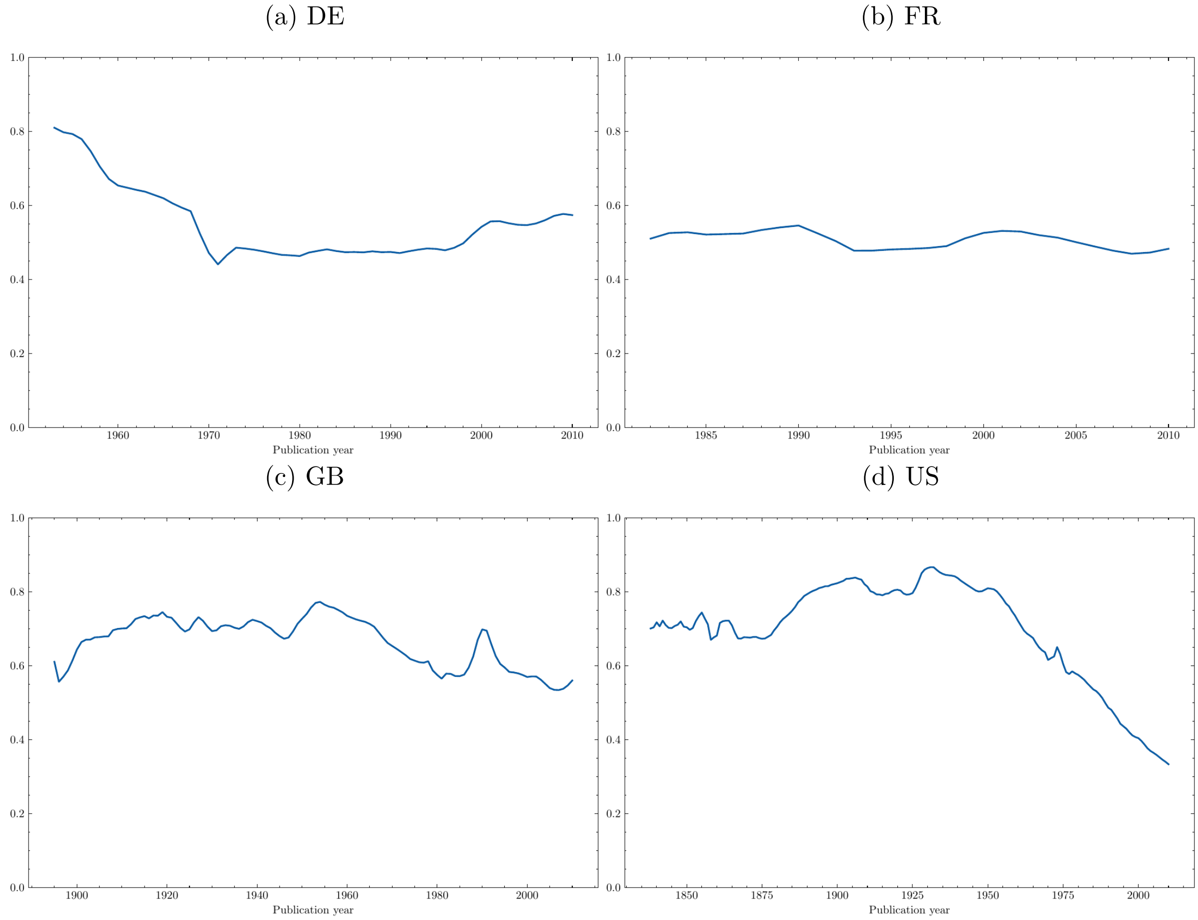
(b) FR

(c) GB

(d) US

Notes: This Figure reports the average number of inventors per patent.

Figure 2.65: Share of patent with only one inventor


(a) DE

(b) FR

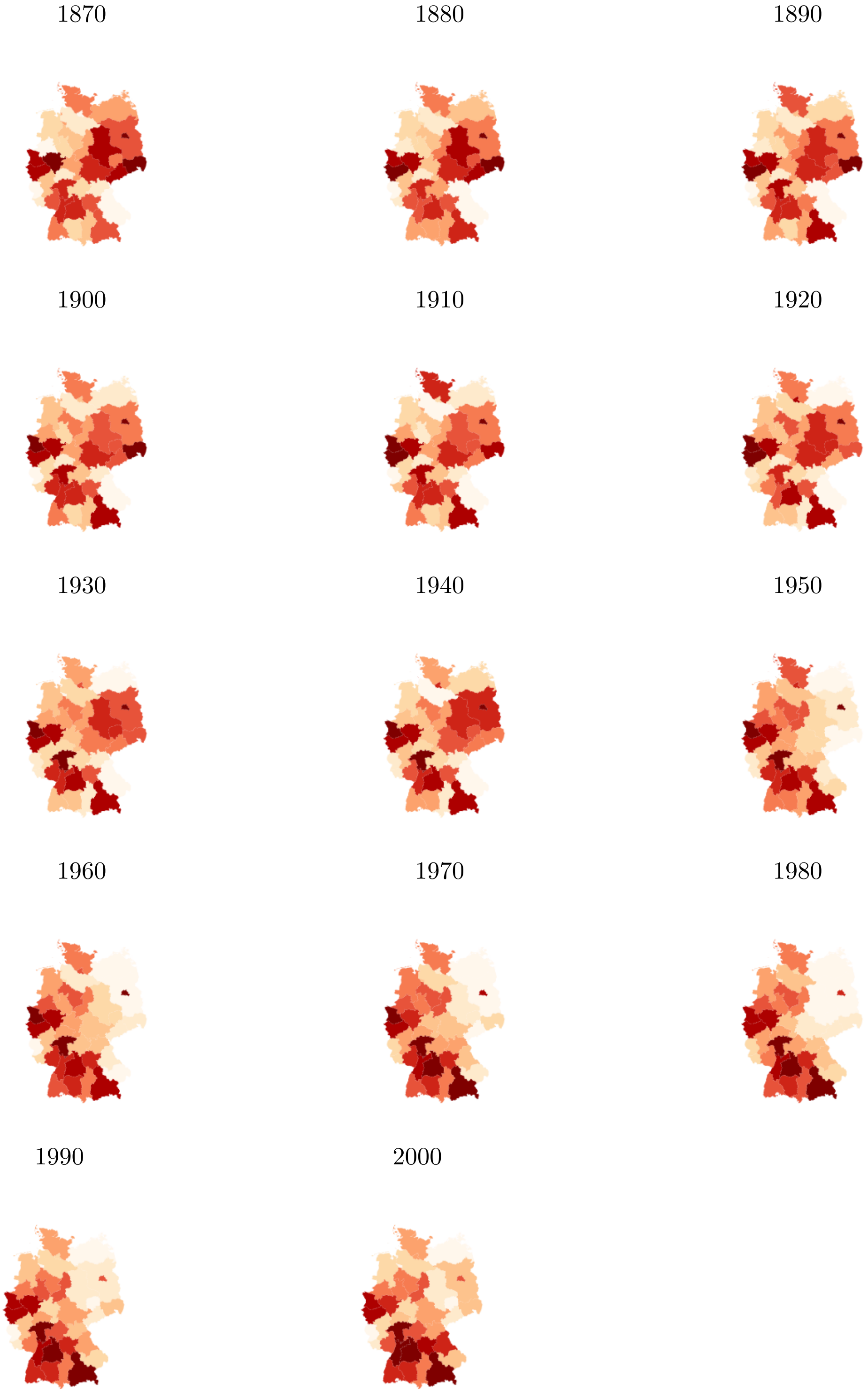
(c) GB

(d) US

Notes: This Figure reports the share of patents with only one inventor.

Maps by decade

Figure 2.67: Patentees location counts in DE (by level 2 statistical area)



1870

1880

1890

1900

1910

1920

1930

1940

1950

1960

1970

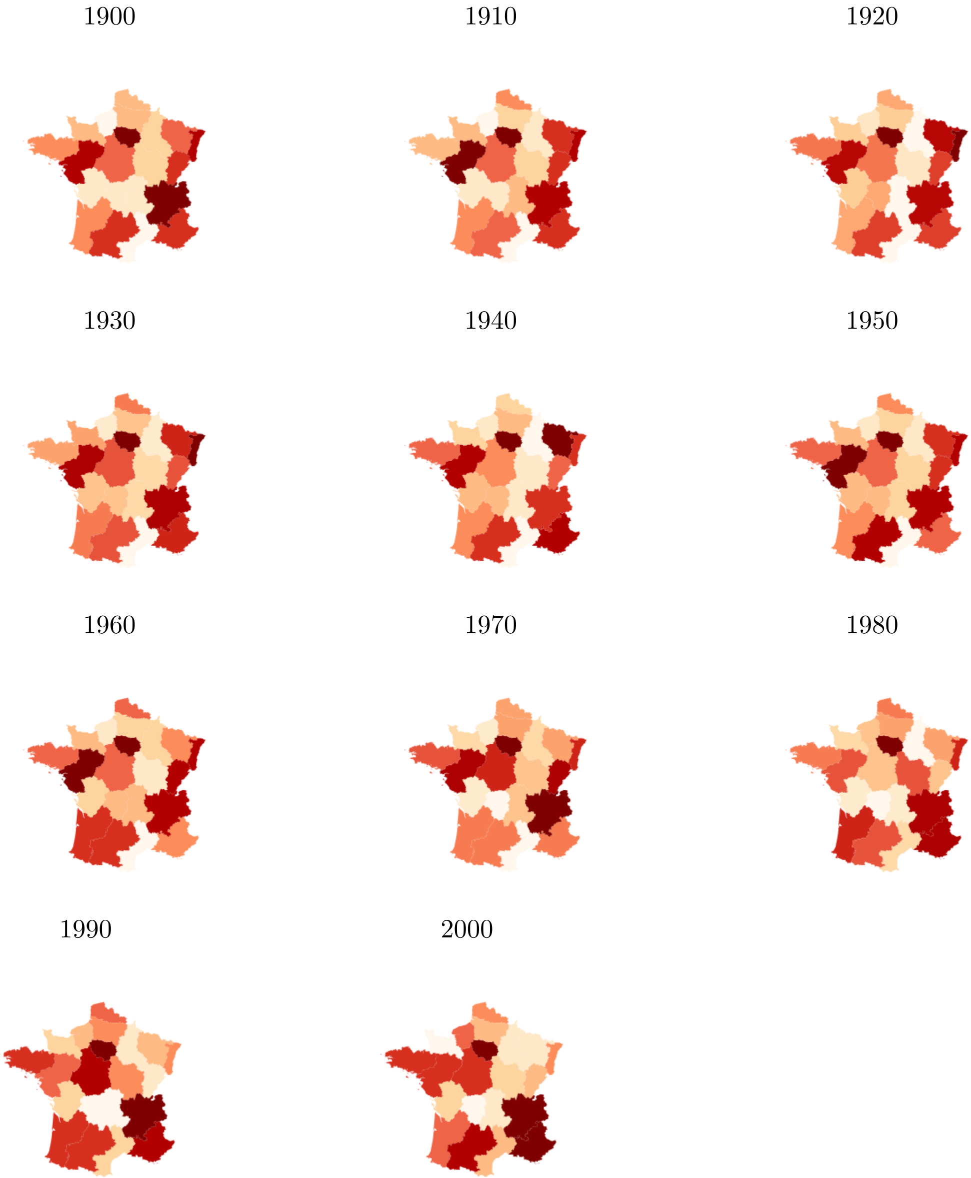
1980

1990

2000

Notes: The warmer, the more patentees (inventors and assignees) are located in a given statistical area (relative to other domestic statistical areas).

Figure 2.69: Patentees location counts in FR (by level 2 statistical area)



1900

1910

1920

1930

1940

1950

1960

1970

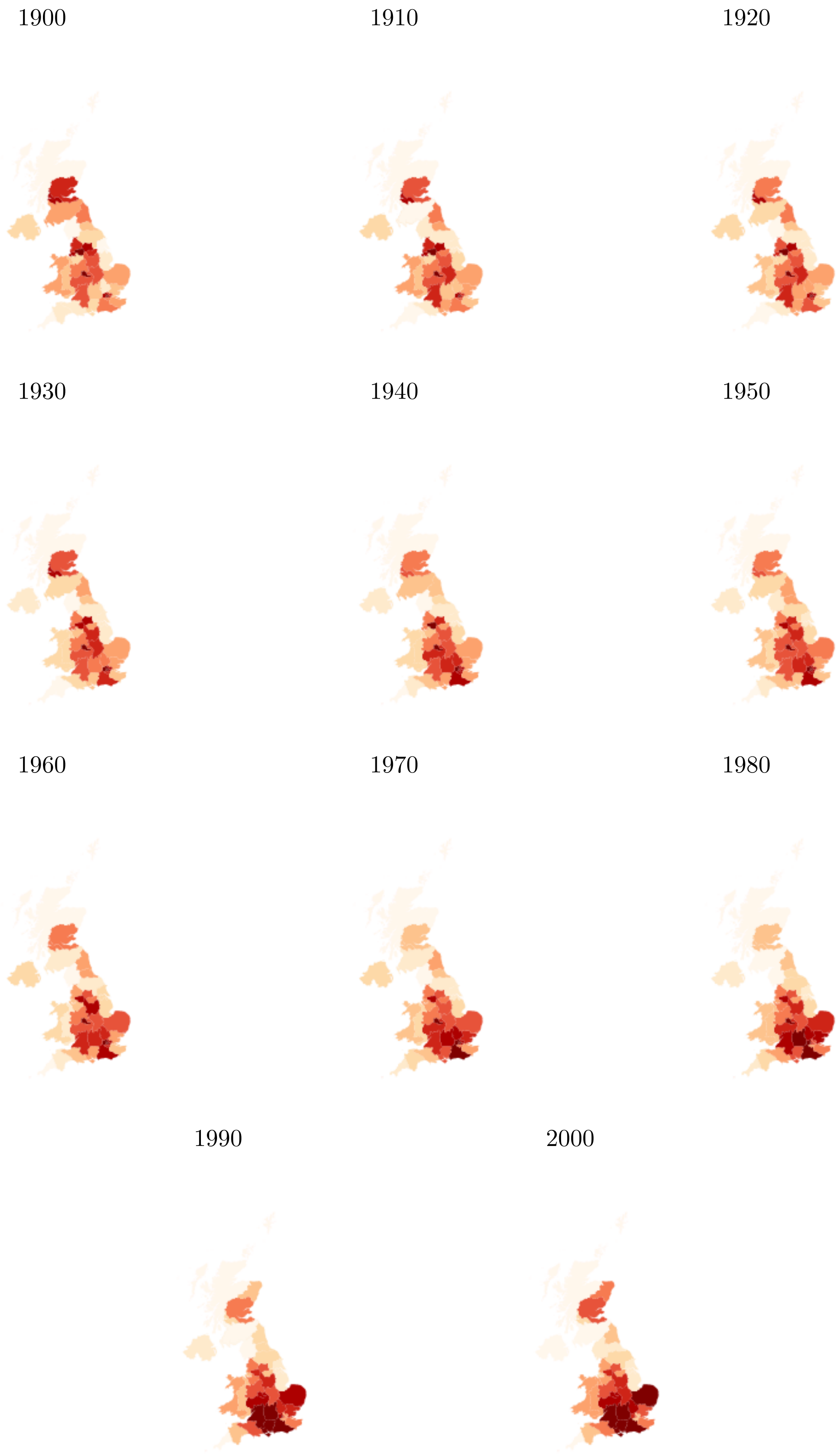
1980

1990

2000

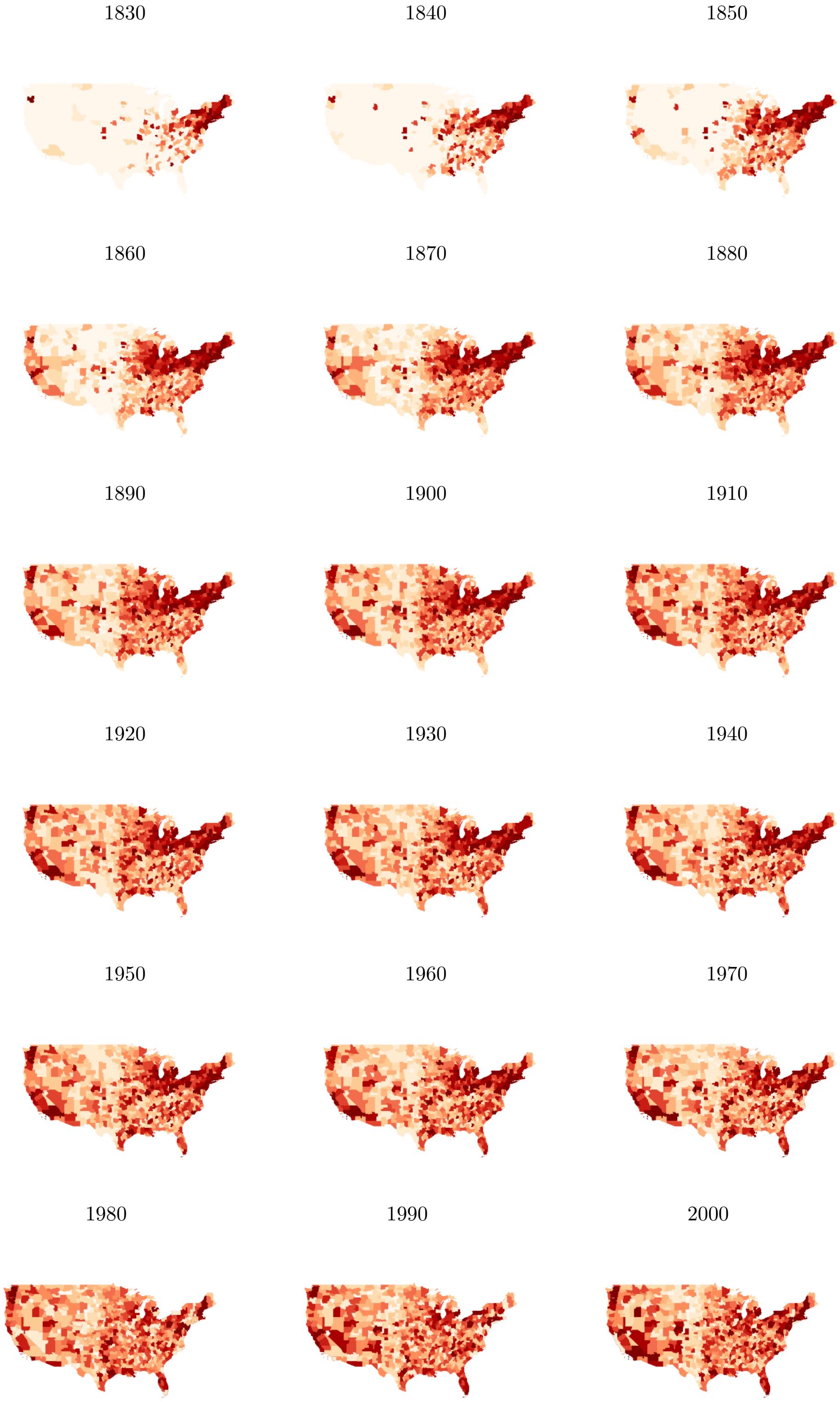
Notes: The warmer, the more patentees (inventors and assignees) are located in a given statistical area (relative to other domestic statistical areas).

Figure 2.71: Patentees location counts in GB (by level 2 statistical area)



Notes: The warmer, the more patentees (inventors and assignees) are located in a given statistical area (relative to other domestic statistical areas).

Figure 2.73: Patentees location counts in US (by level 2 statistical area)



1830

1840

1850

1860

1870

1880

1890

1900

1910

1920

1930

1940

1950

1960

1970

1980

1990

2000

Notes: The warmer, the more patentees (inventors and assignees) are located in a given statistical area (relative to other domestic statistical areas).

Table 2.14: Models' performance by format in DD

Format	Metric	ALL	ASG	INV	LOC	OCC
1	p	0.99	0.99	0.96	0.99	0.99
	r	0.99	0.99	0.96	0.99	1
	f	0.99	0.99	0.96	0.99	0.99
2	p	0.95	0.94	0.95	0.98	0.94
	r	0.94	0.87	0.97	0.95	0.94
	f	0.95	0.91	0.96	0.96	0.94

Notes: Reported performance metrics were computed on the test set - unseen during training. The "Format" column indicates the different models used for the office. For the German office, there was a major shift in the patent information display in 1881 forcing us to train two different models. Performance metrics are reported as follows: precision/recall/F1-score.

Table 2.15: Models' performance by format in DE

Format	Metric	ALL	ASG	CLAS	INV	LOC	OCC
1	p	0.99	0.98	0.99	0.99	1	0.97
	r	0.99	0.99	1	0.96	1	0.98
	f	0.99	0.98	1	0.98	1	0.97
2	p	0.99	0.99	0.99	0.98	0.99	0.97
	r	0.98	0.98	1	0.99	0.98	0.97
	f	0.98	0.98	0.99	0.99	0.98	0.97

Notes: Reported performance metrics were computed on the test set - unseen during training. The "Format" column indicates the different models used for the office. Performance metrics are reported as follows: precision/recall/F1-score.

Models' performance

Table 2.16: Models' performance by format in FR

Format	Metric	ALL	ASG	CLAS	INV	LOC
1	p	0.97	0.99	0.93	0.99	0.99
	r	0.97	0.99	0.93	1	0.99
	f	0.97	0.99	0.93	0.99	0.99
2	p	0.98	0.98	-	0.99	0.99
	r	0.98	0.98	-	0.98	0.99
	f	0.98	0.98	-	0.98	0.99

Notes: Reported performance metrics were computed on the test set - unseen during training. The "Format" column indicates the different models used for the office. Performance metrics are reported as follows: precision/recall/F1-score.

Table 2.17: Models' performance by format in GB

Format	Metric	ALL	ASG	CIT	INV	LOC	OCC
1	p	0.93	0.93	0.96	0.95	0.92	0.9
	r	0.94	0.92	0.96	0.96	0.92	0.86
	f	0.94	0.93	0.96	0.96	0.92	0.88

Notes: Reported performance metrics were computed on the test set - unseen during training. For GB, only one model is used. Performance metrics are reported as follows: precision/recall/F1-score.

Table 2.18: Models' performance by format in US

Format	Metric	ALL	ASG	CIT	INV	LOC
1	p	0.98	0.94	0.98	1	0.98
	r	0.99	0.96	0.98	0.99	0.99
	f	0.99	0.95	0.98	0.99	0.99
2	p	0.98	0.96	0.98	1	0.98
	r	0.99	0.96	0.97	1	0.99
	f	0.98	0.96	0.98	1	0.99
3	p	0.97	0.96	0.97	0.99	0.97
	r	0.97	0.96	0.97	0.98	0.98
	f	0.97	0.96	0.97	0.98	0.98
4	p	0.99	0.99	-	1	0.99
	r	0.99	0.98	-	1	0.99
	f	0.99	0.98	-	1	0.99

Notes: Reported performance metrics were computed on the test set - unseen during training. The "Format" column indicates the different models used for the office. Performance metrics are reported as follows: precision/recall/F1-score.

Chapter 3

The Rise of China's Technological Power: the Perspective from Frontier Technologies

Cyril Verluise. Collège de France & Paris School of Economics
Antonin Bergeaud. HEC Paris & Collège de France

This version: November 2021

Abstract

We use patent data to study the contribution of the US, Europe, China and Japan to frontier technology. To do so, we extend the nascent automated patent landscaping approach. We show that our method generates both accurate and robust results. We find that China's contribution to frontier technology has become *quantitatively* similar to the US in the late 2010s while overcoming the European and Japanese contributions respectively. Although China still exhibits the stigmas of a catching up economy, these stigmas are on the downside. The *quality* of frontier technology patents published at the Chinese Patent Office has leveled up to the quality of patents published at the European and Japanese patent offices. At the same time, frontier technology patenting at the Chinese Patent Office seems to have been increasingly supported by domestic patentees, suggesting the build up of domestic capabilities.

JEL classification: O30, O31, O32, O43
Keywords: Frontier technologies, China, Patent Landscaping

3.1 Introduction

Modern growth theory (Romer, 1990; Aghion and Howitt, 1992b) acknowledges the central role of technological progress in driving long-run economic growth. Yet, the nature of technological progress depends to each country's level of development (Acemoglu et al., 2006). Developing countries typically progressively catch up with frontier economies by making incremental adjustments to technologies previously developed in the latter. As they get closer to advanced economies, growth requires frontier innovation which in turn calls for institutional transformations that include competition policy (Zilibotti, 2017), research education (Vandenbussche et al., 2006; Aghion et al., 2009, 2010), external finance (Diallo and Koch, 2018), and improved management practices (Bloom and Van Reenen, 2007)... Failure to implement a favorable set of institutions to support frontier innovation has shown to be an obstacle to full economic convergence and to maintain developing countries in a "*middle-income trap*". A country's capacity to generate frontier innovation is key for developing countries to join the developed countries club and for developed countries to remain in this club.

Despite the central piece that frontier innovation takes in growth theory, the empirical characterisation of the diffusion of novel technologies remain an empirical challenge. Economists usually rely on patent data to study innovation, but patents do not come with immediate ways to distinguish between catch up innovation – or improvements of well-established technologies – and the development of new frontier technologies, all the more when comparing different countries. In particular, the now well-known and documented dramatic increase in patenting in China (in 2019, the Chinese Patent Office filed 1.4 million patents, that is 43% percent of the world's total applications) is hard to relate to the actual contribution of China to pushing the world's frontier. By deploying a set of new statistical methods to the patent corpus, our work is an attempt to address this question.

From a methodological point of view, researchers might be tempted to use standard technological classes attributed by patent offices to patents in order to delineate patents contributing to a given technology and study frontier innovation. However, the economic literature has long prevented scholars from doing so. Griliches (1990) famously called this question "the patent classification problem". Indeed, the purpose of patent offices' classifications is to ease the search of prior art. Resulting classifications are thus based on techniques, which are not necessarily related to the economists' notion of technology. Schmookler (1966) reports that a subclass related to the dispensing of solids contained patents on both manure spreaders and toothpaste tubes.

In this context, this paper introduces a new methodological approach. We extend the automated patent landscaping literature to accurately and consistently delineate a large set of frontier technologies from the worldwide corpus of patents. Specifically, we extend the algorithm pioneered by Abood and Feltenberger (2018). This algorithm implements a sequence of machine learning operations to emulate human curation at scale. We add a tractable amount of human supervision in the loop in order to improve both accuracy and consistency of our results. We call this approach "automated patent landscaping with humans in the loop". We then apply it to six novel and representative technologies: additive manufacturing, blockchain, computer vision, genome editing, hydrogen storage and self-driving vehicles. These technologies were carefully selected to both cover a large variety of economic sectors and ensure conceptual homogeneity.

In line with our initial question, our approach allows us to study specifically the role and contribution of any country with a patent system in the production of frontier innovation. For the sake of simplicity, in this paper we restrict our attention to three regions that we refer to as "the Triad" (United States, Europe and Japan) and to China which are at the epicentre of frontier technology patenting. All these technologies, although very different in nature, deliver a surprisingly consistent and clear picture. The quantitative contribution of the Chinese Patent office to frontier innovation patenting has been rapidly rising since the early 2000s. China has become the second largest actor of frontier innovation and is quickly catching up with the US. Our results suggest that, although it still exhibits stigmas of a catching up economy, these stigmas are on the downside. The quality of frontier technology patents published at the Chinese Patent Office has been increasing since the 2000s. By the end of the 2010s, it had reached the quality level of patents published at the European and Japanese patent offices. During the same period, frontier technology patenting in China seems to have been increasingly supported by domestic patentees, suggesting the build up of domestic capabilities.

Literature review

Our paper contributes to different strands of the literature.

First, there are a number of papers that measure the rise of China as a new technological power and explore potential explanations. This literature typically reports that China is an important player, if not the leader, in terms of many indicators of innovativeness. China concentrated 159 unicorn companies in 2021 according to CB Insight for a total valuation of more than 500 billion dollars.[1] This is much more than Europe (89 for a valuation of 314 billion dollars, including respectively 32 and 138 billion dollars for the UK alone) and Japan (5 for a valuation of $6.8 billion dollars) but still far behind the US (405 for a valuation of $1,353 billion dollars). In terms of scientific effort, the importance of China has been rising for the past 20 years as measured by the number of top cited articles, which places the country as second scientific powerhouse behind the US.[2]

This catch up in terms of scientific publications is even more dramatic when it comes to Artificial Intelligence research (Baruffaldi et al., 2020). This tends to support the view that China has built the capacity to innovate in the technology of tomorrow, making its catch up more likely to be sustainable, and avoid the middle-income trap (Fan, 2014). This is partly thanks to trade and foreign direct investments (Aghion et al., 2019b; Hu and Jefferson, 2009), but also to subsidies and reforms of property rights (Dang and Motohashi, 2015).[3] However, other strands of the literature argue that China still suffers from stigmas that penalize its capacity to innovate.

[1] See CB Insights for a list.

[2] See the OECD Science, Technology and Industry Scoreboard 2017.

[3] These state subsidies and incentives to file patent applications have led experts to cast some doubt on the relevance and quality of the average Chinese patent (see e.g. (He, 2021)). In the empirical analysis, we take this possibility into account.

Abrami et al. (2014) explain that while China does not lack the number of entrepreneurs, inventors or scientists, its institutions are not well-suited to encourage the development of frontier technologies. For example, every company larger than 50 employees is required to have a Chinese Communist Party (CCP) representative and a party liaison. In 2020, Xi Jinping, the general secretary of the CCP openly opposed the IPO of Ant Group, a large innovative financial company. Aghion et al. (2008) and Aghion et al. (2021b) show that academic freedom and democracy are two critical requirements for the production of original research. Aghion et al. (2021b) thus question the capacity of China to compensate for their lack of freedom with mass investment in R&D.[4]

We contribute to this literature by looking as objectively as possible at the relative importance of China in the development and diffusion of recent frontier technologies that were chosen and identified without any preconceptions.

We also speak to a recent literature that exploits the patent corpus to study the diffusion of frontier technologies. These technologies are typically characterized by their radicalness, novelty, pervasiveness and their capacity to diffuse quickly and to have large impacts but are also highly uncertain and risky (Rotolo et al., 2015). Webb et al. (2018) look at the evolution in the number of patents filed in the US for a number of modern technologies such as Artificial Intelligence, Machine Learning, Semiconductor, Drones. . . They focus on the 1970-2015 period and find that most of these technologies have experienced a boom in the number of patents and inventors in the past decades mostly driven by US and Japanese multinationals. They also report a modest but growing contribution of Chinese inventors and firms to the rise of high tech patenting in the US. In a subsequent work, Bloom et al. (2021) also used patent data to study the diffusion of 29 disruptive technologies and their adoption by firms and labor markets in the US. Their findings suggest that there are long term impacts on the areas that hosted the initial development of these frontier technologies. These two studies focus on the US and attempt to have an overall view on the role and impact of high tech patenting. In contrast, Bessen and Hunt (2007) use patent data to analyse specifically the rise in software patenting in the US and compare the role of increased R&D spending and changes in IP legislation to explain this phenomenon. Other studies typically conducted by patent offices apply a combination of different methods to look at the development of patenting in a specific technology and a specific region.[5] For example IP Australia (2019) has analyzed the significant increase in patenting related to Machine Learning. We contribute to this literature by considering six technologies that cover various subjects and consider patents from the four main global technology contributors. This allows us to compare countries over time since the birth of these technologies.

Finally, we also contribute to a methodological literature which aims at delimiting technologies using patents. Historically, Trajtenberg (1990b) tackled the classification problem by manually curating US patents belonging to the Computed Tomography Scanners technology. This method delivers precise results but is of course too labor-intensive to be extended to a larger corpus

[4] Song et al. (2011); König et al. (2020) use a structural estimation of a dynamic heterogeneous firm model and report that R&D investment in China seems to be less productive than in other countries (namely Taiwan).

[5] For a list of such study, see WIPO (2021).

and multiple technologies. Other studies have used a number of rules combining keywords and Cooperative Patent Classification (CPC) classes to define a technology and constitute groups. This is the methodology applied for by Webb et al. (2018)[6] and by the patent landscaping literature. For example, the European Patent Office (EPO) has published a report on patenting in the field of automated vehicles (EPO, 2018). The rules used in these analysis are typically *ad hoc* and require a high level of expertise. Recently, Abood and Feltenberger (2018) introduced a new methodology that aims at circumventing this difficulty. Their approach, which we present in more details in Section 3.3.2, allows to emulate human-made technology classification using only a small number of representative patents as an input. Related approaches have leveraged Natural Language Processing and clustering algorithms to construct groups of patents (see Bergeaud et al., 2017 for a review). For example, the Fung Institute proposes an application of automatic labeling using machine learning to automated vehicles.[7] Similarly Giczy et al. (2021) have applied a slightly modified version of the Abood and Feltenberger (2018)'s algorithm to identify patents related to AI. These methodological works however do not attempt to measure and compare the diffusion of technologies across countries and time. We build on their method and adapt both the selection process of the imputed set of patents and the way the algorithm expands from this initial seed. Ultimately, our methodology combines a small amount of human work and automated landscaping to select patents related to a given technology with a high degree of precision and with no limitation in the geographical coverage.

The remaining of this paper is organized as follows: section 3.2 details our technology definition and selection procedure; Section 3.3 presents the automated patent landscaping approach and how we extend it; Section 3.4 carefully evaluates the internal and external validity of the results generated by our algorithm on each of the six technologies; section 3.5 documents the rise of China's technological power; Section 3.6 concludes.

3.2 Technology definition and selection

The interpretation of the results we present in this paper are determined by two fundamental questions. First, what do we mean by "technology"? Second, how did we select our set of frontier technologies? We address these two key preliminary questions in this section.

3.2.1 Definition

Technology is a widely used term and can refer to many different concepts. In the economic and innovation literature, we classified its main usages into three categories which we call "technique", "functional application" and "application field". A *technique* is a set of processes sharing a common methodological paradigm. Two distinct techniques can share a common goal. For example, TALENs, Zinc Fingers and CRISPR are all distinct techniques pursuing the same goal of editing the genome. A *functional application* is a high level goal which is directly

[6] See Section 3.2.2 for more details on the selection process.
[7] See the webpage of the Fung Institute Capstone Project.

targeted by one or several techniques in the course of their developments. Examples include computer vision and genome editing. The range of their market applications can vary and usually exceed a single market. Eventually, an *application field* is an existing or newly created economic sector which can leverage functional application to develop new or improve existing products. Examples of application fields include smartphones, nuclear power generation, etc..

In this paper, we work at the *functional application* level. This comes as a natural choice since we are interested in frontier innovation which has the potential to give advanced economies a significant growth momentum. Hence, our focus is on technologies which, like General Purpose Technologies, have the ability to infuse progress in a large range of applications.

3.2.2 Selection

There are two main ways to define a set of technologies of interest: the supervised and unsupervised approaches. The most common approach, the “supervised”, is based on human curation of technology-related documents. This is typically the approach followed by Webb et al. (2018) who define a list of technologies in the high-tech segment from prior knowledge. The second and more recent approach, the “unsupervised”, combines text mining (specifically “topic modelling” techniques) and technology-related corpuses to identify technologies (e.g. topics) without any use of prior knowledge. Such a method is implemented by Bloom et al. (2021) who use earnings conference call transcripts to uncover technologies which are the most frequently cited for their contribution to companies’ momentum.

Although extremely appealing, the unsupervised approach presents two limitations in our context. First, and most importantly, relying on past financial and corporate documents will invariably miss frontier technologies with still nascent market applications. Second, existing topic modeling techniques cannot guarantee that the identified “topics” (here technologies) are conceptually homogeneous. Without any supervision, selected technologies might (and will) include techniques, functional applications and application fields indifferently. Note that this can be partially addressed by adding a manual curation on top of the topic modeling results. Hence, in our specific setting, this however attractive approach appears to be inappropriate.

Although we opted for the supervised approach, we designed a methodology to minimize our own biases and discipline the selection process. In particular, we took great care not to focus on technologies with large media coverage in Western countries which could affect our overall picture. First, we curated a large number of reports and articles published at different time and dedicated to breakthrough technologies. These articles have various sources: international institutions (OECD, 1998, 2016, EPO, 2020) national agencies (Tarasova and Shparova, 2021, Kennedy, 2015), industry associations (BDI, 2011), experts (Review, 2021) and consulting companies (McKinsey, 2021; Deloitte, 2021). We took care to include sources from both developed and developing countries. From those documents, we listed without any *a priori* more than 30 technologies in a broad sense. Then we classified these items into the three aforementioned categories (technique, functional application and application field) and kept only those entering the “functional application” category. Eventually, we reviewed the remaining candidates (goals, recent breakthroughs, expected economic impact, and development stage) with two main objectives in mind: 1) keep only technologies which have already started to have some market applications or are expected to have some in the near future and 2) cover a large number of distinct application fields. From our initial list of technologies, we ended up with six frontier technologies: additive manufacturing, blockchain, computer vision, genome editing, hydrogen storage and self-driving vehicles.

3.3 Automated patent landscaping with humans in the loop

In this section, we introduce the automated patent landscaping approach, how it relates with existing approaches in economics, and what its limitations are. Then, we show how we overcome these limitations.

3.3.1 The traditional approach

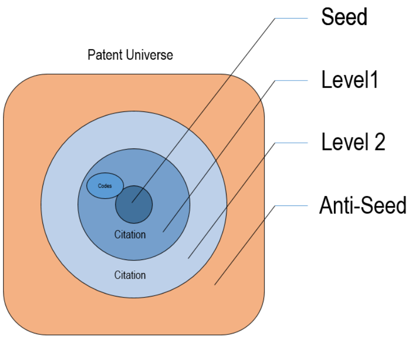
Delineating a technology in the corpus of patents is a long-standing issue. Various approaches have been experimented. The three main instruments that have been leveraged are technological classes, citations and keywords. Although all of these instruments carry some valuable information, they are also affected by a significant degree of noise. In this section, we provide qualitative intuitions on these limitations. Section 3.4 will further quantify them. Technological classes are based on technical principles which are only partially related to the concept of technology we are looking for (functional application). Citations between patents have clear limitations in this case as well. Patent-to-patent citations are generated in order to define the limits of the technological monopoly granted to the patentees. Proximity in the sense of functional application is then just one of the many reasons to generate a citation. Eventually, keywords can help identify patents dealing with a technology. However, language is highly variational. There are many ways to say one thing and a given word can mean many different things. Hence, one can expect neither comprehensiveness nor accuracy from keywords alone. In this context, following Trajtenberg (1990b), manual curation of patents might appear as the only accurate way to delineate a technology from the corpus of patents.[8]

3.3.2 Automated patent landscaping

That is where the *automated* patent landscaping introduced by Abood and Feltenberger (2018) makes an important breakthrough. The authors develop a *semi-supervised* machine learning framework to emulate human-made technology classification. The algorithm only requires a small set of patents as input – the *seed* – which must be representative of the technology of interest. The algorithm then *expands* to “probably related” patents using both technological classes and citations (forward and backward). Specifically, it first expands to technological

[8] Trajtenberg (1990b) manually curated “computed tomography scanners” patents granted in the US to measure the value of citations.

Figure 3.1: Automated patent landscaping: expansion



Notes: From Abood and Feltenberger (2018). Starting from a `Seed`, we first expand to patents belonging to "important" CPC sub-classes (`Codes`) and then proceed to two consecutive levels of citations expansions (`Level 1` and `Level 2`). Citations expansions include both backward and forward patent citations. The `Anti-seed` is randomly drawn from the complementary set of the expansion. We eventually train a `classifier` to distinguish between the `Seed` and the `Anti-seed` and apply it to the expansion set to get rid of "false positives".

classes which are overrepresented in the seed and then it expands twice on citations. Importantly, at this stage, we know that the expansion set includes patents unrelated to the target technology - they are called "false positives". The false positives are *pruned* out using a classification model using the patent abstract (*inter alia*) as input[9], and applied to the expansion set. Specifically, the classification model is trained to distinguish between seed patents and a set of patents randomly drawn from the complementary set of the expansion (so-called *anti-seed*) and therefore "probably unrelated" to the target technology. This approach ultimately returns a group of patents in the target technology at virtually no cost, except for the curation of the seed patents. Importantly, no human intervention is needed to elaborate the set of rules determining whether a patent belongs or not to the target technology. Semantic patterns are learned from the data.

This approach is already highly promising but still exhibits some key limitations. First, the pruning model is trained on "polar" cases while it is to be applied to "intermediary" cases. The seed patents (positive examples) are selected to be at the "core" of the target technology. On the contrary, anti-seed patents (negative examples) are chosen from the complementary of the expansion set, hence far away from the target technology. Even if the algorithm performs well on the validation set[10], there is no guarantee that it will perform well when applied to patents in the expansion set. The expansion set is indeed likely to include a large share of "intermediary" examples, which are neither at the core of the target technology, nor very far from it. Training the model using a vast majority of polar examples has a potential for harming the overall validity of the classification model and algorithm results. Second, the algorithm does not really account for data variation, that is, the impact of variations in the seed on the algorithm outcome. Algorithm robustness is, however, a critical point to assess the degree of

[9] Abood and Feltenberger (2018) use citations and technological classes (CPC) in addition to patent abstracts.

[10] The validation set is typically a 20-50% random split of the learning set (here, the seed and anti-seed patents) which is not used for training the model.

confidence we can place in our results and the overall interpretation.

3.3.3 Extension

Our extended approach addresses these two limitations. First, we *augment* the anti-seed with "harder" examples. These harder examples naturally arise from the human labelling of the seed patents that we performed for each technology. We started by inspecting existing attempts to landscape our technologies of interest using traditional methods. Using this literature and their reported selection rules (usually based on technological classes and/or keywords) we generated a set of candidate patents to be included in the seed for each technology. Table 3.6 to 3.11 (in Appendix) details these rules for each technology. Next, we manually and carefully labelled a random sample of candidates (see annotation guidelines in Appendix, Table 3.12). Importantly, rejected examples provide "hard examples". Although they matched one or more rules used by previous attempts to landscape the technology, human annotators decided to exclude them based on their abstracts. These are typically the "intermediary" examples we want our classification model to learn from. We call this set of examples the *augmented anti-seed.* The model is ultimately trained using both the anti-seed *à la* Abood and Feltenberger (2018) and the augmented anti-seed to constitute the negative examples.

Second, we addressed the data variation question by implementing a series of robustness tests based on random variations in the seed. Specifically, we investigated how variations in the seed affect the expansion and the pruning outcomes. To test the expansion robustness, we drew random subsets from the seed, ran the expansion using each of these subsets and compared the generated expansion sets. Next, we assessed the pruning robustness by iterating over various random train-test splits of the annotated data. Various models were trained on varying sets of training data for each technology. Pruning robustness was ultimately evaluated by looking at models' agreement on a sample of out-of-training patents. Detailed results are reported in Section 3.4.

3.4 Algorithm deployment and validation

In this section we go through the main steps of the actual deployment of the algorithm. Next, we show that our results, in addition to being accurate and consistent, also exhibit patterns in line with technology experts' expectations.

3.4.1 Algorithm deployment

To begin with, it is important to note that contrary to Abood and Feltenberger (2018), we deploy the algorithm at patent family level rather than at patent publication level. A patent family is a collection of patent documents that are considered to cover a single invention. Their technical contents are identical. Hence, considering only one document per family does not imply any loss of information while significantly reducing the total number of items considered.[11] This seemingly minor twist has two important practical advantages. First, it enables us to consider all families with at least one publication having a known abstract. That way, we ultimately cover more than 86% of all publications since 1970, while only 76% of patent publications do have a non-null abstract in our database. Detailed coverage is reported in Figure 3.8 in Appendix. Second, it minimizes the amount of texts to be classified at the pruning stage. Each family is processed only once, even if it includes more than one patent. This improves the overall computational tractability of the algorithm.

Next, we delve into the algorithm deployment itself. As already discussed in Section 3.2, our work starts one step before the algorithm described by Abood and Feltenberger (2018). This first step consists in the definition of rules to identify a set of candidates. These candidates are picked out of patents which match at least one of the rules that we were able to find in the specialized literature. These rules include technological classes, keywords and patent similarity[12]. A random set of candidates are then labeled by humans based on the abstract and detailed annotation guidelines (see Table 3.12 in Appendix). Annotation guidelines guarantee both transparency and reproducibility. In practice, we labeled candidates until we "accepted" at least 300 candidates constituting the technology *seed*. Importantly, rule-based candidates systematically included a large proportion of false positives, which were rejected. This set of rejects constituted the *augmented anti-seed*.

Starting from the seed, the following step is the expansion. Regarding this step, we mainly stick to the Abood and Feltenberger (2018) procedure. We first expand to technological classes that were over-represented in the seed and then expand twice using citations (backward and forward). Note however that we had to adapt at the margin to take into account our choice to work at family level rather than publication level. We had to express citations in terms of the patent family rather than the usual publication format. For each family, we considered all citations received (forward) and sent (backward) by any patent in the family.

Eventually, our pruning stage also differs from Abood and Feltenberger (2018) along 3 dimensions. First comes the composition of the training data. As already discussed, we add an augmented anti-seed to the seed and anti-seed described in their paper. Second, while our predecessors used not only text but also citations and technological classes as input to the classification model, we only restricted to text. In our view, both technological classes and citations imply potential pitfalls at this stage. Using technological classes in both the expansion and the classification model can generate pathological cases. Assuming that all technological classes in the seed are found important, then the anti-seed and the seed have no technological class in common which makes the classification task trivial. Regarding citations, by construction, patents in the second level of the citation expansion (L2) have no citations in common with the seed. Hence, considering citations in the classification task implies a systematic and un-

[11] There are around 120 million patent publications in the CLAIMS dataset versus 70 million patent families.

[12] The specialized literature sometimes specifically reports key patents for a technology. We used the most similar patents as defined in the Google Patents database to include the most similar patents to these key patents in our dataset.

controlled bias against patents in the part of the expansion which we find undesirable. Third comes the model itself. We implement 3 different neural network architectures popular for text classification tasks: the multi-layer perceptron (MLP), the convolutional neural network (CNN) and a transformer, specifically a pre-trained Bert encoder. We provide an overview of these architectures in the following sub-section. The actual pruning is performed using the Transformer model which exhibits both the highest performance and consistency.

3.4.2 Performance and consistency

The most simple architecture we consider is the multi-layer perceptron (MLP). This architecture can be seen as a stack of logistic regressions and treats tokens or groups of tokens independently. Although it can be successful at identifying keyphrases, it is unable to handle context and might eventually be seen as a sophisticated keyphrase matcher. Then, we implemented a Convolutional Neural Network (CNN) model. This architecture leverages the sequential nature of text through the use of feature maps (masks). These feature maps are there to detect sequences of tokens with a common and discriminant "meaning". CNN performances usually dominate those of MLP models thanks to this enriched understanding of language. However, they lack "memory" and cannot handle long context as feature maps typically focus on 3 to 5 token long spans of text. Eventually, the Transformer architecture which was recently introduced has reached state-of-the-art results in many natural language processing (NLP) tasks, including text classification. Transformers rely on a core mechanism called *attention* which enables them to "understand" tokens in the context of neighbouring tokens. Transformers are very large models trained at masked language completion on very large text corpora and eventually fine-tuned on specific tasks (e.g. text classification). This pre-training allows downstream users to start from a model already embodying a large "knowledge" of language. A limited number of examples is then enough to adjust weights and achieve high performances on specific tasks. This is especially well-suited when annotating examples is costly. The main drawback of using Transformers is their high computational costs.[13]

We then train all these models. The task is a standard binary text classification task. Specifically, we train and evaluate each model on ten distinct train-test sets for each technology. We implement this approach as a cross-validation method to have an estimate of the impact of random variations of the training set on both the performance of the model and its out of training sample predictions - later called *consistency.* We first focus on performance and will come back to consistency later. We report the median *precision*, *recall* and *F1-score* for each technology and model architecture in Table 3.1. These metrics were all computed on the test set, that is, on examples unseen by the model at train-time. The precision is the share of texts predicted to be in the seed and which are indeed part of it. The recall is the share of actual seed texts which were indeed predicted to be part of it. The F1-score is the arithmetic mean of the precision and recall. We observe that MLP and CNN architectures tend to exhibit similar

[13] Transformers are almost intractable using traditional Central Processing Unit (CPU) and require Graphics Processing Unit (GPU).

F1-score. However, MLP models have higher precision and lower recall than CNN. This sends us back to the fundamental nature of MLP. As stated earlier, MLP can be seen as a sophisticated keyphrase matcher which usually has high precision but lower recall. In any case, the transformer outperforms both of the models and achieves around 90% of median F1-score for all technologies except for self-driving vehicles (79%).[14]

Table 3.1: Models performance

	MLP			**CNN**			**TRF**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Additive manufacturing	**0.89**	0.79	0.84	0.79	0.85	0.81	0.86	**0.92**	**0.89**
Blockchain	0.90	0.81	0.86	0.83	0.88	0.86	**0.97**	**0.98**	**0.97**
Computer vision	**0.89**	0.81	0.85	0.86	0.87	0.87	0.87	**0.95**	**0.90**
Genome editing	**0.89**	0.87	0.88	0.87	0.91	0.88	0.86	**0.94**	**0.89**
Hydrogen storage	0.86	0.73	0.80	0.76	0.83	0.78	**0.92**	**0.98**	**0.93**
Self-driving vehicle	**0.79**	0.65	0.71	0.69	0.73	0.71	0.75	**0.85**	**0.79**

Notes: Reported performance metrics were computed on the test set - unseen during training. Performance metrics are reported as follows: P for precision, R for recall and F1 for F1-score.

Interestingly, our candidate annotation exercise gives us the opportunity to compare those results with the performance that would have been obtained based on the rules used by existing attempts to landscape our six technologies of interest. Specifically, it enables us to obtain performance metrics for rule-based approaches using technological class, keywords and patent similarity. It clearly appears that our approach to delineate technologies from the corpus of patents not only achieves good performance but also outperforms traditional rule-based methods. Although our approach does not enable us to compute all the performance metrics reported before, we can compute the precision of rule-based approaches. We find that rule based candidate selection delivers both low and variable precision performances across technologies. Specifically, precision from CPC-class rule-based patent selection ranges from 0.01 (blockchain) to 0.34 (additive manufacturing). Precision from keyword rule-based selection goes from 0.09 (blockchain) to 0.89 (genome editing) for an average of 0.32. Precision from patent similarity ranges from 0.02 (additive manufacturing) to 0.57 (genome editing).

As already discussed, although performance *per se* matters, it is also crucial to understand how variations in the seed data can affect the results of the algorithm. We identify two channels. First, data variations can affect the expansion. The latter depends on the seed and has a critical role. It determines the set of documents which will be considered by the pruning model. Second, data variations can affect the pruning itself. The pruning model depends on the seed, the anti-seed and the augmented anti-seed and ultimately determines which documents in the expansion are to enter the technology or not. Robustness to random variations in the data is then crucial to make sure that algorithm results can be exploited rigorously. To investigate the consistency of the expansion, we generated random subsets of the seed. Specifically, we considered 3 different sizes: 90%, 70% and 50% of the initial seed and drew 10 subsets for each

[14] This technology is actually hard to classify even for humans. The very same technology can be used to automate driving or to assist human driving. In the former case, we would accept a patent while in the latter it would be rejected.

size. We then proceeded to the full expansion starting from these distinct seeds and computed the pairwise family overlap of the generated expansion sets for each technology and seed size. Detailed results are reported in Table 3.2. We find that the average pairwise family overlap exceeds 89% in all cases. This remarkably high figure indicates a high level of consistency for the expansion step.

Table 3.2: Median pairwise expansions overlap

	90%	**70%**	**50%**
Additive manufacturing	0.99	0.93	0.89
Blockchain	0.99	0.98	0.96
Computer vision	0.99	0.96	0.92
Genome editing	0.99	0.99	0.98
Hydrogen storage	0.99	0.97	0.95
Self-driving vehicle	0.99	0.97	0.95

Notes: For each size (90%, 70% and 50%), we drew 10 random subsets of the seed and proceeded to an expansion. For each pair, we computed the share of families in the two expansions. We report the median share of overlapping families across all expansion pairs.

Next, we looked at how the pruning stage is affected by variations in the training data. As discussed above, we trained the same architectures on 10 different train-test splits (of respective size 80%-20%) for each technology as a way to emulate natural variations in the data. We then apply these models to a set of 10,000 out-of-training-sample documents randomly drawn from the expansion. For each technology, we then look at the standard deviation of the ten scores (each score ranging between 0 and 1) for each document and report its median in Table 3.3. We find that the standard deviation of the predicted scores is usually very low, most of the time below 0.05. This result supports the consistency of the pruning step.

Table 3.3: Models robustness (Median dispersion in predicted scores)

	MLP	**CNN**	**TRF**
Additive manufacturing	0.029	0.082	0.017
Blockchain	0.008	0.047	0.003
Computer vision	0.015	0.029	0.01
Genome editing	0.003	0.001	0.004
Hydrogen storage	0.015	0.037	0.005
Self-driving vehicle	0.039	0.091	0.011

Notes: For each model architecture, we trained 10 models using distinct random subsets (80%) of the training set. Each model was then applied to a set of 10,000 texts (out of training set). We report the median standard deviation (at the sample level) of the predicted scores across models.

In a nutshell, we find that our approach provides both accurate and consistent results. Next, we show that these results also achieve external validity.

3.4.3 External validation

Going further, we use the output of the algorithm to investigate whether our results make sense. To do so, we compute the top assignees and top inventors as reflected by the total number of patents they hold.[15] We do it for each sudied technology. We then confront these results with prior insights from technology-specialised literature as well as background checks.[16] These lists of top assignees and inventors not only reassure us about the validity of our approach, but also provide insights about the main actors of the different technologies considered.

Top 10 assignees by technology

Table 3.4 reports the top 10 assignees for each technology by the number of patents they were granted worldwide.

Our first observation is that the most famous players in each technology are present. For the sake of brevity, we focus on some remarkable high-ranked agents for each technology and explain why they were indeed expected. Starting with additive manufacturing, Xerox and Hewlett-Packard are two large companies that traditionally developed printers and which naturally developed 3D printing technologies. In the field of blockchain, Alibaba, Intel, nChain and IBM are also in the top list of assignees in the expert-based landscaping of blockchain innovation proposed by Clarke et al. (2020). The most prolific assignees in the field of Computer vision include firms that build and sell electronic devices, including cameras. Genome editing is well known to be pioneered by top Universities like University of California Berkeley, Harvard University and University of Pennsylvania as well as large companies that develop chemistry and pharmaceutical products like Regeneron and Dupont. These findings are consistent with results from an overview of patenting in the genome editing technology field proposed by Benahmed-Miniuk et al. (2017). The field of hydrogen storage technologies is mostly dominated by car manufacturers. This naturally comes from the fact that the main usage of this technology is to propel vehicles using hydrogen. Finally, the field of self-driving cars also includes many traditional car manufacturers, including Toyota and Ford that communicate intensively on their progress in the development of autonomous vehicles. The list of top assignees also includes automotive equipment suppliers such as Bosch and Denso Corp.[17]

On top of very large firms that spread over a large number of different technologies such as IBM and Samsung, we also note the presence of a number of firms that are much more specialized in a specific field. This is notably the case of Air Liquide for hydrogen storage, nChain for blockchain, ASML for additive manufacturing, Regeneron pharma for genome editing and Denso Corp for self-driving cars.

[15] We used the harmonized name of assignees and inventors from the CLAIMS dataset. This harmonization does not always guarantee that two different names of the same entities are actually merged in the same entity (e.g. Toyota Motor Co Ltd and Toyota Motor Corps).

[16] Note that, while the landscaping is done at the family level, analytical results are at the patent publication level.

[17] Toyota, Ford and Bosch are mentioned as the top assignees in the field by WIPO (2019, Chapter 3).

Table 3.4: Top 10 assignees

	Additive manufacturing	**Blockchain**	**Computer vision**	**Genome editing**	**Hydrogen storage**	**Self driving vehicle**
1	Samsung Electronics Co Ltd	Alibaba Group Holding Ltd	Canon KK	Univ California	Toyota Motor Co Ltd	Toyota Motor Co Ltd
2	Hewlett Packard Development Co	IBM	Sony Corp	Pioneer Hi Bred Int	Honda Motor Co Ltd	Bosch Gmbh Robert
3	Xerox Corp	Qualcomm Inc	Samsung Electronics Co Ltd	Du Pont	Nissan Motor	Honda Motor Co Ltd
4	Asml Netherlands BV	Samsung Electronics Co Ltd	Koninkl Philips Electronics NV	Regeneron Pharma	Toyota Motor Corp	Nissan Motor
5	Gen Electric	LG Electronics Inc	Matsushita Electric Ind Co Ltd	Genentech Inc	Matsushita Electric Ind Co Ltd	Ford Global Tech LLC
6	Eastman Kodak Co	Sony Corp	Sharp KK	Monsanto Technology LLC	Sanyo Electric Co	Denso Corp
7	Canon KK	NChain Holdings Ltd	Seiko Epson Corp	Harvard College	Hyundai Motor Co Ltd	Toyota Motor Corp
8	Fujifilm Corp	Huawei Tech Co Ltd	Lg Electronics Inc	Hoffmann La Roche	Air Liquide	Hyundai Motor Co Ltd
9	Siemens AG	Intel Corp	Qualcomm Inc	Univ Pennsylvania	Panasonic Corp	Mitsubishi Electric Corp
10	IBM	Ericsson Telefon Ab L M	IBM	Centre Nat Rech Scient	GM Global Tech Operations Inc	Bayerische Motoren Werke AG

Notes: Assignees are ranked based on the total number of patents for each technology over the whole corpus of patents. The harmonization of assignees' names is taken from the CLAIMS dataset.

Top 10 inventors

Table 3.5 reports the top 10 inventors for each technology by the number of patents they were granted worldwide.

Regarding the top inventors, background checks helped us validate the consistency of our results. As previously, for the sake of brevity we focus on the most emblematic and high-ranked inventors. We note the presence of M. Karczewicz in Blockchain and Computer Vision. She is a prolific inventor working at Qualcomm Technologies, Inc.. Marta. Karczewicz is famous for having developed many technologies related to data compression which facilitates the transfer of important mass of information. The methods she developed are very central for many computer-related technologies such as computer vision and blockchain. As a recognition for her contributions, the EPO named her one of the three finalists for the award of European inventor

Table 3.5: Top 10 inventors

	Additive manufacturing	**Blockchain**	**Computer vision**	**Genome editing**	**Hydrogen storage**	**Self driving vehicle**
1	Silverbrook Kia	Karczewicz Marta	Karczewicz Marta	Murphy Andrew J.	Ovshinsky Stanford R.	Tabata Atsushi
2	Lapstun Paul	Zhang Li	Zhang Li	Macdonald Lynn	Ukai Kunihiro	Shimizu Yasuo
3	Ng Hou T.	Zhang Kai	Nishi Takahiro	Mcswiggen James	Edlund David J.	Nordbruch Stefan
4	Vermeersch Joan	Wright Craig Steven	Kondo Tetsujiro	Zhang Feng	Fetcenko Michael A.	Hayakawa Yasuhisa
5	Van Damme Marc	Qiu Honglin	Wang Ye-kui	Rosen Craig A.	Taguchi Kiyoshi	Lynam Niall R.
6	Lewis Thomas E.	Yang Xinying	Chen Ying	Stevens Sean	Wakita Hidenobu	Watanabe Kazuya
7	Zhao Lihua	Wang Yue	Chen Jianle	Ruben Steven M	Maenishi Akira	Yasui Yoshiyuki
8	Patibandla Nag B.	Liu Hongbin	Yamazaki Shunpei	Wilson James M.	Young Kwo	Liu Jun
9	Ganapathiappan Sivapackia	Wang Zongyou	Kadono Shinya	Ni Jian	Nishio Koji	Breed David S.
10	Ye Jun	Fukushima Shigeru	Sugio Toshiyasu	Gurer Cagan	Reichman Benjamin	Matsuno Koji

Notes: Inventors are ranked based on the total number of patents for each technology over the whole corpus of patents. The harmonization of inventors' names is taken from the CLAIMS dataset.

of the year 2019.[18] Considering additive manufacturing, the most prolific inventor in the field is Kia Silverbrook. He is also a famous inventor who holds more than 9,000 patents worldwide.[19] K. Silverbrook founded Silverbrook Research, a company that developed digital printing and 3D printing technologies, among other inventions. In the field of genome editing, our top inventor is Andrew Murphy. He is the vice president in charge of research of Regeneron, a biotechnology company that develops different drugs and recently made important progress in new therapies using CRISPR (Gillmore et al., 2021). We also note the presence of Feng Zhang, a Professor at MIT and researcher at the Broad Institute. He is well known for his role in the development of optogenetics and CRISPR. He is also famous for his ongoing patent dispute with Chemistry Nobel Prize recipients J. Doudna and E. Charpentier over CRISPR-cas9 human application priority. Next, regarding hydrogen storage, Stanford R. Ovshinsky was a prolific inventor and engineer who contributed enormously to various fields, including energy science. In particular, he developed solid hydrogen storage technologies and founded the company Ovshinsky Innova-

[18] See EPO (2019).
[19] See Wikipedia (2021).

tion LLC at the end of his life to continue to explore alternative sources of power. Finally, in self-driving vehicle technology, Atsushi Tabata is an engineer at Toyota who published several articles related to the automation of driving controls.

3.5 The rise of China

In this section we look at the contribution of the Triad and China to frontier technologies. Specifically, their respective contributions are measured by the count of utility patent first publications (see Table 3.14 in Appendix) at the US (US), European (EP), Japanese (JP) and Chinese (CN) patent offices[20] which belong to one of the frontier technologies we defined earlier. Note that, for each technology , we report only years where the number of published patents exceeded 500. We are fully aware of the limitations of such a direct approach. The two main limitations are i) that all patents are not created equal, meaning that patent count comparisons should be qualified by the quality of patents, and ii) that the patenting office is an imperfect proxy for the actual location of the technology contributors. We address both of these concerns as follows. We will look at the quantity and the quality of patents granted by the four offices and proxy the *origin* of the invention using the office of earliest filing of the patent family.[21]

3.5.1 The bi-polarization of frontier innovation by the US and China

We start by delving into the share of the four patent offices in the patent publication count for each technology. Our results are reported in Figure 3.2. The main trends are surprisingly consistent across technologies.

The most striking fact is the pervasive growth of the share of the Chinese office across all frontier technologies since the 2000s. While it used to be almost insignificant in the early 2000s, at the end of the 2010s, the Chinese office represented around 40% of patent publications for all the frontier technologies we are looking at. This share even exceeded 60% in the case of blockchain.

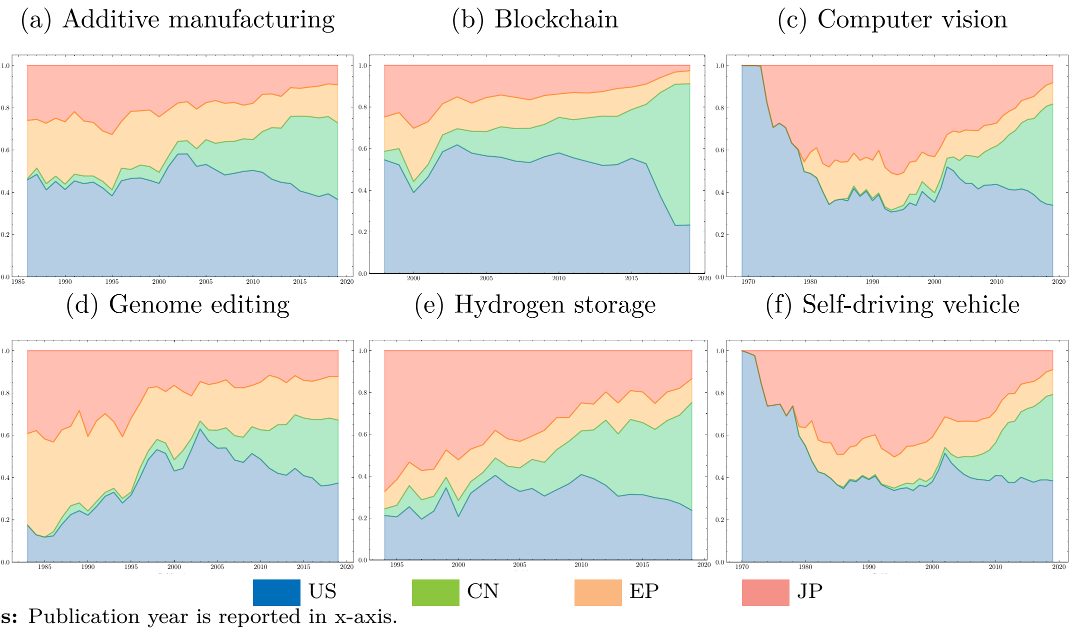
It should also be noted that this relative growth of China's technological power takes place in a setting characterized by a marked heterogeneity of trajectories within the Triad. Although the share of the US remains steady at around 40% since the 1990s, the share of Japan experiences a sustained decline over that period while the European patent office's share remained stable (or slightly decreasing) at around 10%. Japan's decline went from 40% or more of any technology patenting activity in the early 1990s to around 10% at the end of the 2010s.

Overall, the frontier technology landscape which used to be dominated by the Triad is now largely bi-polarized between the US and China. At the end of the 2010s, the US and Chinese patent offices represented together 60% or more of all patents published in any of our frontier technology.

[20] Note that this analysis focuses on the four most important technological contributors but our data enables us to include any country (with a patent office) in our analysis.

[21] The lack of data on the origin of patentees (inventors and assignees) in general and specifically for patents granted at the Chinese office prevent us from a more direct approach.

Figure 3.2: Relative contribution to frontier technologies



(a) Additive manufacturing (b) Blockchain (c) Computer vision

(d) Genome editing (e) Hydrogen storage (f) Self-driving vehicle

Notes: Publication year is reported in x-axis.

3.5.2 The Chinese catch up in quality

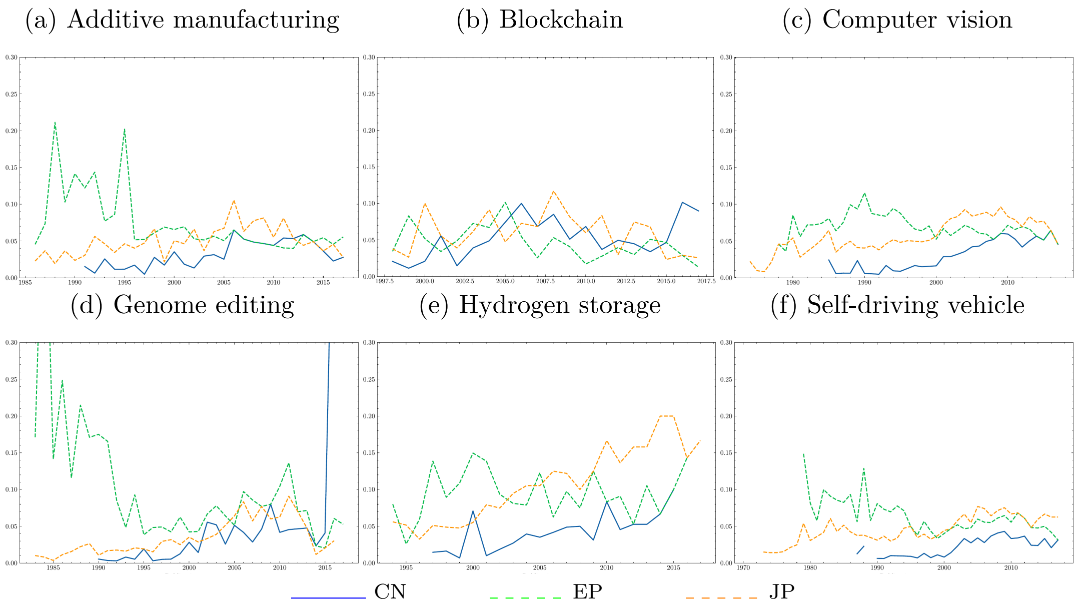
Next, we examine the quality of patents published at the four aforementioned patent offices. As discussed earlier, measuring frontier innovation contributions using only the count of patents published in frontier technologies could be misleading. Patents are not all created equal. In particular, there are good reasons to believe that patent counts, especially in the case of China, might be a noisy signal of technological development undermined by poor patent quality. As discussed by He (2021), patent applications in China reflect diverse incentives which have sometimes little to do with invention. These incentives include government subsidy or job promotion, reputation building for individuals or universities and institutions, or acquiring certification as national high-tech enterprises. In this context, Hudson (2021) *inter alia*, stressed that patent counts is an unreliable methodology to determine 5G leadership. This naturally raises serious questions on China's actual contribution to frontier innovation and the meaning of our first result.

We investigate this by looking at a common measure of patent quality, the number of citations received by patents.[22] We chose to focus on the upper tail of the distribution since the most cited patents are also those which are expected to have the largest impact. Specifically, we look at the ratio between the number of citations received by the 5% most cited patents published at a given office and the number of citations received by the 5% most cited patents published at the US Patent Office (USPTO). Importantly, in order to avoid mixing citations generated by different offices with distinct citation rules, we restrict to citations from USPTO patents. Note that, due to home-bias, the number of citations received by USPTO patents can only be used to compare patents published at other offices between themselves and should *not* be used

[22] Note however that the very notion of patent quality is multi-faceted. The various measures used to apprehend patent quality are generally inconsistent as evidenced by Higham et al. (2021).

to compare USPTO patents with patents published at other offices. Results are reported by Figure 3.4.

Figure 3.4: Citations received by the 5% most cited patents relative to the 5% most cited US patents



(a) Additive manufacturing (b) Blockchain (c) Computer vision

(d) Genome editing (e) Hydrogen storage (f) Self-driving vehicle

Notes: We consider only citations from USPTO patents to avoid mixing citations generated by different offices with distinct citation rules. Due to home-bias, the number of citations received by USPTO patents can only be used to compare patents published at other offices between themselves and should *not* be used to compare USPTO patents with patents published at other offices. Publication year is reported in x-axis.

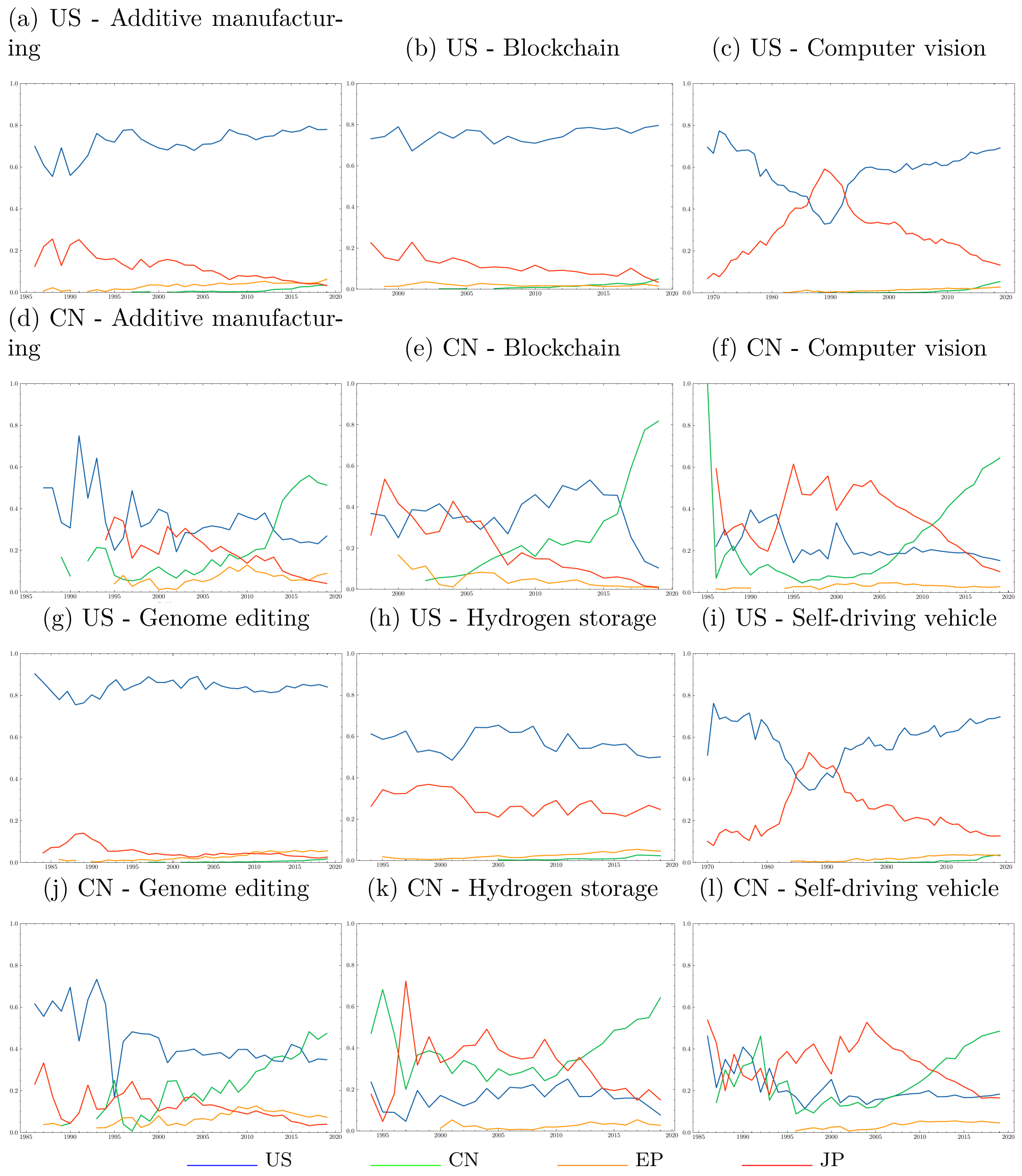
We find that the 5% most cited patents published at the Chinese patent office during the early 2000s used to be less cited than their EP and JP counterparts (except for blockchain). However, the gap has consistently declined since that period. At the end of the period, the 5% top cited patents published at the Chinese patent office are equally cited by US patents as their European and Japanese counterparts. The 5% most cited Chinese patents in the Blockchain and Genome Editing technologies even overtook their European and Japanese counterparts by the number of citations received from patents published at the US Patent Office. Importantly, these results are robust to considering the top 10% of patents published at each patent office. Overall, we find that, although the stock of frontier technology patents published at the Chinese office exhibit a lower quality than their European and Japanese counterparts, this gap has been reducing since the 2000s and is even closing in some technologies (Blockchain and Genome editing).

3.5.3 China's frontier innovation domestic capacities build up

Third, we investigate whether the rise in frontier technology patenting at the Chinese patent office reflects the actual development of domestic innovation capacities. This is a key question for the interpretation of the previous findings. If most of the frontier technology patenting activity at the Chinese patent office comes from foreign patentees (inventors and assignees), then it might simply reflect the attractiveness of the Chinese market for these foreign technology holders. On the contrary, if it is generated by domestic patentees, then it echoes the build up of local frontier innovation capacities which are key to escape the middle-income trap.

Unfortunately, the location of patentees remains largely incomplete in standard datasets, especially for Chinese patents. Hence, we were forced to proxy this data using the office of priority filing, that is the office where the first patent of a family was filed. The underlying rationale is that patentees tend to first file an invention at the patent office of their home country and then file abroad. Using this proxy, we look at the priority filing composition of frontier technology patents published at the US and Chinese patent offices. Results are reported in Figure 3.6.

Figure 3.6: Office of priority filing of patents published at the US and Chinese patent offices



(a) US - Additive manufacturing

(b) US - Blockchain

(c) US - Computer vision

(d) CN - Additive manufacturing

(e) CN - Blockchain

(f) CN - Computer vision

(g) US - Genome editing

(h) US - Hydrogen storage

(i) US - Self-driving vehicle

(j) CN - Genome editing

(k) CN - Hydrogen storage

(l) CN - Self-driving vehicle

We find that the US Patent Office tends to exhibit a consistently high share of US priority filing, usually close to 80%. The mirroring figure was much lower in the early 2000s at the Chinese patent office. Priority filings were mainly coming from the US and Japan by the time. Things started to change in the late 2000s. From this moment on, the share of Chinese priority filings in frontier technology patenting at the Chinese patent office kept rising up to 80% in the late 2010s. This consistent movement seems to reflect the build up of Chinese domestic capacities to generate frontier innovation.

3.6 Conclusion

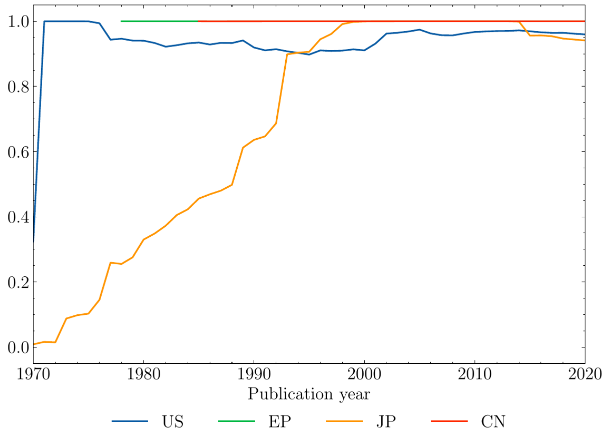
In this paper, we have extended the automated patent landscaping approach from Abood and Feltenberger (2018) to accurately and consistently delineate a group of frontier technologies from the worldwide corpus of patents. Next, we have leveraged these results to investigate the contribution of the US, Europe, Japan and China to the production of frontier innovation. We find that China's technological power is on the rise, representing around 30% of frontier technology patenting activities at the dawn of 2020. The contribution to frontier technologies patenting which used to be dominated by the US, European and Japanese patent offices in the early 2000s is now bi-polarized by the US and Chinese offices. Digging further, we observed that patents published at the Chinese office have long exhibited the stigmas of a catching up economy compared to the most advanced economies. However, the gap seems to be closing and China's domestic capabilities to be building.

So, can China innovate? It seems that the answer is yes and that China is already innovating at the frontier. Yet, two important points should be stressed. First, the Japanese example shows that the innovation capacities of a country can never be taken for granted. What will happen to China's innovative power in the next decades is out of the scope of this paper but China's spectacular take-off since the 2000s does not necessarily foreshadow the next decades. Second, we have shown that China is increasingly contributing to frontier technologies which were pioneered before China's technological take-off. This leaves the question of China's ability to pioneer a new frontier technology untouched.

Despite these obvious limitations we believe that our approach already delivers critical insights to study innovation through the lens of patents. Most importantly, we believe that opens at least two important avenues for research. First, delving into the characteristics (assignees, inventors, patentees locations, etc) of frontier technology patents filed in China and other developing countries appears to be a promising way to better assess the role of the various technology diffusion channels. Second, another promising avenue would be to delve into business dynamics (entry and exit) which take place within technologies themselves and might well have sound implications for the rest of the economy, including the fall of the labor share in the US and the rise of superstar giant innovators, as discussed by Autor et al. (2020) and Aghion et al. (2019a).

3.7 Appendix

Figure 3.8: Patent coverage



Notes: This figure reports the share of patents published in each patent office for which we have an English abstract available, either directly of through patent family expansion.

Table 3.6: Rule based candidates definition - Additive manufacturing

Technological classes (CPC)	**Keywords**	**Patents**
B81C2201/0184, G05B2219/49002, G05B2219/49003, G05B2219/49004, G05B2219/49005, G05B2219/49006, G05B2219/49007, G05B2219/49008, G05B2219/49009, G05B2219/49011, G05B2219/49013, G05B2219/49014, G05B2219/49015, G05B2219/49016, G05B2219/49017, G05B2219/49018, G05B2219/49019, G05B2219/49021, G05B2219/49022, G05B2219/49023, G05B2219/49024, G05B2219/49025, G05B2219/49026, G05B2219/49027, G05B2219/49028, G05B2219/49029, G05B2219/49031, G05B2219/49032, G05B2219/49033, G05B2219/49034, G05B2219/49035, G05B2219/49036, G05B2219/49037, G05B2219/49038, G05B2219/49039, A43D2200/60, A23P2020/253, B29C64/10, C08L101/00, B29C67/00, B22F3/00, G05B2219/49013, G03F7/70416, B28B1/001, B33Y10/00, B23K9/04, B23K10/027, B23K15/0086, B23K11/0013	3d-printing, stereolithography, additive manufacturing, three-dimensional objects, rapid prototyping, additive material manufacturing three dimensional printing material, 3d-printing materials photolithography, fuse deposition mode	US-4575330-A, US-5534104-A, US-6259962-A, US-5204055-A, US-5182056-A, DE-102013205724-A1, FR-3070302-B1, US-10076875-B2, US-8349239-B2, CN-108868141-A, CN-105569344-A, CN-105604327-A, WO-2018229418-A1, KR-101706473-B1, WO-2016111879-A1, US-20180141274-A1, WO-2008061909-A2, US-20170251713-A1, EP-1352619-B1, EP-3319545-B1, EP-3151782-B1, US-10441426-B2, US-9056017-B2

Notes: Candidates include patents matching at least one of the following criteria: 1) the patent's CPC codes include at least one code listed in the "technological classes", 2) the patent's abstract contains at least one of the keywords (or keyphrases) listed in the columns keywords or 3) the patent is highly similar to a patent listed in the columns patents. The latter patents are patents known to be at the core of the technology and the similarity is based on Google Patents embedding.

Table 3.7: Rule based candidates definition - Blockchain

Technological classes (CPC)	**Keywords**	**Patents**
H04L009/08, H04L67/00, H04L009/10, H04L009/12, H04L009/14, H04L009/28, H04L29/06, G06Q20/00, G06F21/00, G06F12/14, G06Q20/06, G06Q20/10, G06Q20/20, G06Q20/32, G06Q20/36, H04L2209/00, G09C001/00, G09C001/02, G09C001/04, G09C001/06, H04L63/00, G06Q30/0619, G06F21/00, G06F021/24, G06F021/00, G06F021/02, G06F012/28, G06F012/14, G06F17/00	blockchain, digital mining, bitcoin, cryptocoin, cryptocurrency, digital wallet, ethereum, smart contracts, record keeping, distributed ledger, distributed node, private ledger, public ledger, intelligent node, full node, digital signatures, public key, user identity, hashing, consensus methodologies, proof of work, proof of stake, deposition based, ripple	EP-3125489-B1, US-9785369-B1, DE-102016104478-A1, US-9853819-B2, US-9842216-B2, US-9855785-B1, US-20180137465-A1, US-9635000-B1, EP-329562-A1, EP-3295350-B1, CN-105719172-A, CN-105701372-B, US-9836908-B2, US-9818092-B2, US-9824031-B1, US-10643202-B2, CN-105844505-A, US-9298806-B1, CN-105790954-B, US-9858781-B1, US-9853977-B1, US-9641338-B2, US-9641342-B2, EP-325719-B1

Notes: See Table 3.6

Table 3.8: Rule based candidates definition - Computer Vision

Technological classes (CPC)	**Keywords**	**Patents**
B25J9/161, G06F17/16, G06N5/003, G06N7/005, G06N7/046, B29C66/965, G08B29/186, F02D41/1405, G01N29/4481, G06F11/1476, G06F17/2282, H02P21/0014, H02P23/0018, H03H2222/04, Y10S128/924, Y10S128/925, B64G2001/247, F05B2270/707, F05B2270/709, F05D2270/709, G10H2250/151, H04L25/03165, H04Q2213/054, H04Q2213/343, B60G2600/1876, B60G2600/1878, B60G2600/1879, E21B2041/0028, F16H2061/0081, F16H2061/0084, G06F2207/4824, G10K2210/3024, G10K2210/3038, H03H2017/0208, B29C2945/76979, G05B2219/33002, G06T2207/20081, G06T2207/20084, G06T2207/20084, H04L2025/03464, H04L2025/03554, H04Q2213/13343, B60W30/06, B60W30/10, B60W30/12, B60W30/14, B60W30/17, G06T9/002, G10L25/30, G06K7/1482, G06T3/4046, B62D15/0285	adaboost, xgboost, bayesian network, decision tree, genetic algorithm, gradient tree boosting, logistic regression, random forest, rankboost, support vector machine, multilayer perceptron, hidden markov model, generalized adversarial network, backpropagation, stochastic gradient descent, supervised training, reinforcement learning, neural network, self learning, semi supervised learning, unsupervised training, transfer learning, overfitting, active learning, clustering, data mining, deep learning, expert system, embedding, machine learning, fuzzy logic, feature selection, objective function, target function, regression model, signal processing, computer vision, machine vision, lidar, character recognition, optical character recognition, handwritten character recognition, image to text, text recognition, face recognition, facial recognition, biometric data, biometrics, mass surveillance, face unlock, traffic cameras, object detection, edge detection, obstacle avoidance, motion tracking	US-8953886-B2, WO-2003023696-A1, US-5881172-A, US-20170024607-A1, US-20170169205-A1, US-20170169303-A1, US-20170235931-A1, US-20200175326-A1, US-10872228-B1

Notes: See Table 3.6

Table 3.9: Rule based candidates definition - Genome Editing

Technological classes (CPC)	Keywords	Patents
A01H4/00, A01K67/00, C12N/1500, C12N1/00, C12N5/00, C12N7/00C12Y, C12N5/10, C12Q1/68, C12Q1/70, G01N33/00, A61K48/00, A61K31/7088, C07K14/00	dna editing, gene editing, genome engineering, recombinant targeting vectors, homologous recombination, double-strand dna break, homology-directed repair, targeted dna sequence, dna cleavage, fok1, sequence-specific nuclease system, zinc finger nuclease, cys2-his2, transcriptional activator-like effector nuclease, talens, clustered regularly interspaced short palindromic repeat, crispr/cas, cas9, pre-crrna, tracrrna, enzyme rnase, single guide rna, crispr-cpf1, ngago, single-stranded dna-guided argonaute endonuclease, natronobacterium gregoryi argonaute	WO-2000041566-A9, WO-2003087341-A3, WO-2010079430-A1, WO-2011072246-A2, US-8440431-B2, US-8440432-B2, US-8450471-B2, US-8566363-B2, WO-2014093661-A2, WO-2013176772-A1, US-20170367280-A1

Notes: See Table 3.6

Table 3.10: Rule based candidates definition - Hydrogen Storage

Technological classes (CPC)	Keywords	Patents
Y02E60/30, Y02E60/32, Y02E60/321, Y02E60/322, Y02E60/324, Y02E60/325, Y02E60/327, Y02E60/328, Y02E60/34, Y02E60/36, Y02E60/362, Y02E60/364, Y02E60/366, Y02E60/368, B01D53/02, C01B3/00-58, F17C2221/012, C22C19/03, C22C22/00, C22C33/00, F25B17/12, H01M4/38, H01M8/06, F17C2221/012, F17C6/00, F17C5/02	hydrogen fuel cells, hydrogen storage, liquid hydrogen, solid-state hydrogen storage, compressed hydrogen storage, dehydrogenation reaction, hydrogen gas, hydrogen fuel, hydrogen storage materials, hydrogen-powered device	US-20080248355-A1, CN-1322266-C, US-7678362-B2, US-7118611-B2, CN-203500844-U, US-7094493-B2, US-10622655-B2, WO-2019239141-A1, US-8871671-B2, JP-6061354-B2, EP-2554694-B1, FR-2939784-A1, CA-2980664-C, CN-103797142-A, US-7678479-B2, JP-6418680-B2, DE-102009016475-B4, US-7093626-B2, DE-102013203892-A, KR-101107633-B1, JP-4849775-B2, JP-3706611-B2, US-6875536-B2, JP-5338903-B2

Notes: See Table 3.6

Table 3.11: Rule based candidates definition - Self driving vehicle

Technological classes (CPC)	**Keywords**	**Patents**
G08G1/02, G08G1/0967, G08G1/0968, G01S7/003, G07B15/063, G07C5/00, G07C5/12, E01F, E01F9/00, E01F9/40, H04W36/00, H04W76/50, B61L3/00, G05D1/0011, G05D1/0027, G05D1/0287, G05D1/0297, G08G1/00, G08G1/01, G08G1/09, G08G1/0968, G08G1/127, G08G1/16, G08G1/164, G08G1/20, G01S13/93, G10S13/931, G01S15/88, G01S15/93, G01S17/88, G01S17/93, G07C5/00, G07C5/01, G07C5/02, G07C5/03, G07C5/04, G07C5/05, G07C5/06, G07C5/07, G07C5/08, E01F9/00, B60L2240/70, B61L25/00, G01S7/00, G01S13/00, G01S15/00, G01S17/00, G01S7/00, G01S7/02, G01S7/52, G01S13/00, G01S13/86, G01S13/87, G01S13/93, G01S15/00, G01S15/025, G01S15/87, G01S15/931, G01S17/00, G06K9/00, G05D1/00, G05D1/0257, B60W2420/52, B60Y2400/3017, B60R19/00, G01S17/023, G01S17/06, G01S17/87, G01S17/88, G01S17/936, G01S7/48, G01S2013/9332, B60W2420/52, G06T1/0007, G06T1/0014, G06T1/20, G06K9/00362, G06K9/00785, G06K9/00791, H04N5/335, B60Y2400/3015, B60W2420/42, B60S1/56, G01C21/00, G01C21/26, G01C21/34, G01S7/52, G01S15/00, G05D1/00, G05D1/0027, G05D1/0088, G05D1/021, G05D1/0212, G05D1/0276, G05D1/0287, G05D1/02, G06T1/0007, G06T1/0014, G06T1/20, G08G1/16, G08G1/161, G08G1/22, H04W4/44, H04W4/46, F16D2500/31, B60L2240/60, B60L2240/62, B60W30/16, B60W2050/008, B60W2550/402, B60W2550/408, B60G17/015, B60G17/016, B60G17/0195, B60G2800/00, B60K28/04, B60W30/00, B60W40/00, F16D2500/508, G05D1/0088, G05D2201/0212, B60W30/095, B60W50/0097, G05D1/0212	self-driving vehicle, autopilot, driverless vehicle, autonomous vehicle, automated vehicles, vehicle connectivity, vehicle-to-vehicle communication, fleet management, vehicle lidar, vehicle sonar, vehicle radar, vehicle camera, object detection, obstacle detection, object classification, cruise control, pedestrian detection, environment mapping, surround view, blind spot detection, park assistance, lane departure, traffic sign recognition, drive assist system, trajectory generation, reactive control, path trajectory planning, manoeuvers planning	US-20050088318-A1, US-9293045-B2, US-9723457-B2, US-10405215-B2, WO-2019052353-A1, US-10089537-B2, US-10564639-B1, DE-112019000049-T5, US-20190384304-A1, DE-112019000122-T5, US-20170030728-A1, US-20190265703-A1, WO-2019094843-A1

Notes: See Table 3.6. The list of technological classes have been splitted into two columns to save space.

Table 3.12: Annotation guidelines

Technology	Options
Additive manufacturing	- Create 3D printable model with computer aided design - Examine stereolithography file for errors and inconsistency - Convert model into a series of thin layers - Manufacture materials for 3D printings - Print 3D model
Blockchain	- Record transactions between two parties - Serve as public transaction ledger of cryptocurrency - Execute or enforce smart contract - Hash tree verification / Verify the authenticity of documents / Proof of work - Analyse transactions in a distributed ledger - Manage Identity System based on the concept of peer-to-peer protocols (IDMS) / Mediate user authentication
Computer vision	- Process digital images - Analyse digital images - Understand digital images
Genome editing	- Target DNA sequence - Break DNA sequence - Edit DNA sequence
Hydrogen storage	- Hydrogen production and compression - Generate power from hydrogen gas - Design vessel containment that is resistant to hydrogen permeation and corrosion (+ thermal management) - Manufacture fuel cell using hydrogen - Provide hydrogen to a hydrogen-powered device (fill, tank)
Self-driving vehicle	- Enable vehicles to make autonomous decisions - Automate vehicle handling - Vehicle-to-vehicle communication - Communication between vehicle and rest-of-the-world

Notes: Human annotator accepts or rejects a candidate patent depending on whether the patent's abstract clearly discusses one or more of the options listed.

Table 3.13: Data source

Variable	Source	Name
Abstract	GPR	abstract
Assignee	PAT	assignee_harmonized.name
Backward citations	PAT	citation.publication_number
Filing date	PAT	filing_date
Forward citations	GPR	cited_by.publication_number
Inventor	PAT	inventor_harmonized.name
Patent family	PAT	family_id
Patent office	PAT	country_code
Publication date	PAT	publication_date
Technological class (CPC)	PAT	cpc.code

Notes: PAT and GPR respectively refer to the patents-public-data:patents.publications and patents-public-data:google_patents_research.publications tables. Data is publicly available from the Google Patents Public Dataset.

Table 3.14: Utility patents first publication kind codes

Country code	**Kind code**
US	A, A1, B1
EP	A1, A2
CN	A
JP	A

Bibliography

Abood, Aaron and Dave Feltenberger, "Automated patent landscaping," *Artificial Intelligence and Law*, Jun 2018, *26* (2), 103–125.

Abrami, Regina M, William C Kirby, and F Warren McFarlan, "Why China can't innovate," *Harvard business review*, 2014, *92* (3), 107–111.

Acemoglu, Daron, Philippe Aghion, and Fabrizio Zilibotti, "Distance to frontier, selection, and economic growth," *Journal of the European Economic association*, 2006, *4* (1), 37–74.

Acs, Zoltan J and David B Audretsch, "Patents as a measure of innovative activity," *Kyklos*, 1989, *42* (2), 171–180.

Adams, Stephen, "The text, the full text and nothing but the text: Part 1–Standards for creating textual information in patent documents and general search implications," *World Patent Information*, 2010, *32* (1), 22–29.

Agarwal, Rajshree, Martin Ganco, and Rosemarie H Ziedonis, "Reputations for toughness in patent enforcement: Implications for knowledge spillovers via inventor mobility," *Strategic Management Journal*, 2009, *30* (13), 1349–1374.

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

_ and _ , "A Model of Growth Through Creative Destruction," *Econometrica*, 1992, *60* (2), 323–351.

_ , Antonin Bergeaud, Timo Boppart, Peter J Klenow, and Huiyu Li, "A theory of falling growth and rising rents," Technical Report, National Bureau of Economic Research 2019.

_ , _ , Timothee Gigout, Mathieu Lequien, and Marc Melitz, "Spreading Knowledge across the World: Innovation Spillover through Trade Expansion," 2019. manuscript, Harvard University.

_ , _ , Timothée Gigout, Matthieu Lequien, and Marc Melitz, "Exporting ideas: How trade spills over to knowledge," 2021. Mimeo Harvard.

_ , Celine Antonin, David Stromberg, and Xueping Sun, "Democracy and Sciences," 2021. manuscript, London School of Economics.

_ , Leah Boustan, Caroline Hoxby, and Jerome Vandenbussche, "The causal impact of education on economic growth: evidence from US," *Brookings papers on economic activity*, 2009, *1* (1), 1–73.

_ , Mathias Dewatripont, and Jeremy C Stein, "Academic freedom, private-sector focus, and the process of innovation," *The RAND Journal of Economics*, 2008, *39* (3), 617–635.

_ , _ , Caroline Hoxby, Andreu Mas-Colell, and André Sapir, "The governance and performance of universities: evidence from Europe and the US," *Economic policy*, 2010, *25* (61), 7–59.

_ , Ufuk Akcigit, Ari Hyytinen, and Otto Toivanen, "The Social Origins of Inventors," CEP Discussion Papers dp1522, Centre for Economic Performance, LSE December 2017.

Akcigit, Ufuk, John Grigsby, and Tom Nicholas, "Immigration and the rise of american ingenuity," *American Economic Review*, 2017, *107* (5), 327–31.

_ , _ , and _ , "Immigration and the Rise of American Ingenuity," *American Economic Review, Papers and Proceedings*, 2017, *107*, 327–331.

_ , _ , and _ , "The Rise of American Ingenuity: Innovation and Inventors of the Golden Age," NBER Working Papers 23047, National Bureau of Economic Research, Inc January 2017.

_ , _ , _ , and Stefanie Stantcheva, "Taxation and Innovation in the 20th Century," Working Paper 24982, National Bureau of Economic Research September 2018.

Albert, Michael B, Daniel Avery, Francis Narin, and Paul McAllister, "Direct validation of citation counts as indicators of industrially important patents," *Research policy*, 1991, *20* (3), 251–259.

Alcacer, Juan and Michelle Gittelman, "Patent citations as a measure of knowledge flows: The influence of examiner citations," *The Review of Economics and Statistics*, 2006, *88* (4), 774–779.

Almeida, Paul, "Knowledge sourcing by foreign multinationals: Patent citation analysis in the US semiconductor industry," *Strategic management journal*, 1996, *17* (S2), 155–165.

_ and Bruce Kogut, "Localization of knowledge and the mobility of engineers in regional networks," *Management Science*, 1999, *45* (7), 905–917.

Andersson, David E and Fredrik Tell, "Dependent Invention and Dependent Inventors," 2018. Uppsala University mimeo.

Andrews, Michael, "Comparing historical patent datasets," 2019. Mimeo University of Iowa.

_ , "Historical patent data. A practitioner's guide," *Available at SSRN: https://ssrn.com/abstract=3415318*, 2020.

Andrews, Michael J and Alexander Whalley, "150 years of the geography of innovation," *Regional Science and Urban Economics*, 2021, p. 103627.

Andrews, Mike, "How do institutions of higher education affect local invention? Evidence from the establishment of US colleges. Evidence from the Establishment of US Colleges," 2020. Mimeo University of Maryland.

Arkolakis, Costas, Sun Kyoung Lee, and Michael Peters, "European immigrants and the United States' rise to the technological frontier," 2020. mimeo Yale.

Arts, Sam, Bruno Cassiman, and Juan Carlos Gomez, "Text matching to measure patent similarity," *Strategic Management Journal*, 2018, *39* (1), 62–84.

_ , Jianan Hou, and Juan Carlos Gomez, "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures," *Research Policy*, 2020, p. 104144.

Arundel, Anthony and Isabelle Kabla, "What percentage of innovations are patented? Empirical estimates for European firms," *Research policy*, 1998, *27* (2), 127–141.

Audretsch, Bruce, "Agglomeration and the location of innovative activity," *Oxford Review of Economic Policy*, 1998, *14* (2), 18–29.

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

Autor, David, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen, "The fall of the labor share and the rise of superstar firms," *The Quarterly Journal of Economics*, 2020, *135* (2), 645–709.

Babina, Tania, Asaf Bernstein, and Filippo Mezzanotti, "Crisis Innovation," Working Paper w27851, National Bureau of Economic Research 2020.

Bahar, Dany, Prithwiraj Choudhury, and Hillel Rapoport, "Migrant inventors and the technological advantage of nations," *Research Policy*, 2020, *49* (9).

Baruffaldi, Stefano, Brigitte van Beuzekom, Hélène Dernis, Dietmar Harhoff, Nandan Rao, David Rosenfeld, and Mariagrazia Squicciarini, "Identifying and measuring developments in artificial intelligence: Making the impossible possible," Working Paper 2020/05, OECD, Sciences, Technology and Innovation Directorate 2020.

Bazzi, Samuel, Martin Fiszbein, and Mesay Gebresilasse, "Frontier culture: The roots and persistence of "rugged individualism" in the United States," *Econometrica*, 2020, *88* (6), 2329–2368.

BDI, "Germany 2030. Future perspectives for value creation," Technical Report 2011. https://espas.secure.europarl.europa.eu/orbis/document/germany-2030-future-perspectives-value-creation.

Belenzon, Sharon and Mark Schankerman, "Spreading the word: Geography, policy, and knowledge spillovers," *Review of Economics and Statistics*, 2013, *95* (3), 884–903.

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

Benahmed-Miniuk, Fairouz, Mat Kresz, Jitendra K Kanaujiya, and Christopher D Southgate, "Genome-editing technologies and patent landscape overview," *Pharmaceutical patent analyst*, 2017, *6* (3), 115–134.

Benson, Christopher L and Christopher L Magee, "Quantitative determination of technological improvement from patent data," *PloS one*, 2015, *10* (4), e0121635.

Bergeaud, Antonin, Gilbert Cette, and Rémy Lecat, "Productivity trends in advanced countries between 1890 and 2012," *Review of Income and Wealth*, 2016, *62* (3), 420–444.

_ , Yoann Potiron, and Juste Raimbault, "Classifying patents based on their semantic content," *PloS one*, 2017, *12* (4), e0176310.

Berkes, Enrico, "Comprehensive universe of US patents (CUSP): data and facts," *Unpublished, Ohio State University*, 2018.

_ , "Comprehensive Universe of U.S. Patents (CUSP): Data and Facts," 2018. Mimeo Ohio State University.

_ and Ruben Gaetani, "The geography of unconventional innovation," 2019. Mimeo Ohio State University.

Bernstein, Shai, Rebecca Diamond, Timothy McQuade, Beatriz Pousada et al., "The contribution of high-skilled immigrants to innovation in the United States," Technical Report 3748 2018.

Bessen, James and Robert M Hunt, "An empirical look at software patents," *Journal of Economics & Management Strategy*, 2007, *16* (1), 157–189.

Bloom, Nicholas and John Van Reenen, "Measuring and explaining management practices across firms and countries," *The quarterly journal of Economics*, 2007, *122* (4), 1351–1408.

_ , Charles I Jones, John Van Reenen, and Michael Webb, "Are ideas getting harder to find?," *American Economic Review*, 2020, *110* (4), 1104–44.

_ , Tarek Alexander Hassan, Aakash Kalyani, Josh Lerner, and Ahmed Tahoun, "The Diffusion of Disruptive Technologies," Technical Report w28999, National Bureau of Economic Research 2021.

Bolt, Jutta and Jan Luiten van Zanden, "Maddison style estimates of the evolution of the world economy. A new 2020 update," Maddison-Project Working Paper WP-15, University of Groningen, Groningen, The Netherlands 2020.

Borjas, George J and Kirk B Doran, "The collapse of the Soviet Union and the productivity of American mathematicians," *The Quarterly Journal of Economics*, 2012, *127* (3), 1143–1203.

Breschi, Stefano and Francesco Lissoni, "Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows," *Journal of Economic Geography*, 2009, *9* (4), 439–468.

Bryan, Kevin A, Yasin Ozcan, and Bhaven Sampat, "In-text patent citations: A user's guide," *Research Policy*, 2020, *49* (4), 103946.

Buzard, Kristy, Gerald A Carlino, Robert M Hunt, Jake K Carr, and Tony E Smith, "The agglomeration of American R&D labs," *Journal of Urban Economics*, 2017, *101*, 14–26.

Candia, Cristian, C Jara-Figueroa, Carlos Rodriguez-Sickert, Albert-László Barabási, and César A Hidalgo, "The universal decay of collective memory and attention," *Nature Human Behaviour*, 2019, *3* (1), 82–91.

Carlino, Gerald A, Satyajit Chatterjee, and Robert M Hunt, "Urban density and the rate of invention," *Journal of Urban Economics*, 2007, *61* (3), 389–419.

Carpenter, Mark P, Francis Narin, and Patricia Woolf, "Citation rates to technologically important patents," *World Patent Information*, 1981, *3* (4), 160–163.

Chien, Colleen V and Jiun Ying Wu, "Decoding Patentable Subject Matter," *Patently-O Patent Law Journal 1, Santa Clara University Legal Studies Research Paper*, 2018, *1.*

Chiu, Jason PC and Eric Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Transactions of the Association for Computational Linguistics*, 2016, *4*, 357–370.

Choudhury, Prithwiraj, Evan Starr, and Rajshree Agarwal, "Machine learning and human capital complementarities: Experimental evidence on bias mitigation," *Strategic Management Journal*, 2020, *41* (8), 1381–1411.

Clarke, Nigel S, Björn Jürgens, and Victor Herrero-Solana, "Blockchain patent landscaping: An expert based methodology and search query," *World Patent Information*, 2020, *61*, 101964.

Coe, David T and Elhanan Helpman, "International r&d spillovers," *European economic review*, 1995, *39* (5), 859–887.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, 2011, *12* (ARTICLE), 2493–2537.

Corsino, Marco, Myriam Mariani, and Salvatore Torrisi, "Firm strategic behavior and the measurement of knowledge flows with patent citations," *Strategic Management Journal*, 2019, *40* (7), 1040–1069.

Cotropia, Christopher A, "The folly of early filing in patent law," *Hastings Law Journal*, 2009, *61*, 65–129.

Dang, Jianwei and Kazuyuki Motohashi, "Patent statistics: A good indicator for innovation in China? Patent subsidy program impacts on patent quality," *China Economic Review*, 2015, *35*, 137–155.

de Rassenfosse, Gaétan, Adam Jaffe, and Emilio Raiteri, "The procurement of innovation by the US government," *PloS one*, 2019, *14* (8), e0218927.

_ , Jan Kozak, and Florian Seliger, "Geocoding of worldwide patent data," *Scientific data*, 2019, *6* (1), 1–15.

de Rassenfosse, Gaétan, Jan Kozak, and Florian Seliger, "Geocoding of worldwide patent data," *Nature - Scientific Data*, 2019, *6* (260).

Delgado, Mercedes, Michael E Porter, and Scott Stern, "Clusters and entrepreneurship," *Journal of economic geography*, 2010, *10* (4), 495–518.

Deloitte, "Future of the Tech Sector in Europe," https://www2.deloitte.com/uk/en/pages/technology-media-and-telecommunications/articles/future-of-tech-in-europe.html#- 2021. Accessed: 2021-05-26.

Diallo, Boubacar and Wilfried Koch, "Bank concentration and Schumpeterian growth: theory and international evidence," *Review of Economics and Statistics*, 2018, *100* (3), 489–501.

Eaton, Jonathan and Samuel Kortum, "Measuring technology diffusion and the international sources of growth," *Eastern Economic Journal*, 1996, *22* (4), 401–410.

_ **and** _ , "Trade in ideas Patenting and productivity in the OECD," *Journal of International Economics*, 1996, *40* (3-4), 251–278.

_ **and** _ , "Engines of growth: Domestic and foreign sources of innovation," *Japan and the World Economy*, 1997, *9* (2), 235–259.

_ **and** _ , "International technology diffusion: Theory and measurement," *International Economic Review*, 1999, *40* (3), 537–570.

Eckert, Fabian, Andrés Gvirtz, Jack Liang, and Michael Peters, "A Method to Construct Geographical Crosswalks with an Application to US Counties since 1790," Working Paper 26770, National Bureau of Economic Research February 2020.

Eddy, Sean R, "Hidden markov models," *Current opinion in structural biology*, 1996, *6* (3), 361–365.

Egger, Peter H and Nicole Loumeau, "The economic geography of innovation," Technical Report DP13338, CEPR Discussion Paper 2018.

EPO, "Patents and self-driving vehicles. The inventions behind automated driving," Report, EPO 2018.

_ , "Honouring a prolific inventor's dedication to advancing video compression: Marta Karczewicz named European Inventor Award 2019 finalist," https://www.epo.org/news-events/press/releases/archive/2019/20190507n.html 2019. Accessed: 2021-05-26.

_ , "Top 10 Emerging Technologies 2020," https://www.weforum.org/reports/top-10-emerging-technologies-2020 2020. Accessed: 2021-05-26.

Etzioni, Oren, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates, "Unsupervised named-entity extraction from the web: An experimental study," *Artificial intelligence*, 2005, *165* (1), 91–134.

Fan, Peilei, "Innovation in China," *Journal of Economic Surveys*, 2014, *28* (4), 725–745.

Feldman, Maryann P and Dieter F Kogler, "Stylized facts in the geography of innovation," in "Handbook of the Economics of Innovation," Vol. 1, Elsevier, 2010, pp. 381–410.

Fohlin, Caroline, "The Venture Capital Divide: Germany and the United States in the Post-War Era," *Available at SSRN 2849237*, 2016.

Fouquin, Michel and Jules Hugot, "Two Centuries of Bilateral Trade and Gravity Data: 1827-2014," Working Paper N°2016-14, CEPII 2016.

Freilich, Janet, "Prophetic Patents," *UC Davis Law Review*, 2019, *53*, 663–731.

Galibert, Olivier, Sophie Rosset, Xavier Tannier, and Fanny Grandry, "Hybrid Citation Extraction from Patents.," in "Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010" 2010, pp. 17–23.

Giczy, Alexander V, Nicholas A Pairolero, and Andrew Toole, "Identifying artificial intelligence (AI) invention: A novel AI patent dataset," Technical Report 2021.

Gillmore, Julian D, Ed Gane, Jorg Taubel, Justin Kao, Marianna Fontana, Michael L Maitland, Jessica Seitzer, Daniel O'Connell, Kathryn R Walsh, Kristy Wood et al., "CRISPR-Cas9 in vivo gene editing for transthyretin amyloidosis," *New England Journal of Medicine*, 2021, *385* (6), 493–502.

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

Guellec, Dominique and Bruno van Pottelsberghe de la Potterie, “The internationalisation of technology analysed with patent data,” *Research Policy*, 2001, *30* (8), 1253–1266.

Gyourko, Joseph, Christopher Mayer, and Todd Sinai, “Superstar cities,” *American Economic Journal: Economic Policy*, 2013, *5* (4), 167–99.

Hall, Bronwyn H, Adam Jaffe, and Manuel Trajtenberg, “Market value and patent citations,” *RAND Journal of economics*, 2005, pp. 16–38.

Hall, Bronwyn H. and Dietmar Harhoff, “Recent Research on the Economics of Patents,” *Annual Review of Economics*, July 2012, *4* (1), 541–565.

Hanlon, Walker, “British Patent Technology Classification Database: 1855-1882,” 2016.

Hausman, Naomi, “University innovation and local economic growth,” *Review of Economics and Statistics, Forthcoming*, 2020.

He, Alex, “What Do China’s High Patent Numbers Really Mean?,” 2021.

Higham, Kyle, Gaétan De Rassenfosse, and Adam B Jaffe, “Patent quality: towards a systematic framework for analysis and measurement,” *Research Policy*, 2021, *50* (4), 104215.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd, “spaCy: Industrial-strength Natural Language Processing in Python,” 2020.

Hsieh, Chang-Tai, Erik Hurst, Charles I Jones, and Peter J Klenow, “The allocation of talent and us economic growth,” *Econometrica*, 2019, *87* (5), 1439–1474.

Hu, Albert Guangzhou and Gary H Jefferson, “A great wall of patents: What is behind China’s recent patent explosion?,” *Journal of Development Economics*, 2009, *90* (1), 57–68.

Huang, Zhiheng, Wei Xu, and Kai Yu, “Bidirectional LSTM-CRF models for sequence tagging,” *arXiv preprint arXiv:1508.01991*, 2015.

Hudson, Institute, “5G Technological Leadership,” 2021.

IP Australia, “Machine Learning Innovation. A Patent Analytics Report,” Report, IP Australia 2019.

Jaffe, Adam and Gaetan de Rassenfosse, “Patent citation data in social science research: Overview and best practices,” *Journal of the Association for Information Science and Technology*, 2017, *68* (6), 1360–1374.

Jaffe, Adam B and Manuel Trajtenberg, “International knowledge flows: Evidence from patent citations,” *Economics of innovation and new technology*, 1999, *8* (1-2), 105–136.

_ , _ , and Michael S Fogarty, “Knowledge spillovers and patent citations: Evidence from a survey of inventors,” *American Economic Review*, 2000, *90* (2), 215–218.

_ , _ , and Rebecca Henderson, “Geographic localization of knowledge spillovers as evidenced by patent citations,” *the Quarterly journal of Economics*, 1993, *108* (3), 577–598.

_ , Michael S Fogarty, and Bruce A Banks, “Evidence from patents and patent citations on the impact of NASA and other federal labs on commercial innovation,” *The Journal of Industrial Economics*, 1998, *46* (2), 183–205.

Jones, Benjamin F, “The burden of knowledge and the “death of the renaissance man”: Is innovation getting harder?,” *The Review of Economic Studies*, 2009, *76* (1), 283–317.

Kaplan, Sarah and Keyvan Vakili, “The double-edged sword of recombination in breakthrough innovation,” *Strategic Management Journal*, 2015, *36* (10), 1435–1457.

Kay, Anthony, “Tesseract: An Open-Source Optical Character Recognition Engine,” *Linux J.*, July 2007, *2007* (159), 2.

Keller, Wolfgang, "International technology diffusion," *Journal of economic literature*, 2004, *42* (3), 752–782.

_ **and Stephen R Yeaple**, "Multinational enterprises, international trade, and productivity growth: firm-level evidence from the United States," *The Review of Economics and Statistics*, 2009, *91* (4), 821–831.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy, "Measuring Technological Innovation over the Long Run," Working Paper 25266, National Bureau of Economic Research November 2018.

Kelly, Morgan, Joel Mokyr, and Cormac Ó Gráda, "Precocious Albion: a new interpretation of the British industrial revolution," *Annu. Rev. Econ.*, 2014, *6* (1), 363–389.

Kennedy, Scott, "Made in China 2025," https://www.csis.org/analysis/made-china-2025 2015. Accessed: 2021-05-26.

Kerr, William R, "Ethnic scientific communities and international technology diffusion," *The Review of Economics and Statistics*, 2008, *90* (3), 518–537.

Klenow, Peter J and Andres Rodriguez-Clare, "Externalities and growth," *Handbook of economic growth*, 2005, *1*, 817–861.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman, "Technological Innovation, Resource Allocation, and Growth*," *The Quarterly Journal of Economics*, 03 2017, *132* (2), 665–712.

König, Michael, Zheng Michael Song, Kjetil Storesletten, and Fabrizio Zilibotti, "From imitation to innovation: Where is all that Chinese R&D going?," Technical Report w27404, National Bureau of Economic Research 2020.

Krugman, Paul R, *Geography and trade*, MIT press, 1991.

Kuhn, Jeffrey, Kenneth Younge, and Alan Marco, "Patent citations reexamined," *The RAND Journal of Economics*, 2020, *51* (1), 109–132.

Lafferty, John, Andrew McCallum, and Fernando CN Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

Lamoreaux, Naomi R. and Kenneth L. Sokoloff, "Location and technological change in the American glass industry during the late nineteenth and early twentieth centuries," *NBER Working paper*, 1997, (w5938).

_ **and** _ , "The Geography of Invention in the American Glass Industry, 1870-1925," *The Journal of Economic History*, 2000, *60* (3), 700–729.

Lampe, Ryan, "Strategic citation," *Review of Economics and Statistics*, 2012, *94* (1), 320–333.

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer, "Neural Architectures for Named Entity Recognition," in "Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies" Association for Computational Linguistics San Diego, California June 2016, pp. 260–270.

Lanjouw, Jean O and Mark Schankerman, "Patent quality and research productivity: Measuring innovation with multiple indicators," *The Economic Journal*, 2004, *114* (495), 441–465.

Lerner, Joshua, *Architecture of Innovation*, Harvard Business Review Press, 2012.

Li, Jing, Aixin Sun, Jianglei Han, and Chenliang Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, 2020, pp. 1–1.

Lopez, Patrice, "Automatic extraction and resolution of bibliographical references in patent documents," in "Information Retrieval Facility Conference" Springer 2010, pp. 120–135.

Machin, Nathan, “Prospective Utility: A New Interpretation of the Utility Requirement of Section 101 of the Patent Act,” *California Law Review*, 1999, *87*, 421–456.

Mansfield, Edwin, “Patents and innovation: an empirical study,” *Management science*, 1986, *32* (2), 173–181.

Marx, Matt and Aaron Fuegi, “Reliance on Science by Inventors: Hybrid Extraction of In-text Patent-to-Article Citations,” *National Bureau of Economic Research, Working Paper 27987*, 2020.

Mayer, Thierry and Soledad Zignago, “Notes on CEPII’s distances measures: The GeoDist database,” Working Paper N°2011-25, CEPII 2011.

McKinsey, “The top trends in tech,” https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-top-trends-in-tech 2021. Accessed: 2021-05-26.

Meisenzahl, Ralf and Joel Mokyr, “The Rate and Direction of Invention in the British Industrial Revolution: Incentives and Institutions,” Working Paper 16993, National Bureau of Economic Research April 2011.

Meyer, Martin, “What is special about patent citations? Differences between scientific and patent citations,” *Scientometrics*, 2000, *49* (1), 93–123.

Miguélez, Ernest and Andrea Morrison, “Migrant Inventors as Agents of Technological Change,” *AQR–Working Papers, 2021, AQR21/05*, 2021.

Montani, Ines and Matthew Honnibal, “Prodigy: A new annotation tool for radically efficient machine teaching,” *Artificial Intelligence*, 2018, *to appear.*

Moser, Petra, “How do patent laws influence innovation? Evidence from nineteenth-century world’s fairs,” *American economic review*, 2005, *95* (4), 1214–1236.

_ , Alessandra Voena, and Fabian Waldinger, “German Jewish émigrés and US invention,” *American Economic Review*, 2014, *104* (10), 3222–55.

_ and Tom Nicholas, “Was electricity a general purpose technology? Evidence from historical patent citations,” *American Economic Review*, 2004, *94* (2), 388–394.

Nicholas, Tom, “The role of independent invention in US technological development, 1880–1930,” *The Journal of Economic History*, 2010, *70* (1), 57–82.

Nuvolari, Alessandro and Michelangelo Vasta, “The geography of innovation in Italy, 1861–1913: evidence from patent data,” *European Review of Economic History*, 2017, *21* (3), 326–356.

_ and Valentina Tartari, “Bennet Woodcroft and the value of English patents, 1617–1841,” *Explorations in Economic History*, 2011, *48* (1), 97–115.

_ , Gaspare Tortorici, and Michelangelo Vasta, “British-French technology transfer from the Revolution to Louis Philippe (1791-1844): evidence from patent data,” CEPR Discussion Papers 15620, C.E.P.R. Discussion Papers 2020.

_ , Valentina Tartari, and Matteo Tranchero, “Patterns of innovation during the industrial revolution: a reappraisal using a composite indicator of patent quality,” *Explorations in Economic History*, 2021, p. 101419.

OECD, *21st Century Technologies* 1998.

_ , *Future technology trends* 2016.

Orford, Scott, Danny Dorling, Richard Mitchell, Mary Shaw, and George Davey Smith, “Life and death of the people of London: a historical GIS of Charles Booth’s inquiry,” *Health & place*, 2002, *8* (1), 25–35.

Owen-Smith, Jason, Massimo Riccaboni, Fabio Pammolli, and Walter W Powell, “A comparison of US and European university-industry relations in the life sciences,” *Management science*, 2002, *48* (1), 24–43.

Packalen, Mikko and Jay Bhattacharya, "Cities and Ideas," Working Paper 20921, National Bureau of Economic Research January 2015.

Pakes, Ariel and Zvi Griliches, "Patents and R&D at the firm level: A first report," *Economics letters*, 1980, *5* (4), 377–381.

Peri, Giovanni, "Determinants of knowledge flows and their effect on innovation," *Review of Economics and Statistics*, 2005, *87* (2), 308–322.

Perlman, Elisabeth R et al., "Dense enough to be brilliant: patents, urbanization, and transportation in nineteenth century America," 2016. Working Paper, Boston Univ.

Peters, Matthew E, Waleed Ammar, Chandra Bhagavatula, and Russell Power, "Semi-supervised sequence tagging with bidirectional language models," *arXiv preprint arXiv:1705.00108*, 2017.

Petralia, Sergio, Pierre-Alexandre Balland, and David L Rigby, "Unveiling the geography of historical patents in the United States from 1836 to 1975," *Scientific data*, 2016, *3* (160074).

Plasseraud, Yves and François Savignon, *Paris 1883: genèse du droit unioniste des brevets*, LITEC, 1983.

Porter, Alan L, Jan Youtie, Philip Shapira, and David J Schoeneck, "Refining search terms for nanotechnology," *Journal of nanoparticle research*, 2008, *10* (5), 715–728.

Review, MIT Technology, "10 Breakthrough Technologies 2021," *MIT Technology Review*, 2021.

Righi, Cesare and Timothy Simcoe, "Patent examiner specialization," *Research Policy*, 2019, *48* (1), 137–148.

Roche, Emmanuel and Yves Schabes, *Finite-state language processing*, MIT press, 1997.

Romer, Paul M., "Endogenous Technological Change," *Journal of Political Economy*, 1990, *98* (5, Part 2), S71–S102.

Rosenthal, Stuart S and William C Strange, "Geography, industrial organization, and agglomeration," *review of Economics and Statistics*, 2003, *85* (2), 377–393.

Rosés, Joan Ramón and Nikolaus Wolf, *The economic development of Europe's regions: A quantitative history since 1900*, Routledge, 2018.

Rotolo, Daniele, Diana Hicks, and Ben R Martin, "What is an emerging technology?," *Research policy*, 2015, *44* (10), 1827–1843.

Sarada, Sarada, Michael J Andrews, and Nicolas L Ziebarth, "Changes in the demographics of American inventors, 1870–1940," *Explorations in Economic History*, 2019, *74*, 101275.

Schmookler, Jacob, *Invention and Economic Growth*, Harvard U.P., 1966.

Sekine, Satoshi and Chikashi Nobata, "Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy.," in "LREC" Lisbon, Portugal 2004.

Sokoloff, Kenneth L, "Inventive activity in early industrial America: evidence from patent records, 1790-1846," *Journal of Economic History*, 1988, pp. 813–850.

Song, Zheng, Kjetil Storesletten, and Fabrizio Zilibotti, "Growing like china," *American economic review*, 2011, *101* (1), 196–233.

Squicciarini, Mara P and Nico Voigtländer, "Human capital and industrialization: Evidence from the age of enlightenment," *The Quarterly Journal of Economics*, 2015, *130* (4), 1825–1883.

Sutton, Charles and Andrew McCallum, "An introduction to conditional random fields for relational learning," *Introduction to statistical relational learning*, 2006, *2*, 93–128.

Tarasova, Nina N. and Polina Shparova, "Top 15 Digital Technologies in Manufacturing Industry," https://issek.hse.ru/en/news/494926896.html 2021. Accessed: 2021-09-05.

Trajtenberg, Manuel, "A penny for your quotes: patent citations and the value of innovations," *The RAND Journal of Economics*, 1990, pp. 172–187.

_ , "A Penny for Your Quotes: Patent Citations and the Value of Innovations," *The RAND Journal of Economics*, 1990, *21* (1), 172–187.

Vandenbussche, Jérôme, Philippe Aghion, and Costas Meghir, "Growth, distance to frontier and composition of human capital," *Journal of economic growth*, 2006, *11* (2), 97–127.

Verluise, Cyril and Gaétan de Rassenfosse, "PatCit: A Comprehensive Dataset of Patent Citations (Version 0.15) [Data set]," *Zenodo. http://doi.org/10.5281/zenodo.3710994*, 2020.

Wagner, Stefan, Karin Hoisl, and Grid Thoma, "Overcoming localization of knowledge?the role of professional service firms," *Strategic Management Journal*, 2014, *35* (11), 1671–1688.

Webb, Michael, Nick Short, Nicholas Bloom, and Josh Lerner, "Some Facts of High-Tech Patenting," Technical Report, National Bureau of Economic Research 2018.

Weston, Jason, Samy Bengio, and Nicolas Usunier, "Wsabie: Scaling Up To Large Vocabulary Image Annotation," in "Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI" 2011.

Wikipedia, "List of Prolific Inventors," https://en.wikipedia.org/wiki/List_of_prolific_inventors 2021. Accessed: Sept. 2021.

WIPO, "The Geography of Innovation: Local Hotspots, Global Network," Technical Report 2019.

_ , "Patent Landscape Reports," https://www.wipo.int/patentscope/en/programs/patent_landscapes/ 2021. Accessed: 2021-05-26.

Younge, Kenneth A and Jeffrey M Kuhn, "Patent-to-patent similarity: a vector space model," *Available at SSRN: https://ssrn.com/abstract=2709238*, 2016.

Zhang, Shaodian and Noémie Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," *Journal of biomedical informatics*, 2013, *46* (6), 1088–1098.

Zilibotti, Fabrizio, "Growing and slowing down like China," *Journal of the European Economic Association*, 2017, *15* (5), 943–988.