World Patent Information 57 (2019) 1–7

Contents lists available at ScienceDirect

World Patent Information

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

SQL query to increase data accuracy and completeness in PATSTAT

Francesco Pasimeni[a,b,*]

[a] European Commission, Joint Research Centre (JRC), Petten, Netherlands
[b] University of Sussex, Science Policy Research Unit (SPRU), Brighton, United Kingdom

ARTICLE INFO

Keywords:
PATSTAT
Data accuracy
Data completeness
Patent data
Data cleaning
Climate change mitigation technology

ABSTRACT

PATSTAT is the worldwide patent statistical database created and maintained by the European Patent Office. Many methods and techniques have been developed to increase its accuracy and completeness. This paper contributes to this body of research. It proposes an allocation procedure which reduces by 44% the number of empty entries concerning the residence country of patentees, and, at the same time, it increases by 22% the accuracy of country code allocation. The procedure consists of a replicable SQL query to be run in PATSTAT. An application of this procedure illustrates that patent analyses based on raw data underestimate the role of China and Japan in the area of climate change mitigation technologies.

1. Introduction

Patent data provide important empirical evidences to science and technology studies [1–3]. Together with a detailed description of the technical progress achieved, patent data provide additional information on the process of technical change and on the actors involved. The bibliographic data contained in every patent application indicate the date when it is filed and in which patent office, the name(s) of applicant(s) and inventor(s) and their residence country, the classification code(s) indicating which technology field is tackled and prior patent applications on which the current patent is based, if any.

Given this rich set of information, the scientific community has developed guidelines on how to use patent data [4] and on how to build proxies and indicators regarding technological change [5–7]. Accordingly, patent data are used to study and evaluate knowledge and technology transfer [8,9], technology international diffusion [10], network of collaborations [11,12], firms' tangible (technology) and intangible (know-how) assets [13,14], trajectories of technological change [15] and innovation performances [16,17].

Patent data are easily available through PATSTAT, which is the worldwide patent statistical database created and maintained by the European Patent Office (EPO). It contains bibliographical data and legal status information of all patent applications from leading industrialised and developing countries. The database is updated twice a year, early spring and early autumn [18]. According to de Rassenfosse et al. [19], PATSTAT is the most prominent patent data source, and it is widely used. The structured query language (SQL) permits to interrogate the database, which is accessible to scientists and to other interested users via a very user friendly online interface.

PATSTAT collects patent data directly from the European Patent Office and from other sources, such as national and supranational patent authorities. However the incomplete provision of data from national authorities generates lack of accuracy and completeness, for which the EPO does not assume any legal liability or responsibility [18, p. 16]. Since this issue is well-known in the scientific community, several methods and techniques have been proposed in order to increase quality of research outcomes and to avoid distort information [11,20–23].

This paper contributes to this body of research and presents a simple way to increase data accuracy and completeness in PATSTAT. Available approaches to data cleaning and data harmonisation are proven to be very effective in producing better and more reliable outcomes. However, these are often based on advanced techniques which apply complex algorithms or artificial intelligence to manage big data. Consequently it is difficult to reuse or replicate these approaches. Contrarily, the procedure proposed in this paper consists of a simple query to be run in PATSTAT and, therefore, it is easily replicable.

The paper is structured as follows. Section 2 discusses the role of identifiers in PATSTAT and describes inconsistencies regarding the allocation of country code to applicants and inventors. Subsequently, the country allocation procedure is presented: it is a simple query which permits to reduce by 44% the number of *blank* entries and to increase by 22% country code allocation. By using the same rationale, it is then shown that the proposed allocation procedure increases the accuracy and completeness of two other attributes assigned to patentees: regional code and sector. In section 3 the proposed allocation procedure is tested

---

[*] European Commission, Joint Research Centre (JRC), P.O. Box 2, NL-1755 ZG Petten, the Netherlands.
E-mail address: Francesco.PASIMENI@ec.europa.eu.

https://doi.org/10.1016/j.wpi.2019.02.001
Received 27 June 2018; Received in revised form 19 December 2018; Accepted 10 February 2019

against raw data in PATSTAT by comparing the outcomes of a patent analysis conducted in the case of climate change mitigation technologies. It is shown that higher data accuracy and completeness provides better and more realistic research outcomes. Section 4 concludes.

2. The allocation procedure

The PATSTAT Data Catalog [18] describes in detail the structure of the database, how it is built and the logic behind tables and attributes (or fields). Table *tls201_appln* contains bibliographical data concerning all patent applications and table *tls206_person* gives information on applicants and inventors. These two tables are linked to each other via another table, *tls207_pers_appln*, which allows the identification and distinction between patent applicant(s) and inventor(s). The use of these three tables is essential to study the geographical provenience of actors (or entities) involved in any patenting activity. The residence country of inventors indicates where the inventive activity is undertaken, while the country of applicants indicates the location of the owners of the invention [19].

In PATSTAT, information on the residence country is in table *tls206_person*, specifically in the field *person_ctry_code*. *person_id* is the primary key of this table, and, differently from what could be thought, does not represent a unique entity in the database. Instead, it is a surrogate key for all combinations of three other fields: *person_name*, *person_address* and *person_ctry_code*. This implies that several *person_id* may indicate the same entity. In order to provide harmonised information, the standardisation procedure occurring in the DOCDB, the EPO's master bibliographic database, defines an additional identifier and name, which group several *person_id* under a unique entity. These are the fields *doc_std_name* and its related *doc_std_name_id*, both contained in table *tls206_person*. Therefore all *person_id* grouped under one *doc_std_name_id* represent the same entity.[1] PATSTAT users would expect that only one country code is assigned to these identifiers and, therefore, that only one *person_ctry_code* is assigned to each *doc_std_name_id*. However, this is not the case as acknowledged by the EPO [18, pp. 47, 280].

In order to show an example of these inconsistencies, Table 1 summarises the result of the query run in PATSTAT Online (2018 spring version) that searches and retrieves all *person_id*, and the related *person_ctry_code*, that have *doc_std_name* = 1. This identifier represents the Finnish Nokia Corporation and groups together 174 different entries. 130 of them are associated correctly with the country code *'FI'*, 20 of them with the United States, 10 do not have any code and the remaining are associated with several other countries (Table 1). Despite this lack of accuracy, it is worth noting that a country code occurs more frequently than the others. In this example, about 75% of *person_id* are assigned correctly to Finland. Therefore, it is reasonable to assume that also the remaining *person_id*, grouped under the *doc_std_name_id* = 1, can be assigned to the same country.

The allocation procedure proposed in this paper is based on this rationale. It is assumed that the *person_ctry_code* associated more frequently to one *doc_std_name_id* is the correct one, and that can be automatically assigned to all *person_id* grouped under the *doc_std_name_id* itself. There are more than 56 million distinct *person_id* in PATSTAT, and these are harmonised by more than 25 million *doc_std_name_id*. One simple query permits the detection of the country most frequently assigned to each *doc_std_name_id* and its consequent automatic allocation to the connected *person_id*, as shown in Fig. 1. The result of this query is the list of all *person_id* in table *tls206_person*, grouped under the relative *doc_std_name_id* and *doc_std_name*, to which a unique *person_ctry_code* is assigned.

Table 1
person_ctry_code assigned to Nokia Corporation: *doc_std_name_id = 1*.

person_ctry_code	Country	Count of *person_id*
FI	Finland	130
US	United States of America	20
(blank)	unknown	10
FR	France	4
CN	China	2
GB	United Kingdom	2
CA	Canada	1
ID	Indonesia	1
IN	India	1
KI	Kiribati	1
NL	Netherlands	1
SG	Singapore	1
		174

The proposed allocation procedure consists of three embedded queries. The first sub-query counts, for each *doc_std_name_id*, all records in table *tls206_person* (that is the number of *person_id*) grouped by *person_ctry_code* and ranks these values in descending order. In this query, it is important to notice two elements. The first one concerns the *WHERE* condition which limits the search only to *person_ctry_code* with values, that is those that are not *blank*. In this way it is possible to eliminate the risk to pick up *person_ctry_code = blank* when it is assigned more frequently than another code. The second regards the fact that, when two or more codes have been assigned to the same number of *person_id* under a *doc_std_name_id*, these codes are ordered alphabetically. This condition could add a bias to the allocation procedure only when two or more *person_ctry_code* are assigned to the same number of *person_id* under the same *doc_std_name_id* and, simultaneously, they are ranked first. Nevertheless, it has been calculated that this combination of events occurs only to 1.6% of the *doc_std_name_id*, meaning that there is almost always one country code that appears more frequently than others.

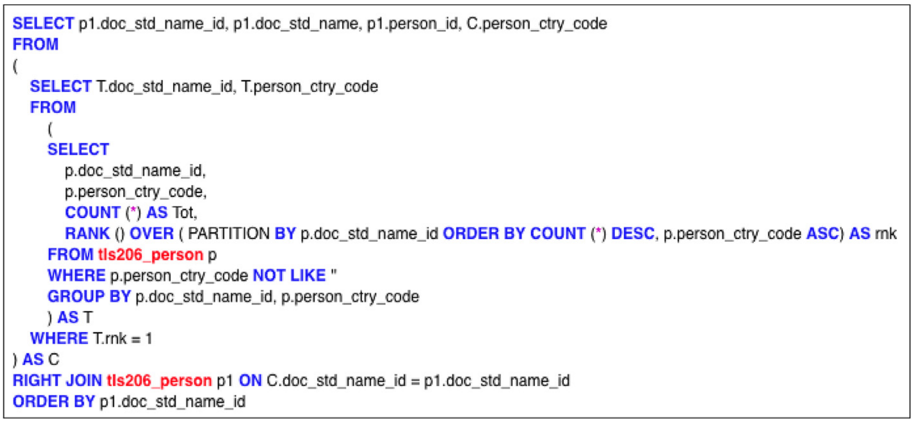
The second sub-query selects from the result of the first one all *doc_std_name_id* and assigns to these only the relative *person_ctry_code* ranked first. The last embedded query considers again table *tls206_person* which is right-joined with the second sub-query in order to not miss those *doc_std_name_id* for which the only available *person_ctry_code* is *blank*. This final query, therefore, lists all *person_id* in table *tls206_person* and the relative country code which is allocated consistently among all records standardised under the same *doc_std_name_id* and *doc_std_name*.

This simple query reduces by 44% the number of *person_id* without *person_ctry_code* (those that are *blank*) compared to raw data in PATSTAT (Fig. 2). Simultaneously, the proposed procedure increases by 22%, on average, the accuracy of country allocation. Currently 34% of *person_id* in table *tls206_person* do not have a country code assigned to them. By means of the allocation procedure, this share decreases to 19%, meaning that the geographical residence is assigned to more than 8 million *person_id* in table *tls206_person*. Furthermore, the number of applicants or inventors increases substantially in many countries: the improvement is between 20% and 40%, as shown in Fig. 2.

Table *tls206_person* provides two additional sets of harmonised information [24]. The first one is the result of a method developed by K.U.Leuven and Eurostat which harmonises patentees' names and assigns a sector classification to them [25,26]. This method generates another identification number, *psn_id*, which is added to PATSTAT, and concerns about 98% of the total *person_id* in table *tls206_person*. Therefore, also this additional identifier groups several *person_id* under the same entity. However, as for the case of *doc_std_name_id*, these additional sets of harmonised information present the same type of inconsistencies, as shown in Table 2. Consequently, the allocation procedure presented in Fig. 1 can be replicated by using this additional

---

[1] To be noted that "It is not 100% certain that the DOCDB standardised names are always linked with the correct person name, in particular if the person information came from a source other than DOCDB" [18, pp. 144].



```
SELECT p1.doc_std_name_id, p1.doc_std_name, p1.person_id, C.person_ctry_code
FROM
(
  SELECT T.doc_std_name_id, T.person_ctry_code
  FROM
    (
    SELECT
      p.doc_std_name_id,
      p.person_ctry_code,
      COUNT (*) AS Tot,
      RANK () OVER ( PARTITION BY p.doc_std_name_id ORDER BY COUNT (*) DESC, p.person_ctry_code ASC) AS rnk
    FROM tls206_person p
    WHERE p.person_ctry_code NOT LIKE ''
    GROUP BY p.doc_std_name_id, p.person_ctry_code
    ) AS T
  WHERE T.rnk = 1
) AS C
RIGHT JOIN tls206_person p1 ON C.doc_std_name_id = p1.doc_std_name_id
ORDER BY p1.doc_std_name_id
```

Fig. 1. Query to assign automatically country to *person_id*.

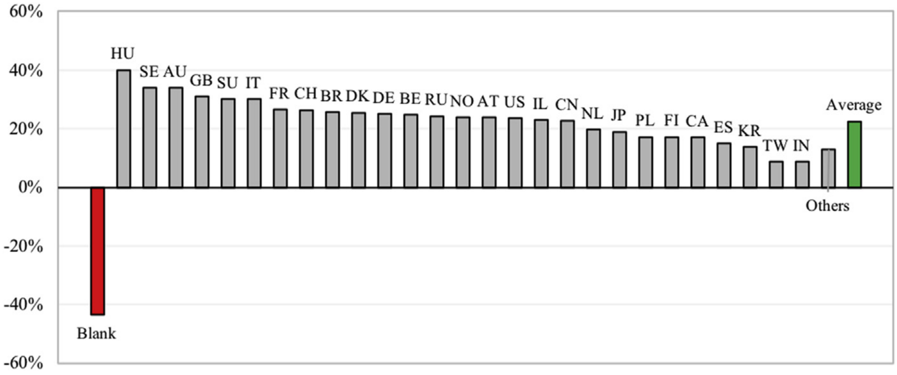


Fig. 2. Increase of data accuracy and completeness.

Table 2
person_ctry_code assigned to Nokia Corporation: *psn_id* = 20816957.

person_ctry_code	Country	Count of *person_id*
FI	Finland	315
(blank)	unknown	86
US	United States of America	61
IE	Ireland	6
DE	Germany	4
FR	France	4
CA	Canada	1
CN	China	1
FL	NULL	1
GB	United Kingdom	1
ID	Indonesia	1
IR	Iran	1
JP	Japan	1
KI	Kiribati	1
LI	Liechtenstein	1
NL	Netherlands	1
SG	Singapore	1
		487

identifier as main standardised reference, hence by replacing *doc_std_name_id* with *psn_id*.

The second information set is developed by the OECD Task Force on Patent Statistics [27,28]. Similarly to the previous one, consolidated and harmonised names are provided under an additional identifier, *han_id*, that groups together several *person_id*. This method is designed to eliminates inconsistencies in table *tls206_person*. In fact, Nokia Corporation is assigned correctly to Finland to all *person_id* grouped by the identifier *han_id* = 2196902. It is, therefore, a standardisation procedure that results in a more accurate set of information. However, this method is applied only to patent applicants which are resident in 40 countries [18, p. 177]. This implies that all patent inventors and many other countries are not affected by this procedure: in fact, less than 10% of the total *person_id* in table *tls206_person* are harmonised. Raw information are copied to the remaining entries in the table and, in order to avoid inconsistent information, a unique *han_id* is generated as a surrogate key for all combinations of the name and country. Consequently, the allocation procedure presented in Fig. 1 applied to this identifier does not produce any improvement since, by construction, there are not inconsistencies.

The table *tls206_person* in PATSTAT provides additional information relative to each *person_id*, such as *psn_sector* and *nuts*. However, there are inconsistent allocations also for these fields: there is more than one sector (e.g. company, university, individual, etc.) assigned to the same entity as well as there is more than one regional code assigned to same European applicant or inventor. Table 3 shows the lack of data accuracy for the case of Nokia Corporation. As discussed above, this specific case illustrates that the field *psn_sector* is allocated correctly to the identifier *psn_id* and to *han_id*. Instead, all identifiers presents wrong allocations regarding the regional code.

Considering that these additional inconsistencies occur in table *tls206_person*, the allocation procedure in Fig. 1 can be adapted and implemented in order to increment data accuracy and completeness in these fields. Table 4 summarises the reduction of *blank* entries in PATSTAT concerning three important information relative to patentees: country and region of residence and their sector classification. The allocation procedure is applied to these fields based on the three available standardised identifiers: *doc_std_name_id*, *psn_id* and *han_id*. The allocation procedure based on the first identifier produces, in all fields, a higher reduction of empty entries compared to the second, while, the third one, *han_id* does not increase data completeness since it is explicitly designed to avoid inconsistencies.

Table 3
Nokia Corporation: total *person_id* assigned to *psn_sector* and *nuts*.

		doc_std_name_id = 1	*psn_id* = 20816957	*han_id* = 2196902
psn_sector	Company	166	487	111
	Individual	6	–	–
	University	1	–	–
	unknown	1	–	–
		174	487	111
nuts	FI1B1	18	32	17
	FI197	5	6	4
	FI1C1	1	1	1
	DEA11	–	1	–
	IE021	–	1	–
	(blank)	150	446	89
		174	487	111

Table 4
Reduction of *blank* or *unknown* entries in PATSTAT.

	doc_std_name_id	*psn_id*	*han_id*
person_ctry_code	44%	33%	0%
psn_sector	43%	18%	0%
nuts	13%	8%	1%

In conclusion, given the fact that this allocation procedure guarantees a better level of data accuracy and completeness, it is interesting to answer the follow question: how does this new approach change patent statistics? Next section addresses this question and illustrates its relevance, since the use of raw data could lead to misleading outcomes.

3. A more accurate patent analysis

In order to check to what extent the proposed allocation procedure modifies patent statistics, this section compares the outcome of a patent analysis conducted by using two different set of data. The first one uses raw data as they are available in PATSTAT, while the second analyses data for which the country code is assigned to patent applicants as presented in the previous section, using the identifier *doc_std_name_id* as the standardised reference. The analysis is limited to climate change mitigation technologies (CCMT), which are identified by means of CPC codes (Cooperative Patent Classification) under the Y02 classification scheme [29,30]. The objective is to measure, in the period 2000–2015, how many times a country has participated in patent applications concerning CCMT, regardless the number of distinct applicants. In other words, this analysis wants to assess the level of participation of a country, hence its contribution to the technological progress concerning climate change mitigation technologies (query in Fig. 3). The focus of the analysis is on the four major countries contributing to the development of CCMT: China (CN), Germany (DE), Japan (JP) and United States (US).

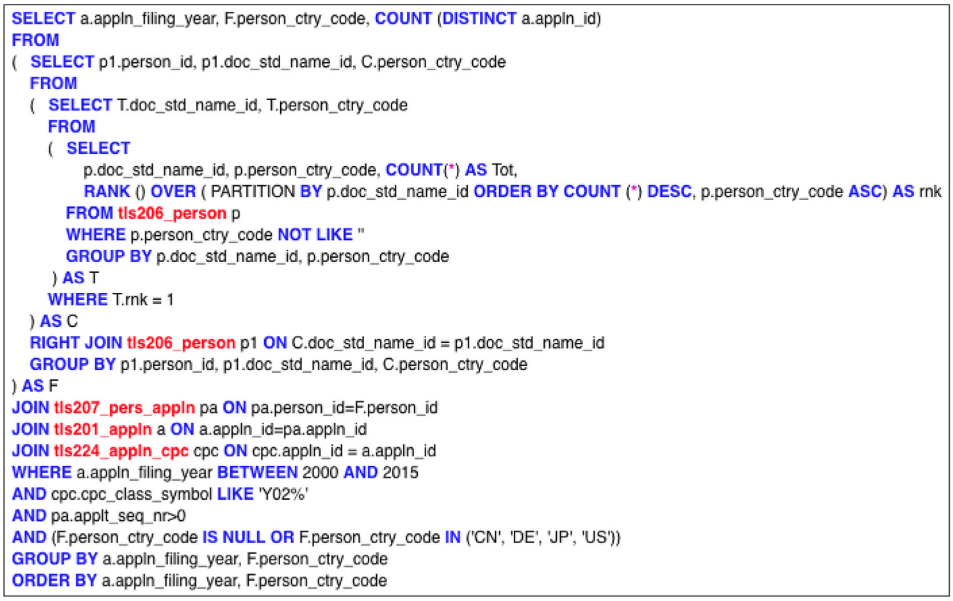
Fig. 4 compares the two analyses: on the left, with dotted lines, it shows the outcome using raw data in PATSTAT, while on the right, the solid lines show the outcome using the country allocation procedure proposed in this paper. The axis on the right in both charts indicates the number of participations for which the country code is not assigned (grey lines). It is clear that the proposed procedure substantially reduces the number of unknown participations by around 50% compared to the analysis using raw data. It is also clear that there are important differences in the result generated by the two analyses. When raw data are used, US is the country that, over the years, contributes more in this sector. JP and DE have a lower level of participation, showing a very similar trend over time, while CN catches up these two countries in 2006, but falls behind rapidly right after 2008. The outcome of the analysis conducted via the procedure presented in this paper (on the right in Fig. 4) changes considerably. JP is the leading country in CCMT since early 2000's followed by the US that in 2009 reaches the same level of contribution. In recent years, CN overcomes the two leading countries, as the result of an extraordinary growth trend. It is therefore evident that raw data generate an outcome which underestimates the role of Japan in CCMT and almost completely ignores the recent contribution that China is giving to this sector.

Similar to the case of country contribution, it is possible to assess the regional participation to CCMT and to see how the outcome changes when data analysed have higher accuracy and completeness. Fig. 5 shows the top 20 European regions, labeled by means of their *nuts* codes (level 3). Overall, the outcome of the analysis based on raw data in PATSTAT has a very low coverage in terms of *nuts* allocation to applicant. In the period 2000–2015, the proposed allocation procedure increases by about half million the number of participations for which the regional code is assigned (see the size of black bars of the top 20 *nuts* code compared to the grey bars). This means that the set of data used to analyse regional contribution to CCMT has higher level of completeness. Consequently, it produces more accurate representation of regional contribution to CCMT-related patent activity. By comparing the position of the top 20 regions with their ranking based on raw data (number in brackets) it is clear that the performance of many regions is greatly affected. For example, the Danish region *DK041*, which is the residence area of important multinational corporations active in the sector of climate change mitigation technologies, results being in the 94th position if raw data are analysed, neglecting its specialisation that, instead, emerges from the analysis of more complete and accurate patent dataset.

Usually patentees file their first patent application domestically and, depending on their market strategy, they also seek protection for their invention internationally, at a later stage [31]. One way, yet imperfect, to test the validity of the proposed allocation procedure is to check whether raw data confirm this tendency or if a cleaned-up dataset provides a more realistic outcome. As emerged from the analysis in Fig. 4, the data clean-up process mostly affects the outcome for Japan and China. Therefore, it could be expected an improvement in relation to the share of first domestic patent applications filed to the respective national authority, namely, the State Intellectual Property Office of the People's Republic of China (SIPO) and the Japan Patent Office (JPO). For example, about 85% of the priority patent applications concerning CCMT are filed to the European Patent Office (EPO) by domestic applicants.[2] This share is about 80% for priority patent applications filed

---

[2] Patent filed to the European Patent Offices (EPO) protects the invention in 38 different European countries by means of one patent application only. The following countries are EPO members: Albania (AL), Austria (AT), Belgium



```
SELECT a.appln_filing_year, F.person_ctry_code, COUNT (DISTINCT a.appln_id)
FROM
(  SELECT p1.person_id, p1.doc_std_name_id, C.person_ctry_code
   FROM
   (  SELECT T.doc_std_name_id, T.person_ctry_code
      FROM
      (  SELECT
            p.doc_std_name_id, p.person_ctry_code, COUNT(*) AS Tot,
            RANK () OVER ( PARTITION BY p.doc_std_name_id ORDER BY COUNT (*) DESC, p.person_ctry_code ASC) AS rnk
         FROM tls206_person p
         WHERE p.person_ctry_code NOT LIKE ''
         GROUP BY p.doc_std_name_id, p.person_ctry_code
      ) AS T
      WHERE T.rnk = 1
   ) AS C
   RIGHT JOIN tls206_person p1 ON C.doc_std_name_id = p1.doc_std_name_id
   GROUP BY p1.person_id, p1.doc_std_name_id, C.person_ctry_code
) AS F
JOIN tls207_pers_appln pa ON pa.person_id=F.person_id
JOIN tls201_appln a ON a.appln_id=pa.appln_id
JOIN tls224_appln_cpc cpc ON cpc.appln_id = a.appln_id
WHERE a.appln_filing_year BETWEEN 2000 AND 2015
AND cpc.cpc_class_symbol LIKE 'Y02%'
AND pa.applt_seq_nr>0
AND (F.person_ctry_code IS NULL OR F.person_ctry_code IN ('CN', 'DE', 'JP', 'US'))
GROUP BY a.appln_filing_year, F.person_ctry_code
ORDER BY a.appln_filing_year, F.person_ctry_code
```

Fig. 3. Query to count country participation in CCMT.

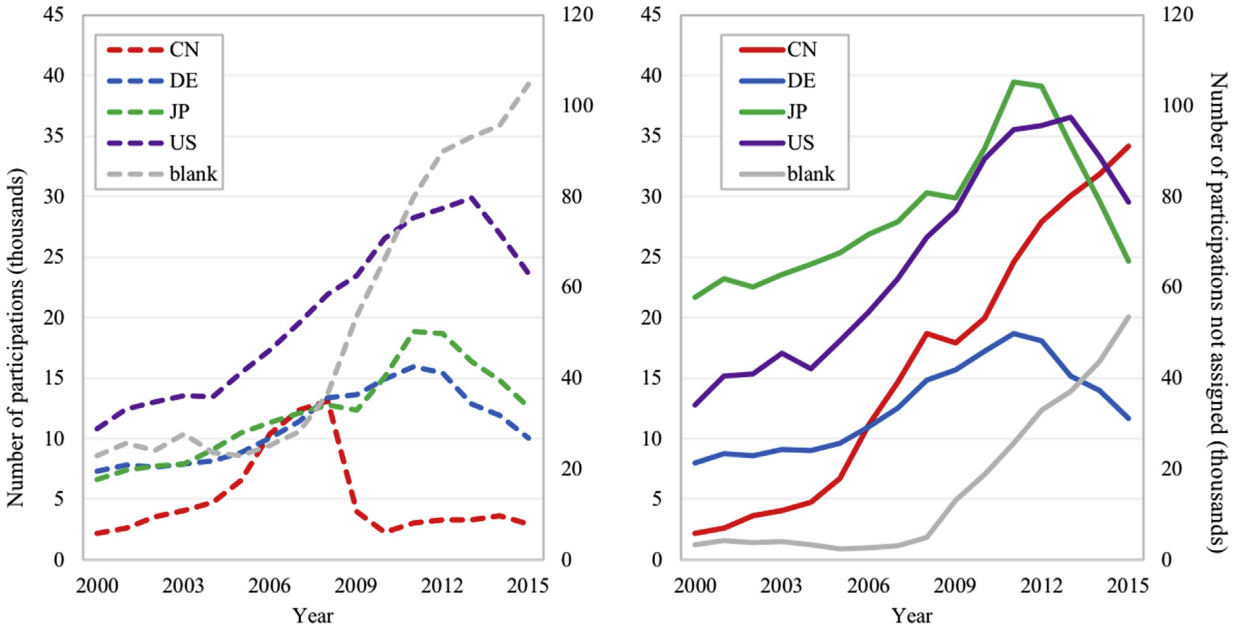


Fig. 4. Patent analysis with raw data (left) and after country allocation (right).

to the United States Patent and Trademark Office (USPTO) by domestic applicants. These values remain nearly unvaried whether raw data are used or the allocation procedure is implemented to improve country code imputation to patentees. This is due to the fact that data provision from these patent offices is very accurate.

On the contrary the proposed imputation practice substantially modifies the analysis for the two Asian countries. The use of raw data shows that, on a yearly average, about 52% of priority applications related to CCMT are filed to SIPO by domestic applicants, while for the remaining 48% the country of the patentee is unknown. These values become, respectively, 72% and 24% when country code is better allocated, being now in line with the global tendency, also seen at EPO and USPTO. The improvement is even larger when this analysis is run for the JPO. Because of the incomplete provision of information [18, p. 280], raw data shows that applicants' country is always unknown for patent filed to the Japanese patent authority. Instead, the use of more accurate data shows that about 90% of priority applications are filed to JPO by domestic applicants. Therefore, the proposed allocation procedure permits to increase results reliability.

Another way to assess the validity of the proposed allocation procedure is to compare its patent statistics with those provided by commercial sources, whose outcomes are reliable. The hypothesis here is

---

(*footnote continued*)
(BE), Bulgaria (BG), Switzerland (CH), Cyprus (CY), Czech Republic (CZ), Germany (DE), Denmark (DK), Estonia (EE), Spain (ES), Finland (FI), France (FR), United Kingdom (UK), Greece (EL), Croatia (HR), Hungary (HU), Ireland (IE), Iceland (IS), Italy (IT), Liechtenstein (LI), Lithuania (LT), Luxembourg (LU), Latvia (LV), Monaco (MC), Former Yugoslav Republic of Macedonia (MK), Malta (MT), Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT), Romania (RO), Serbia (RS), Sweden (SE), Slovenia (SI), Slovakia (SK), San Marino (SM), Turkey (TR).

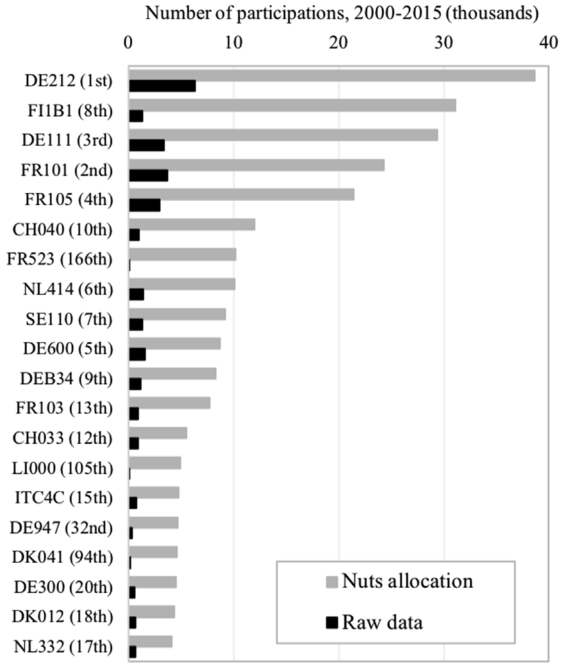


Fig. 5. Top 20 *nuts*. Ranking based on raw data in brackets.

that the use of raw data in PATSTAT might bring to results that are not directly comparable to those from other sources, and that a higher accuracy and completeness could improve comparability. In order to test this proposition, the country allocation procedure presented in section 2 is applied before performing two different patent analysis. These replicate the analyses conducted by two commercial sources, whose nome are not displayed in order to maintain anonymity. For the same reason, also the focus of the analysis is not made explicit and results for four countries are shown as shares of the total (Fig. 6). In both cases, the use of raw data underestimates the patent share in China and Japan, while the proposed allocation procedure makes outcomes more comparable.

4. Conclusion

This paper has proposed an allocation procedure which increases data accuracy and completeness in PATSTAT. It consists of a simple query that can be run directly in PATSTAT, and it is easily replicable. Its objective is not to substitute more complex and detailed methods or techniques already developed by the scientific community to harmonise information and to reduce missing entries. Instead, it is considered as a preliminary step to be applied to those approaches. For example, de Rassenfosse and colleagues have developed an algorithm to recover the missing information by considering priority filings, applicant's and inventor's country of residence, patent families and priority offices [21, Appendix B]. This is a very accurate and iterative data-recovery process which only starts if country information is missing for patentees. By preliminarily reducing missing information through the allocation procedure proposed in this paper, the method in Ref. [21] would reduce substantially its computational operations and could only focus on cases where information is not available and not recoverable differently.

The proposed allocation procedure reduces by 44% the empty entries regarding the residence country of inventors and applicants, and it increases by 22%, on average, accuracy of country allocation. It also reduces by 43% and 13% the empty entries regarding, respectively, the sector of patentees and their regional residence. The importance to have more accurate and complete set of patent data has been demonstrated by means of an example. The proposed allocation procedure has been adopted in the patent analysis concerning countries' contribution to the development of climate change and mitigation technologies (CCMT). It is shown that a more refined set of data avoids misleading conclusion. For example, when patent data are used as they are available in PATSTAT, the important role that Asian countries, particularly Japan and China, are playing in patenting CCMT-related inventions is underestimated.

Patent data are crucial information to carry out empirical research in the context of science and technology, which may also support policymakers' strategic decisions [32]. Studying patenting activity provides an early signal of technological progress since patents are the first publicly available information on new products or processes. And, given the fact that technological progress is often the result of successful R&D activity, patents also offer an indication of its quality and effectiveness. Consequently, it is very important to have datasets on patents that are as more accurate and complete as possible, thus reducing biases in outcome of the analyses. This paper has highlighted the need and the importance of data clean-up process in PATSTAT, which is one of the most prominent patent data source. However, patent data are made available by other commercial sources, for which the lack of data accuracy and completeness remains an issue. Therefore, it may be beneficial for commercial tools to provide more instruments for data clean-up.

Disclaimer

The opinions expressed in this paper are the author's alone and cannot be attributed to the European Commission.

Acknowledgement

The paper has benefited from discussion with participants at the 10th EMAEE conference (Strasbourg, 2017), EC-EPO expert meeting (Brussels, 2016), and from valuable comments from JRC colleagues. I am grateful to the editor and to two anonymous reviewers whose comments have improved the final version of this paper. The author declares no conflict of interest. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-

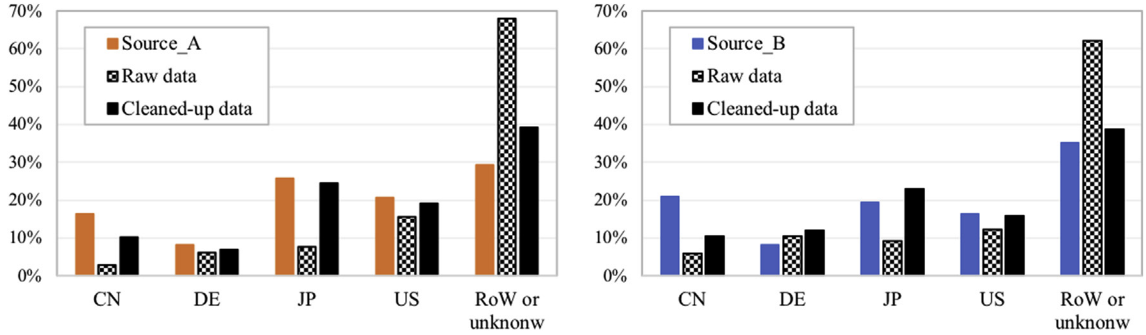


Fig. 6. Patent country share: comparison between PATSTAT data (raw and cleaned-up) and two different datasets. RoW = Rest of the World.

profit sectors.

References

[1] Z. Griliches, Market value, R&D, and patents, Econ. Lett. 7 (2) (1981) 183–187.
[2] Z. Griliches, R&D, Patents, and Productivity, The University of Chicago Press, Chicago/London, 1984.
[3] B.L. Basberg, Patents and the measurement of technological change: a survey of the literature, Res. Pol. 16 (2–4) (1987) 131–141.
[4] OECD, "OECD Patent Statistics Manual," Tech. Rep, OECD, Paris, 2009.
[5] Z. Griliches, Patent statistics indicators as economic indicators: a survey, J. Econ. Lit. 28 (4) (1990) 1661–1707.
[6] H. Ernst, Patent information for strategic technology management, World Patent Inf. 25 (3) (2003) 233–242.
[7] M. Squicciarini, H. Dernis, C. Criscuolo, Measuring Patent Quality: Indicators of Technological Economic Value, Tech. Rep. OECD, Paris, 2013.
[8] U. Schmoch, Tracing the knowledge transfer from science to technology as reflected in patent indicators, Scientometrics 26 (1) (1993) 193–211.
[9] A.B. Jaffe, M. Trajtenberg, M.S. Fogarty, Knowledge spillovers and patent citations: evidence from a survey of inventors, Am. Econ. Rev. 90 (2) (2000) 215–218.
[10] D. Guellec, B. Van Pottelsberghe De La Potterie, The internationalisation of technology analysed with patent data, Res. Pol. 30 (8) (2001) 1253–1266.
[11] M. Balconi, S. Breschi, F. Lissoni, Networks of inventors and the role of academia: an exploration of Italian patent data, Res. Pol. 33 (1) (2004) 127–145.
[12] M.A. Schilling, C.C. Phelps, Interfirm collaboration networks: the impact of large-scale network structure on firm innovation, Manag. Sci. 53 (7) (2007).
[13] B.H. Hall, A. Jaffe, M. Trajtenberg, Market value and patent citations, Rand J. Econ. 36 (1) (2005) 16–38.
[14] B.-w. Lin, J.-s. Chen, Corporate technology portfolios and R&D performance measures: a study of technology intensive firms, R&D Management 35 (2) (2005) 157–170.
[15] R. Fontana, A. Nuvolari, B. Verspagen, Mapping technological trajectories as patent citation networks. An application to data communication standards, Econ. Innovat. N. Technol. 18 (4) (2009) 311–336.
[16] K. Pavitt, Patent statistics as indicators of innovative activities: possibilities and problems, Scientometrics 7 (1–2) (1985) 77–99.
[17] D. Archibugi, M. Pianta, Measuring technological change through patents and innovation surveys, Technovation 16 (9) (1996) 451–468.
[18] EPO, Data Catalog 2018 Spring Edition - Version 5.11, Tech. Rep. European Patent Office, 2018.
[19] G. de Rassenfosse, H. Dernis, G. Boedt, An Introduction to the patstat database with example queries, Aust. Econ. Rev. 47 (3) (2014) 395–408.
[20] M. Trajtenberg, G. Shiff, R. Melamed, "The "names game": harnessing inventors, patent data for economic research, Annals of Economics and Statistics/Annales d'Économie et de Statistique 93 (94) (2009) 79–108.
[21] G. de Rassenfosse, H. Dernis, D. Guellec, L. Picci, B. Van Pottelsberghe De La Potterie, The worldwide count of priority patents: a new indicator of inventive activity, Res. Pol. 42 (3) (2013) 720–737.
[22] F. Lissoni, Academic patenting in europe: a reassessment of evidence and research practices, Ind. Innov. 20 (5) (2013) 379–384.
[23] F. Alkemade, G. Heimeriks, A. Schoen, L. Villard, P. Laurens, Tracking the internationalization of multinational corporate inventive activity: national and sectoral characteristics, Res. Pol. 44 (9) (2015) 1763–1772.
[24] B. Kang, G. Tarasconi, PATSTAT revisited: suggestions for better usage, World Patent Inf. 46 (2016) 56–63.
[25] M. du Plessis, B. Van Looy, X. Song, T. Magermen, Data Production Methods for Harmonized Patent Statistics: Patentee Sector Allocation, Tech. Rep. EUROSTAT, 2009.
[26] T. Magerman, B.V. Looy, X. Song, Data Production Methods for Harmonised Patent Statistics: Patentee Name Harmonisation, Tech. Rep. European Commission, 2006.
[27] G. Thoma, S. Torrisi, Creating powerful indicators for innovation studies with approximate matching algorithms. A test based on PATSTAT and amadeus databases, Conference on Patent Statistics for Policy Decision Making, No. September, (2-3 October 2007, Venice), 2007.
[28] G. Thoma, S. Torrisi, A. Gambardella, D. Guellec, B. Hall, D. Harhoff, Harmonizing and Combining Large Datasets - an Application to Firm-Level Patent and Accounting Data, NBER Working Paper Series, 2010.
[29] I. Rudyk, G. Owens, A. Yolpe, R. Ondhowe, A. Dechezleprêtre, Climate Change Mitigation Technologies in Europe - Evidence from Patent and Economic Data, Tech. Rep. The United Nations Environment Programme (UNEP) and the European Patent Office (EPO), 2015.
[30] V. Veefkind, J. Hurtado-Albir, S. Angelucci, K. Karachalios, N. Thumm, A new EPO classification scheme for climate change mitigation technologies, World Patent Inf. 34 (2) (2012) 106–111.
[31] H. Dernis, M. Khan, "Triadic Patent Families Methodology," OECD Science, Technology and Industry Working Papers 2004/02 OECD Publishing, Paris, 2004, pp. 1–33.
[32] A. Fiorini, A. Georgakaki, J. Jimenez Navarro, A. Marmier, F. Pasimeni, E. Tzimas, Energy R&I Financing and Patenting Trends in the EU: Country Dashboards 2017 Edition, JRC Science for Policy Report, 2017 EUR 29003 EN, JRC109654.

Francesco Pasimeni is an industrial engineer specialised in science, technology and innovation policy with particular expertise on clean energy technologies. In 2015 he joined the Joint Research Centre (JRC) of the European Commission. As policy analyst in the Directorate of Energy, Transport and Climate, he contributes to Commission's Energy Union Strategy by monitoring research, innovation and competitiveness. He is currently PhD candidate at SPRU (University of Sussex, UK) and studies impact and policy implication of fractional ownership and shared consumption using agent-based models. He holds BSc and MSc in industrial engineering from University of Salento (Italy) and MSc in science and technology from SPRU, University of Sussex (UK).