Research Policy 53 (2024) 104903

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

A new dataset to study a century of innovation in Europe and in the US

Antonin Bergeaud [a,*], Cyril Verluise [b]

[a] HEC Paris, CEPR and CEP-LSE, 1 Rue de la Libération 78350 Jouy-en-Josas, France
[b] Collège de France and PSE, 3 Rue d'Ulm 75005, Paris, France

ARTICLE INFO

JEL classification:
N90
O30
Y10

Keywords:
History of innovation
Patent
Text as data

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 PatentCity, 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. This dataset has been constructed with the view of facilitating the exploration of the geography of innovation and includes additional information on citizenship and occupation of inventors.

1. Introduction

The availability of new quantitative data has led to numerous studies that analyze the social and economic implications of innovation activities and the enabling environment for strengthening innovation (see Hall and Harhoff, 2012 for a review). Most of these studies use patent documents as a means of measuring innovation across time and space. While patents are imperfect and incomplete proxies for innovation due to the heterogeneity in patenting propensity across countries, time, sectors, and firm size (see Arundel and Kabla, 1998; Mansfield, 1986), they are widely used in economic literature because of the rich quantity of information they contain. Additionally, despite their limitations, evidence shows that patents as a measure of innovation provide a relevant signal (they are, in particular, well-correlated with R&D activities, see Pakes and Griliches, 1980; Acs and Audretsch, 1989).

Patent system has been in place for a very long time. It is commonly acknowledged that the first British patent was granted to John of Utynam in 1449 (Plasseraud and Savignon, 1983) and patent publications exist since the 19*th* century in many countries (see Appendix C for a short history of patent systems and references). Yet no comprehensive structured datasets were available to researchers for patent filed prior to the 1980s. One important exception is the United States Patent and Trademark Office (USPTO) which consistently published patents since 1836 and made them publicly available.[1] 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; Lamoreaux and Sokoloff, 2000; Sokoloff, 1988) which were nonetheless limited by the quantity of USPTO documents. Recent improvements in large data handling and text data processing have stimulated a renewed interest in large scale use of historical patents (see in particular Packalen and Bhattacharya, 2015; Petralia et al., 2016; Akcigit et al., 2017; Berkes, 2018; 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.[2]

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 1980s. 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 (EPO, 2023) and IFI Claims before the 1980s. However, most historical

* Corresponding author at: HEC Paris.
E-mail address: bergeaud@hec.fr (A. Bergeaud).

[1] USPTO patent publication texts are publicly available for bulk download from the USPTO website and the Google Patents public dataset. USPTO publications existed before 1836 but a fire burned an unknown number of them.

[2] Several studies have explored patents published before the 19*th* century in other countries than the US, such as Hanlon (2016); Nuvolari and Tartari (2011); Nuvolari et al. (2020, 2021). However, these works typically concentrate on specific characteristics of a select group of patents—either constrained by region or time period—rather than aiming to glean comprehensive information from the broadest set of patent documents available.

https://doi.org/10.1016/j.respol.2023.104903
Received 11 July 2022; Received in revised form 28 September 2023; Accepted 11 October 2023
Available online 4 November 2023

patent documents are available as scanned images. Starting from these images and using a 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. Specifically, we used raw images of patent documents as our input, extracted and structured the embedded information and produced a relational database covering patents published in Germany (including East Germany), France, the United Kingdom, and the US since the 19th century.

To the best of our knowledge, our database PatentCity is the largest of its kind, both in terms of time-space coverage and scope of applications. We make it open access with open-source tools to help the community build on/extend our work.[3] 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 for other countries. We have also made the database as interoperational as possible. Each patent and its 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 and to check whether quantitative stylized facts about innovation on the long-run (such as those presented in Feldman and Kogler, 2010), essentially based on US data, are also true in Europe.

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 efforts by Lamoreaux and Sokoloff (1997), Lamoreaux and Sokoloff (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 (2018); Berkes and Gaetani (2019); Akcigit et al., 2017, Akcigit et al., 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, de Rassenfosse et al. (2019) 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.

In addition to providing information on the name and location of inventors and assignees, we also extract additional details such as the occupation and citizenship of the inventors when applicable. These are often available in the text of patents, especially for British publications and can be used to extend our understanding of who are the actors of innovation over the 20*th* century. This relates directly to a recent literature that has looked at how innovative activities have changed over time (see e.g., Akcigit et al., 2017; Berkes, 2018; Babina et al., 2020). Using information drawn from the census, Akcigit et al. (2017) and Sarada et al. (2019) 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. These studies also emphasize the role of foreign inventors. By gathering data on the citizenship of inventors, our dataset offers a complementary perspective on the global spread of innovative talents. Much of the existing literature has focused on immigrant inventors, drawing primarily from census data, as demonstrated by Akcigit et al. (2017) and Arkolakis et al. (2020).[4] Rather than serving as a direct proxy for migration, citizenship data offers a nuanced insight into an inventor's legal and socio-political ties to a country. For instance, foreign citizens living in the US could potentially be recent migrants and these data can thus be used when analyzing the impact and efficiency of shifts in citizenship laws, especially during eras characterized by extensive migration (see e.g. Diodato et al., 2022 for an analysis considering the country of origin of inventors.).

From a data perspective, our work borrows extensively 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 generalize 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 recognize 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 discusses the main steps of the construction of the dataset and we refer the reader to the Online Appendix and to the GitHub repository for more technical details. Section 3 provides an overview of the dataset and Section 4 sketches some potential applications for economic analysis. Section 5 concludes.

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

[4] For example, Arkolakis et al. (2020) found that European immigrants contributed to more radical innovations than their domestic counterparts. Similarly, Akcigit et al. (2017) noted that the unique expertise of immigrants from the 1880–1940 period led to an uptick in patenting within those specific domains from 1940 to 2000.

2. Data

We now detail the construction of the database. The key steps are the following. We start by collecting the patent document images. We convert these documents into text data using Optical Character Recognition (OCR). We then leverage modern Named Entity Recognition 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 or her location). Finally, we enrich the dataset by converting extracted natural language text spans into harmonized attributes. In particular, we geocode the locations and provide administrative codes to facilitate the interoperability of the database with other sources. Fig. A13 summarizes the workflow that we describe in detail in this section.[5]

2.1. Data collection and coverage

Contrary to the USPTO, patent publications from the German, French and British intellectual property offices are not publicly available for bulk download in text format.[6] 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. Appendix A.1 reports the list of kind codes selected as referring to utility patents for each patent office.[7] For the sake of brevity, we refer to utility patents 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 1973–1976), 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. (2019) which reports the patentees location for a very large corpus of patents, including publications from the patent 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.[8] Our dataset comprehensively[9] spans over the following periods: 1877–1980 for German patents, 1950–1972 and 1977–1992 for East German patents,[10] 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. (2019)'s which provides data up until 2013 included.

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 that resolves the relations between extracted entities. Both layers are crucial to fully exploit the potential of patent texts.

2.2.1. Main challenges

Constructing structured data from patent text presents a significant challenge due to the vast variability in document formats. One of the main difficulties lies in establishing a strategy that can effectively extract relevant information, such as the inventor's name or geographical location, which is often presented in varying formats across different patent offices and countries, and even over time. In the case of the US, Berkes (2018) and Petralia et al. (2016) discuss in details how the changing structure of patent documents requires to adapt the rules used to extract information. In our case, patent document formats can vary greatly across different countries, which makes it inefficient to use rules that rely on the structure of the document (Table 2 provides some examples).

2.2.2. 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 1 and discussed in Appendix A. 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' nationality and location provides their citizenship 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

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

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

[7] 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.

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

[9] 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.

[10] To our knowledge, digitized copies of East German patent documents published between 1973 and 1976 are not available. However, recent efforts have been made to bridge this gap, see Hipp et al. (2022).

Table 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 A 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. (2019) 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.

any company related to the invention.[11]

Table 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 A.4 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 Fig. A7.

2.2.3. 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 preamble or in the header of the document. Table 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 our specific case, they also present the advantage to have considerable generalization abilities based on a relatively small number of examples making them robust to typographical errors 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 released manually annotated data which supports rigorous and transparent performance evaluation and future extensions.[12]

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 A.2). More details are provided in Appendix D.

In Table 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' 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.

2.2.4. 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 essentially a "bag" of entities. For example, assuming that we have extracted one inventor, one assignee and two locations, we still don't know how these entities are related to one another. Such relationships can be extremely detrimental to the analysis. For instance, if we want to know whether an inventor is an non-citizen, we need to link their name to a citizenship and to a location. This case of multiple patentees in a given publication is a well identified additional difficulty to the conversion of unstructured patent documents into a set of entities (see Berkes, 2018). For this reason, we go one step further and reconstruct the latent relationships between our different entities. That is what we call relationship prediction.

In our case, there are three different kinds of relationships: the *location* which relates the patentee to her address, the *occupation* which relates the patentee to her occupation, or academic title and the *citizenship* which relates the patentee to 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 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 former. As

[11] This is a necessary but arbitrary point which has important implication for comparability across countries (see also Section 3.3 for more on this point). 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 A.3 for more details. Additionally, some patents mention additional individuals such as the name of the patent attorney, the representative of the inventor, witnesses etc… In theory, the NER model has been trained in such a way that these entities are not labeled as either inventors or assignees. As a result, the model should not mistakenly classify these entities as inventors.

[12] 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`.

Table 2
Example of patent documents with embedded entities.

Country	Example	Source
DD	*Erfinder: Wilhem Ulrig, 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

Examples of patent document for each of the five 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.
Note to publisher: ideally the color in the notes (red, purple, olive, brown and blue) would be written with the corresponding color.

Table 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 A.2). Performance metrics are reported as follows: precision/recall/F1-score. Model by model performance for each patent offices can be found in Appendix D.

Table 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 A.2). 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 D.

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.

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 of 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 as well. In this section, we detail how we overcame these limitations and the resulting data enrichment process.

2.3.1. Location geocoding

Our first task is to turn natural language attributes into high quality and harmonized 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 typographical errors that can be found in natural language addresses and the size of the universe of

Table 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.

worldwide addresses. In our case, there are the additional difficulties of multiple languages and changing names and borders over the time span considered. 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 of historical entities which have changed names (e.g., "Karl-Marx Stadt" in East Germany now known as "Chemnitz"). This is however sometimes done 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 these specificities 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.[13] The two outputs were relatively straightforward to align in a common data structure.

Table 5 presents the share of matched locations together with the level of quality of the geocoding (conditional on match). The 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.

2.3.2. 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.[14]

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 typographical errors. Confronting the output with a set of manually annotated citizenship values, 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.

2.3.3. 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,[15] 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.

3. Overview of the dataset

Having described the construction of PatentCity, we now describe how the dataset can be used and and emphasize the importance of paying attention to differences between countries and over time before using the data to make comparisons.

3.1. 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 A.7. 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 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, Commuting Zone (CZ) and state codes for the US.

The database version we offer includes all patents that we collected with a kind code specified in Table A1, which pertains to utility patents.

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

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

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

This database features over 16 million unique publication numbers; however, it contains multiple duplicates because a single patent can have several publications (e.g., first publication, second publication, reissue, and so on). Researchers who want to study patents at a specific stage might prefer to restrict the dataset to a corresponding set of kind codes. In general, most users of the database will likely need to eliminate duplicates and retain only one observation per patent, typically the earliest one. In Appendix A.1, we provide a simple procedure to accomplish this task.

3.2. Coverage

Benchmarking with commercial datasets In order to study the coverage of PatentCity, we compare it to two standard patent databases that are typically used in the literature: PATSTAT and IFI Claims (which is often referred to as Google Patent). Fig. 1 reports the share of yearly observations in PatentCity compared to IFI Claims for publications that fall within the criteria defined in Table A1 (utility patents). Fig. A11 in the Appendix provides a similar comparison with Patstat. Given that these two datasets exhibit very similar coverage for the period we considered, the following discussion will primarily focus on our benchmarking with IFI Claims.

Overall, our coverage of patent documents is very high, and in some cases, exceeds that of IFI Claims. For example, in Germany, we have been able to recover missing publication dates before 1920 and for East Germany between 1952 and 1989, resulting in a higher number of observations. Some documents are however still missing, in particular after 1980 in France and Germany due to the data provided by patent offices to de Rassenfosse et al. (2019). In Appendix A.4, we provide additional details on the coverage of our dataset and in particular the share of patents for which we detect at least one inventor, and similarly for all the entities that we extract. In particular, Figs. A3 show that not all patents are associated with a location. This is generally due to the fact that during some subperiods,

geographical information can be often missing from the patent publications (for example in France during the 1970–1980 period).

Benchmarking with other research datasets When assessing the extent of our coverage at the USPTO, it can be useful to compare our results to those of existing datasets. As discussed, as a benchmark, we primarily rely on the Histpat dataset (Petralia et al., 2016), which is readily available for download and widely recognized for its quality. Histpat provides detailed information on both domestic and foreign patentees for USPTO patents up until 1975 and has been compared to other similar projects in Andrews (2019). Our analysis indicates that the coverages of PatentCity and Histpat are nearly identical for the period spanning from 1836 to 1975, differing by less than 0.1 %. This is not surprising given that both datasets are sourced from the same USPTO Bulk Data Storage System. Furthermore, the two datasets are highly consistent in their classification of inventors and assignees by country, with approximately 92 % of patents having the same location in both datasets.

For a more detailed assessment, we manually examined 350 patents that were filed with the USPTO and present in both PatentCity and Histpat, as well as in Berkes (2018)'s CUSP dataset.[16] While a comprehensive evaluation of the geocoding accuracy of these three datasets would require further investigation, we can report that the quality of CUSP, PatentCity, and Histpat appears to be very comparable. Specifically, we found that for more than 90 % of the 350 patents examined, all three datasets identified the county-level location.

with consistency. This number appears consistent with the one reported in Gross and Sampat (2023) (see their Appendix B). Notably, our examination of the 350 patents revealed that errors in the three datasets seemed to have varying origins. For Histpat, the most common source of errors arose from inconsistent pairs of county-city. In contrast, CUSP frequently experienced issues with homonymous cities located in different states, while PatentCity primarily encountered problems with incorrectly geocoded entities. These different types of errors can be explained by the different strategies considered to construct each dataset. Histpat uses a large dictionary of locations to identify all geographical entities from the text of a patent and then trains a statistical model to select the best candidates using a set of manually encoded patents. CUSP employs a complex mixture of rules targeting specific locations in the patent publication and relies on the consistency of the USPTO publications' format for given subperiods. By contrast, PatentCity does not use a priori rules or a location dictionary but relies on a named entity recognition algorithm trained on manually labeled patents. In spite of these differences, the overlap between Histpat, CUSP and PatentCity showcases an very high precision rate (exceeding 99.9 %) when limited to the patents on which two of the datasets concur (see Gross and Sampat, 2023, Appendix B). This suggests that that amalgamating these diverse approaches could potentially enhance precision, a strategy that is discussed in Abramitzky et al. (2020) to link historical data with each other. Finally, note that a manual review of the 10 % of locations with discrepancies suggests that both CUSP and PatentCity generally have a slightly lower error rate than Histpat.

3.3. Comparison across countries and over time

The final format of PatentCity facilitates easy comparisons of patent numbers across various patent offices over time, as illustrated in Fig. A11. Additionally, it enables comparisons across diverse technologies, offering insights into their long-term development. It is worth noting that even though historical patents typically have their own classifications, we chose not to incorporate them into PatentCity. The reason is that patent offices often retroactively assign technological classes based on modern classification systems, namely the CPC or IPC and we decided to employ these contemporary classifications that are consistent over time.

Fig. 2 serves as an example of how this data can be melded. It shows the distribution of patents among different patent offices by their 1-digit IPC technological classes. From this Figure, we can clear see that the USPTO has a different composition than European patent offices. Specifically, it leans more towards patents in Physics and Electricity, while European countries show a pronounced focus on performing operations and mechanical engineering. Going further, the IPC classification provides a detailed lens, allowing comparisons of technological evolution. We illustrate this using the example of Germany and consider trends in the patent shares of three representative technologies: Weaving (which has seen a decline), Combustion Engine (dominant until the 1990s), and Computing and Calculating (which is on an upward trajectory). These findings are presented in Fig. 3.

Lastly, Fig. 4 showcases the integration of this data with the geographical dimension of PatentCity. It presents a map highlighting the average share of patents associated with at least one IPC class linked to Combustion Engines (F02). Distinct regional clusters emerge from this visual, notably in areas like Stuttgart, Bavaria, and Wolfsburg.

These exercises could enable a deeper understanding of the life cycle of different technologies, allowing comparisons across time and regions. The geographical and technological details can be useful in pinpointing the emergence and demise of regional industrial clusters. However, using patent data to compare countries and periods presents inherent challenges. The definition of a patent, its legal scope, and the barriers to obtaining one can differ markedly from one decade to another, or from one country to another. Comparing patent counts on an international scale using historical data is particularly difficult, given that patent laws were not as harmonized in the past as they are in the present (Moser, 2013). A notable distinction lies in the cost of obtaining a patent and the

[16] We are grateful to Enrico Berkes for providing information on these 350 patents. For a deeper dive and evaluation of earlier datasets, we direct readers to Andrews (2019).

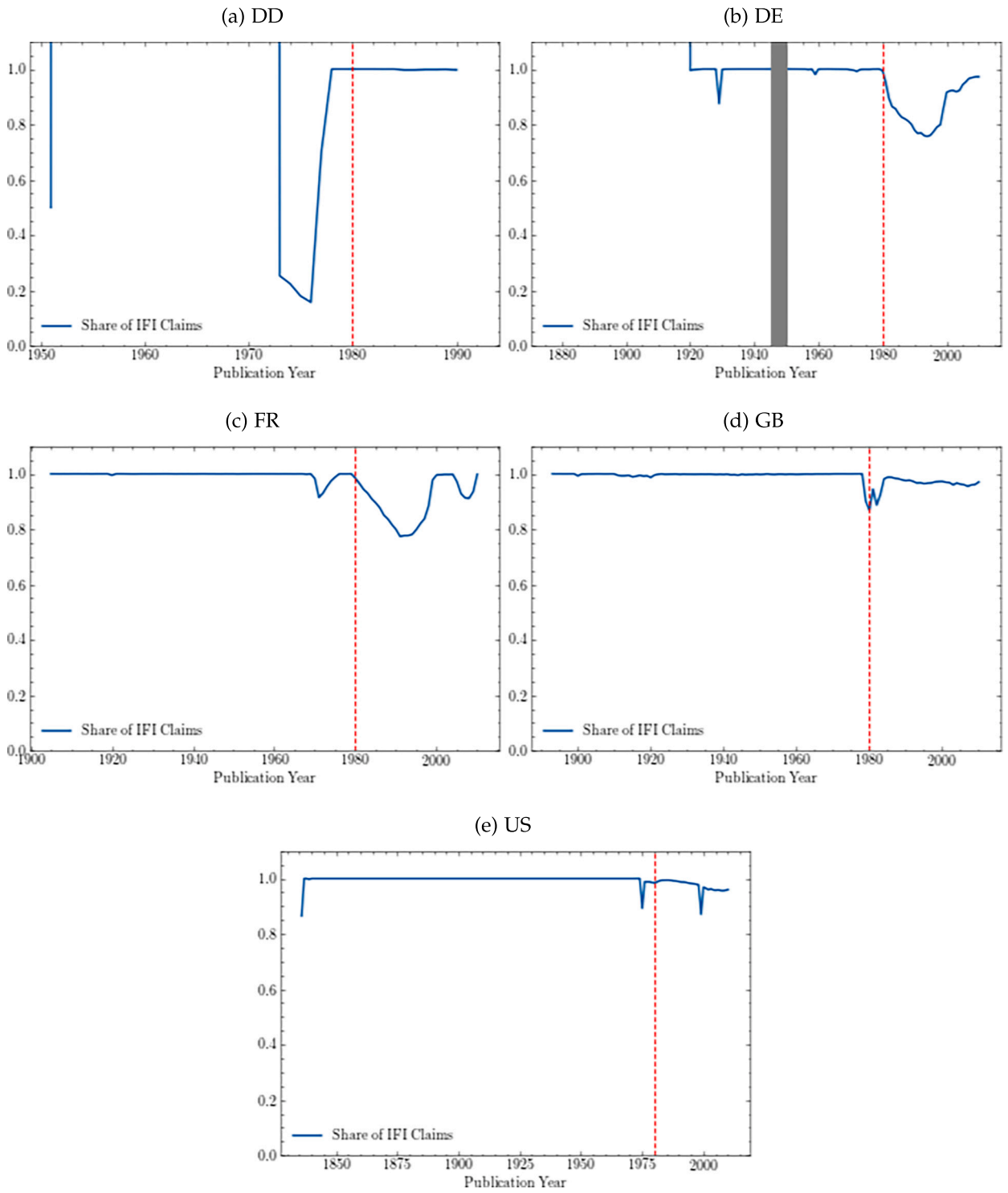


Fig. 1. Number of patents in PatentCity compared as a share of IFI Claims (Google Patents). **Notes:** these Figures report the share of patents included in PatentCity as a share of the number of patents included in IFI Claims. The vertical line indicates 1980, the beginning of the switch from PatentCity to de Rassenfosse et al. (2019).

rigor of the intellectual property system, both of which can fluctuate significantly over time and across various patent offices. Khan and Sokoloff (2001) describe the different philosophies of various patent offices at the time of their creation and gives specific examples to caution against drawing hasty conclusions. For instance, they emphasize that while the German system was influenced by the USPTO it was generally stricter, resulting in fewer patent grants but with a likely higher average quality. They also presents the USPTO as being guided with the general policy of keeping patent fees particularly low compared to France or the United Kingdom in the 19*th* century, and even for a large part of the 20*th* century (De Rassenfosse and van Pottelsberghe, 2013). And even in a given country, these fees can suddenly collapse, resulting in a higher propensity to patent and in a quick increase in the number of new publications (Nicholas, 2011). For example, in Britain in 1852, the fee for obtaining a patent was dramatically reduced from 100 pounds - the average annual wage of a skilled worker - to 25 pounds, and then further reduced to 4 pounds in 1883, and finally to 1 pound in 1905 (Van Dulken, 1999). Appendix C provides a summary of the significant

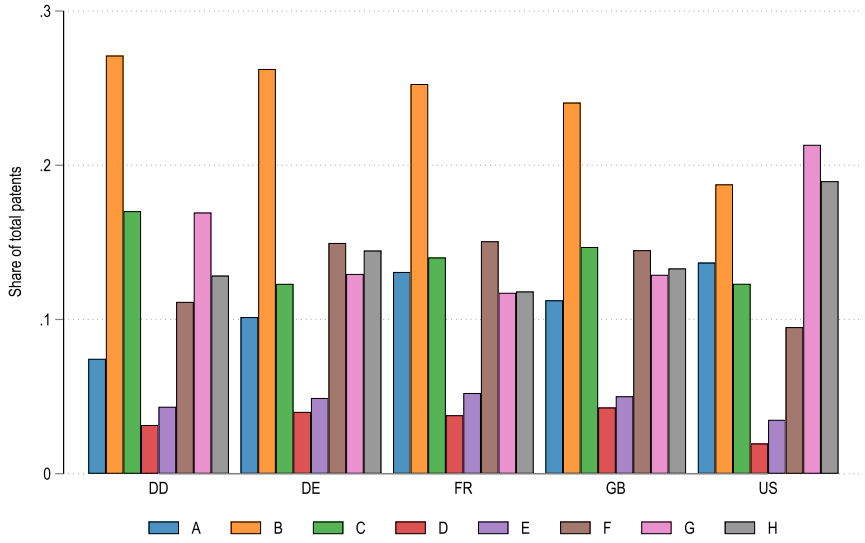


Fig. 2. Share of each 1-digit IPC in each patent office. **Notes:** This figure displays the share of each 1-digit IPC technological class among all patents filed in each of the five patent offices over the entire period covered by PatentCity. When a patent has multiple IPC codes, we allocate a fractional share accordingly. IPC codes correspond to: A Human Necessities; B Performing Operations /Transporting; C Chemistry/Metallurgy; D Textiles /Paper; E Fixed Construction; F Mechanical Engineering /Lighting /Heating/Weapons /Blasting; G Physics; H Electricity.

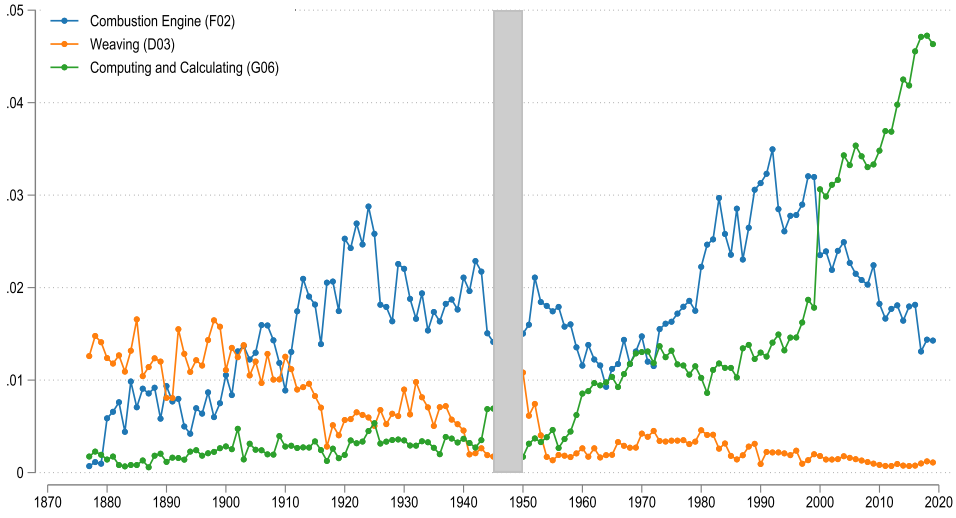


Fig. 3. Share of patent in 3 IPC classes in the German patent office. **Notes:** Share of patents in three 3-digit IPC technologies (Combustion Engine, F02; Weaving, D03; and Computing and Calculating, G06) for each publication date. When a patent has multiple IPC codes, we allocate a fractional share accordingly. All patents filed in the German patent offices (either East or West Germany) are included.

changes in patent laws in various patent offices during the 19*th* and 20*th* centuries. Patent offices have followed their own paths from their initial stages which were deeply rooted in domestic social, philosophical, and economic characteristics, to progressively converged around 1980 to more harmonized procedures and definitions.

Besides this challenge, comparisons of the extracted entities across countries can also be tricky. Patents do not always include the same information or level of detail. One illustration of this is the nature of the patentee. In the case of the US, inventors and assignees are clearly separated and declared as such. Almost all patents have an inventor, and this has been consistent over time (see Fig. A1). In contrast, assignees are very rarely mentioned before 1924 and became then increasingly common (see Fig. A2). This simple distinction is not as straightforward in other patent offices, and our definition of an assignee or an inventor has been adapted accordingly (see Appendix A.3). Another important distinction to note is the definition of "occupation" in German and British patents (see also Section 4.2). In British patents, inventors sometimes explicitly state their occupation in the patent's preamble. In contrast, German inventors often indicate their education and field of study through an academic title preceding their name. Although these provide different types of information, we have labeled both as occupations in PatentCity.

In summary, researchers interested in using PatentCity to produce a comparative analysis of patenting over time, technology or space should be cautious about correcting for these differences. For example, to allocate a patent to a region, it might be useful to first consider the location of the inventor, and in case the inventor is missing, to look at the assignee.

In the next sections, we look at three different possible uses of the database in more details exploiting the various entities extracted from the patent documents.

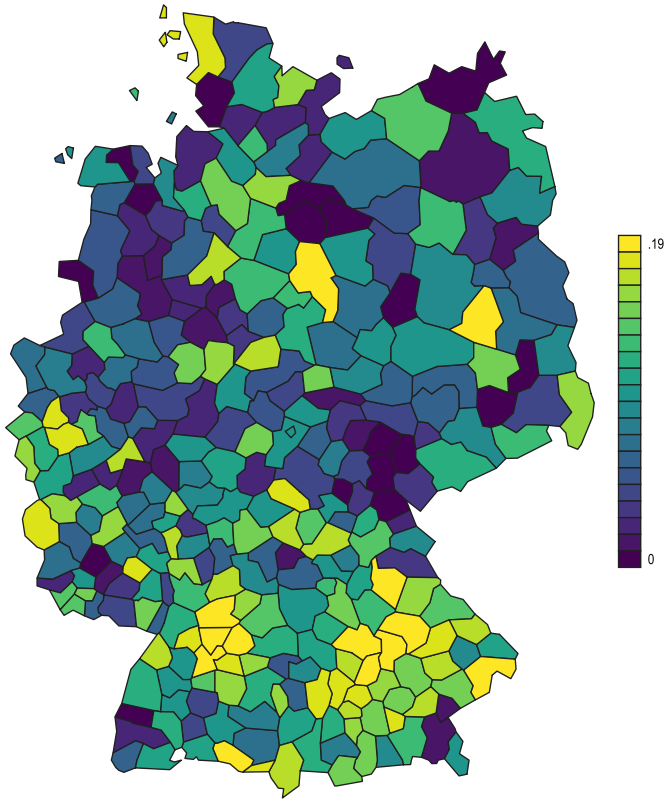


Fig. 4. Share of "Combustion Engine" patents by region. **Notes:** Share of patents with at least one IPC-3 digit code equal to F02 (Combustion Engine) in each German's NUTS 3 region. The share is calculated over the full period 1877–2014. All patents filed in the German patent offices (either East or West Germany) are included.

4. Overview of the entities

We now discuss in more details the three types of entities extracted from the patent files: location, occupation and citizenship and give simple example applications.

4.1. Geographic distribution of patents

The first and more natural usage of PatentCity is to analyze the geography of innovation. Numerous studies have looked a this question and have usually reported that innovative activities are highly concentrated, even when population density is taken into account (see Feldman, 1994; Feldman and Kogler, 2010 for a comprehensive review). With geographic information provided for each patentee, PatentCity can aid in identifying potential variations in the spatial distribution of innovation different periods and in different countries which could open doors for future investigations into the concentration of innovation over time and space.

Section 2.3.1 details the level of granularity achieved through our geocoding process (see also Fig. A7 in Appendix A). However, this Table 5 includes all patentees, whether domestic or foreign. Restricting the dataset to domestic inventors and assignees increases the average granularity significantly as foreign patentees sometimes only report their country. More than 99 % of domestic patentees are located at least at the county level (counties in the US and NUTS3 regions in other countries), with the exceptions of East Germany (98 %) and France (90 %).

This "county" level of aggregation is particularly useful as it usually allows to confront the number of patents with other information, drawn from example from the Census, such as population, income, education etc… To illustrate this, we constructed local population estimates from various sources and for each of our four countries (this required some minor border adjustments, see Appendix A.5 for more details and sources). From this we report in Fig. 5 the logarithm of the number of patentees (regardless on whether they are assignees or inventors) and the logarithm of total population all taken as an average over time and for each of the 4 countries (pulling together East and West Germany).[17] It shows that a well-known result for the US, that inventors and assignees are mostly located around large urban areas (Audretsch and Feldman, 1996; Feldman and Kogler, 2010), is also true in other European countries. For example, the urban area of Paris accounts for 45 % of all domestic patentees over the period 1900–2014, but only little more than 10 % of the country's population in 2014. In the US, the six counties that make up the Silicon Valley account for 10 % of all patentees over the same period for less than 1 % of the population. This is also true for the UK as Inner London counts 27 % of patentees for 5 % of the population. The innovation in Germany is more uniformly distributed but large cities like Berlin or Munich concentrated an important share of the country's innovation activity over the 20*th* century. To show more clearly the joint distribution of population and patenting across regions, we also map the average value of patentees per capita in Fig. 5.[18] These maps spotlight regions that significantly outperform in terms of innovation (at least when compared to what their population would predict). Conversely, they also draw attention to underperforming areas: East Germany, Northern England and Scotland, the Southern US, and Northern France. An investigation into the root causes of these differences across countries is beyond the scope of this paper, however Figs. B.

County level analysis already provides a very granular picture of the geography of innovation which is likely to be sufficient for many analysis. However, the level of precision can be much finer in the case of British patents and 85 % of patentees are located at the street or even house number level. This offers a very micro perspective on the location of inventors or assignees. We illustrate this in Fig. 6 which reports the exact location of patentees in London. This Figure shows that most of the assignees are located in central London while inventors' locations are more widespread. Information at this level of precision can be useful for researchers interested in studying the role of the development of infrastructure to foster innovation, local technological clusters or the link between wealth and innovation.

4.2. Occupation of inventors

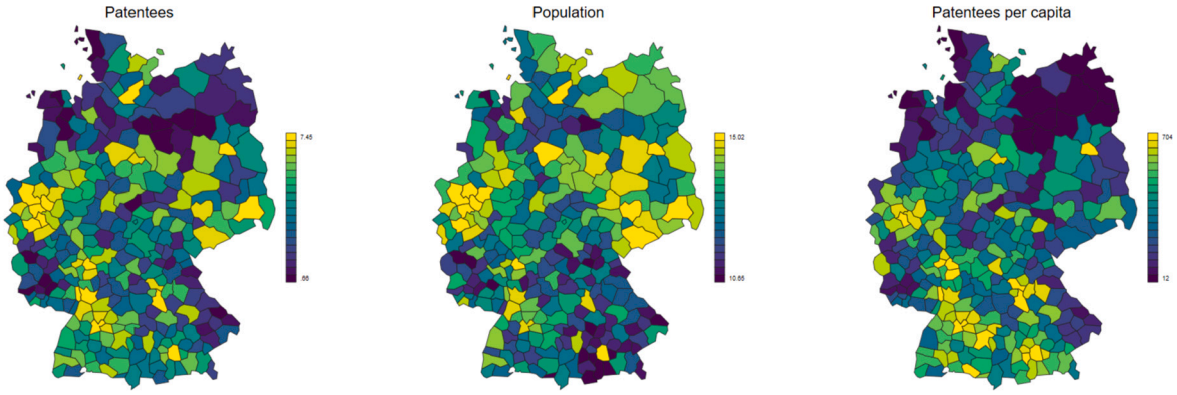
Patents filed in the UK patent office at the beginning of the 20*th* century frequently report the occupation of the inventor. This represents a new source of information to document the professional activities of inventor and how this evolves over a 30-year window.

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 Jeweler", "Cemetery mason" or "Artificial limb manufacturer", or vaguer 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 includes 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", "gentlemen" or "inventor".

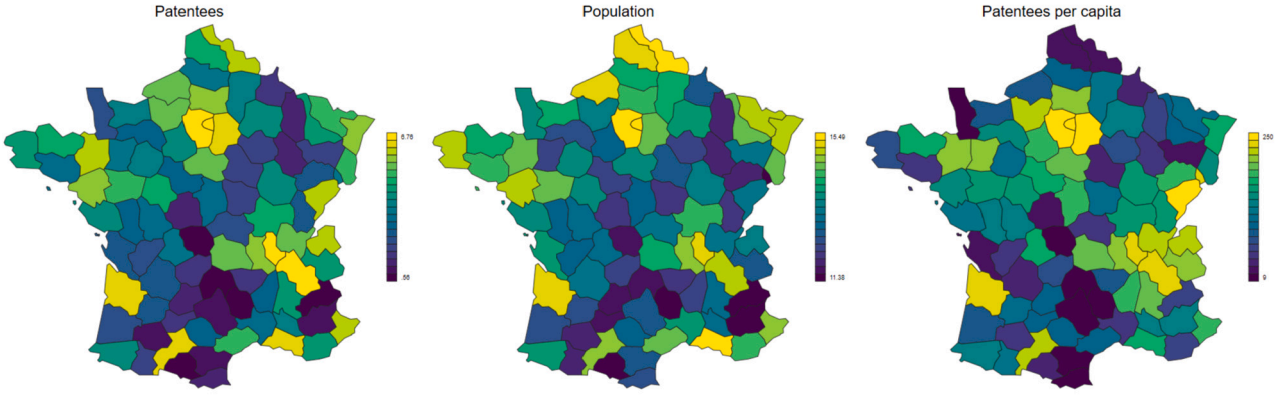
[17] To draw these maps, we have assigned the same weight to any patentee regardless of the number of inventors and assignees in the patent. Using a fractional count (i.e. only counting a fraction of the patent equal to 1 over the number of patentees) does not affect the results meaningfully.

[18] While all these Figures consider the data without any restriction on the year of publication of the patent, one advantage of PatentCity is that it offers enough historical depth to study the evolution of these pictures over time. This is what we do in Appendix B for every decade.

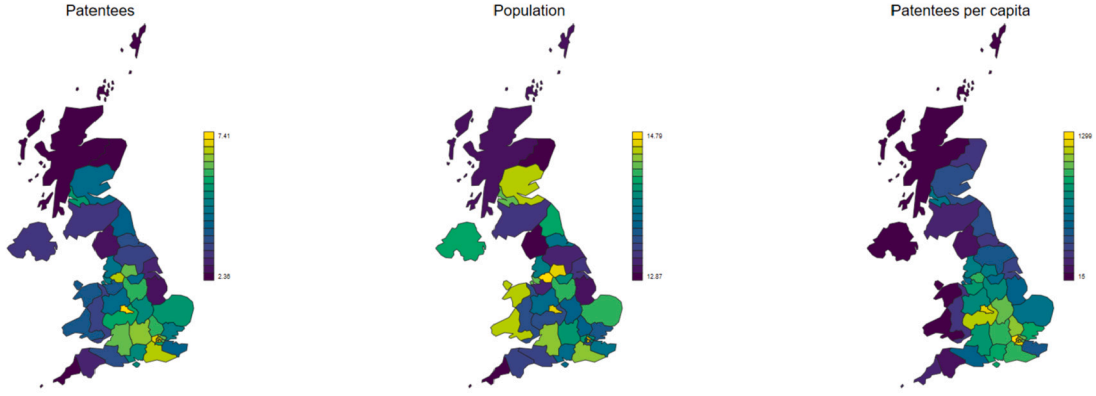
(a) Germany



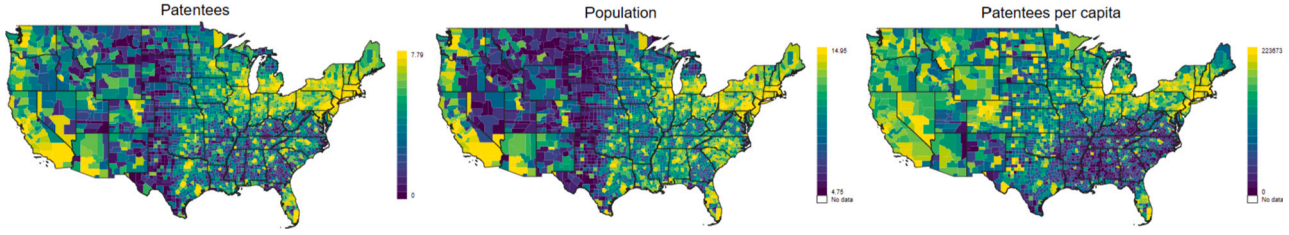
(b) France



(c) United Kingdom



(d) United States



(caption on next page)

Fig. 5. Patentee location and population at the county level. **Notes:** these figures maps the number of patentees (whether assignees or inventors), in log, total population in log and the number of patentees divided by population (in millions) for each county in Germany, France, the UK and the US. In Germany and France, a county is a NUTS3 region with minor border adjustments explained in Appendix A.5. In the UK, we used NUTS 2 regions. All variables are taken as yearly averages over the full period (see Table A1). West and East Germany are taken together as a single patent office when applicable. The number of patentees is taken as a total over the full set of domestic patentees that are located at least at the county level without restriction on the time period. One patentee is given the same weight regardless of the number of inventors and assignees in the patent. Using a fractional count does not affect the results meaningfully.

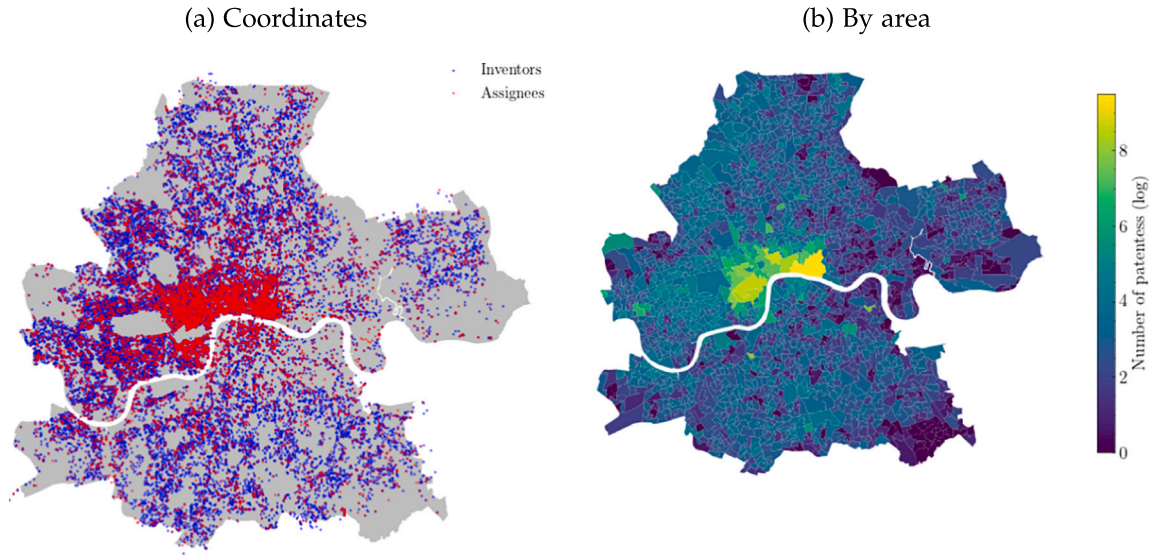


Fig. 6. Location of patentees in London. **Notes:** these figures maps the location of inventors and assignees of the UK patent office that are located in Inner London and for which the geocoding has been done at the street or house number level. Left-hand side map shows the coordinate of the house number reported or the centroid of the street. Right-hand side shows the number patentees (in log) by Lower Super Output Area.

4.2.1. *Reporting of occupation*

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 Fig. A5 in Appendix A.4. There is no legal obligation to disclose one occupation (Van Dulken, 1999, chapter 4.7) and this seems to be a practice inherited from earlier patents (MacLeod, 2002) which stopped around 1920. As explained in Van Dulken (1999), the occupation is often consistent with the nature of the innovation patented or the company's name.

Regarding inventors who did not choose to disclose their occupation, we follow Hanlon (2022) and characterize these occupations as "unknown". The corresponding patents do not seem to differ from others: the correlation between the relative weights of each technological class[19] in the groups of patents where inventors disclose their occupations and the other group over the period 1894–1920 is 96 %. Similarly, the correlation in the relative weights of NUTS 3 regions is 88 % for domestic patentees.

4.2.2. *The rise of engineers?*

Hanlon (2022) recently examined British patents from 1700 to 1854 to explore the increasing importance of engineers in the country's technological landscape. With occupation information available in PatentCity, we can conduct a similar analysis for various occupation groups over a later time period.

Fig. 7 shows the share of patents with at least one inventor declaring an occupation in the following groups: engineer, manager, manual worker, and gentleman. The data indicates that the percentage of patents involving an engineer increased from approximately 20 % to over 30 % between 1895 and 1920. During the same period, fewer patents involved at least one manual worker. While the share of patents with an inventor reporting "gentleman" as an occupation decreased from 4 % to 2 %, the share of patents with a manager increased from 2 % to 5 %, albeit at a much lower level.

4.2.3. *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 some higher education.[20] Fig. 8 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, Fig. 8 shows that the share of patents involving an inventor who reports 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.

[19] We use the International Patent Classification system at the 3-digit level, which counts 114 different categories.

[20] As already explained, this information does not directly relate to the occupation of the inventor but rather to its education level. Since many scholars consider occupation to construct a measure of the skill of workers, we chose to label this entity as occupation, i.e. under the same category as actual occupations reported in British patents.

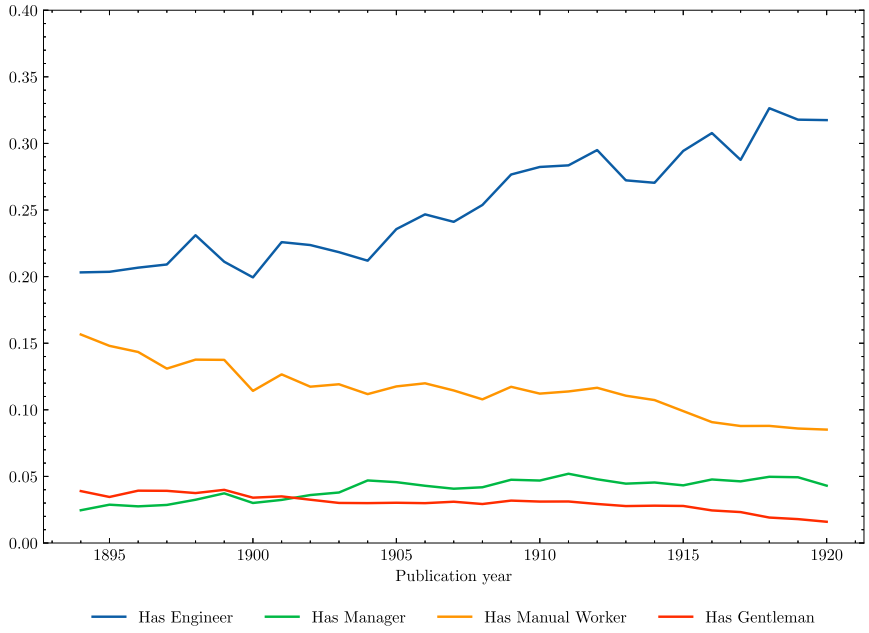


Fig. 7. Occupation of inventors in the United Kingdom. **Notes:** This figure reports the share of patents involving at least one engineer (Has engineer), one manager (Has manager), one manual worker (Has manual worker) or one gentleman (Has gentleman) in terms of the occupation of the inventor reported in the text. Time period: 1894–1920.

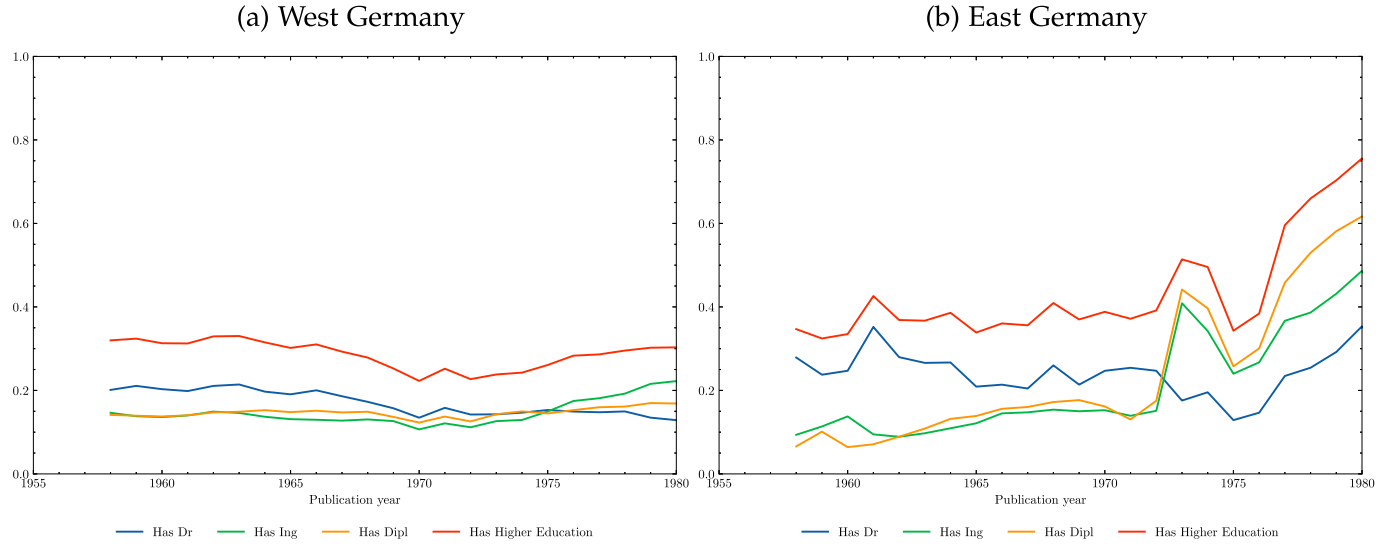


Fig. 8. Share of inventors with an academic title in Germany. **Notes:** This figure reports the share of patents with at least one inventor declaring an academic title: Doctor (Has Dr), Ingenior (Has Ing), Diploma (Has Dipl). We also define "Has Higher Degree" as the union of the previous variables. Time period: 1958–1980 for West Germany and 1965–1980 for East Germany.

4.3. Citizenship

Inventors typically demonstrate a significant degree of international mobility (Akcigit et al., 2016). Through PatentCity, we extract information about the citizenship of inventors from US and British patents. This offers a unique window into two distinct periods—1920-1950 for the UK and 1880–1925 for the US—during which patent documents explicitly stated both the citizenship and location of certain inventors.

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. For the US, this share is around 37 % between 1880 and 1925 but is closer to 45 % after 1910 (see Appendix Fig. A6).

We find that between 4 % and 5 % of inventors who report an address in the US but are *not* Americans. In a recent work focusing on the USPTO, Diodato et al. (2022) reports a similar order of magnitude. In the United Kingdom, this share is lower, between 1 % and 2 %, at any point in time between 1920 and 1950. In Fig. 9, we report this share every year for the two countries. We can see that the US experienced a sizeable increase in the share of non-citizen inventors during the 1910s. The United Kingdom experienced a similar upswing during the 1940s.

4.3.1. Citizenship in PatentCity and immigration

By nature, the citizenship status in PatentCity deviates from the Census-based definition of an immigrant, which classifies an immigrant

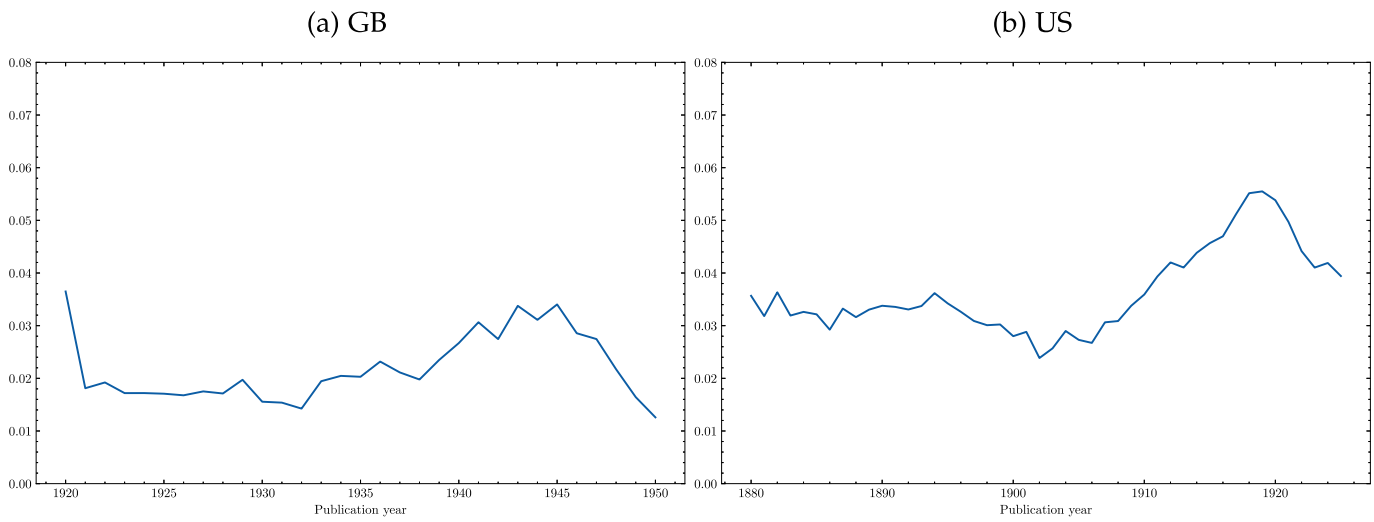


Fig. 9. Share of non-citizen inventors over time. **Notes:** The share of non-citizen inventors is computed as the ratio of the number of inventors who report a non-domestic citizenship different over the number of inventors reporting a domestic address. Time periods: 1920–1950 (GBR) and 1880–1925 (USA).

as an individual born abroad. This distinction is particularly apparent in studies such as those conducted by Sarada et al. (2019); Arkolakis et al. (2020); Akcigit et al. (2017) for at least three reasons. First, an inventor residing in the US or the UK but declaring to be a citizen of another country may only be a temporary visitor without any plan to settle permanently.

Second, individuals who have obtained citizenship prior to filing their patents may not be identified as having a foreign citizenship in our data but would be considered as immigrant in many studies. This second point is in particular critical in the US as the period during which USPTO patents sometimes report the citizenship status of the inventor corresponds to the "age of mass migration" during which naturalization was relatively easy to get in the US (typically after five year of residency). This could account for the lower shares of immigrant inventors reported in PatentCity or Diodato et al. (2022) compared to Akcigit et al. (2017) or Sarada et al. (2019). This also means that the number of foreign citizens reported in PatentCity might be influenced by the stringency of citizenship laws in the US and in the UK and not only by changed in immigration rates.

Third, our method for determining citizenship is based solely on whether the patentee disclosed their citizenship within the patent's text. From our review of regulations overseeing the USPTO from 1880 to 1925 (e.g. Khan and Sokoloff, 2001) and the British patent offices in the 1920s (Van Dulken, 1999), we did not identify any specific requirements or stipulations pertaining to inventor citizenship, such as the imposition of higher fees or other specific mandates.[21] This has led previous attempt (e.g. Diodato et al., 2022; Campo et al., 2020) to assume that a non-reported nationality in the USPTO corresponds to a US citizen. To check this, we proceed as in Section 4.2. Specifically, we compared the distribution of technological classes and regions for domestic patents without disclosed citizenship (A) with patents filed by migrant inventors (B) and national citizen inventors (C) for the USPTO from 1880 to 1925 and the British patent office from 1920 to 1950. The correlation between A and C was higher than 99 % in the US and equal to 97 % in the UK for technological classes, and 97 % and 95 %, respectively, for regional shares. The correlations between A and B were 94 % in the US and 95 % in the UK for technological classes and 92 % and 90 %, respectively, for regional shares. Thus, patents without reported citizenship appeared to be closer to those filed by national citizens than those filed by immigrants.

In summary, the entity reporting the citizenship in PatentCity should be interpreted as evidence that the inventor has recently settled in their country of residence or have lived there without having the nationality. This is therefore a measure of the citizenship status of the inventor at the time of the patent publication. This definition differs from the conventional approach used in existing literature for identifying immigrant inventors. However, it offers the potential to yield novel perspectives and insights. For instance, inventors who have not acquired citizenship may exhibit even lower degrees of integration into social networks of natives relative to other immigrants who have secured citizenship. Furthermore, our methodology presents the advantage to circumvent the necessity for implementing intricate matching procedures with external data to determine immigration status, a process typically reliant on the inventor's name and geographical location.

Finally, note that the citizenship entity can also be linked to assignees when a firm declares to be established under the commercial laws of a given country (see Appendix A.3) an information that is absent from other datasets.

4.3.2. *Innovation and citizenship status*

Fig. 10 reports the evolution of the composition of these inventors by country of citizenship for the 10 most frequent nationalities respectively in the United Kingdom and the US. As expected, Europeans constituted the bulk of inventors (consistently between 70 % and 90 %) in the US.[22] 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 citizens represented up to around 40 % of immigrant inventors.

5. Conclusion

In this paper, we have presented a novel dataset constructed from an

[21] Although such rules have applied in some cases in earlier periods, see Appendix C.

[22] The list of the most represented citizenship is similar to the one shown in Diodato et al. (2022)

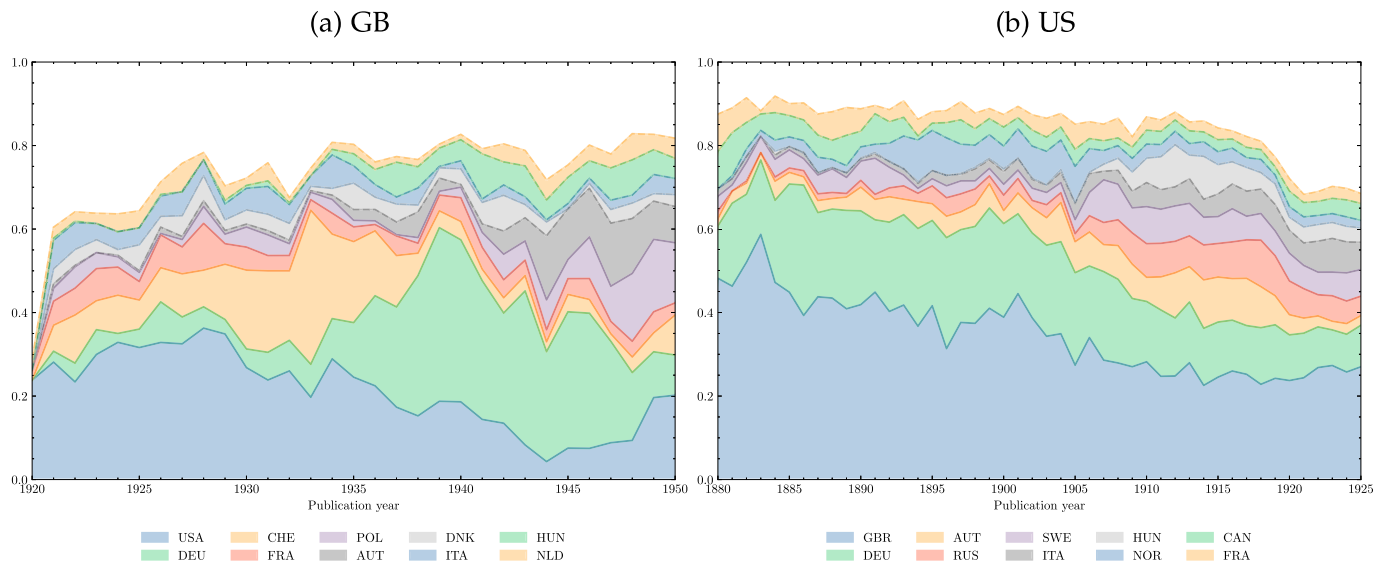


Fig. 10. Composition of inventors' citizenship. **Notes:** Each area represents the share of top 10 most frequent citizenship in the set of detected non-citizen inventors in US (left-hand side) and British (right-hand side) patents. The remaining (blank) area represent the remaining citizenship. Time periods: 1920–1950 (GBR) and 1880–1925 (USA).

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 and we presented three examples of future avenues using information on the geography of the patent, the occupation of the patentee and its citizenship.

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 many places that are important R&D actors: Japan, Sweden, Switzerland… The methodology presented in this paper has been designed with the goal of limiting future efforts to apply it to new patent corpus. We also hope that making the codebase open source will support a collective data design and continuous improvement momentum.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Antonin Bergeaud reports financial support was provided by Google for Education.

Data availability

Data are openly available

Acknowledgement

We are especially grateful to the Farhi Innovation Centre at Collège de France without which this project would not have existed. We are also indebted to Benjamin David and Aymann Mhammedi for outstanding research assistance and we thank Juliette Coly and Francesco Gerotto for their help during the first steps of the project. The Banque de France and Michel Juillard provided computational resources and technical support. We acknowledge financial support from Google for Education and Google Maps. The project also benefited from insightful comments and help from Philippe Aghion, Jérôme Baudry, Enrico Berkes, Nick Bloom, Gaia Dossi, Gaétan de Rassenfosse and John van Reenen.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.respol.2023.104903.

References

Abramitzky, Ran, Mill, Roy, Pérez, Santiago, 2020. Linking individuals across historical sources: a fully automated approach. Historical Methods: A Journal of Quantitative and Interdisciplinary History 53 (2), 94–111.
Acs, Zoltan J., Audretsch, David B., 1989. Patents as a measure of innovative activity. Kyklos 42 (2), 171–180.
Akcigit, Ufuk, Baslandze, Salomé, Stantcheva, Stefanie, 2016. Taxation and the international mobility of inventors. Am. Econ. Rev. 106 (10), 2930–2981.
Akcigit, Ufuk, Grigsby, John, Nicholas, Tom, 2017. Immigration and the rise of American ingenuity. Am. Econ. Rev. Pap. Proc. 107, 327–331.
Akcigit, Ufuk, Grigsby, John, Nicholas, Tom, Stantcheva, Stefanie, 2018. Taxation and Innovation in the 20th Century. Working Paper 24982,. National Bureau of Economic Research September.
Andersson, David E., Tell, Fredrik, 2018. Dependent Invention and Dependent Inventors. Uppsala University mimeo.
Andrews, Michael, 2019. Comparing Historical Patent Datasets. Mimeo University of Iowa.
Arkolakis, Costas, Lee, Sun Kyoung, Peters, Michael, 2020. European Immigrants and the United States' Rise to the Technological Frontier. mimeo Yale.
Arundel, Anthony, Kabla, Isabelle, 1998. What percentage of innovations are patented? Empirical estimates for European firms. Res. Policy 27 (2), 127–141.
Audretsch, David B., Feldman, Maryann P., 1996. R&D spillovers and the geography of innovation and production. Am. Econ. Rev. 86 (3), 630–640.
Babina, Tania, Bernstein, Asaf, Mezzanotti, Filippo, 2020. Crisis Innovation. Working Paper w27851,. National Bureau of Economic Research.
Berkes, Enrico, "Comprehensive Universe of U.S. Patents (CUSP): Data and Facts," 2018. Mimeo Ohio State University.
Berkes, Enrico, Gaetani, Ruben, 2019. The Geography of Unconventional Innovation. Mimeo Ohio State University.
Campo, Francesco, Mendola, Mariapia, Morrison, Andrea, Ottaviano, Gianmarco I.P., 2020. Immigrant inventors and diversity in the age of mass migration. In: Discussion Paper 1700. Centre for Economic Performance.
Chiu, Jason P.C., Nichols, Eric, 2016. Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics 4, 357–370.
Collobert, Ronan, Weston, Jason, Bottou, Léon, Karlen, Michael, Kavukcuoglu, Koray, Kuksa, Pavel, 2011. Natural language processing (almost) from scratch. J. Mach. Learn. Res. 12 (ARTICLE), 2493–2537.
De Rassenfosse, Gaetan, de la Potterie, Bruno van Pottelsberghe, 2013. The role of fees in patent systems: theory and evidence. J. Econ. Surv. 27 (4), 696–716.
Diodato, Dario, Morrison, Andrea, Petralia, Sergio, 2022. Migration and invention in the age of mass migration. J. Econ. Geogr. 22 (2), 477–498.
Eddy, Sean R., 1996. Hidden markov models. Curr. Opin. Struct. Biol. 6 (3), 361–365.
EPO, "PATSTAT, Spring 2023 version," https://www.epo.org/searching-for-patents/business/patstat.html 2023.

Etzioni, Oren, Cafarella, Michael, Downey, Doug, Popescu, Ana-Maria, Shaked, Tal, Soderland, Stephen, Weld, Daniel S., Yates, Alexander, 2005. Unsupervised named-entity extraction from the web: An experimental study. Artif. Intell. 165 (1), 91–134.
Feldman, Maryann P., 1994. The Geography of Innovation, vol. 2. Springer Science & Business Media.
Feldman, Maryann P., Kogler, Dieter F., 2010. Stylized facts in the geography of innovation. In: Handbook of the Economics of Innovation, vol. 1. Elsevier, pp. 381–410.
Gross, Daniel P., Sampat, Bhaven N., 2023. America, Jump-Started: World war II R&D and the Takeoff of the U.S. Innovation System. In: Working Paper 27375, National Bureau of Economic Research June 2023.
Hall, Bronwyn H., Harhoff, Dietmar, July 2012. Recent research on the economics of patents. Annual Review of Economics 4 (1), 541–565.
Hanlon, W. Walker, 2022. The Rise of the Engineer: Inventing the Professional Inventor During the Industrial Revolution. Technical Report w29751,. National Bureau of Economic Research.
Hanlon, Walker, 2016. British Patent Technology Classification Database: 1855–1882.
Hipp, A., Fritsch, M., Greve, M., Günther, J., Lange, M., Liutik, C., Pfeifer, B., Shkolnykova, M., Wyrwich, M., 2022. Comprehensive Patent Data of the German Democratic Republic 1949–1990. Jahrbücher für Nationalökonomie und Statistik. https://doi.org/10.1515/jbnst-2022-0058.
Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020.
Huang, Zhiheng, Wei Xu, and Kai Yu, "Bidirectional LSTM-CRF models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.
Kay, Anthony, July 2007. Tesseract: an open-source optical character recognition engine. Linux J. 2007 (159), 2.
Khan, B. Zorina, Sokoloff, Kenneth L., 2001. History lessons: the early development of intellectual property institutions in the United States. J. Econ. Perspect. 15 (3), 233–246.
Lafferty, John, McCallum, Andrew, Pereira, Fernando C.N., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, pp. 282–289. ICML.
Lamoreaux, Naomi R., Sokoloff, Kenneth L., 1997. Location and technological change in the American glass industry during the late nineteenth and early twentieth centuries. Working Paper 5938, National Bureau of Economic Research w5938.
Lamoreaux, Naomi R., Sokoloff, Kenneth L., 2000. The geography of invention in the American glass industry, 1870-1925. The Journal of Economic History 60 (3), 700–729.
Lample, Guillaume, Ballesteros, Miguel, Subramanian, Sandeep, Kawakami, Kazuya, Dyer, Chris, June 2016. 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 TechnologiesAssociation for Computational Linguistics San Diego, California, pp. 260–270.
Li, J., Sun, A., Han, J., Li, C., 2020. A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering 34 (1), 50–70.
MacLeod, Christine, 2002. Inventing the Industrial Revolution: The English Patent System, 1660–1800. Cambridge University Press.
Mansfield, Edwin, 1986. Patents and innovation: an empirical study. Manag. Sci. 32 (2), 173–181.
Montani, Ines, Honnibal, Matthew, 2018. Prodigy: a new annotation tool for radically efficient machine teaching. Artif. Intell. (to appear).
Moser, Petra, 2005. How do patent laws influence innovation? Evidence from nineteenth century world's fairs. Am. Econ. Rev. 95 (4), 1214–1236.
Moser, Petra, 2013. Patents and innovation: evidence from economic history. Journal of Economic Perspectives 27 (1), 23–44.
Nicholas, Tom, 2010. The role of independent invention in US technological development, 1880–1930. J. Econ. Hist. 70 (1), 57–82.
Nicholas, Tom, 2011. Cheaper patents. Res. Policy 40 (2), 325–339.
Nuvolari, Alessandro, Tartari, Valentina, 2011. Bennet Woodcroft and the value of English patents, 1617–1841. Explorations in Economic History *48* (1), 97–115.
Nuvolari, Alessandro, Vasta, Michelangelo, 2017. The geography of innovation in Italy, 1861–1913: evidence from patent data. Eur. Rev. Econ. Hist. 21 (3), 326–356.
Nuvolari, Alessandro, Tortorici, Gaspare, Vasta, Michelangelo, 2020. British-French technology transfer from the Revolution to Louis Philippe (1791–1844): evidence from patent data. In: CEPR Discussion Papers 15620. C.E.P.R. Discussion Papers.
Packalen, Mikko, Bhattacharya, Jay, January 2015. Cities and Ideas. Working Paper 20921,. National Bureau of Economic Research.
Pakes, Ariel, Griliches, Zvi, 1980. Patents and R&D at the firm level: a first report. Economics letters 5 (4), 377–381.
Perlman, Elisabeth R., et al., 2016. Dense Enough to Be Brilliant: Patents, Urbanization, and Transportation in Nineteenth Century America. Boston Univ. Working Paper,.
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, Balland, Pierre-Alexandre, Rigby, David L., 2016. Unveiling the geography of historical patents in the United States from 1836 to 1975. Scientific data 3 (160074).
Plasseraud, Yves, Savignon, François, 1983. Paris 1883: genèse du droit unioniste des brevets. LITEC.
de Rassenfosse, Gaétan, Kozak, Jan, Seliger, Florian, 2019. Geocoding of worldwide patent data. Nature-Scientific Data 6 (260).
Sarada, Sarada, Andrews, Michael J., Ziebarth, Nicolas L., 2019. Changes in the demographics of American inventors, 1870–1940. Explor. Econ. Hist. 74, 101275.
Sekine, Satoshi, Nobata, Chikashi, 2004. Definition, dictionaries and tagger for extended named entity hierarchy. In: LREC. Lisbon, Portugal.
Sokoloff, Kenneth L., 1988. Inventive activity in early industrial America: evidence from patent records, 1790–1846. J. Econ. Hist. 813–850.
Van Dulken, Stephen, 1999. British Patents of Invention, 1617–1977: A Guide for Researchers. British Library, Science Reference & Information Service.
Zhang, Shaodian, Elhadad, Noémie, 2013. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. J. Biomed. Inform. 46 (6), 1088–1098.