Centre for Economic Performance

Discussion Paper ISSN 2042-2695

No.1850

April 2022

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

Antonin Bergeaud
Cyril Verluise

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

UKRI Economic and Social Research Council

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 19th 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.

Keywords: history of innovation, patent, text as data

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

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 from Philippe Aghion, Jérôme Baudry, Nick Bloom, Gaétan de Rassenfosse, John van Reenen and seminar participants from the Summer School on Data and Algorithms for ST&I studies, the I3 working group and the Collège de France & CREST seminar.

Antonin Bergeaud, Bank of France and Centre for Economic Performance, London School of Economics. Cyril Verluise, Paris School of Economics and Collège de France.

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

1 Introduction

Endogenous growth theories (Romer, 1990; Aghion and Howitt, 1992) 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 enabling environment for strengthening 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 quantity of information they contain. In addition, despite their known limitations, evidence shows that the use of patents as a measure of 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. It is commonly acknowledged that the first British patent was granted to John of Utynam in 1449, see Plasseraud and Savignon (1983). Yet only limited information is available before the 1980s and actual publications did not systematically exist before the end of the 19^{th} century in most countries. 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, 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., 2017a; 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

[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] 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). These studies however do not focus on the geography of patentees.

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 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 19^{th} 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 and 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. 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.

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 (2018); Berkes and Gaetani (2019); Akcigit et al. (2017a, 2018) and Petralia et al. (2016). These datasets follow different purposes. For example Akcigit et al. (2018) use

[3]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 `https://cverluise.github.io/patentcity/`.

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 location of inventors and assignees, we also extract additional details such as the occupation of the inventors and their citizenship when applicable. These are often available in the text of patents, especially for British publications and bring interesting insight on 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., 2017a; Berkes, 2018), in particular in time of crisis (Babina et al., 2020). Akcigit et al. (2017b) 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. By collecting information on the citizenship of inventors, our dataset can also speaks to the literature on the relation between immigration and innovation. This literature typically finds that immigration is a privileged vehicle for importing knowledge.[4] In terms of historical trends, Akcigit et al. (2017a) and Arkolakis et al. (2020)

[4]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

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. (2017a) 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 these different studies, information on the citizenship and occupation of inventors are usually the results of a complex matching of patent publication data with different vintages of the census. Our database offer an alternative perspective by looking at the information directly reported in patent publications.

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 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).

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 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 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. Figure 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 East German, 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 as patents thereafter. As previously mentioned, we focus on patents published by the East German, German, French,

[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 of them 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.

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 which role is to resolve the relations between the extracted entities. Both layers are crucial to fully exploit the potential of patent texts.

2.2.1 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

[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.

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 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 Figure A7.

2.2.2 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 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 amount of examples - making them

[11] 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 A.3 for more details.

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.

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 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 C.

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,

[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 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.

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.

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 C.

2.2.3 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. 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 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 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.

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

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 C.

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.

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 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 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). 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 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.

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

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 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.

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

[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.

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.

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 A.6. 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 code, Commuting Zone code and state code for the US.

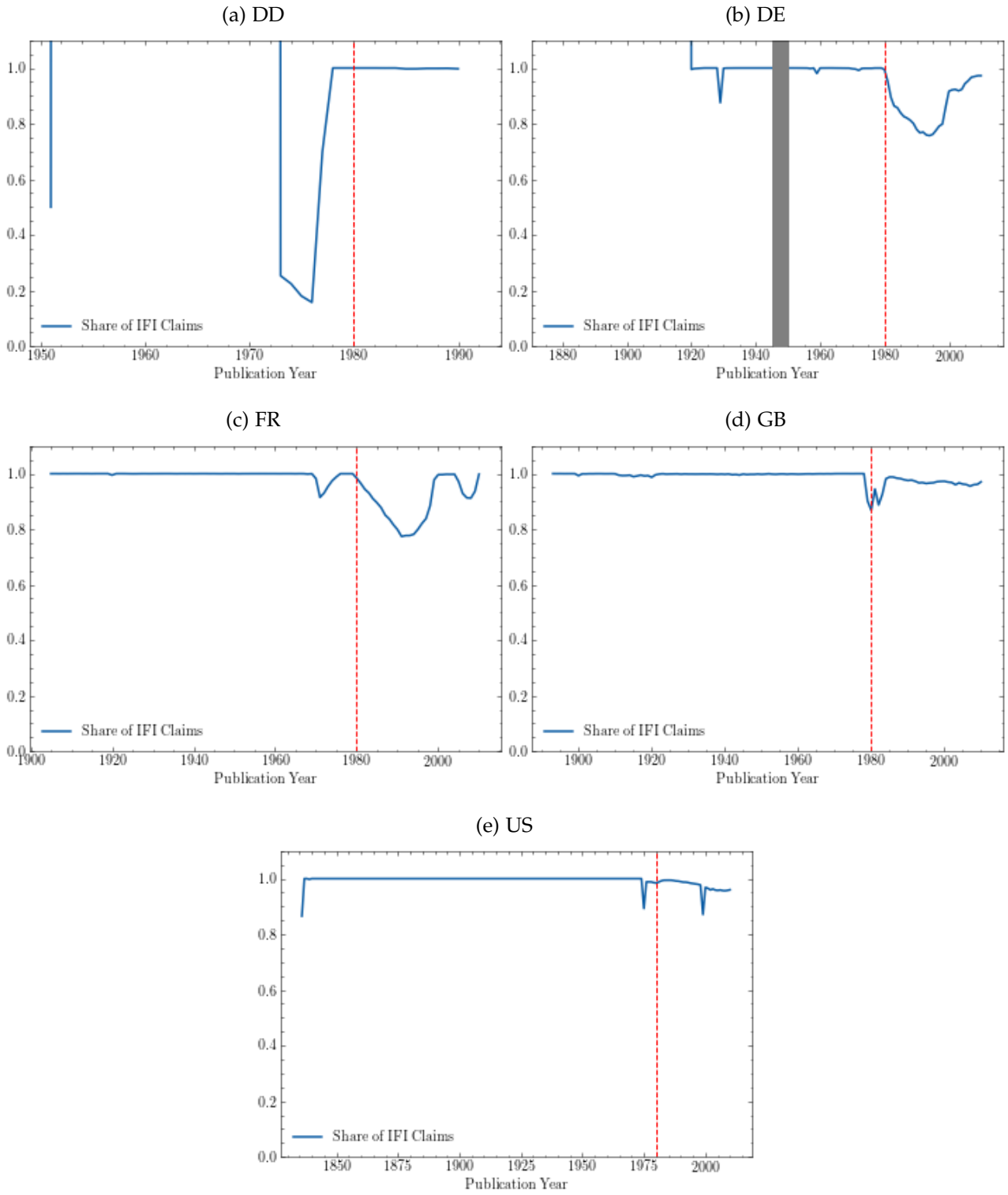
Finally, the version of the database that we provide contains all patents that we retrieved conditional on having a kind code in the list described in Table A1, i.e. utility patents. This database contains a little more than 16 millions different publication numbers but includes many duplicates as a same patent can have several publications at different stages of its life. Researchers interested in studying patents at a given stage may want to restrict to a specific set of kind codes. However, most uses of the database are likely to require deleting duplicates and keeping only one observation per patent. We describe a simple procedure to do so in Appendix A.1.

3 Overview of the dataset

3.1 Coverage

We show simple results taken from the database to give a sense of its content. First, Figure 1 compare the coverage of our dataset to a benchmark that we take as the IFI Claims database. Specifically, we report for each year the number of publications that fall within the criteria defined in Table A1 (utility patents) divided by the same number in IFI.

Figure 1: Number of patents in Patentcity compared as a share of IFI Claims



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).

Overall, our coverage is very high and even exceeds these of Claims in the case of Germany due to our effort to recover the missing dates of publication before 1920 and for East Ger-

many. 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 give additional details on the coverage of our dataset and in particular the share of patents for which we detect one inventor at least and similarly for all the entities that we extract. In particular, Figures A3 show that not all patents are associated a location. This is generally due to the fact that during some subperiods, geographical information can be missing from the patent publications (for example in France during the 1970-1980 period).

3.2 Geographic distribution of patents

The geography of innovation is the subject of numerous studies that have in particular put forward its very large degree of concentration (even when controlling for population density, see Feldman and Kogler, 2010 for a review).

One of the advantage of PatentCity is that it provides a geographic information for each patentee that could help illustrate possible differences in the spatial distribution of innovation across years and countries. Section 2.3.1 details the level of granularity that we achieve with our geocoding (see also Figure A7 in Appendix A). It is however important to note that Table 5 includes all patentees, whether domestic or foreign. Restricting to domestic inventors and assignees increase significantly the average granularity of the dataset. More than 99% of patentees are located at least at the county level (counties in the US and NUTS3 regions in other countries) except for East Germany (98%) and France (90%).

Figure 2 uses this level of aggregation to map the number of patentees for each county in the 4 countries (pulling together East and West Germany). It shows that as expected inventors and assignees are mostly located around large urban areas. 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.

County level analysis already provides a very granular picture of the geography of innovation. However, the level of precision is even 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. This is illustrated in Figure 3

which plots the exact location of patentees in London. This Figure shows that most of the assignees are located in central London while inventors' location 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[16] or the link between wealth and innovation.

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 Figure 2 for every decade (see Figures B1, B2, B3 and B4).

3.3 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.[17] 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".

Figure 4 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 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%.

[16] We recall that patent data includes a list of technological class.

[17] 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 A5 in Appendix A.4.

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 5 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 5 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.

3.4 Citizenship

Most of existing studies that focus on the link between immigration and innovation (see e.g. Arkolakis et al., 2020 and Akcigit et al., 2017a) use external data to identify immigrants, for example different vintages of the US Census or registers of inventors. PatentCity offers a complementary approach by using the information on citizenship included in the text of the patent publications in the US and in the United Kingdom. This is mostly possible during two distinct subperiods (respectively 1920-1950 for the United Kingdom and 1880-1925 for the US) for which the patent documents directly report the citizenship and the location of some inventors[18] which allows us to classify them as "immigrant". Of course 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.

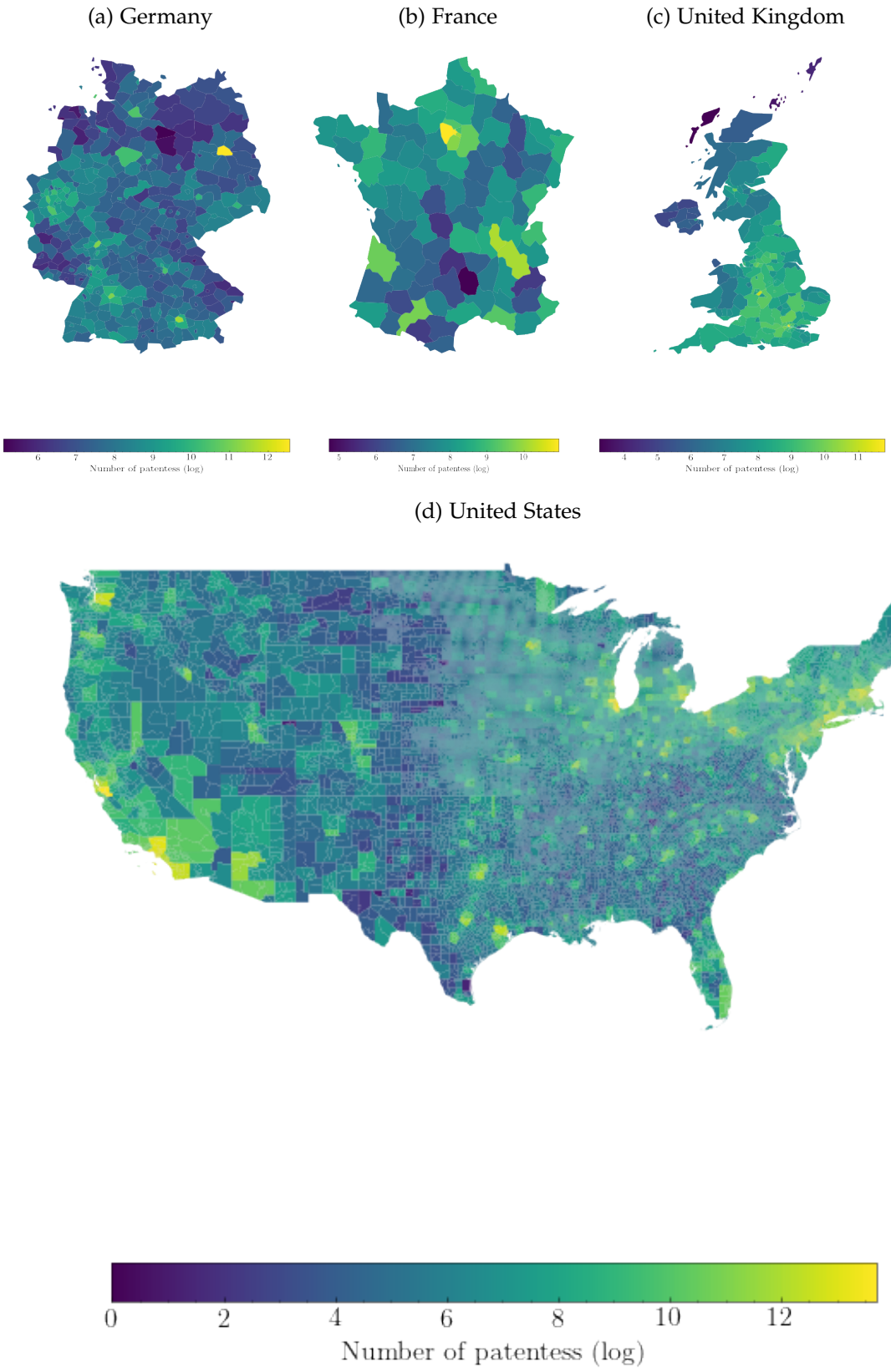
[18]Not all patentees declare a citizenship even during these subperiods. Among the set of patentee that are located in the United Kingdom, 87% report a citizenship for patents filed between 1920 and 1950. During the period 1950-1980, around 20% of inventors filing a British patent did declare their citizenship. For the US, this share is around 37% between 1880 and 1925 but is closer to 45% after 1910 (see Appendix Figure A6).

We find that between 4% and 5% of inventors who report an address in the US but are *not* American.[19] In the United Kingdom, this share is lower, between 1% and 2%, at any point in time between 1920 and 1950. In Figure 6, 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 1940s.

Figure 7 reports the evolution of the composition of these immigrants 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 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.

[19] These numbers are lower than those reported by Akcigit et al. (2017a) 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. (2017a) 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.

Figure 2: Patentee location at the county level

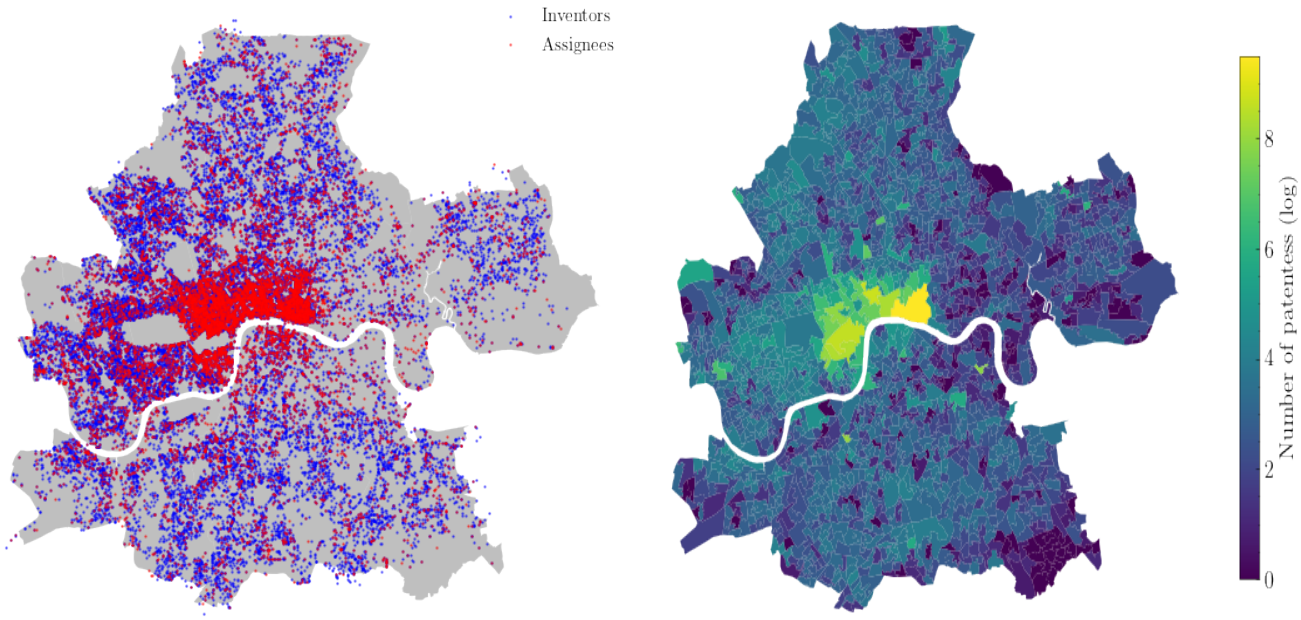


Notes: these Figures maps the total number of patentees (whether assignees or inventors), in log, for each county in Germany, France, the UK and the US. In the three European countries, a county is a NUTS3 region. The number of patentees is taken as a total over the full set of domestic patentess that are located at least at the county level without restriction on the time period.

Figure 3: Patentee location in London

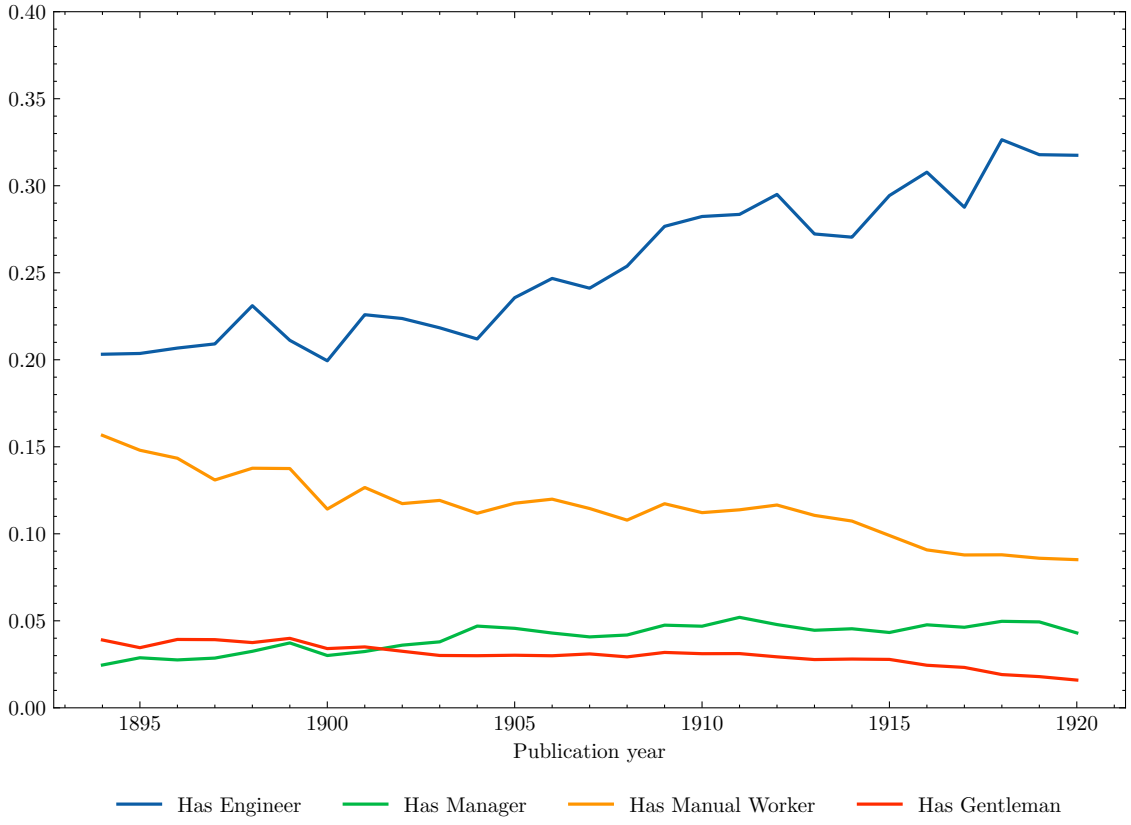
(a) Coordinates

(b) By area



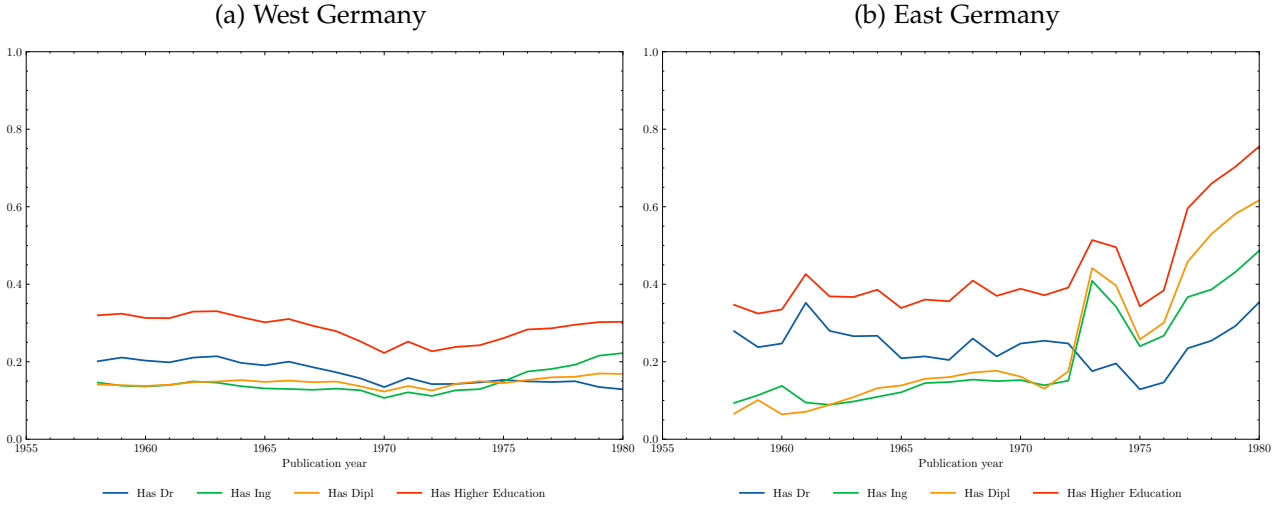
Notes: these Figures maps the location of inventors and assignees of the UK patent office that are located in Inner London and for which the geocoding has been done at the street or house number level. Left-hand side map shows the coordinate of the house number reported or the centroid of the street.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

Figure 4: Occupation of inventors in the United Kingdom



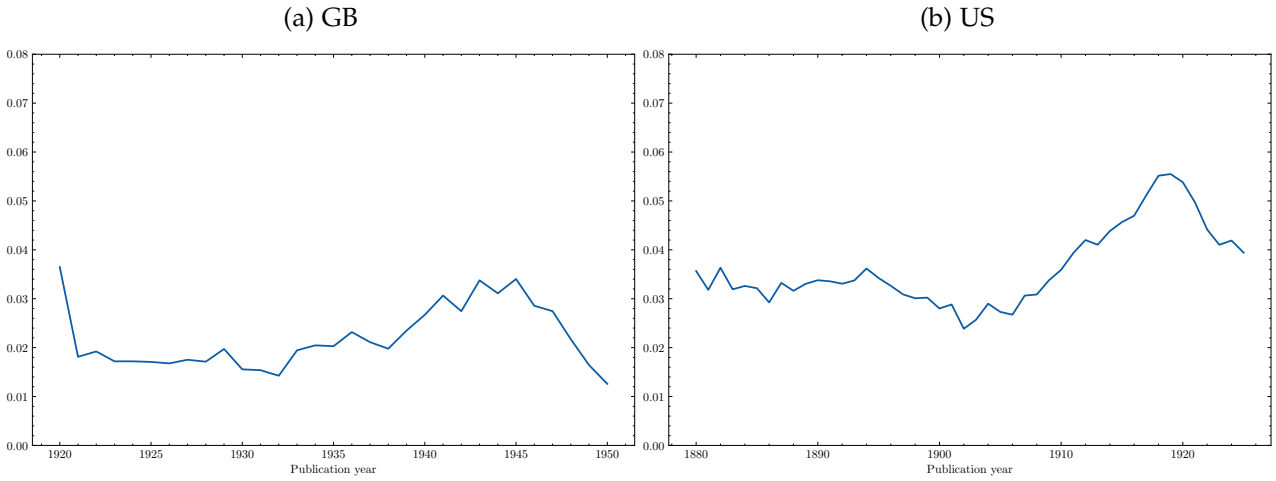
Notes: This Figure reports the share of patents involving at least one engineer (Has engineer), one manager (Has manager), one manual worker (Has manual worker) or one gentleman (Has gentleman) in terms of the occupation of the inventor reported in the text. Time period: 1894-1920.

Figure 5: 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.

Figure 6: Share of immigrant inventors over time



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 7: Composition of immigrant inventors' citizenship

Notes: Each area represents the share of top 10 most frequent citizenship in the set of detected immigrant 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).

4 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 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.

References

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

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

Akcigit, Ufuk, John Grigsby, and Tom Nicholas, "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.

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.

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

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

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).

Berkes, Enrico, "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.

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.

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

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.

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

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

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.

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.

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.

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

Lafferty, John, Andrew McCallum, and Fernando CN Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in "Proceedings of the 18th International Conference on Machine Learning" ICML 2001, pp. 282–289.

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.

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.

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.

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

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.

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.

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.

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.

Romer, Paul, "Endogenous Technological Change," *Journal of Political Economy*, 1990, *98* (5), S71–102.

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.

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.

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.

Online Appendix

A Data Appendix

A.1 Selection of utility patents

Intellectual property offices publish many documents that are called "patents". For example, the USPTO publishes "plant patents", "defensive publications", "reissue patents". To select the patents that most closely match the idea of *patents of invention* and to avoid double counting, we select documents based on the kind codes. Our goal is to mimic as closely as possible the concept of "first publication of granted patent". We start with a list of kind code that corresponds to utility patents which we enumerate in Table A1. From this list, we make additional selections to remove non granted patents and to keep only first publications. Formally, we made the following selections:

- **United States:** We keep documents with kind code A (Granted patents prior to 2001), B1 (Granted patent published without an application), B2 (Granted patent published with an application, after 2001).
- **Germany:** We keep publications with kind code C (*"Patentschrift"*) or B (*"Auslegeschrift"*) without conditions. These corresponds to first publications of patents before 1970. After 1970, the publication process changed and a given patents could have several publication. We keep kind code A1 (*"Offenlegungsschrift"*) whenever a given patent (identified by the publication number without the kind code) has more than one publication. We then keep kind code C2 if the patent does not have a A1 publication but has more than one additional publication (on top of the C2). Finally, we keep all patents that have only one publication, except if this publication is a A1 document.
- **France:** We keep publications with kind code A and A5 without conditions (*"Brevet d'invention"*). We then keep kind code A1 (*"Demande de Brevet d'Invention"*) if there is only one publication for a given patent and if the publication year is earlier than 1971. Otherwise we keep publications with kind code A1 if there are more than one publication for the patent.
- **United Kingdom:** We keep documents with kind code A (Patent Application) if the publication number is lower than 2000000 or if the publication year is earlier than 1921. Otherwise, we keep A if there are more than one publication for the patent.

These rules are governed by the fact that the patent systems change over time. Typically in earlier years, all patent publications correspond to the one and only document that served as the final granted patent. In the most recent decades, patent offices published the patent applications along with other subsequent documents if the granting process was successful. Simply counting all patent applications would result in the inclusion of patents that have not been granted and to overestimate the number of patents in the most recent period. Note that we also release a version of the database where we did not make these restrictions and include all utilities patents, whose kind code are summarized in Table A1.

Table A1: 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.

A.2 Formats

The structure of a patent document can change over time as the patent office modernizes its publications and processes. We tracked these changes and adapted the statistical model that we used to each cases. Table A2 shows the different formats for each patent offices and the first and last patents of each format.

Table A2: 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: Format numbers are for internal usage only. A patent format corresponds to a span of patents exhibiting similar information and displayed in a similar way.

A.3 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

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.

A.4 Data coverage

This Section presents the coverage of each entities as a share of patents for the five patent offices considered. Precisely:

- Figure A1 plots the yearly share of patents with at least one inventor
- Figure A2 the yearly share of patents with at least one assignee
- Figure A3 plots the yearly share of patents with at least one location

- Figure A4 the share of patentees that are matched with a location
- Figure A5 plots the yearly share of inventors with at least one entity occupations
- Figures A6 plots the yearly share of inventors with at least one entity citizenship
- Figure A7 shows the relative share of each level of geographical matching.
- Figure A8 reports the composition of the geocoding by source: either using commercial geocoding supplier: HERE or GMAPS or manually

Finally, Figures A9, A10 report the number and share of patent publications by source (either from PatentCity, from de Rassenfosse et al., 2019 or from the expansion (that is, we expand the entities and relationships to all patents of the same family when information is missing). Figures A11 and A12 compare the coverage of the PatentCity database with the coverage of the Claims database that we take as the universe of patents.

A.5 Additional annotation guidelines

Tables A3 and A4 present additional representative examples of the rules we used to label the patents. See Section 2 and the detailed guidelines for East Germany, Germany, France, the United Kingdom and the United States.

A.6 Structure of the dataset

The dataset is publicly available both as a csv file and in SQL. The unit of observation is the patent, identifiable from the DOCDB publication number. Each patent is associated with a set of patentees (inventors or assignees) which have nested attributes: name, citizenship, location and occupation. The structure of the dataset is presented in Table A5.

Figure A1: Share of patents with at least one inventor

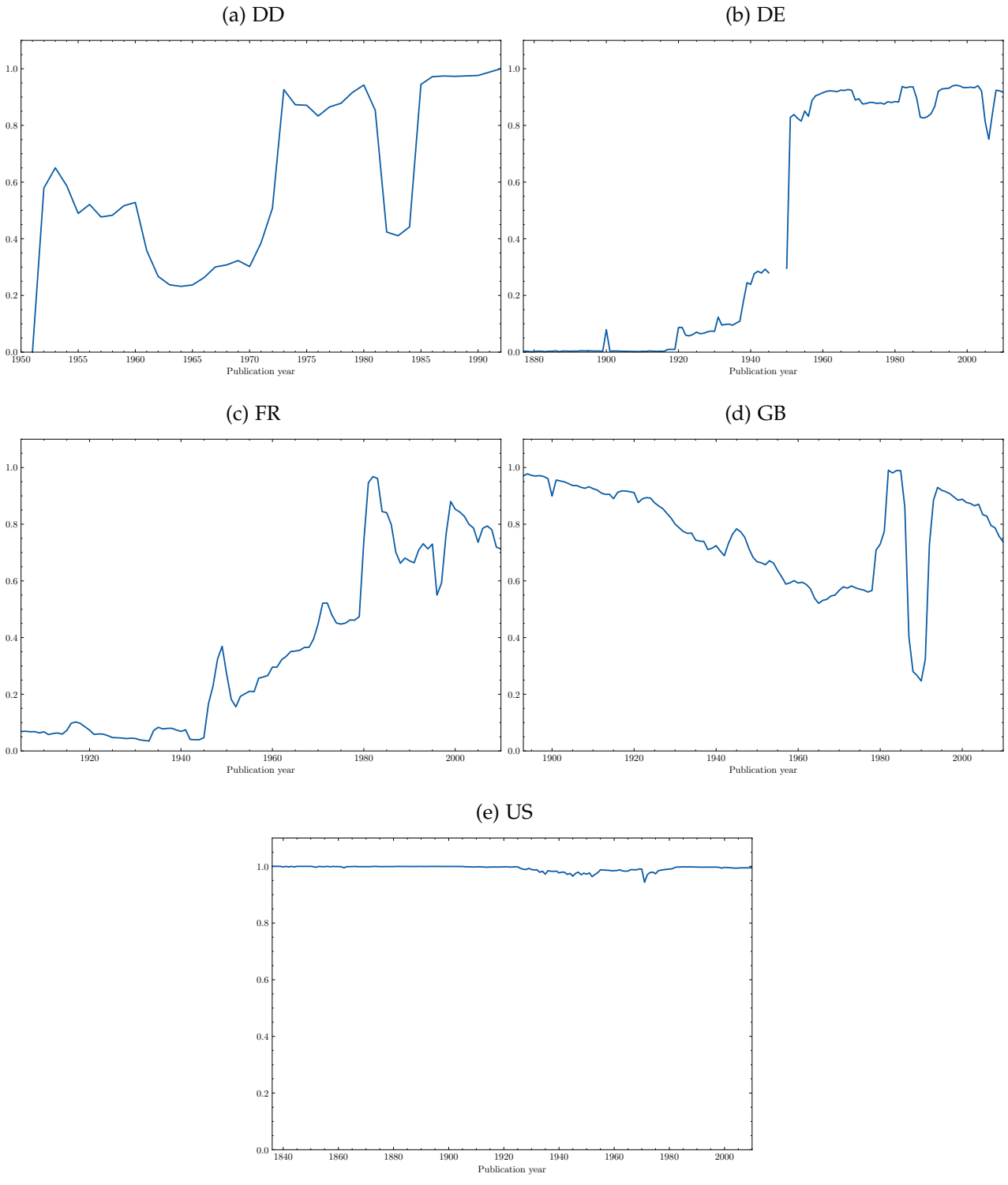


Figure A2: Share of patents with at least one assignee

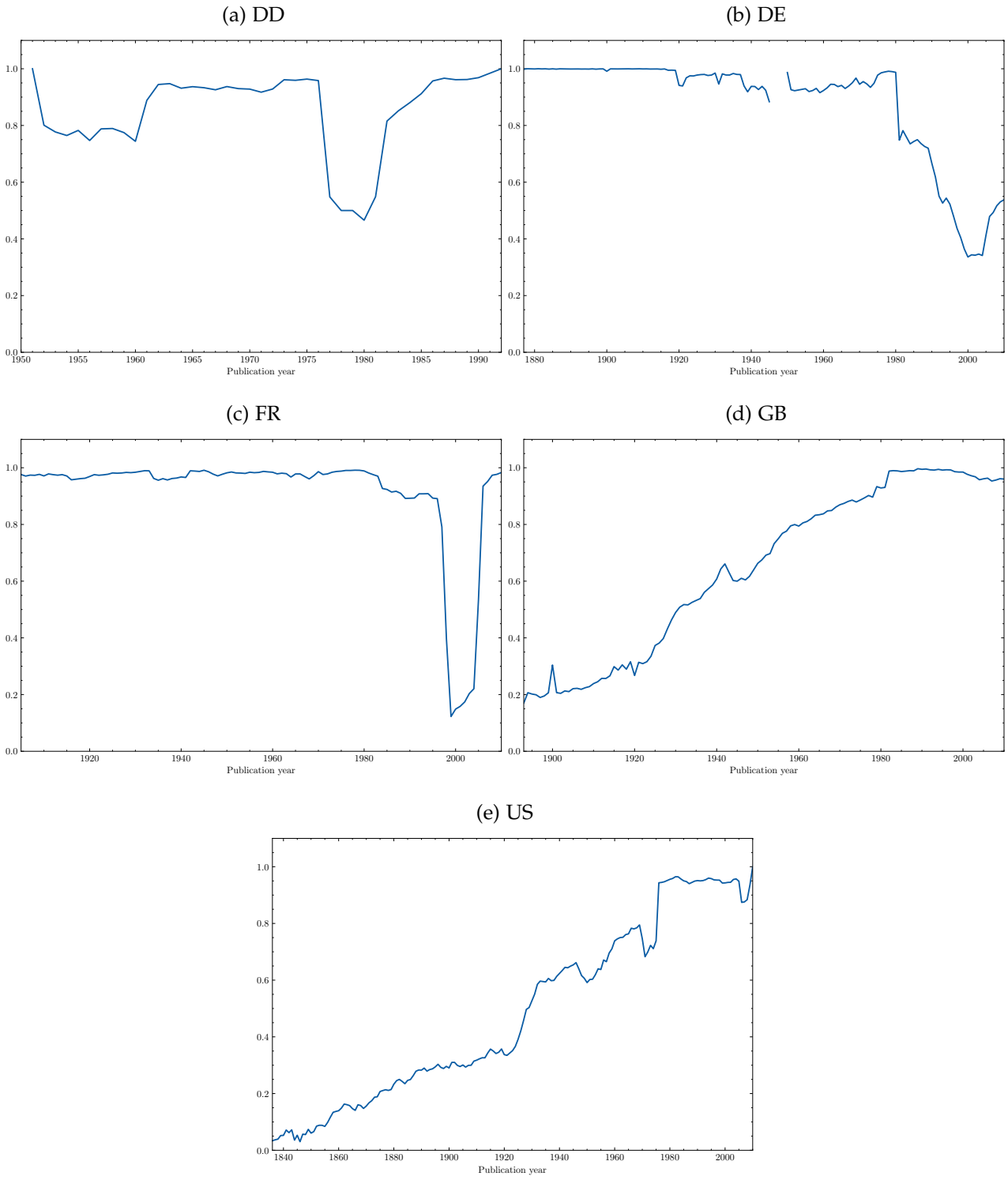


Figure A3: Share of patents with at least one location

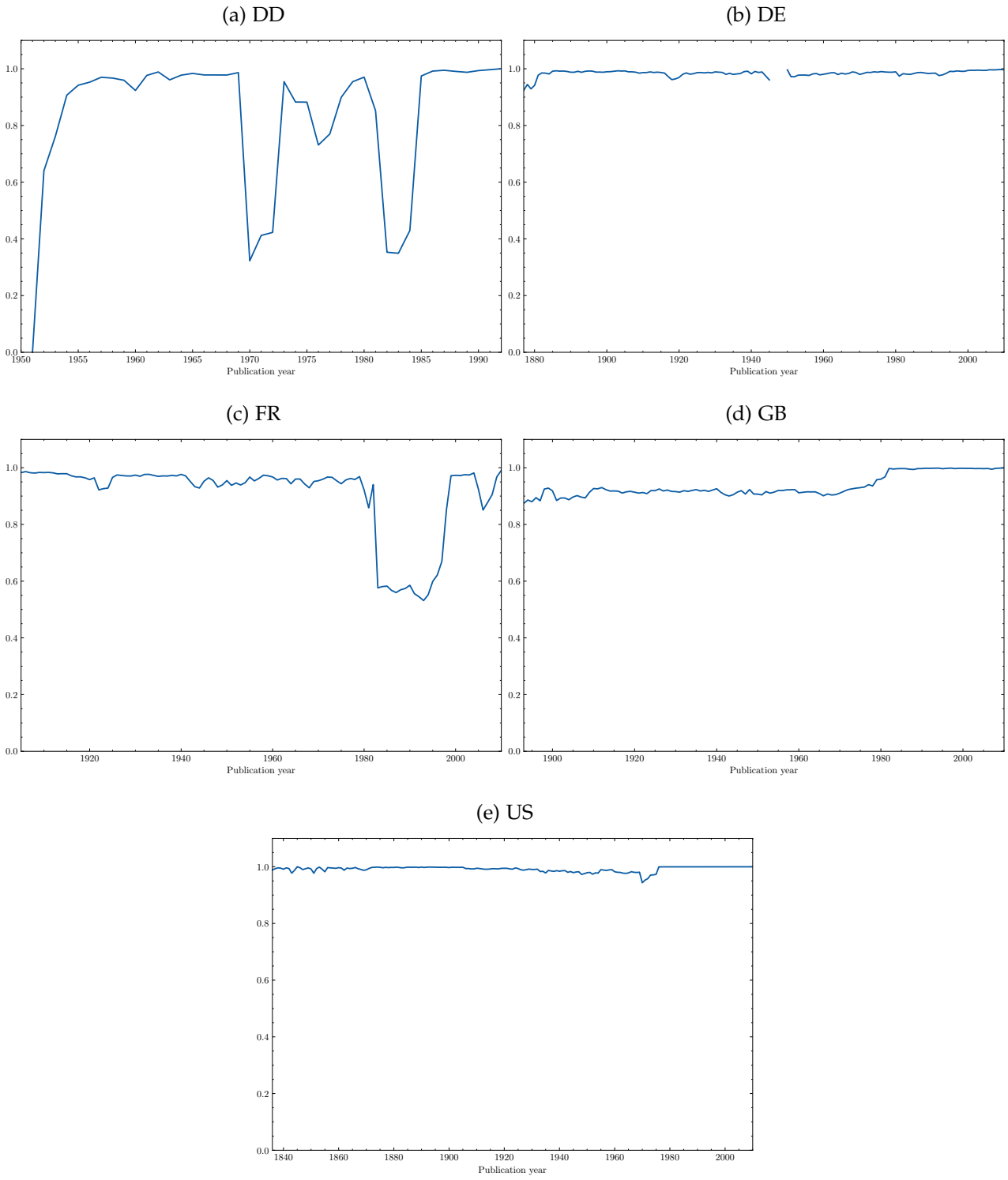


Figure A4: Share of patentees with a detected location

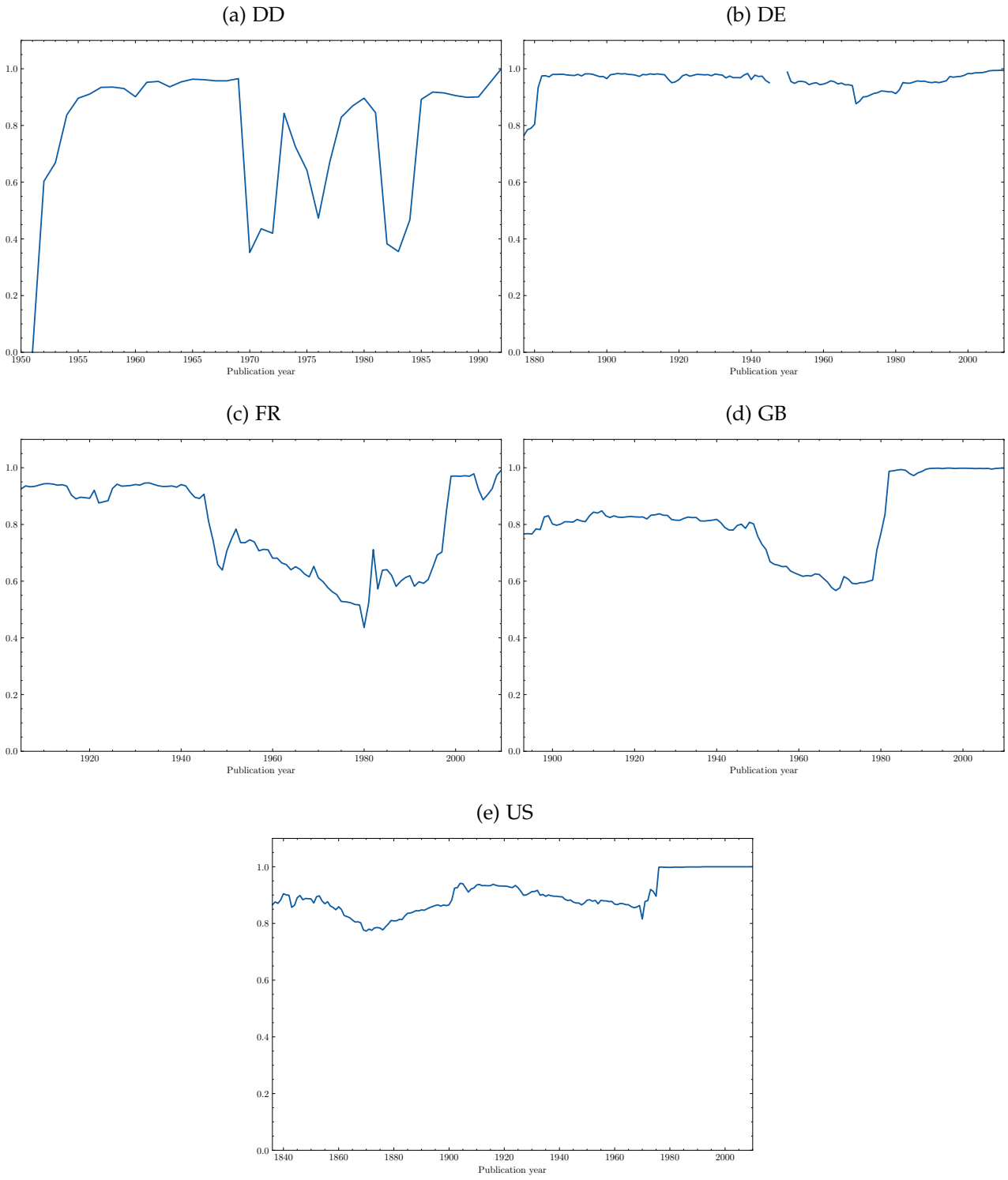
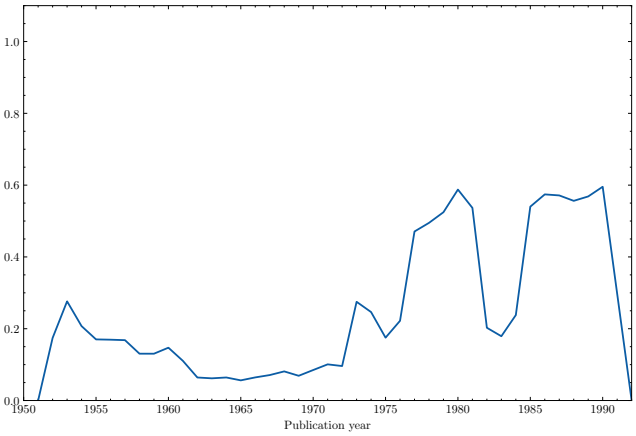
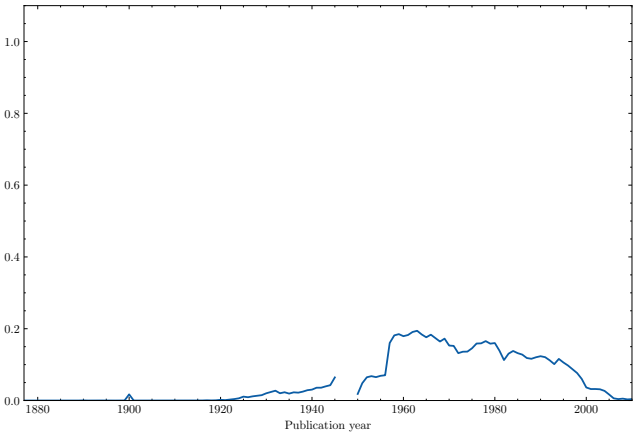


Figure A5: Share of inventors with a detected occupation

(a) DD



(b) DE



(c) GB

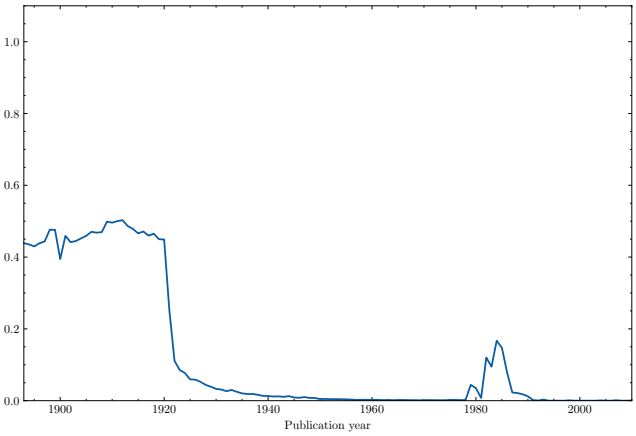
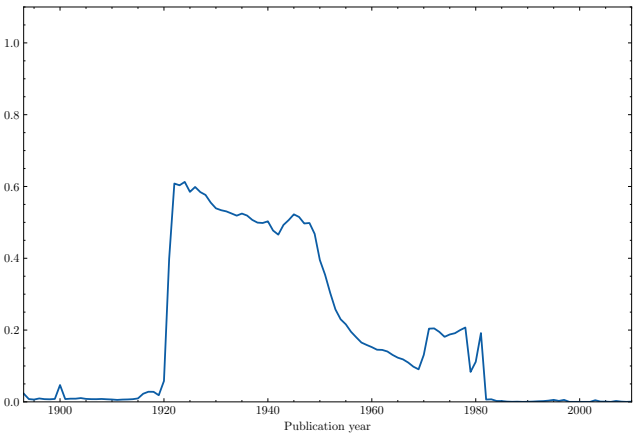


Figure A6: Share of inventors with a detected citizenship

(a) GB



(b) US

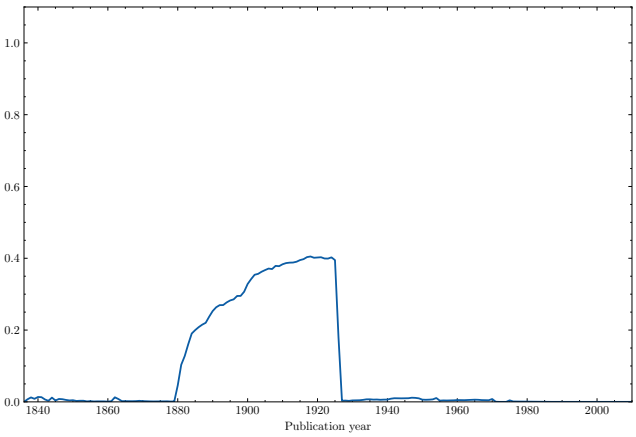
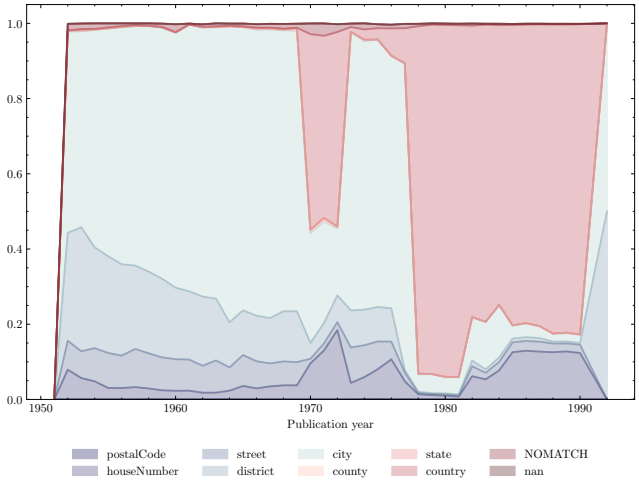
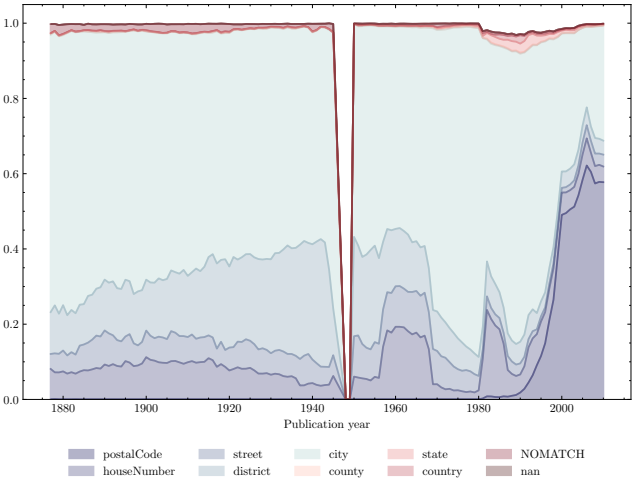


Figure A7: Composition of the most detailed level of geocoding

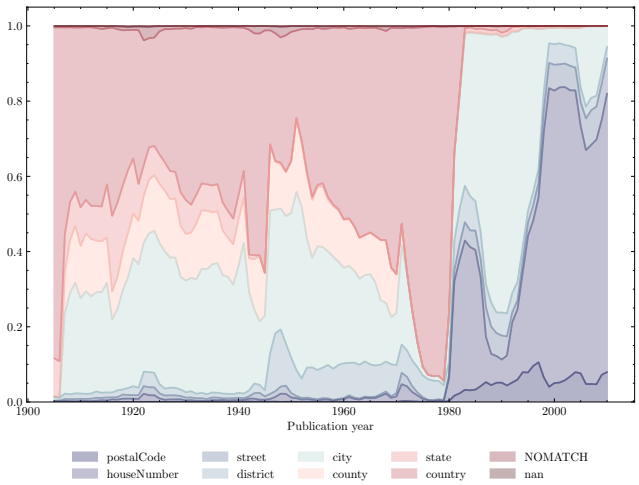
(a) DD



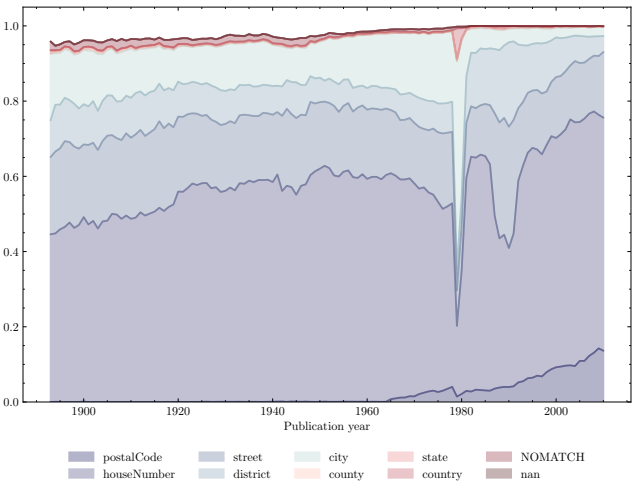
(b) DE



(c) FR



(d) GB



(e) US

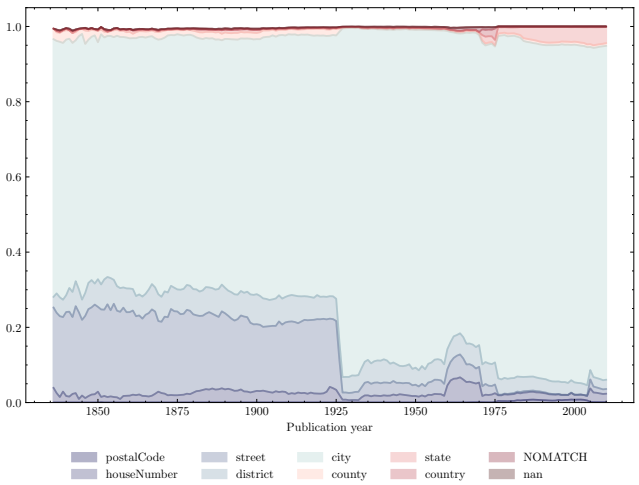
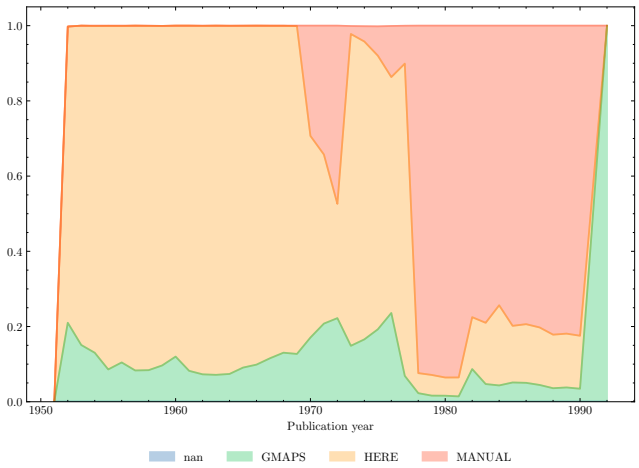
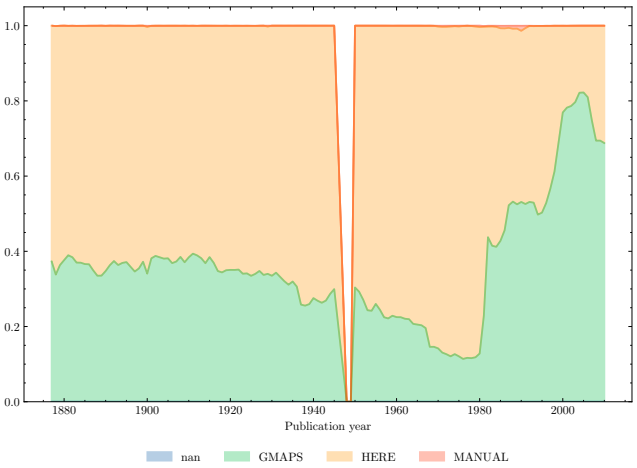


Figure A8: Composition of geocoding by geocoding source

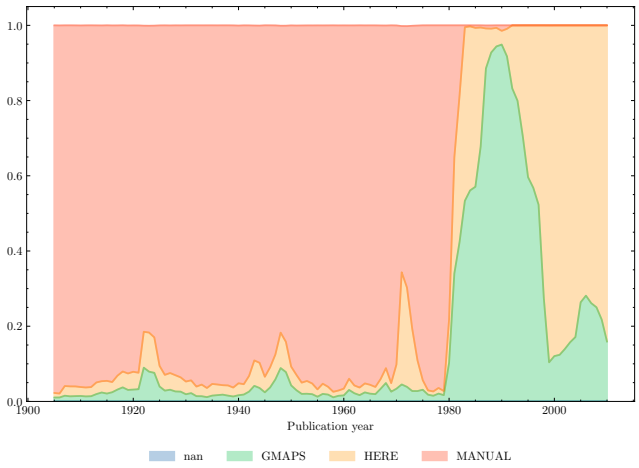
(a) DD



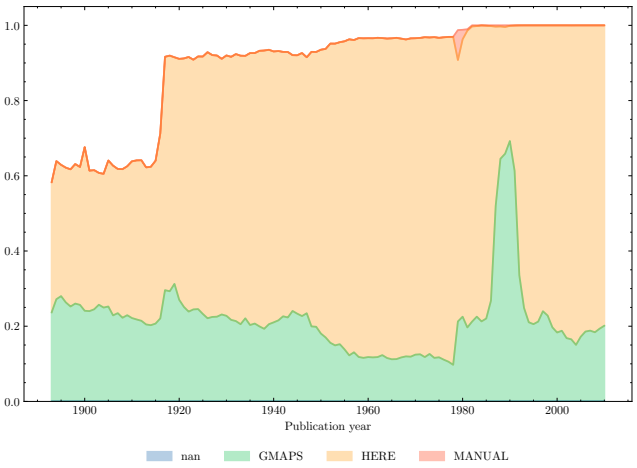
(b) DE



(c) FR



(d) GB



(e) US

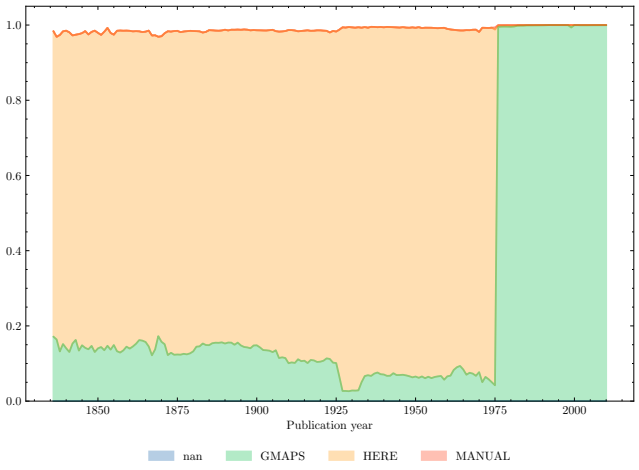
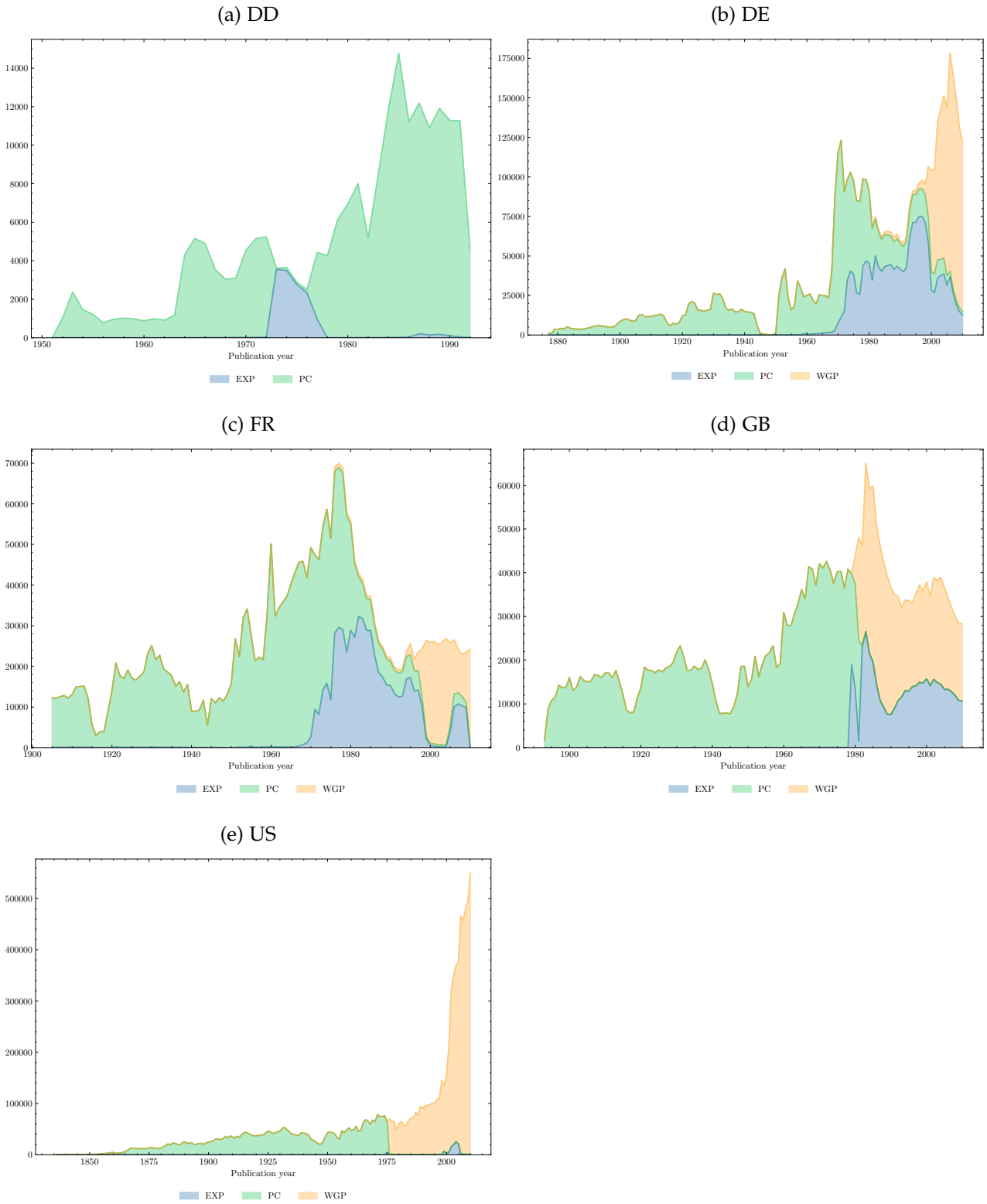
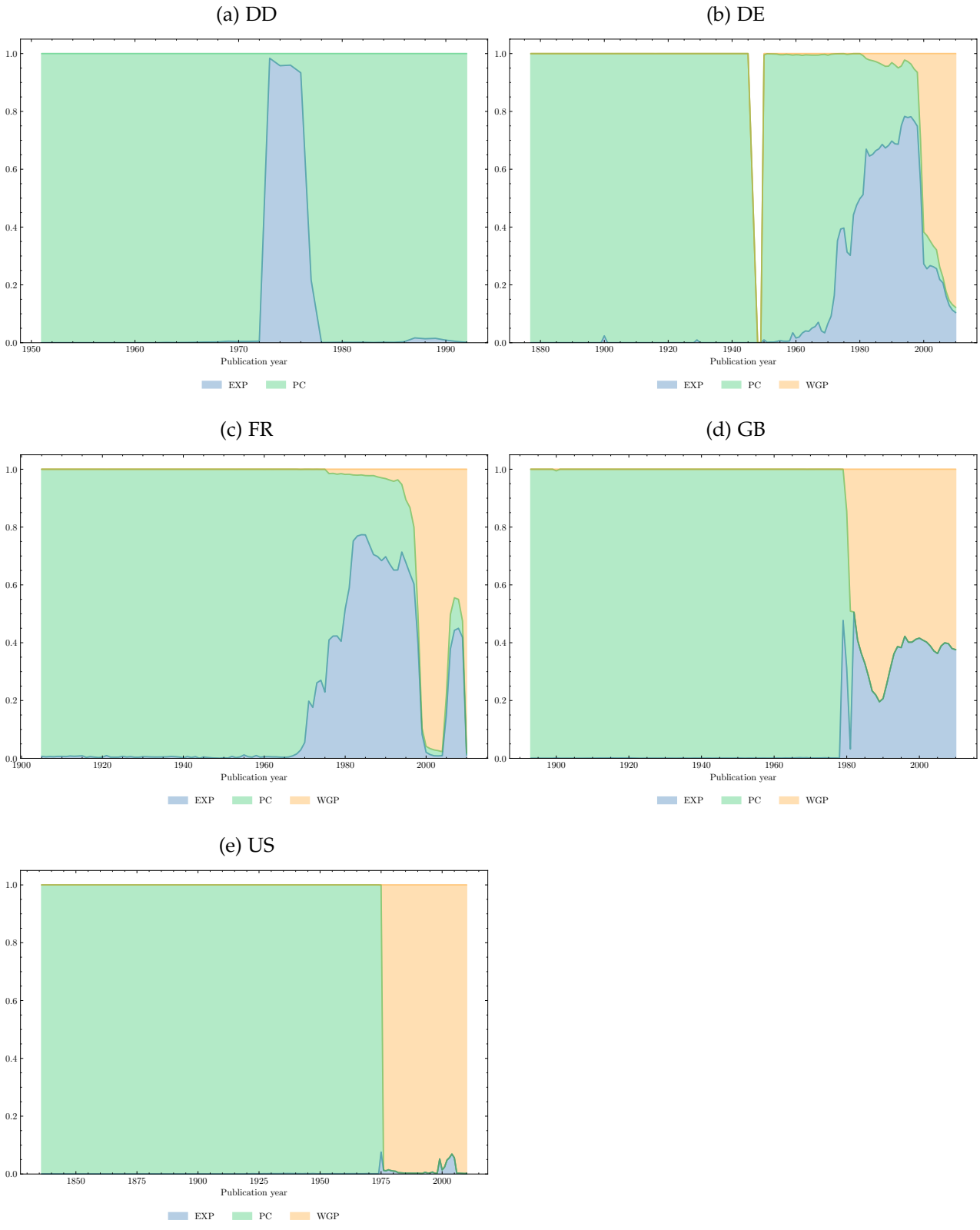


Figure A9: Database composition by source (number of patents)



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

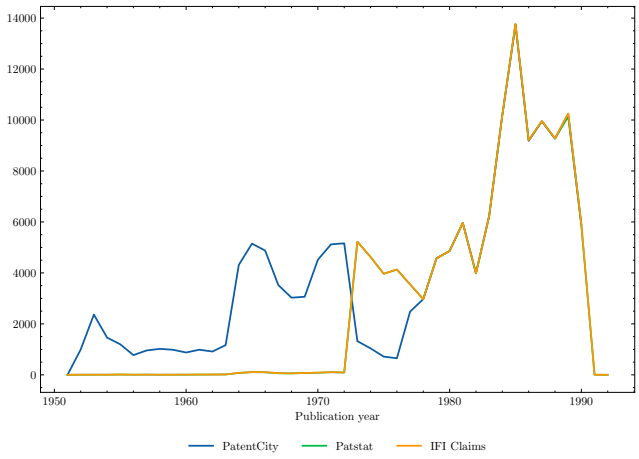
Figure A10: Database composition by source (in share)



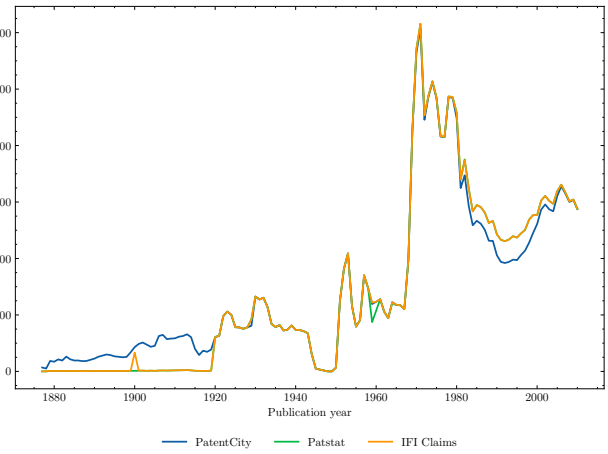
Notes: PC refers to PatentCity data, WGP refers to de Rassenfosse et al. (2019) data and EXP refers to data collected from family expansion from patents included in either PC or WGP.

Figure A11: Database coverage by office and publication year (in absolute values)

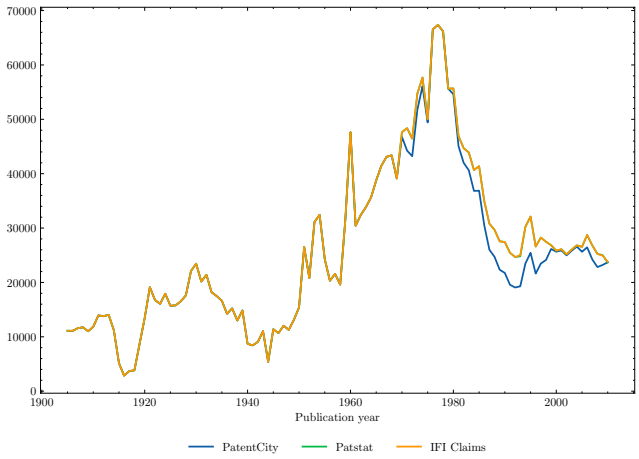
(a) DD



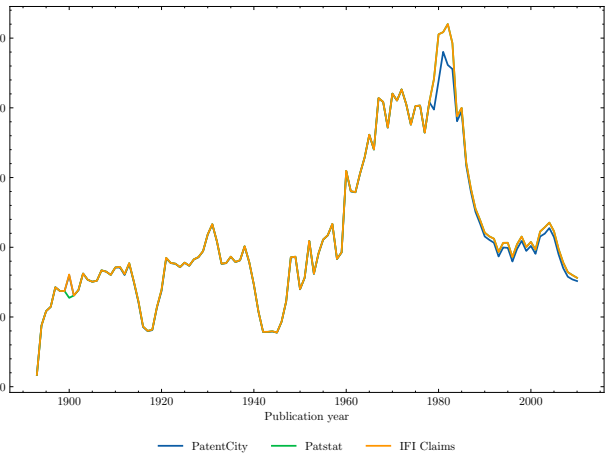
(b) DE



(c) FR



(d) GB



(e) US

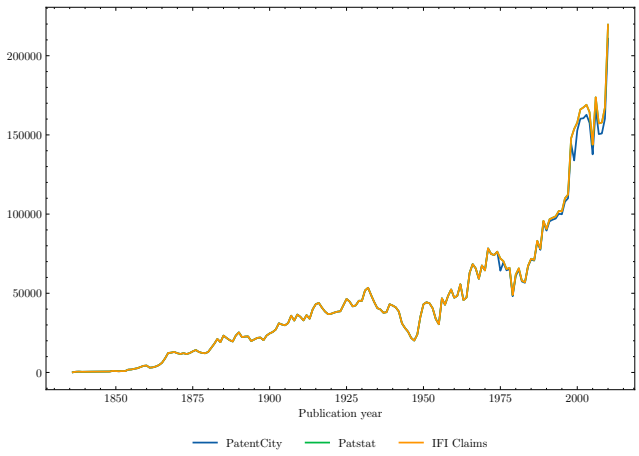
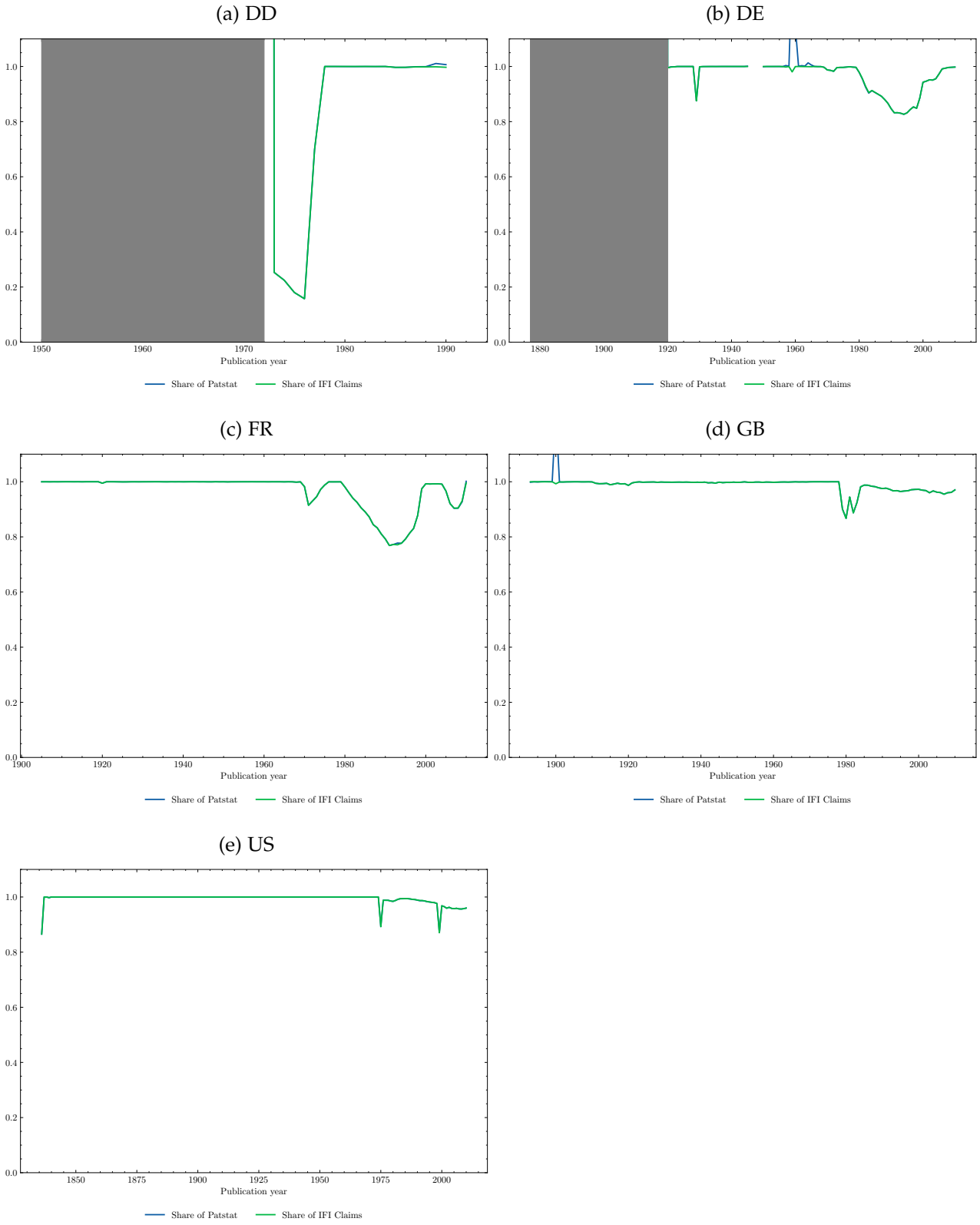


Figure A12: Database coverage by office and publication year (in share of the Claims database coverage



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.

Table A3: 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 SAARBRUCKEN
	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 A4: 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)
GB	CITIZENSHIP	Links an ORG/PERS to its CIT	Maxim Hanson Hersey ⟶ CITIZENSHIP ⟶ subject of the king of Great Britain and Ireland
	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
US	CITIZENSHIP	Links an INV/ASG to its CIT	WILLIAM H. BAKER ⟶ CITIZENSHIP ⟶ citizen of the United States
	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.

Table A5: 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
__.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

Continued on next page

__.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
__.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

Continued on next page

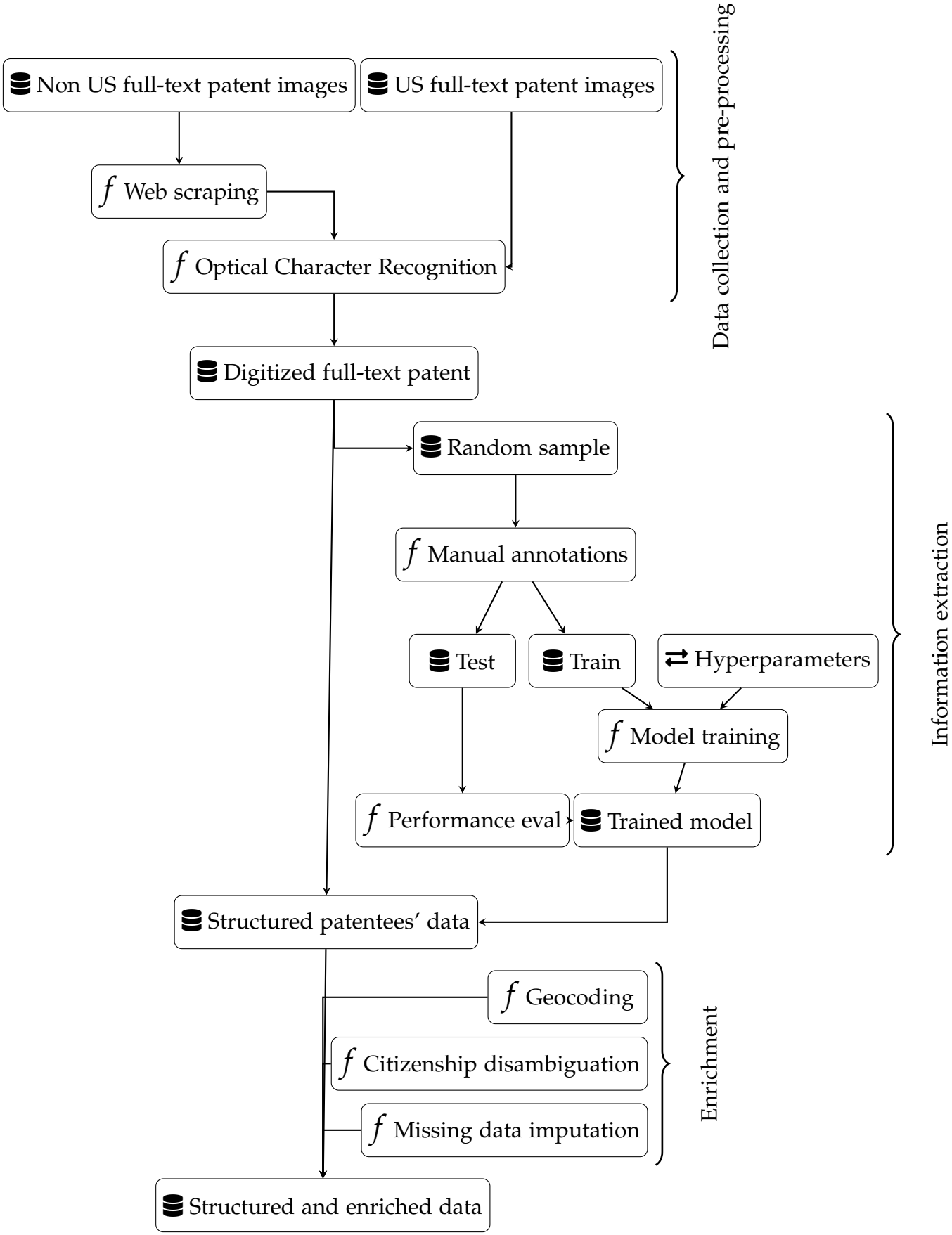
__.loc_matchQualityPostalCode	Same at postalCode level.	FLOA	147,862
__.loc_matchQualityStreet	Same at street level.	FLOA	2,452,802
__.loc_matchQualityHouseNumber	Same at houseNumber level.	FLOA	1,034,844
__.loc_matchQualityBuilding	Same at building level.	FLOA	410
__.loc_key	Key used for statistical 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
__.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.

A.7 Pipeline

We summarize the full pipeline from the raw documents to the structured and enriched database in Figure A13.

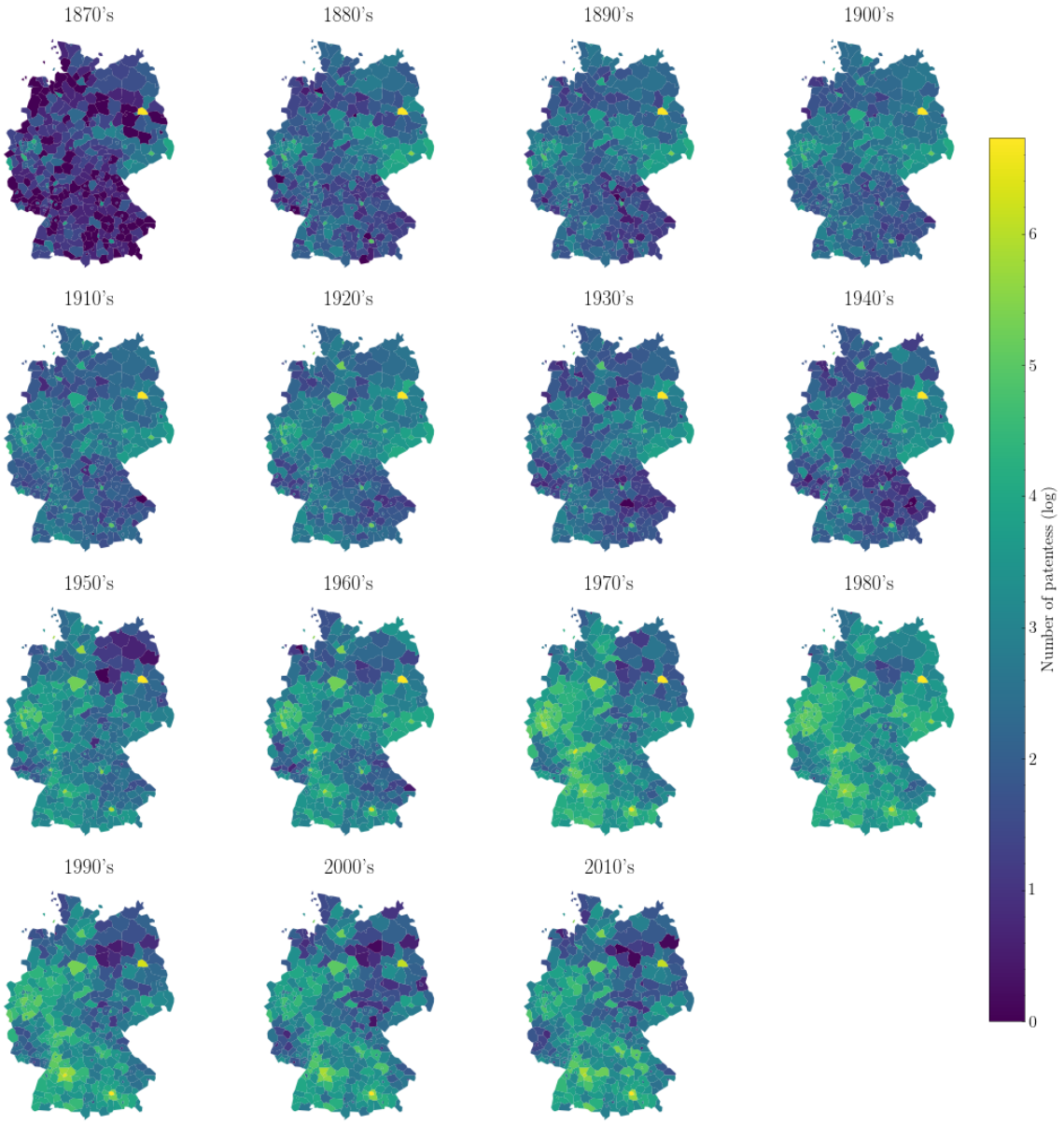
Figure A13: Workflow pipeline



B Additional Maps

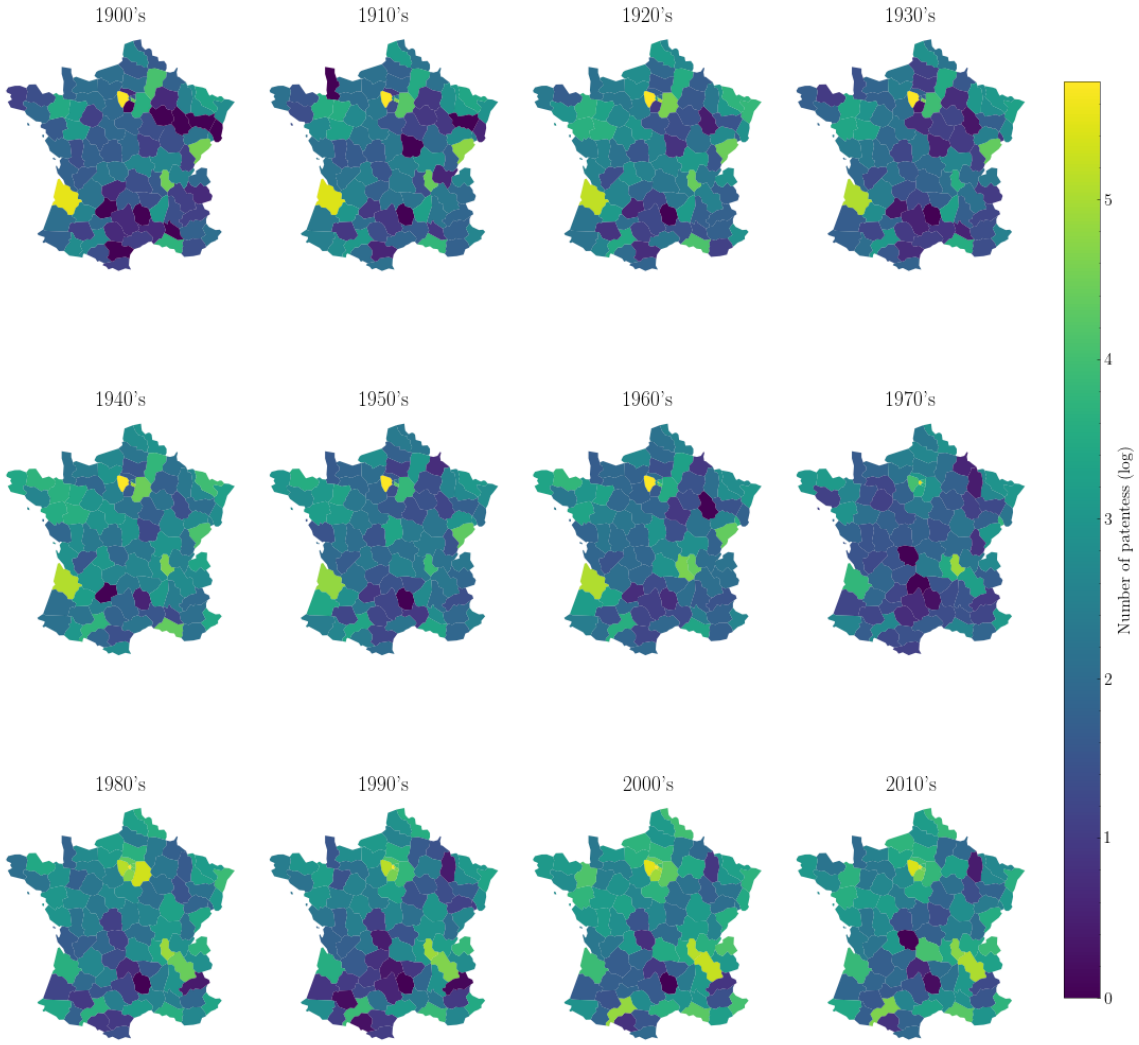
Figures B1, B2, B3 and B4 map the number of patentees by regions NUTS 3 (commuting zones in the US) by decade.

Figure B1: Patentees by regions and decade - Germany



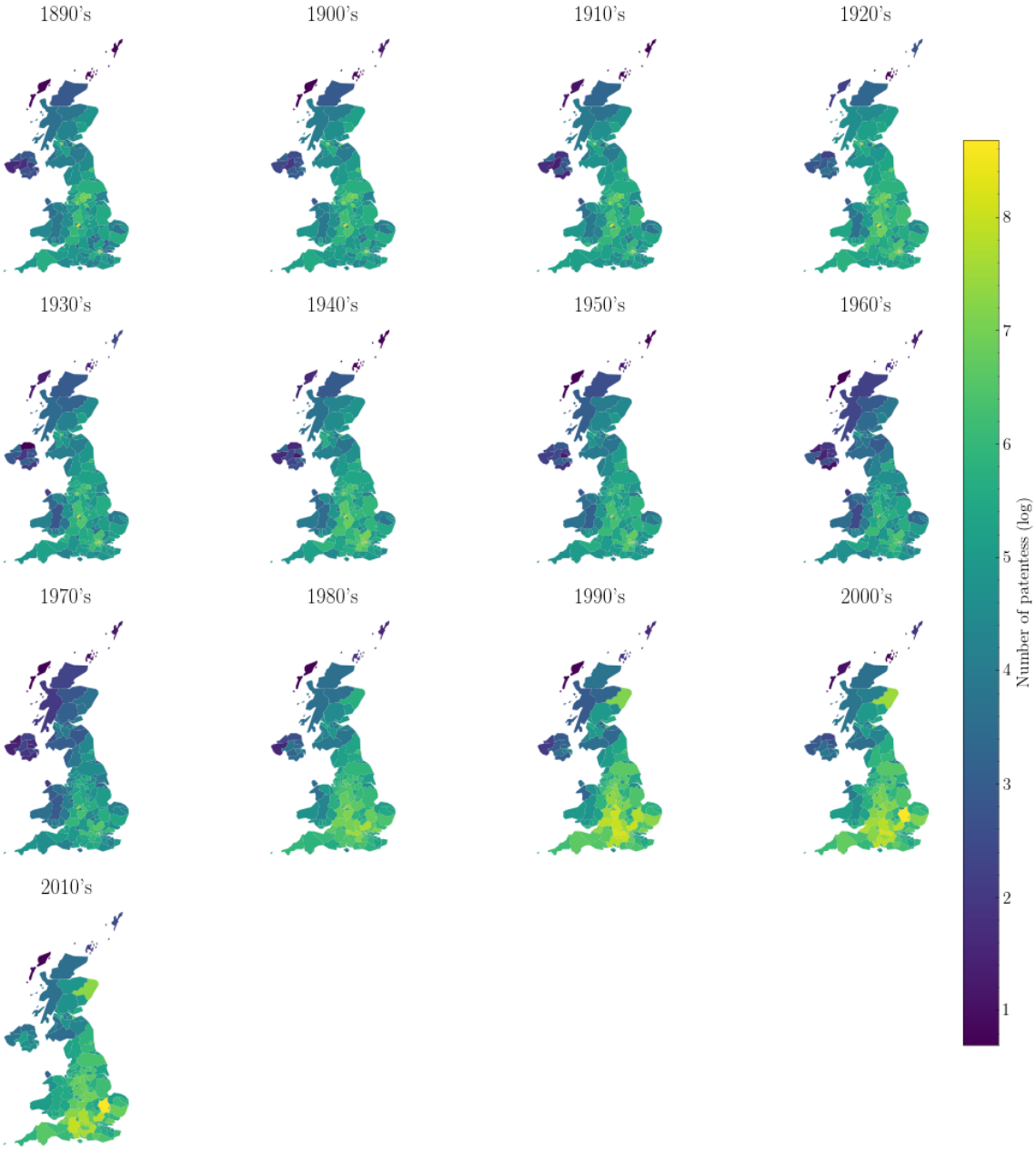
Notes: this Figure maps the total number of patentees (whether assignees or inventors), in log, for each county in Germany (*Kreise*) for each decade. The number of patentees is taken as a total over the full set of domestic patentees that are located at least at the county level.

Figure B2: Patentees by regions and decade - France



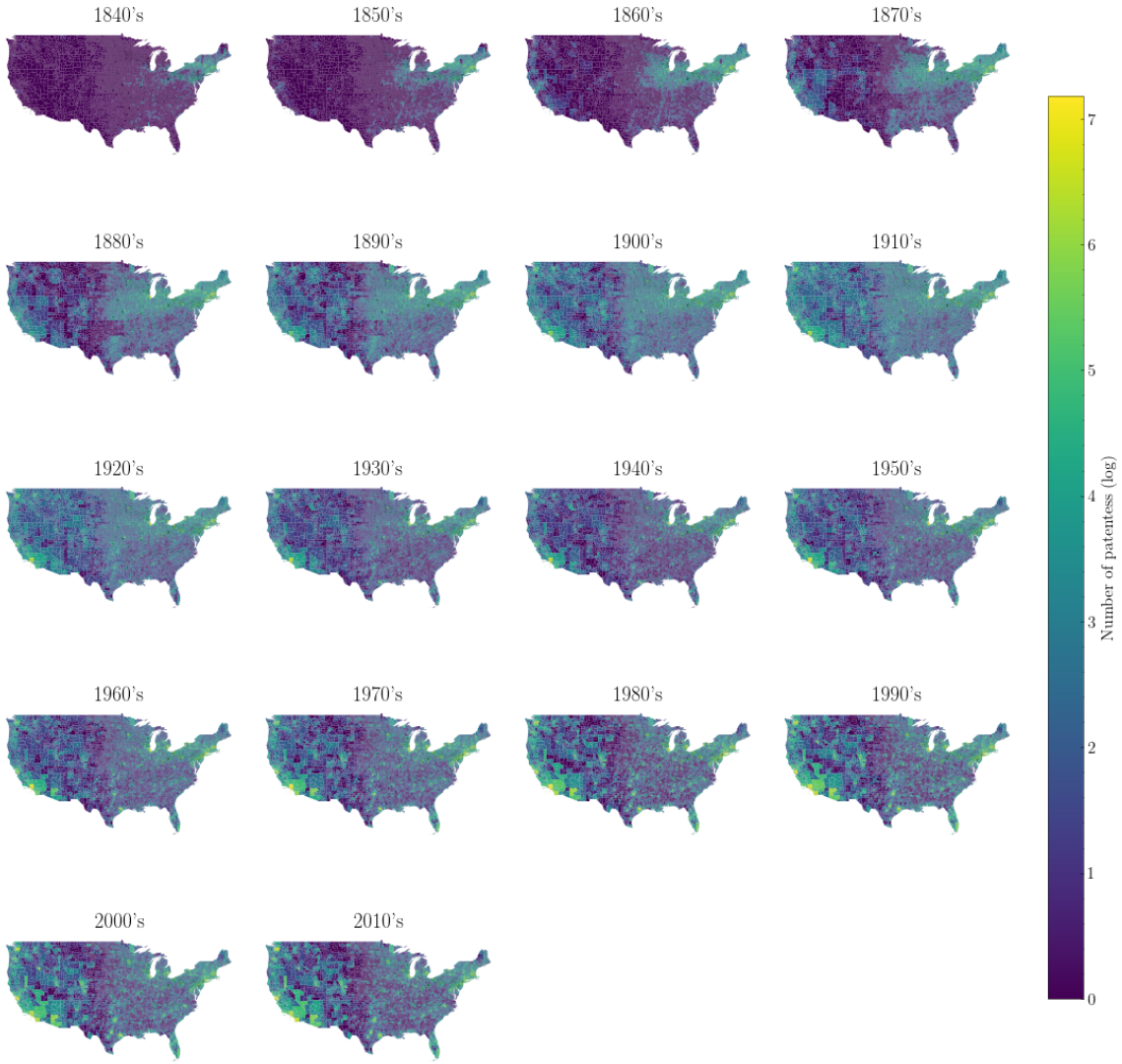
Notes: this Figure maps the total number of patentees (whether assignees or inventors), in log, for each county in France (*département*) for each decade. The number of patentees is taken as a total over the full set of domestic patentees that are located at least at the county level.

Figure B3: Patentees by regions and decade - United Kingdom



Notes: this Figure maps the total number of patentees (whether assignees or inventors), in log, for each county in the UK (NUTS 3 regions) for each decade. 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.

Figure B4: Patentees by regions and decade - United States



Notes: this Figure maps the total number of patentees (whether assignees or inventors), in log, for each county in the USA for each decade. 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.

C Model Cards

Details on the performance of the model are given in the website of the project. Specifically:

- Model cards for DD
- Model cards for DE
- Model cards for FR
- Model cards for GB
- Model cards for US

Model performance are also summarized in Tables C1, C2, C3, C4 and C5, respectively for East Germany, Germany, France, the United Kingdom and the United States.

Table C1: Models' performance by format in DD

Format	Metric	ALL	ASG	INV	LOC	OCC
	p	0.99	0.99	0.96	0.99	0.99
1	r	0.99	0.99	0.96	0.99	1
	f	0.99	0.99	0.96	0.99	0.99
	p	0.95	0.94	0.95	0.98	0.94
2	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 C2: Models' performance by format in DE

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

Table C3: Models' performance by format in FR

Format	Metric	ALL	ASG	CLAS	INV	LOC
	p	0.97	0.99	0.93	0.99	0.99
1	r	0.97	0.99	0.93	1	0.99
	f	0.97	0.99	0.93	0.99	0.99
	p	0.98	0.98	-	0.99	0.99
2	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 C4: Models' performance by format in GB

Format	Metric	ALL	ASG	CIT	INV	LOC	OCC
	p	0.93	0.93	0.96	0.95	0.92	0.9
1	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 C5: Models' performance by format in US

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

CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

1849	Jo Blanden Matthias Doepke Jan Stuhler	Education inequality
1848	Martina Manara Tanner Regan	Ask a local: improving the public pricing of land titles in urban Tanzania
1847	Rebecca Freeman Kalina Manova Thomas Prayer Thomas Sampson	Unravelling deep integration: UK trade in the wake of Brexit
1846	Nicholas Bloom Scott W. Ohlmacher Cristina J. Tello-Trillo Melanie Wallskog	Pay, productivity and management
1845	Martin Gaynor Adam Sacarny Raffaella Sadun Chad Syverson Shruthi Venkatesh	The anatomy of a hospital system merger: the patient did not respond well to treatment
1844	Tomaz Teodorovicz Raffaella Sadun Andrew L. Kun Orit Shaer	How does working from home during Covid-19 affect what managers do? Evidence from time-use studies
1843	Giuseppe Berlingieri Frank Pisch	Managing export complexity: the role of service outsourcing
1842	Hites Ahir Nicholas Bloom Davide Furceri	The world uncertainty index
1841	Tomaz Teodorovicz Andrew L. Kun Raffaella Sadun Orit Shaer	Multitasking while driving: a time use study of commuting knowledge workers to access current and future uses

1840	Jonathan Gruber Grace Lordan Stephen Pilling Carol Propper Rob Saunders	The impact of mental health support for the chronically ill on hospital utilisation: evidence from the UK
1839	Jan Bietenbeck Andreas Leibing Jan Marcus Felix Weinhardt	Tuition fees and educational attainment
1838	Jan De Loecker Tim Obermeier John Van Reenen	Firms and inequality
1837	Ralph De Haas Ralf Martin Mirabelle Muûls Helena Schweiger	Managerial and financial barriers during the green transition
1836	Lindsay E. Relihan	Is online retail killing coffee shops? Estimating the winners and losers of online retail using customer transaction microdata
1835	Anna D'Ambrosio Vincenzo Scrutinio	A few Euro more: benefit generosity and the optimal path of unemployment benefits
1834	Ralf Martin Dennis Verhoeven	Knowledge spillovers from clean and emerging technologies in the UK
1833	Tommaso Sonno Davide Zufacchi	Epidemics and rapacity of multinational companies
1832	Andreas Teichgraeber John Van Reenen	A policy toolkit to increase research and innovation in the European Union

The Centre for Economic Performance Publications Unit
Tel: +44 (0)20 7955 7673 Email info@cep.lse.ac.uk
Website: http://cep.lse.ac.uk Twitter: @CEP_LSE