INSTITUTE OF DEVELOPING ECONOMIES

IDE Discussion Papers are preliminary materials circulated
to stimulate discussions and critical comments

IDE DISCUSSION PAPER No. 527

PATSTAT revisited

Gianluca Tarasconi* and Byeongwoo Kang**

April 2015

Abstract

This study provides a comprehensive summary of and guidance for using the EPO Worldwide Patent Statistical Database (PATSTAT), one of the most widely used patent databases for researchers. We highlight the three most important issues that PATSTAT users must consider when performing patent data analyses and suggest ways to deal with those issues. Although PATSTAT is chosen in this study, the issues that we discuss are also applicable to other patent databases.

Keywords: PATSTAT, patent data analysis, innovation studies
JEL classification: O39, Y20, Z00

* Database Architect, Bocconi University, Italy, (gianluca.tarasconi@unibocconi.it)
** Researcher, Inter-disciplinary Studies Center, IDE (Byeongwoo_Kang@ide.co.jp)

The Institute of Developing Economies (IDE) is a semigovernmental, nonpartisan, nonprofit research institute, founded in 1958. The Institute merged with the Japan External Trade Organization (JETRO) on July 1, 1998. The Institute conducts basic and comprehensive studies on economic and related affairs in all developing countries and regions, including Asia, the Middle East, Africa, Latin America, Oceania, and Eastern Europe.

INSTITUTE OF DEVELOPING ECONOMIES (IDE), JETRO
3-2-2, WAKABA, MIHAMA-KU, CHIBA-SHI
CHIBA 261-8545, JAPAN

1. Introduction

While there are many approaches in economic and social studies, these can be roughly divided into two main types, namely the theoretical approach and the empirical approach. Neither of them can stand alone. Improved sophisticated models using theoretical analysis help better explain empirical analyses, and new findings arising from empirical analysis help set new theoretical models. Findings and discussions coming from theoretical and empirical analyses are used for various purposes including policy formation. Thus, both approaches must be used appropriately.

Lately, scholars have widely used patent data for empirical economic and social science research. One advantage of using patent data is that it provides useful information that enables us to understand the technological innovation process. An example of the first page of a patent application is illustrated in Figure 1, depicting the type of information to be found. From the example, we can identify (1) the patent office from which the patent was applied (the US in this example), (2) its title, (3) inventor names and addresses, (3) assignee name and address, (4) application number, (5) publication number, (6) publication date, (7) other related patent applications, (8) foreign application priority data, (9) patent classification, (10) abstract, and (11) best mode figure. As the patent is often regarded as an output of R&D, analysis of the information acquired from patent documents reveal how R&D is conducted and how technological innovation is derived from inventions. Direct use of such raw information is one way to use patent data for economic and social studies. Another way is to use statistical information retrieved from a large quantity of patent data called patent statistics. Dozens of patent statistics have been proposed by scholars for effective analysis of patent data (Jaffe & Trajtenberg, 2002; Nagaoka, Motohashi, & Goto, 2010; Lerner & Seru, 2015). Patent statistics are used in various fields such as science and technology, social sciences, and economics. In addition, empirical studies employing patent statistics have significantly increased in recent years. Table 1 presents the examples of frequently used patent statistics.

US 20120082102A1

(19) **United States**

(12) **Patent Application Publication**

Kang et al.

(10) **Pub. No.: US 2012/0082102 A1**

(43) **Pub. Date: Apr. 5, 2012**

(54) **METHOD FOR INDICATING PRECODING MATRIX INDICATOR IN UPLINK MIMO SYSTEM WITH BASED ON SC-FDMA**

(75) Inventors: **Byeong Woo Kang**, Anyang-si (KR); **Joon Kui Ahn**, Anyang-si (KR); **Dong Youn Seo**, Anyang-si (KR); **Jung Hoon Lee**, Anyang-si (KR); **Yu Jin Noh**, Anyang-si (KR); **Byoung Hoon Kim**, Anyang-si (KR); **Suck Chel Yang**, Anyang-si (KR); **Bong Hoe Kim**, Anyang-si (KR); **Dae Won Lee**, Anyang-si (KR)

(73) Assignee: **LG ELECTRONICS INC.**, Seoul (KR)

(21) Appl. No.: **13/148,886**

(22) PCT Filed: **Feb. 19, 2010**

(86) PCT No.: **PCT/KR2010/001039**

§ 371 (c)(1), (2), (4) Date: **Nov. 14, 2011**

Related U.S. Application Data

(60) Provisional application No. 61/161,049, filed on Mar. 17, 2009, provisional application No. 61/157,206, filed on Mar. 4, 2009, provisional application No. 61/153,974, filed on Feb. 20, 2009.

(30) **Foreign Application Priority Data**

Jul. 7, 2009 (KR) 10-2009-0061699

Publication Classification

(51) **Int. Cl.**
H04W 74/04 (2009.01)
H04W 72/04 (2009.01)

(52) **U.S. Cl.** **370/329**

(57) **ABSTRACT**

A method of transmitting PMI (precoding matrix indicator) information in an uplink MIMO system is disclosed. The present invention includes the steps of receiving channel information from a user equipment and transmitting information on a resource allocated to the user equipment in uplink transmission and PMI information indicating a precoding matrix to apply to a region of the resource among a plurality of precoding matrices to the user equipment based on the received channel information, wherein the resource allocated to the user equipment is allocated by a bundle unit of a prescribed number of subcarriers, wherein each of a plurality of the precoding matrices are applied to regions generated from dividing a whole frequency band into a prescribed number of regions, respectively, and wherein the precoding matrix applied to the resource among a plurality of the precoding matrices has a maximum area resulting from overlapping a frequency band occupied by the allocated resource with a frequency band having the precoding matrix applied thereto.

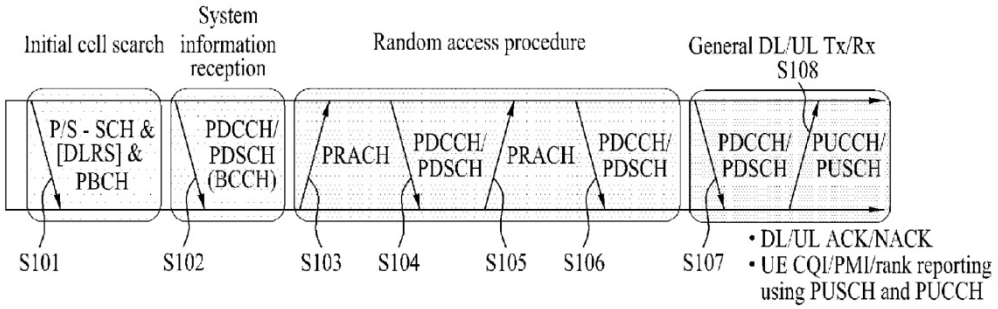

Figure 1. Example of patent applications (first page)

Table 1. Examples of the patent statistics

Analysis	Patent statistics	Purposes
Citation analysis	Forward citations	Measure technological value
	Backward citations	Find knowledge source
Patent counts analysis	Patent counts	Observe patent portfolio
	RTA (Revealed Technology Advance) PS (Patent Share)	Identify core technological competence
Technology class analysis	Generality	Measure endogenous applicability to different technological fields
	Originality	Measure knowledge absorption from different technological fields
Inventor analysis	Inventor counts	Measure invention quality Measure absorptive capability
	Inventor	Identify specific inventors' info such as star engineers Follow mobility of R&D personnel

This study provides a comprehensive summary of and guidance for using the EPO Worldwide Patent Statistical Database (PATSTAT). As discussed, patent information has many uses, and therefore, guidance on using one of the most widely used patent databases will be useful to researchers in many fields including economics and social sciences for several reasons. First, although scholars are performing increasing numbers of patent data analyses, some researchers still face difficulties in performing patent data analyses, while others unintentionally perform patent data analyses inappropriately. Most knowledge regarding patent data analysis has been obtained through learning by doing or personal consultation with skilled users. Such methods are very time intensive. Providing a comprehensive summary and guidance will help young researchers and new users save time and effort in accustoming themselves in performing patent data analyses. Second, PATSTAT has become one of the most widely used patent databases for scholars. Patent data is increasingly being used for various purposes. As a response to increased demand, basics regarding a patent database with example SQL queries have been provided by de Rassenfosse, Dernis, & Boedt (2014). We supplement this information by addressing some issues that users must consider to perform patent data analyses correctly. This study identifies issues arising in patent data analyses and providing solutions to these issues. These issues are

also applicable to other patent databases.

The rest of this study is organized as follows. Section 2 introduces PATSTAT, a raw patent database that is widely used and provides great degrees of research freedom. Section 3 discusses general issues that potential users must consider and ways to deal with them. Section 4 concludes the study.

2. PATSTAT

This section reviews PATSTAT and provides merits and demerits of its use.

2.1. Basic information on PATSTAT[1]

The Organisation for Economic Co-operation and Development (OECD) is leading the Patent Statistics Task Force, members of which are the World Intellectual Property Organisation (WIPO), the European Patent Office (EPO), the Japanese Patent Office (JPO), the Korean Intellectual Property Office (KIPO), the US Patent and Trademark Office (USPTO), the US National Science Foundation (NSF), and the European Commission (EC).Upon request by the Task Force, the EPO created PATSTAT.

PATSTAT comprises four components (As the core of PATSTAT is the PATSTAT raw data, users refer to this element when they say “PATSTAT.” This study also uses the term PATSTAT to refer to the PATSTAT raw data elsewhere in this study, excluding this paragraph). The first component is PATSTAT raw (patent) data, which are extracted bibliographic information from patent documents. Much of the raw data is extracted from the EPO’s master bibliographic database, called DOCDB. The second component is legal event data for PATSTAT. This contains information on legal events that occurred during the life of a patent, either before or after it being granted, such as requests for examination, payment of renewal fees, lapse of the patent, change of ownership, withdrawal of the application, patents entering the national phase, patents that have been opposed or revoked, and so on. The third one is the PATSTAT online extension. This database contains additional tables and attributes that are either derived from PATSTAT raw data or additional data taken from freely available sources. The last one is the European Patent Register for PATSTAT. This database contains bibliographic,

[1] Knowledge explained in this subsection is retrieved from the PATSTAT catalogue. If readers want to know more details regarding PATSTAT, we recommend accessing the PATSTAT catalogue from its webpage:
http://www.epo.org/searching/subscription/raw/product-14-24.html
http://www.epo.org/searching/subscription/raw/product-14-24-1.html
http://www.epo.org/searching/subscription/raw/product-14-24-2.html

legal, and procedural information on published European patent applications and on published patent applications according to the PCT for which the EPO is a designated office.

As the diagram of PATSTAT illustrates (Figure 2), it contains information regarding applications, publications, applicants, inventors, citations, patent families, (technological) categories, priorities, and so on. First, application information includes the patent authorities from whom the patent of interest was applied, patent numbers, patent type (patent, utility model, design patent, etc), dates of patent applications, titles, and abstracts. Second, publication information includes similar information as contained in the application information: patent authorities from whom the patent application of interest was published, publication numbers, publication types attributed by the patent authority issuing the publications (such as publication of patent, reexamination, reissue, etc[2]), publication dates, publication languages, and the number of claims in the given publication. Third, information regarding applicants and inventors contains their names, country codes, and address. Fourth, information regarding citations gives details on patent citations, non-patent citations (such as journal and conference papers, books, symposiums and workshop reports, technical reviews from magazines, and websites), provenance of citations, and country codes identifying the patent authority performing the international search reports. Fifth, patent families can be searched via two approaches; DOCDB and INPADOC patent families.

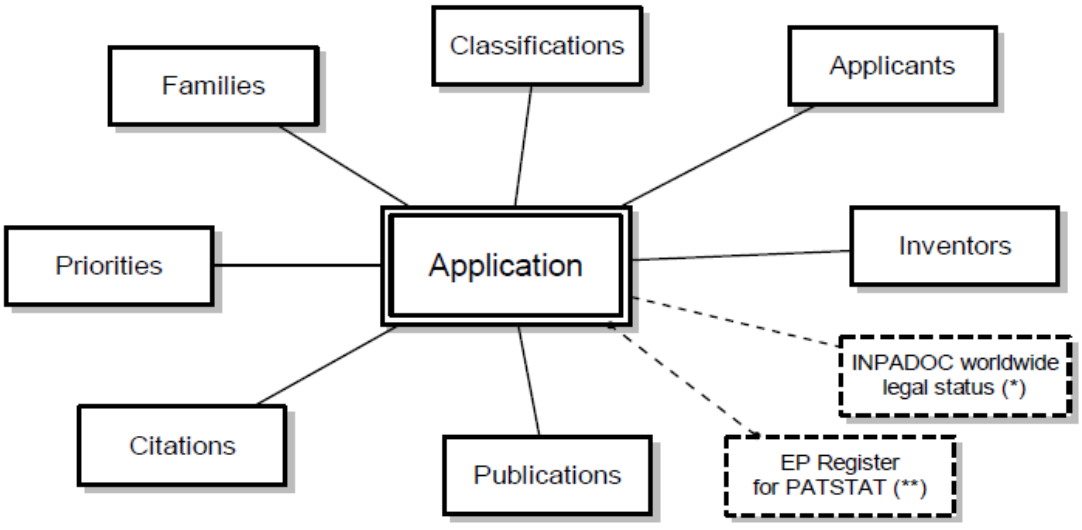

Figure 2. EPO PATSTAT Diagram

[2] Each authority has different publication types. For more information, please find the "Kind code concordance list," which is available at http://www.epo.org/searching/data/data/tables/regular.html

(Source: PATSTAT Catalogue – Edition Spring 2014)

2.2. Merits and demerits of PATSTAT

PATSTAT has many merits. First, by providing raw patent data, PATSTAT provides great degree of freedom for researchers. For example, national analyses are possible by using the county origins of patent applicants and inventors. Industrial analyses can be performed using technological classifications, searching key words in titles and abstracts, and/or looking at companies in the sector of interest. Specific companies of interest can be analyzed by searching for those firms. In addition, individual analyses can be performed by searching for inventors of interest. Setting time windows enables one to track historical changes in those analyses.

Second, PATSTAT offers real data. The real data here refers to the data that are not sampled. One limitation in using sampled data is that the amount of samples decreases when one tries to analyze subsets. As the sample size correlates with confidence interval and error range, using subsets of sampled data is not preferred.

Third, since PATSTAT is becoming a *de facto* standard among patent databases, databases based on and linkable to PATSTAT have been produced by other institutions. Sometimes, common IDs are used in PATSTAT and these other databases, and hence, users can easily link to them. Databases linkable to PATSTAT include ECOOM–EUROSTAT–EPO PATSTAT Person Augmented Table (EEE-PPAT) Database (Du Plessis et al., 2009; Magerman et al., 2009; Peeters et al., 2009), OECD Harmonised Applicant Names (HAN) Database (http://www.oecd.org/sti/inno/oecdpatentdatabases.htm), APE-INV (Lissoni et al., 2009), EPO worldwide legal status database (http://www.epo.org/searching/subscription/raw/product-14-11.html), OECD REGPAT Database (Maraut et al., 2008), OECD Triadic Patent Families Database (Dernis & Khan, 2004), and NACE-IPC concordance table (Schmoch et al., 2003). These databases enable users to extend their analyses beyond patent data analyses. They also help overcome some of PATSTAT's limitations, which will be discussed in the next section.

Unfortunately, PATSTAT also has drawbacks, despite its many advantages. First, the data were originally collected for use by patent examiners. The main data source of PATSTAT is DOCDB bibliographic data. DOCDB is the EPO's master bibliographic database, which contains IPCs, citations, titles, and all bibliographic data. Accordingly, the data contain an examiner-oriented bias, e.g., data that are useful for or originated from the examination process (for instance, priorities or citations) and have higher quality than the other information (see applicant and inventors addresses).

Although improvement in addresses, for example, has been undertaken in response to user community requests, the coverage still remains very poor for most offices (Table 2).

Table 2. Missing inventor geographic data by inventor number (top 15)

APPLN_AUTH	no zip	no country	no address	no city
US	98%	21%	97%	25%
EP	100%	0%	1%	1%
DE	100%	33%	100%	100%
JP	100%	98%	99%	100%
CN	100%	2%	100%	100%
CA	100%	45%	100%	100%
AU	100%	98%	100%	100%
SU	100%	41%	100%	100%
AT	100%	29%	100%	100%
KR	100%	14%	100%	100%
FR	100%	98%	99%	100%
GB	100%	70%	65%	100%
RU	100%	29%	100%	100%
CH	100%	11%	100%	100%
BR	100%	89%	100%	100%

Second, PATSTAT has a European-centered bias: data from national authorities are exchanged with EPO on the basis of conventions that may change over time, sometimes leaving gaps unfilled for certain IP offices in terms of missing applications, citations, or applicants/inventors. EPO provides a table giving coverage by application authority but this unfortunately covers only applications in terms of absolute numbers, not percentages of total applications issued by a patent office, and no benchmark for other information is provided. Thus, when operating outside main Western patent authorities, data should always be checked in terms of coverage.

3. Issues with PATSTAT

In this section, we will discuss three most critical issues that users must recognize when planning to use PATSTAT.

3.1. Applicant and inventor names

The first issue is the need to harmonize names of applicants and inventors. These are two of the most important pieces of information when analyzing patent data. One can use applicant and inventor names to find the number of patents filed by the applicants of interest and the number of inventions by different inventors. Performing such analyses currently requires extensive procedures for several reasons.

First, applicant and inventor names remain blank in PATSTAT even though this information is available for most patents (Table 3).[3] There is no practical solution to this issue. Accordingly, there might be unobserved samples when users retrieve data using applicant and inventor names.

Table 3. Missing applicant and inventor information

Table 3-1. Top 15 by number of applications with no inventors

APPLN_AUTH	No. of applications	%
JP	6928740	39%
US	2809584	20%
DE	2648915	37%
GB	2374229	69%
FR	1921790	61%
CA	981211	31%
CN	805171	10%
SE	738164	68%
ES	583890	54%
BE	562939	87%
NL	545512	88%
KR	430353	15%
CH	403005	38%
AT	399333	34%
IT	316618	43%

Table 3-2. Top 15 by number of applications with no applicant

APPLN_AUTH	No. of applications	%
JP	6311086	36%
US	3118448	23%
DE	1274894	18%
FR	1223831	39%
CA	965209	30%
CN	762959	10%
SE	734953	68%
GB	581375	17%
BE	458848	71%
KR	428405	15%
NL	386835	63%
CH	347467	33%
RU	299288	42%
SU	204684	15%
AU	156645	10%

Second, one entity can appear with different names in patent documents. As a result, one single entity might have tens or hundreds of IDs (in the worst case) due to several

[3] In autumn 2014, 76.6% of applications have at least one valid applicant and 69% have one valid inventor.

reasons in nature of the patent filing process: (1) as most patent offices receive patent applications via legal patent attorneys, they have no need to maintain an applicant or inventor database; (2) in many cases, applicants prefer to avoid competitor business intelligence on their patents or even head hunting of their inventors, (3) for international patent applications, patent documents are translated to local languages, but these translations can be based on different rules, e.g., local grammar or sounds, (4) different character sets, (5) simple typos, and so on. For example, Toyota Motor Corporation appears as "Toyota Motor Corporation," "Toyota Motor Co," "Toyota Jidosha Kabushiki Kaisha" ("Jidosha" and "Kabushiki Kaisha" mean a car and a corporation in Japanese, respectively), "Toyota Jidosha K. K.," and so on. Similarly, "Yılmaz" and "Şahin," which are Turkish common names, become "Yilmaz" or "Sahin" in the US patent office due to different character sets for documents written in English. All these reasons make it difficult for database managers to identify identical entities and their patents as well as make it difficult for users to find them.

There are several methods to avoid overcounting or undercounting. One method for a researcher is to personally harmonize applicants and inventors names of interest. This involves searching for applicants and inventors name in all possible patterns and correcting name errors. Another method is to use a unique ID table in PATSTAT. EPO tried to make this easier by introducing in *table 206_persons* an identifier (*DOC_STD_NAME_ID* in PATSTAT) that should assign a unique ID to the same entity. Nevertheless, this identifier has not been fully developed, rendering it less than fully reliable. The other method is to use external databases. Some researchers have exploited text mining techniques and offered the community the result of their efforts to create a unique identifier for applicants and inventors. We will now describe three of these free and publicly available databases.

I. EEE-PPAT

This database was developed by ECOOM (KU Leuven) and EUROSTAT (Peeters et al., 2009), and mainly comprises a table linkable by *person_id* to PATSTAT where the patent assignee name has been harmonized and a sector allocation (i.e., identifying whether patentees are private business enterprises, universities/higher education institutions, governmental agencies, or individuals) has been performed. The methodology involved more than 4000 text cleaning patterns to remove most common misspellings and to impose uniform character sets (replacing accents, umlauts and non-common characters with their plain version). Eventually, a hand check was performed to improve precision and recall of the dataset. To obtain the EEE-PPAT table, as well as the papers describing the underlying methodologies, requests should be

addressed to TechnoInfo@ecoom.be.

II.OECD HAN (Harmonised Applicants' Names) Database

The OECD HAN database provides a dictionary of applicants' names, which have been elaborated with business register data to be easily matched by all users. It results from a three-step data processing procedure:

(1) Identify business organizations, non-business organizations, and individuals among patent holders;

(2) Clean the company names; (step 1 and 2 based on KU Leuven algorithm–see above); and

(3) Consolidate the cleaned names by matching patent data with other databases, e.g., business registers.

Data are organized by *person_id* to easily link it to PATSTAT. For more details, see Thoma & Torrisi (2007). Compared to EEE-PPAT, the OECD HAN Database has a more limited scope (EU only), but a more extensive cleaning process.

III. APE-INV

APE-INV is a project funded by the European Science Foundation to clean and standardize inventors' names and addresses, as well as to match them to academic scientists' names and affiliations for EPO patents. The methodology, also described in Lissoni et al. (2007), begins from a disambiguation algorithm (Pezzoni, 2014) that uses all information in common among similar names (for example address, applicant name, technological classes) to collapse distinct *person_id*s into the same inventor.

Data also are validated and enriched through a collaborative network where dataset users can provide a feedback that is moderated by an arbitration system allowing improvements to the disambiguation produced. Data can be accessed at http://www.esf-ape-inv.eu

Additional necessary works after name harmonization include reflecting changes in corporate names, M&A, parent–subsidiary relationships, and so on.

Even if PATSTAT data obtained through all name harmonization efforts mentioned above may still differ from the reality, many concerns will have been eliminated. Statistical analyses may help further reduce concerns regarding missing observations if the number of missing observation is minimal.

3.2. Patent family

The next issue is the patent family. Patents are territorial. Because patent laws and

examination processes vary between countries, patents must be filed in countries of interest in methods established by authorities of those countries. As a result, a patent application for an invention in one country is filed either in the same format or in different formats in other countries. For instance, the DOCDB family ID is 3818534, to which the WO patent number 0140926 belongs, is compounded by 214 applications, issued in patent offices as shown in Table 4.

Table 4. Example of the size of one patent family (DOCDB family ID 3818534)

APPLN_AUTH	Count
AT	3
AU	27
CA	11
CN	9
DE	5
DK	1
EP	11
IL	15
JP	14
MX	7
SG	11
US	100

If a user does not count international patent applications without considering the patent family, then the user will obtain exaggerated counts. The patent family is defined to bundle the same invention in different patent documents. For example, US PTO defines a patent family as “the same invention disclosed by a common inventor(s) and patented in more than one country.”

Although the definition is well-defined, how to account for patent families has been an important and complex issue in practice, mainly due to the existence of different types of patent families. For example, WIPO (2013) defines six types of patent families: simple, complex, extended, national, domestic, and artificial.[4] The two most commonly accepted types are the simple and the extended patent families. The simple patent family means a patent family relating to the same invention, each member of which has for the basis of its “priority right” exactly the same originating application or

[4] For a wider analysis of family types and limitation, we suggest Martínez (2010), Martínez (2011), and Harhoff et al. (2003).

applications (WIPO, 2013). This implies that the simple patent family indicates patents in which all documents have the exact set of priorities (this type of family contains equivalent documents). In contrast, the extended patent family means a patent family relating to one or more inventions, each member of which has for the basis of its "priority right" at least one originating application in common with at least one other member of the family (WIPO, 2013). In other words, an extended patent family contains all documents relating in any way to the root document. Table 5 offers an example showing how the same set of documents can be classified using simple or extended family criteria.

Table 5. Simple and extended patent families.

Simple patent family	**Extended patent family**	**Document**	**Priorities**		
S1	E1	Document D1	Priority P1		
S2	E1	Document D2	Priority P1	Priority P2	
S2	E1	Document D3	Priority P1	Priority P2	
S3	E1	Document D4		Priority P2	Priority P3
S4	E1	Document D5			Priority P3

PATSTAT is released with two family tables (*tls218_docdb_fam* and *tls219_inpadoc_fam*); while the latter allows extended patent families to be built, the former is supposed to be more of an *examiners' technology based family* as it includes applications sharing the same set of priorities "adding new technological content." The DOCDB family, thus, may be described as a subset of simple families, excluding USPTO continuation and divisional applications. In reality, the family is built by linking applications that have exactly the same Paris Convention priorities in table *TLS204_appln_prior*. To give a practical example of usage, we provide here the SQL code for count of the average number of citations for the INPI (French patent office) by year, in terms of patent families by application.

```
Select
   t01.APPLN_AUTH, Year(t01.APPLN_FILING_DATE) as year,
   Count(Distinct t18.DOCDB_FAMILY_ID)/Count(Distinct t01.APPLN_ID) as avg
From
   tls218_docdb_fam t18 Inner Join
   tls211_pat_publn t11a On t18.APPLN_ID = t11a.APPLN_ID Inner Join
   tls212_citation t12 On t11a.PAT_PUBLN_ID = t12.PAT_PUBLN_ID Inner Join
   tls211_pat_publn t11b On t12.CITED_PAT_PUBLN_ID = t11b.PAT_PUBLN_ID
   Inner Join      tls201_appln t01 On t11b.APPLN_ID = t01.APPLN_ID
Where
   t01.APPLN_AUTH = 'FR'
Group By
   t01.APPLN_AUTH, Year(t01.APPLN_FILING_DATE)
```

Another clear problem is that the patent family is not defined by laws but within the domain of the database being used. This means that in a situation where an invention was first filed, for instance, in the US then extended to EP, MX, and CA, if the database does not contain MX data, the derived patent family will not contain such a patent. If the following edition includes Mexican data, the examined patent family would contain such a patent.

3.3. Technological Patent Classifications

The final issue discussed in this study is technological classifications. Technological classifications of inventions are recorded as International Patent Classification (IPC) of patents in PATSTAT. The first version of the IPC system entered into force in 1975, after the Strasbourg agreement (1971), and it comprises eight sections (indicated by a letter), followed by two digits indicating the class and a letter for subclass. The subclass is followed by one to three digits (group number) and two more digits separated by a backslash (subgroup). Such a system can currently identify 129 classes, 639 subclasses, 7,314 main groups, and 61,397 subgroups. The main sections include (A) Human Necessities, (B) Performing Operations, Transporting, (C) Chemistry, Metallurgy, (D) Textiles, Paper, (E) Fixed Constructions, (F) Mechanical Engineering, Lighting, Heating, Weapons, (G) Physics, and (H) Electricity. IPCs have been widely used by PATSTAT users to find patents in specific technological fields. Sometimes, although technological classifications are not necessarily the same as industrial codes, IPCs are also used to find patents in specific industries. However, users

must specifically focus on using IPCs for three reasons.

First, there are several criteria involved in IPC assignment. IPCs are primarily assigned to help patent examiners and users easily search for prior arts (WIPO, 2014). IPCs can be assigned based on invention information as well as non-invention information including categories of subject matter, places in the classification for an invention's technical subjects, function-oriented and application-oriented places, and classification of an invention's technical subjects. Accordingly, a list of patents retrieved by an IPC search may not fall into the same domain. Furthermore, some patents retrieved by an IPC search may not be originally aimed at technologies different from the IPCs searched. One method to overcome this issue is to use the primary IPC. Some authorities define the concept of the primary technological classifications that best describes the inventive information of the patent. For example, in the US patent documents, such classifications appear in bold and in the first position (Figure 3). PATSTAT reflects such information in *IPC_POSITION*. Using the primary IPC can help retrieve patents in the primary technologies of interest.

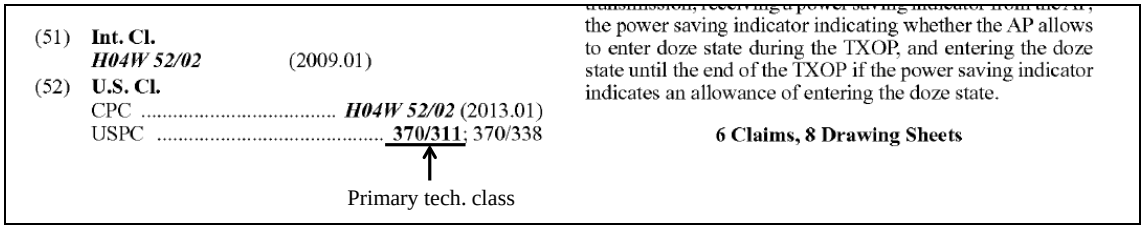

Figure 3. Example of primary technological class

Second, IPCs are technological classifications. An issue arises when trying to find a list of patents in a specific industry. Technological classifications do not exactly correspond to industrial classifications as technologies become more complex and technological convergence continues. One method to deal with this problem was to use concordance schemes. For example, a concordance file was provided by the US Patent Office and Trademark Office (US PTO). The file contained information regarding the technological classifications in US Patent Classification (USPC) that match industrial classifications in the Standard Industry Classification. However, this concordance table is no longer available.[5] A similar effort was performed by KU Leuven (Schmoch et al., 2003; van Looy et al., 2014). They made a concordance table between IPCs and industrial classifications as defined in Statistical Classification of Economic Activities in the European Community (NACE).

[5] As of 10/03/2015.

Third, many IPC assignment cases are wrong. There is no practical solution to overcome these typos, except by users checking them individually. Nonetheless, this concern becomes less significant when a patent is assigned several IPCs.

4. Conclusion

Demand for and interests in patent data analyses have been significantly increasing in recent years. Even if many scholars have used patent data for their studies, many other people, including scholars, managers, researchers, and policy makers, are interested in patent data analysis regardless of their experience and discipline background. This study provided a comprehensive summary of and guidance for using PATSTAT. Specifically, we highlighted the three most important issues that PATSTAT users must consider and the ways to deal with them. We chose PATSTAT because it has become one of the most widely used patent databases for scholars. However, our discussion is also applicable to other patent databases.

References

Dernis, H. & Khan, M. (2004) Triadic Patent Families Methodology, OECD Science, Technology and Industry Working Papers, 2004/02, OECD Publishing. http://dx.doi.org/10.1787/443844125004

Du Plessis, M., Van Looy, B., Song, X & Magerman, T. (2009) Data Production Methods for Harmonized Patent Indicators: Assignee sector allocation, EUROSTAT Working Paper and Studies, Luxembourg.

Harhoff, D., Scherer,F., & Vopel, K. (2003) Citations, family size, opposition and the value of patent rights, *Research Policy* 32(8), 1343–1363.

Jaffe, A. B., & Trajtenberg, M. (2002) *Patents, Citations & Innovations: A Window on the Knowledge Economy*. MIT Press.

Lerner, J., & Seru, A. (2015) The use and misuse of patent data: Issues for corporate finance and beyond.

Lissoni, F., Coffano, M., Maurino, A., Pezzoni, M., & Tarasconi G. (2010) APE-INV's "name game" algorithm challenge: A guideline for benchmark data analysis and

reporting, Technical Report, Academic Patenting in Europe - APE-INV

vanLooy, B., Vereyen, C., & Schmoch, U. (2014) Patent Statistics: Concordance IPC V8 – NACE Rev. 2, EUROSTAT.

Magerman, T, Grouwels, J., Song, X. & Van Looy, B. (2009) Data Production Methods for Harmonized Patent Indicators: Patentee Name Harmonization. EUROSTAT Working Paper and Studies, Luxembourg.

Maraut, S., Dernis, H., Webb, C., Spiezia, V., & Guellec, D. (2008) The OECD REGPAT Database: A Presentation, OECD Science, Technology and Industry Working Papers, 2008/02, OECD Publishing. http://dx.doi.org/10.1787/241437144144

Martínez, C. (2010) Insight into Different Types of Patent Families, OECD Science, Technology and Industry Working Papers 2010/2, OECD Publishing.

Martínez, C. (2011) Patent families: when do different definitions really matter?, *Scientometrics* 86(1), 39–63.

Nagaoka, S., Motohashi, K., & Goto, A. (2010) Patent Statistics as an Innovation Indicator, in Hall, B.H, Rosenberg, N., (Eds), *Handbook of the Economics of Innovation*, Volume 2, Academic Press, 1083-1128.

Peeters, B., Song X., Callaert J., Grouwels, J., & Van Looy, B. (2009) Harmonizing harmonized patentee names: an exploratory assessment of top patentees. EUROSTAT working paper and Studies, Luxembourg

Pezzoni M., Lissoni F., & Tarasconi G. (2014) How to kill inventors: testing the Massacrator© algorithm for inventor disambiguation, *Scientometrics* 101(1), 477-504.

deRassenfosse, G., Dernis, H. & Boedt, G., (2014) An introduction to the Patstat database with example queries, *Australian Economic Review* 47(3), 395-408.

Schmoch, U., Laville, F., Patel, P., & Frietsch, R. (2003) *Linking technology areas to industrialsectors: Final report to the European Commission*, DG research,

Bruxelles.

Thoma, G. & Torrisi, S.(2007) Creating Powerful Indicators for Innovation Studies with Approximate Matching Algorithms: A test based on PATSTAT and Amadeus databases, KITeS Working Papers 211, KITeS, Centre for Knowledge, Internationalization and Technology Studies, Universita' Bocconi, Milano, Italy, revised Dec 2007

WIPO (2014) *International Patent Classification* (Version 2014).

WIPO (2013) Handbook on industrial property information and documentation,