Lee Kuan Yew School of Public Policy

Asia Competitiveness Institute

ACI Research Paper #07-2022

Imputation of Missing Information in PATSTAT Database: A Re-assessment

Yixuan GE

Taojun XIE

Chi ZHANG

First Version: 8 June 2022
Current Version: 16 June 2022

Please cite this article as:

Ge, Yixuan, Taojun Xie and Chi Zhang, "Imputation of Missing Information in PATSTAT Database: A Re-assessment", Research Paper #07-2022, *Asia Competitiveness Institute Research Paper Series (June 2022)*

Imputation of Missing Information in PATSTAT Database: A Re-assessment

Yixuan Ge Taojun Xie Chi Zhang *

June 16, 2022

Abstract

We impute the missing information on country codes and technology classifications at patent family level using the Spring 2022 version of the PATSTAT database. Given that a patent family is a collection of patents protecting the same invention in multiple countries, missing information on a patent family can be inferred from information on the patents within the same family. Where the information is missing, we follow Rassenfosse and Seliger (2021) and infer the information using relevant data available at the three earliest filings within the family, with earlier filing taking priority. The imputation delivers significant increase in the information coverage. We also examine the adequacy using three earliest filings within the family. Results show that most countries reach 100% coverage for country code among patent families with size greater than 1. We compare the patent quality rankings with and without imputing the missing information, and find that some countries' rankings improve with the imputation.

Keywords: PATSTAT, Patent family, Patent, Imputation, Country code, Technical field, IPC, NACE2 code

*Asia Competitiveness Institute, Lee Kuan Yew School of Public Policy, National University of Singapore. Ge: corresponding author. Email: yixuan.g@nus.edu.sg

1 Introduction

The PATSTAT database is a global database of patent statistics published by the European Patent Office on behalf of the Organization for Economic Co-operation and Development (EPO, 2020). It is regarded as the backbone data set for patent statistics research. According to the Spring 2022 edition (version 5.19), the PATSTAT contains over 100 million patents from 196 patent offices between 1782 and 2021, including patent bibliographic data, citation information, legal events, inventor and applicant data, etc.

Due to the high variety of the sources of data, an outstanding issue of this database lies in the missing information such as the country code and technology classification to which the patents belong. Researchers often find it hard to use certain data points when such information is missing (Pasimeni, 2019; Rassenfosse and Seliger, 2021). For example, as shown in Table 1, almost all patent applications received by the Japanese Patent Office do not contain the country codes of applicants and inventors. For the Intellectual Property Office of the United Kingdom, not only is the country code coverage rate at a low level, 0.94% in 1990, but also about half of the technology classification information are missing. Given the large volume of patent applications at these patent offices, missing key information in the data likely leads to biased country- or technology-specific analyses.

Various approaches have been proposed to address the issue of missing information. Pasimeni (2019) proposes to group *person_id* according to *doc_std_name_id* and assign the most frequently occurring non-empty *person_ctry_code* in each group of *doc_std_name_id* to the whole group of *person_id*. In the meantime, Pasimeni compares the performance of conducting the allocation procedure using fields *doc_std_name_id*, *psn_id* and *han_id* from the table *tls206_person*. The results show that using *doc_std_name_id* performs best, reducing blank or unknown *person_ctry_code* by 44%, while using *han_id* does nothing to increase data completeness. However, not all *doc_std_names* go through the standardization process, and not all DOCDB standardised names are linked to the correct inventors/applicants, especially for patents from the USPTO (EPO, 2020).

Among the latest ones, Rassenfosse and Seliger (2021) propose the perspective of patent families as an entry point. Patent protection is territorial, and usually, applicants file the

same patent in different patent offices to obtain protection in the respective country/region. These patents with the "same" contents form an international patent family. The feature allows a patent with missing information to obtain the corresponding information from other patents in the same family.

In this article, we apply Rassenfosse and Seliger (2021)'s approach to impute the missing information in the PATSTAT database. Despite the clear documentations provided by Rassenfosse and Seliger, there are several factors making it imperative to re-assess the performance of their approach. Firstly, the database structure has been changed significantly since Autumn 2019. The version of the database used by Rassenfosse and Seliger was Spring 2019, implying that these latest changes in the data were not captured in Rassenfosse and Seliger's work. For instance, in the Autumn 2019 version, there were significant changes to tables pertaining to inventors' and applicants' information. We assess the performance of the method in imputing missing information to the Spring 2022 edition of the database, including the *person_ctry_code* of the inventor, the *person_ctry_code* of the applicant, and the technology classification of the patent. Secondly, Rassenfosse and Seliger's imputation method did not go beyond 3 sources in a patent family, while in actual filing, applicants tend to file as many applications as they need. Therefore, we explore the effectiveness of missing information imputation after three rounds of source browsing to decide whether to continue browsing for more rounds.

Our re-assessment of the Rassenfosse and Seliger's imputation methodology finds the following. First, there is a more severe lack of country and technology information in the pre-imputation data, particularly among those filed at the patent offices in Japan and the the Great British. Second, Rassenfosse and Seliger' method proves to be still useful in filling the missing information. Third, as we make use of the 3 sources on country code specified by Rassenfosse and Seliger, we are able to fill the missing information in 100% of the patent families in the major patent offices.

The remainder of this paper is organized as follows. Section 2 introduces the detailed knowledge related to patents and the PATSTAT database. The imputation algorithm is presented in Section 3. Section 4 shows the results, and the last section concludes.

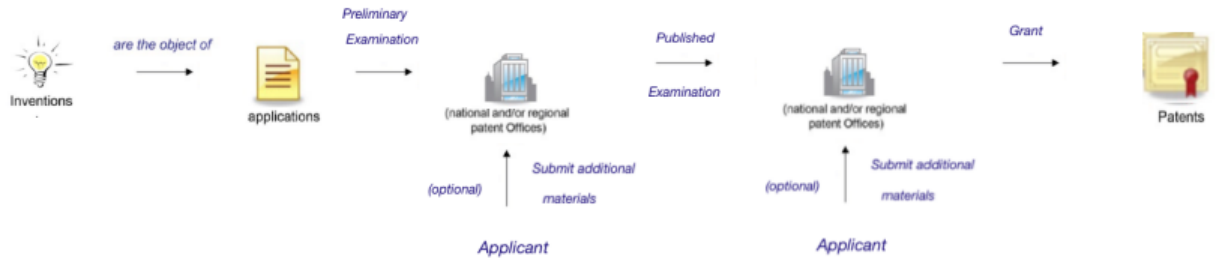
2 Background

In this section, we introduce a few key concepts in patent applications and the PATSTAT database.

2.1 General Patent Application Process

The inventor may seek patent protection when an invention is created by applying to one or several patent offices. Different patent offices implement different patent application processes, and the general steps that a patent goes through from application to grant are presented in Figure 1.

Figure 1: General Patent Application Process



Source: Created by Author

First, the applicant needs to determine the kind of application (e.g., invention patent, utility model, design patent, plant patent, and so on). Next, the applicant submits an application form to the patent office containing a technical and accurate description of the patent and information about the inventor(s) and applicant(s)[1]. Once the patent office receives the application form, it will conduct a preliminary examination of the patent. The applicant may be requested to submit additional materials or receive comments on modifications during this period. If all filing requirements are met, the applicant(s) will receive a filing date, and the patent office will publish the patent. Then, the patent will be requested for another examination, and the patent office will grant the application as a patent after the examination is successful. If the examination is unsuccessful, the applicant may appeal to examine the application again.

1. Some patent applications provide only inventor information or applicant information (EPO, 2020).

2.2 Kinds of Patent Application

Patents are subject to territorial limits. There are several ways in which you can have your invention protected in multiple countries — (1) by direct application, (2) by the Patent Cooperation Treaty, and (3) by the Paris Convention.

2.2.1 Direct Application

Applicant(s) can file separate patent applications at the same time with several patent offices in countries or regions where they would like to seek protection for the invention.

2.2.2 Patent Cooperation Treaty (PCT)

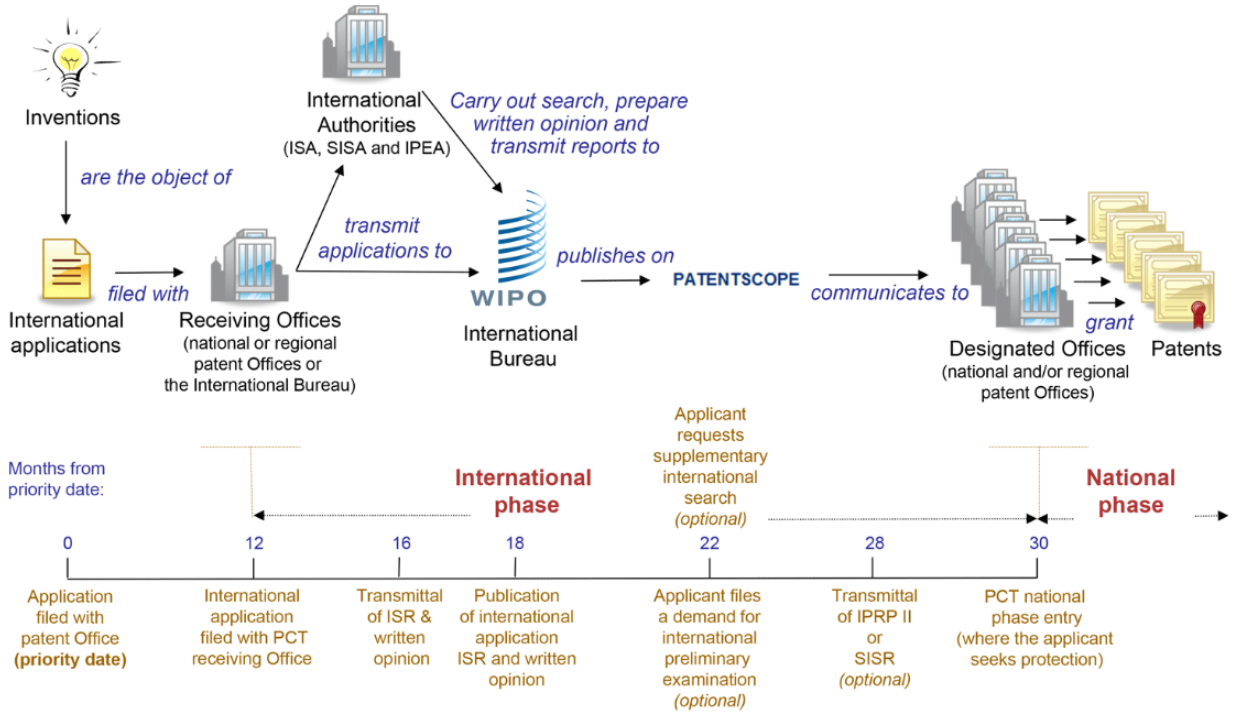
The PCT is an international treaty that makes it possible to provide protection for same invention in multiple countries simultaneously by filing one "international" patent application[2]. Usually, it can be requested either from the national or regional patent offices of the contracting states of which the applicant is a national or resident, or from WIPO. In PATSTAT, all patents filed through the PCT are marked with a "W" for the kind of application. An overview of the Patent Cooperation Treaty application process is shown in Figure 2.

2.2.3 Paris Convention

Since 1883, the Paris Convention has been adopted to apply industrial property in its broadest sense. One of the critical elements of the Convention is the priority of patents, which means that a patent applicant can apply for patent protection in one or more other contracting states within 12 months of a regular initial application in one contracting state. In addition, these subsequent applications will not interfere with anything that happens during the interval, such as disclosure of the invention. The Paris Convention provides applicants with 12 months to consider in which countries they would like their invention to

2. As of March 2022, the Patent Cooperation Treaty has 155 contracting states, the list of the states can be found on the website of WIPO at https://www.wipo.int/pct/en/pct_contracting_states.html. (Accessed March 14, 2022)

Figure 2: Overview of the Patent Cooperation Treaty (PCT) System



Source: WIPO (2020)

be protected. The filing steps and corresponding times for the Paris Convention and PCT are shown in Figure 3.

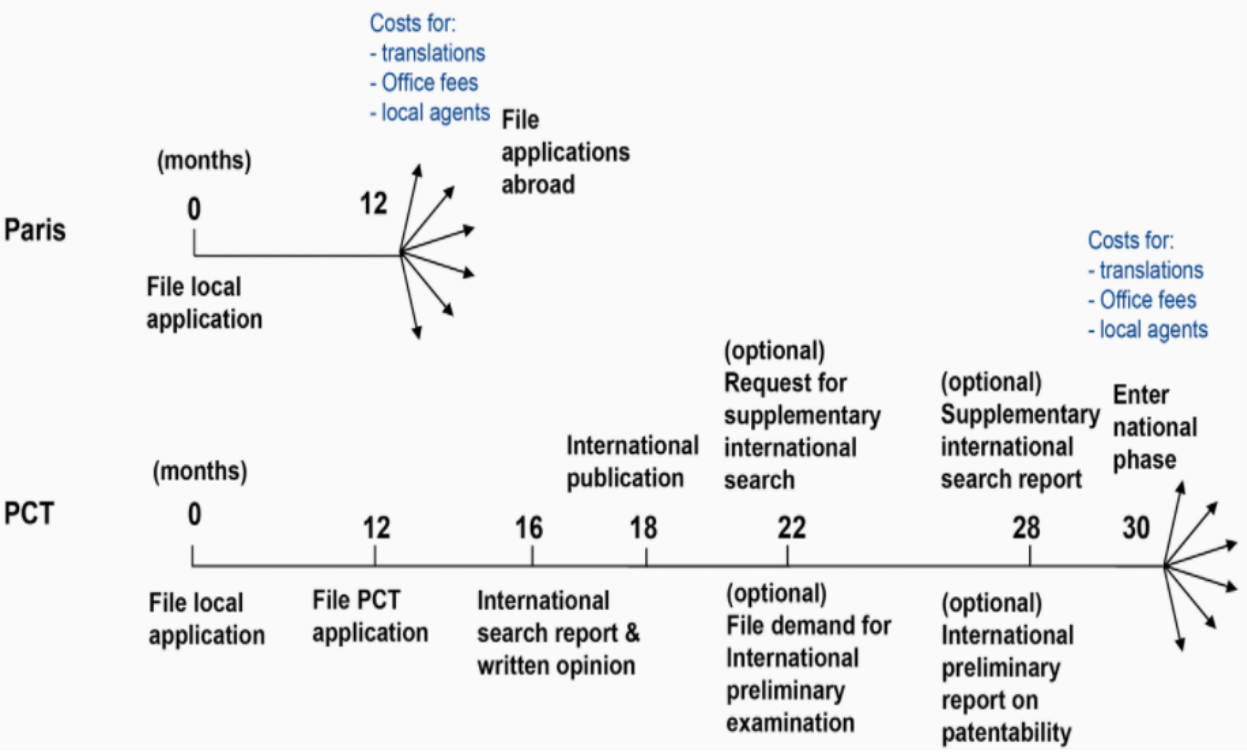
2.3 Priorities of Patent Application

In PATSTAT, the priority of a patent application is not only the "Paris Convention priority", but also includes other types of priority that link an application to a prior application, such as earlier PCT international application, technical relation, and continuation (EPO, 2020).

2.3.1 Earlier PCT international application

In the table *tls201_appln*, the field *internat_appln_id* stores the *appln_id* of the international application phase of the PCT (EPO, 2020). In the case of *internat_appln_id* is greater than 0, this application has an earlier PCT application, if it is 0, then there is no earlier PCT application.

Figure 3: The Comparison of Paris Convention and Patent Cooperation Treaty Route



Source: WIPO (2020)

2.3.2 Technical Relations

Technical relations are a technical solution for linking older applications without priority information to families. These relations are detected by European Patent Office (EPO) examiners or bibliographic data specialists when there are no other priority-like relations between applications. In addition, these relations are symmetric, which means if PATSTAT has a technical relation record "application A - application B", it will also include a record "application B - application A".

2.3.3 Continuations

Similar to the priority in the Paris Convention, the links between parent and child applications for various types of relations such as continuation (in part), divisional applications, internal priorities are defined via the table. Continuation (in part) is generally only applicable to US patent applications.

2.4 Patent Family

Applications with the same or similar invention are grouped into patent families. In PATSTAT, there are two kinds of patent families. One is a simple family, also called the DOCDB family. All patent applications within the same simple family have the same priority, and the technical content of these family members is regarded as almost identical. The other is the extended family, also called the INPADOC family. All patent applications within the extended family are directly or indirectly linked to the same root priority application. Usually, the INPADOC family members are related to the same technical invention, but their content may differ (EPO, 2020).

Every application belongs to exactly one DOCDB family and exactly one INPADOC family. The INPADOC family is potentially "broader" than the simple family, which means each INPADOC family contains the applications of one or more DOCDB families. We believe that applications within a DOCDB family share the same inventor(s)/applicant(s) and belong to the same IPC (Kang and Tarasconi, 2016). Thus, to infer missing information on country code and technology field, in this paper, we focus on imputation within DOCDB families.

2.5 International Patent Classification (IPC)

The IPC is a hierarchical technology classification established by the Strasbourg Agreement of 1971, which divides technology into eight sections and approximately 80,000 subdivisions (WIPO, 1971). In the Spring 2022 version of PATSTAT, the field *ipc_class_symbol* in table *tls209_appln_ipc* provides the classification code of the patent. In order to keep up with the development of technology, the new version of IPC takes effect on January 1 every year (WIPO, 1971).

2.6 Technical field

In PATSTAT, 35 technical fields are a more aggregated classification of technology derived from the IPC, patents without IPCs are not assigned to technical fields. In order to better measure the connection between the patents and the technical fields, PATSTAT

provides the field *weight* in the table *tls230_appln_techn_field*, which is a real number between 0 and 1. If a patent belongs to only one technical field "A", the weight of the technical field "A" of this patent is 1. If a patent belongs to multiple technical fields, taking two as an example, "A" and "B", the weights of the technical fields of this patent are a and b respectively, and the sum of a and b is always 1. The higher the weight, the closer the relationship between the patent and the corresponding technical field.

2.7 NACE2 code

According to EUROSTAT (2008), NACE is the acronym of "**N**omenclature générale des **A**ctivités économiques dans les **C**ommunautés **E**uropéennes", which is derived from the International Standard Industrial Classification (ISIC), and was used in the European Union since 1970. The Spring 2022 version of PATSTAT adopts the latest version of the NACE code (i.e., NACE Rev. 2). The 2-4 digit industry classification code is present in field *nace2_code* of table *tls229_appln_nace2* and *weight* indicates the corresponding weight. Since the table *tls229_appln_nace2* is calculated from *tls902_ipc_nace2*, applications without IPCs are not assigned to NACE2 code (EPO, 2020). Each application belongs to one or more industries, and the weight is assigned in a similar way to the weight of the technical field.

2.8 Country code

Patenting activity is territorial, with the applicant's *person_ctry_code* reflecting where the invention belongs and the inventor's *person_ctry_code* reflecting where the invention originated (De Rassenfosse, Dernis, and Boedt, 2014; Kang and Tarasconi, 2016). Depending on the research objectives, the *person_ctry_code* of the inventor or applicant is used as the reference country for the patent (De Rassenfosse et al., 2013). In order to obtain as informative and complete data as possible, the *person_ctry_code* of both the applicant and the inventor is considered in this paper. For patents where neither the inventor nor the applicant's *person_ctry_code* is available, the country in which the patent office where the application is filed is considered the patent country code (Rassenfosse and Seliger, 2021).

3 Method

We follow Rassenfosse and Seliger (2021)'s methodology to impute missing information on country code, international patent classification (IPC), technical field, and NACE2 of patents. The choice of patent offices is consistent with Rassenfosse and Seliger (2021)'s selection, i.e., patent offices from all OECD countries, all EU countries, Switzerland, Norway, the BRICS countries, the EPO and WIPO. The size of patent applications received by these 52 patent offices covers almost the total size of the world, which is 92%[3]. For the kinds of earliest patent applications, as we describe in section 2.2, including "Paris Convention", "Technical relations", and "Continuations" relations. There is no restriction on the application year of the patent, that is, all patents filed before or equal to 2021 included in the 2022 Spring version of PATSTAT are imputed.

3.1 The Imputation Algorithm for Country Code

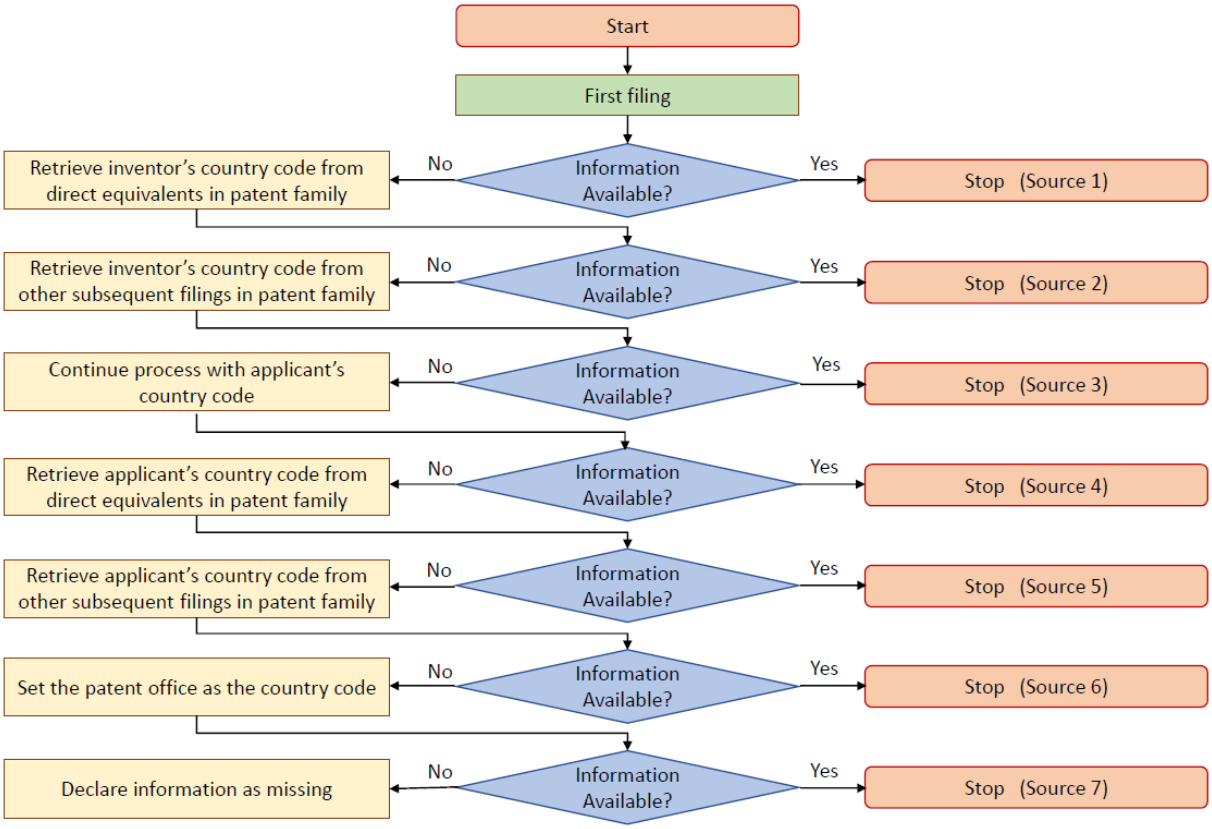
The flowchart of the imputation algorithm for country code is present in Figure 4. First, we create a table containing all earliest patent applications within a patent family as source 1, with or without the desired country code information.

Second, we focus on patents whose information is not available from source 1. In order to obtain information about these patents, we need to look at the availability of relevant information at earliest-second filed patents within the same patent family. These subsequent patents are patents filed in other patent offices after the earliest patent application. Therefore, we first prepare a pool containing all subsequent applications, from which we select the earliest subsequent application within each patent family. If there is a situation where there are multiple earliest subsequent patents, only the information relating to the patent with the smallest *appln_id* is selected. The filtered pool of subsequent applications constitutes source 2.

Third, we consider obtaining information from other subsequent patent applications for patents that still have missing information after the first two steps. The flow and requirement are conducted similarly to the second step, with the imputed results stored as source 3.

3. See Table A1 for list of detailed selected patent offices.

Figure 4: The Flowchart of the Imputation Algorithm for Country Code



Source: Created by Author

As shown in the flowchart Figure 4, in addition to using the inventor's country code as the country code of the patent, we consider the applicant's country code. The imputation process for the applicant's country code is similar to the inventor's and saves the results as sources 4-6.

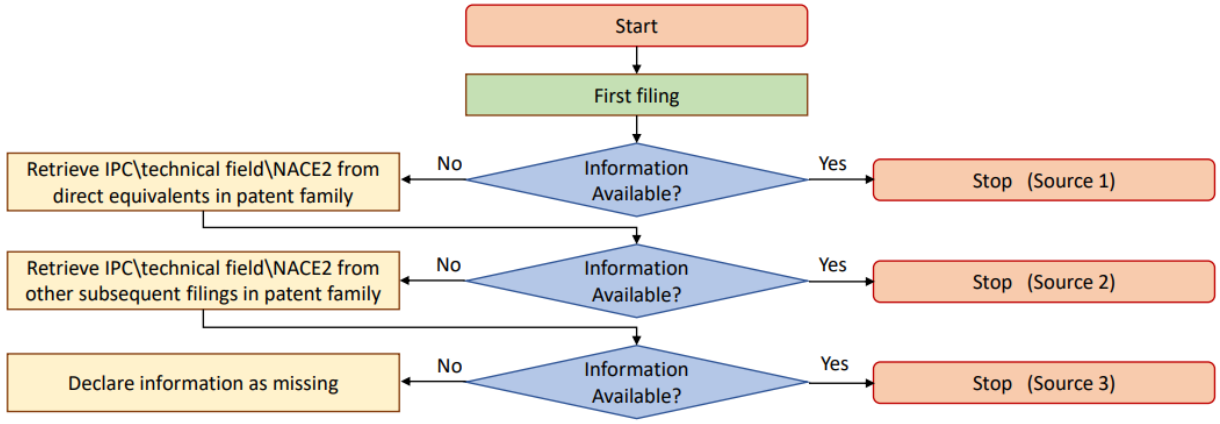
If neither the inventor's nor the applicant's country code is available, Rassenfosse and Seliger (2021) sets the code of the national patent office as the country code of the patent, saving the result as source 7. We propose that, in this case, the receiving patent office (considering only the national/regional patent office) of a patent filed at WIPO through the PCT can also be included in source 7[4]. For patents filed under the PCT, the receiving office is considered their country code.

4. The patent offices (including receiving offices) in source 7 do not include the organizational patent offices: the European Patent Office (EPO), the African Regional Intellectual Property Organization (ARIPO), the Eurasian Patent Organization (EAPO), the Patent Office of the Cooperation Council for the Arab States of the Gulf (GCCPO), the African Intellectual Property Organisation (OAPI), the World Intellectual Property Organisation (WIPO).

3.2 The imputation algorithm for IPC, Technical Field or NACE2

The flowchart of the imputation algorithm for IPC, technical field or NACE2 is present in Figure 5. Similar to the method of imputing country code, the first three steps are performed in sequence to obtain Sources 1-3 respectively. For patents whose classification information is still missing after three steps, the classification information on these patents is declared missing.

Figure 5: The Flowchart of the Imputation Algorithm for IPC, Technical Field or NACE2



Source: Created by Author

4 Results

We conduct the imputation from 1782 to 2021 for 52 patent offices. In what follows, we display the imputation result of ten major patent offices in three selected years — 1990, 2000 and 2010, for illustration purpose.

4.1 Imputation of Country Codes

Overall, as shown in Table 1, the imputation method effectively improved the country code coverage of first filing patents. In particular, after imputation of inventor and applicant country codes, the Intellectual Property Office of the United Kingdom (1990) and the Intellectual Property Office in France (1990, 2000) witness a significant increase in coverage, with an average increase of 80%. The Japan Patent Office (1990, 2000, 2010) and the Chinese Patent Office (2010), where country code coverage is almost zero before imputation, increased but still at a low level (between 2.88% and 20.66%) after browsing sources 1-6. Given that sources 2-3 and sources 5-6 make use of the information provided by subsequently filed applications within a patent family, these steps can only be conducted if the patent family size is more than one.

Patent families with size equal to 1 are usually of small market value. And most likely, the origin countries of these patent families are in line with the patent offices where they are applied. To infer the country code of patent families with size equal to 1, we fill in the patent office code as the country code as source 7. After step 7, all selected patent offices achieved 100% coverage of the country code for the first filing patents received in 1990, 2000, and 2010.

Imputation performance for patents with patent family sizes more than one is presented in Table 2. For the selected patent offices and years, 90% of the patent applications received by the patent offices in China (1990, 2000, 2010), Japan (1990, 2000), and Korea (1990, 2000) are filed in the original patent office only. Prior to imputation, early patent applications filed at the Chinese patent office contained complete applicant country codes, e.g., 1990 and 2000, but by 2010 this coverage was only 0.08%. For the Japan patent office (1990, 2000, 2010), the coverage of country codes is always lower than 0.5%. After the first three steps (sources

Table 1: The coverage of available information on country code before and after imputation

Patent Office	Year	Inventor country before imputation Source 1 (%)	Inventor country after imputation Sources 1-3 (%)	Inventor and applicant country after imputation Sources 1-6 (%)	Imputation after all steps Sources 1-7 (%)
CA	1990	99.61%	99.61%	99.75%	100.00%
CN	1990	99.68%	99.68%	99.89%	100.00%
DE	1990	74.00%	85.90%	88.81%	100.00%
EP	1990	99.37%	99.41%	100.00%	100.00%
FR	1990	1.10%	49.76%	83.02%	100.00%
GB	1990	0.94%	27.55%	80.23%	100.00%
JP	1990	0.05%	8.94%	9.02%	100.00%
KR	1990	99.92%	99.92%	100.00%	100.00%
US	1990	92.07%	92.09%	92.20%	100.00%
WO	1990	99.95%	99.95%	100.00%	100.00%
CA	2000	98.94%	99.56%	99.76%	100.00%
CN	2000	99.88%	99.88%	99.95%	100.00%
DE	2000	99.96%	99.97%	99.99%	100.00%
EP	2000	99.53%	99.65%	100.00%	100.00%
FR	2000	1.05%	60.14%	99.99%	100.00%
GB	2000	99.64%	99.80%	99.92%	100.00%
JP	2000	0.03%	12.97%	13.00%	100.00%
KR	2000	99.11%	99.17%	99.17%	100.00%
US	2000	95.27%	95.48%	97.27%	100.00%
WO	2000	83.00%	88.37%	100.00%	100.00%
CA	2010	99.61%	99.73%	100.00%	100.00%
CN	2010	0.00%	2.84%	2.88%	100.00%
DE	2010	99.99%	100.00%	100.00%	100.00%
EP	2010	99.59%	99.76%	100.00%	100.00%
FR	2010	99.99%	99.99%	100.00%	100.00%
GB	2010	99.29%	99.73%	99.98%	100.00%
JP	2010	0.01%	20.27%	20.66%	100.00%
KR	2010	100.00%	100.00%	100.00%	100.00%
US	2010	100.00%	100.00%	100.00%	100.00%
WO	2010	97.07%	98.52%	99.98%	100.00%

1-3), almost all patent offices have 95% and higher country code coverage. Moreover, the coverage rate theoretically increases with the rounds of imputation implemented. Table 2 indicates that three rounds of searching is sufficient for almost all countries we cover here.

4.2 Imputation of International Patent Classification (IPC), Technical Field, and NACE2 Code

The result of the imputation for the IPC, technical field, and NACE2 code of first filings patents is present in Table 3. Before imputation, over 80% of patent offices' coverage is more than 98%. The only patent office with less than half coverage was the Intellectual Property Office of the United Kingdom, with 46.8% in 1990 and 33.4% in 2010. After browsing sources

Table 2: The coverage of available information on country code before and after imputation (patent family size more than one)

Patent Office	Year	Percentage of first filings with family size >1	Inventor country before imputation Source 1 (%)	Inventor country after imputation Sources 1-3 (%)	Inventor and applicant country after imputation Sources 1-6 (%)	Imputation after all steps Sources 1-7 (%)
CA	1990	23.86%	100.00%	100.00%	100.00%	100.00%
CN	1990	0.79%	100.00%	100.00%	100.00%	100.00%
DE	1990	44.49%	72.49%	97.66%	99.20%	100.00%
EP	1990	86.37%	99.39%	99.44%	100.00%	100.00%
FR	1990	48.06%	2.00%	97.38%	99.81%	100.00%
GB	1990	32.32%	2.02%	81.07%	98.75%	100.00%
JP	1990	4.94%	0.32%	97.05%	98.12%	100.00%
KR	1990	6.27%	99.89%	100.00%	100.00%	100.00%
US	1990	32.37%	99.46%	99.49%	99.49%	100.00%
WO	1990	71.33%	100.00%	100.00%	100.00%	100.00%
CA	2000	37.76%	98.20%	99.84%	99.95%	100.00%
CN	2000	2.39%	99.78%	99.78%	99.89%	100.00%
DE	2000	52.18%	99.99%	100.00%	100.00%	100.00%
EP	2000	67.52%	99.59%	99.76%	100.00%	100.00%
FR	2000	57.99%	1.47%	98.43%	100.00%	100.00%
GB	2000	49.79%	99.67%	99.97%	100.00%	100.00%
JP	2000	9.99%	0.11%	98.48%	98.67%	100.00%
KR	2000	7.65%	99.16%	99.79%	99.79%	100.00%
US	2000	46.59%	99.22%	99.53%	99.87%	100.00%
WO	2000	83.25%	86.23%	92.58%	100.00%	100.00%
CA	2010	36.84%	99.79%	100.00%	100.00%	100.00%
CN	2010	3.00%	0.08%	86.85%	88.19%	100.00%
DE	2010	53.83%	99.99%	100.00%	100.00%	100.00%
EP	2010	73.80%	99.55%	99.78%	100.00%	100.00%
FR	2010	60.68%	100.00%	100.00%	100.00%	100.00%
GB	2010	50.76%	99.17%	100.00%	100.00%	100.00%
JP	2010	19.25%	0.03%	93.19%	94.92%	100.00%
KR	2010	15.95%	99.98%	100.00%	100.00%	100.00%
US	2010	50.07%	100.00%	100.00%	100.00%	100.00%
WO	2010	56.16%	97.27%	99.85%	100.00%	100.00%

2 and 3, the coverage increases to 54.0% and 59.5%, respectively.

In PATSTAT, technical fields and NACE2 codes are a more aggregated classification derived from the IPC, so each patent office's coverage of technical fields and NACE2 code is similar to the IPC, both before and after imputation.

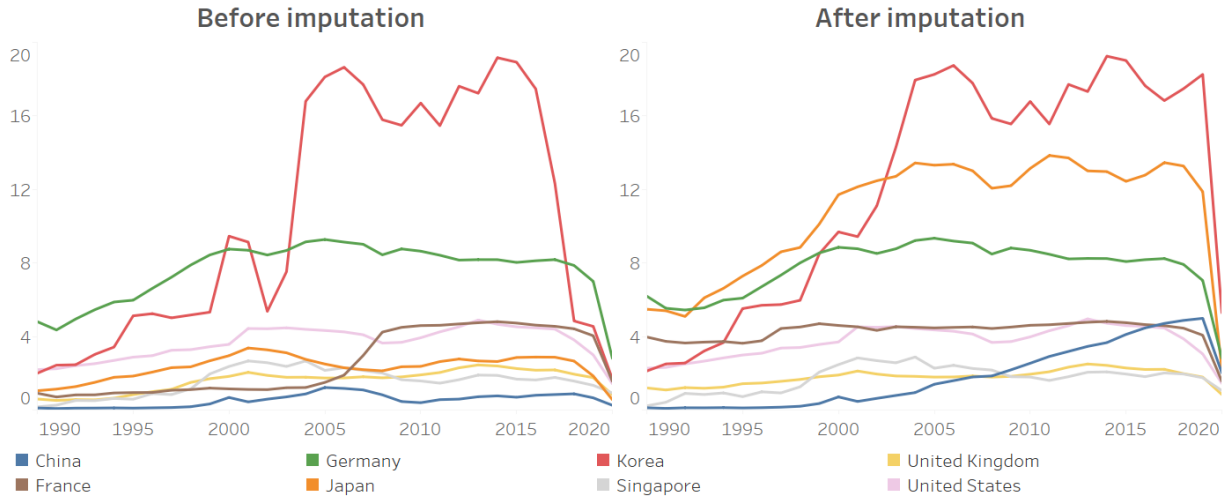
4.3 Comparative analysis before and after imputation

Figure 6 presents the results of the comparison of patent quality for selected countries before and after imputation. We use number of priority applications, conditioning on the patent family size being greater than 1, as a measure of patent quality. The absence of the inventor's or applicant's country code has the greatest impact on the measurement of patent quality in Japan, jumping from a low ranking before imputation to the top two levels (Japan ranked first in the period 1992-2002 and second in the period 2003-2020). The rates

Table 3: The coverage of available information on classification before and after imputation

Patent office	Year	IPC before imputation Source 1 (%)	IPC after imputation Sources 1-3(%)	Technical field before imputation Source 1 (%)	Technical field after imputation Sources 1-3(%)	NACE2 before imputation Source 1 (%)	NACE2 after imputation Sources 1-3(%)
CA	1990	95.77%	96.07%	95.77%	96.09%	95.75%	96.07%
CN	1990	97.79%	97.82%	97.79%	97.82%	97.72%	97.75%
DE	1990	92.01%	99.70%	92.01%	99.70%	92.00%	99.68%
EP	1990	93.85%	99.98%	93.85%	99.98%	93.85%	99.98%
FR	1990	98.74%	99.28%	98.74%	99.30%	98.73%	99.29%
GB	1990	46.22%	53.42%	46.17%	53.57%	46.22%	53.58%
JP	1990	97.01%	99.72%	97.01%	99.73%	97.01%	99.72%
KR	1990	76.57%	78.13%	76.57%	78.14%	76.57%	78.14%
US	1990	66.50%	91.60%	66.50%	91.60%	66.50%	91.60%
WO	1990	98.79%	99.89%	98.79%	99.89%	98.79%	99.89%
CA	2000	93.62%	94.68%	93.62%	94.72%	93.62%	94.72%
CN	2000	98.54%	99.33%	98.54%	99.33%	98.32%	99.11%
DE	2000	91.48%	99.54%	91.48%	99.55%	91.46%	99.53%
EP	2000	77.69%	99.89%	77.69%	99.92%	77.69%	99.92%
FR	2000	97.56%	99.16%	97.56%	99.16%	97.56%	99.16%
GB	2000	50.13%	59.20%	50.10%	59.58%	50.12%	59.58%
JP	2000	90.18%	97.53%	90.18%	97.53%	90.17%	97.53%
KR	2000	74.41%	76.01%	74.41%	76.01%	74.39%	75.99%
US	2000	58.32%	92.59%	58.32%	92.61%	58.32%	92.61%
WO	2000	98.29%	99.95%	98.29%	99.95%	98.29%	99.95%
CA	2010	66.79%	67.41%	66.79%	67.41%	66.74%	67.34%
CN	2010	99.04%	99.92%	99.04%	99.92%	98.99%	99.87%
DE	2010	91.26%	99.79%	91.26%	99.80%	91.13%	99.71%
EP	2010	70.28%	99.79%	70.28%	99.85%	70.26%	99.84%
FR	2010	97.70%	99.55%	97.70%	99.58%	97.67%	99.58%
GB	2010	33.34%	59.30%	33.34%	59.66%	33.29%	59.66%
JP	2010	87.33%	99.08%	87.33%	99.08%	87.31%	99.07%
KR	2010	95.74%	99.31%	95.74%	99.31%	95.73%	99.30%
US	2010	55.53%	91.81%	55.53%	91.85%	55.52%	91.84%
WO	2010	97.50%	99.92%	97.50%	99.92%	97.46%	99.89%

Figure 6: Number of priority patent applications with patent family size greater than 1 per 1 billion of GDP (in US$) per priority year



Source: Created by Author

in France and China also show a marked change. Whereas the former's patent quality was underestimated in the first 15 years due to the absence of country codes, after imputation, France has largely maintained a stable situation of four priority applications with a patent

family size greater than one per billion US dollars of GDP over the 30-year period from 1990 to 2020. For the latter, the imputed data better captures China's progress in the R&D, from overtaking the UK and Singapore in 2009 to overtaking the US and France in the top four in 2017.

5 Conclusion

In this paper, we impute the missing country code and technology information in the PATSTAT database. We apply the method proposed by (Rassenfosse and Seliger, 2021), with some important changes in the database in mind, to the Spring 2022 version of the database. We find that the pre-imputation data exhibit a more severe lack of information than the version used by Rassenfosse and Seliger. After applying Rassenfosse and Seliger's method, and with additional sources included, the information can be imputed in almost all applications in the major patent offices in the world. We also briefly show that the imputation improved some countries' ranking in patent quality.

References

De Rassenfosse, Gaètan, Helene Dernis, Dominique Guellec, Lucio Picci, and Bruno van Pottelsberghe de la Potterie. 2013. "The worldwide count of priority patents: A new indicator of inventive activity." *Research Policy* 42 (3): 720–737.

De Rassenfosse, Gaétan, Hélène Dernis, and Geert Boedt. 2014. "An introduction to the Patstat database with example queries." *Australian economic review* 47 (3): 395–408.

EPO. 2020. *Data Catalog PATSTAT Global - 2020 Autumn Edition.* Available: https://www.epo.org/searching-for-patents/business/patstat.html.

EUROSTAT. 2008. *NACE Rev. 2 — Statistical classification of economic activities in the European Community.* https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF.pdf/dd5443f5-b886-40e4-920d-9df03590ff91?t=1414781457000. Accessed April 19, 2022.

Kang, Byeongwoo, and Gianluca Tarasconi. 2016. "Patstat revisited: Suggestions for better usage." *World patent information* 46:56–63.

Pasimeni, Francesco. 2019. "SQL query to increase data accuracy and completeness in PATSTAT." *World Patent Information* 57:1–7.

Rassenfosse, Gaétan de, and Florian Seliger. 2021. "Imputation of missing information in worldwide patent data." *Data in Brief* 34:106615.

WIPO. 1971. *Summary of the Strasbourg Agreement Concerning the International Patent Classification (1971).* https://www.wipo.int/treaties/en/classification/strasbourg/summary_strasbourg.html. Accessed April 19, 2022.

———. 2020. *Protecting your Inventions Abroad: Frequently Asked Questions About the Patent Cooperation Treaty (PCT).* https://www.wipo.int/pct/en/faqs/faqs.html. Last modified April 2020, Accessed April 18, 2022.

A Appendix

A.1 Code availability

The imputation of missing information on country code and technology classification was implemented in Microsoft SQL Server Management Studio 18. The data extraction, transformation and analysis were done in Python 3.7.3. All the SQL code produced for this project can be accessed via our GitHub[5].

A.2 List of Patent Offices to Explore

Table A1: List of Patent Offices to Browse

No	Code of patent office	Patent office	No	Code of patent office	Patent office
1	AL	Albania	27	IN	India
2	AT	Austria	28	IS	Iceland
3	AU	Australia	29	IT	Italy
4	BE	Belgium	30	JP	Japan
5	BG	Bulgaria	31	KR	Republic of Korea
6	BR	Brazil	32	LT	Lithuania
7	CA	Canada	33	LU	Luxembourg
8	CH	Switzerland	34	LV	Latvia
9	CL	Chile	35	MK	North Macedonia
10	CN	China	36	MT	Malta
11	CY	Cyprus	37	MX	Mexico
12	CZ	Czechia	38	NL	Netherlands
13	DE	Germany	39	NO	Norway
14	DK	Denmark	40	NZ	New Zealand
15	EE	Estonia	41	PL	Poland
16	EP	European Patent Office (EPO)	42	PT	Portugal
17	ES	Spain	43	RO	Romania
18	FI	Finland	44	RS	Serbia
19	FR	France	45	RU	Russian Federation
20	GB	United Kingdom	46	SE	Sweden
21	GR	Greece	47	SI	Slovenia
22	HR	Croatia	48	SK	Slovakia
23	HU	Hungary	49	SM	San Marino
24	IB	International Bureau of the World Intellectual Property Organization (WIPO)	50	TR	Turkey
25	IE	Ireland	51	US	Uruguay
26	IL	Israel	52	ZA	South Africa

5. https://github.com/Yixuan-Ge/Imputation-of-Missing-Information-in-PATSTAT-Database-A-Re-assessment.