Scientometrics (2016) 106:187–211
DOI 10.1007/s11192-015-1786-0

Patent citation indicators: One size fits all?

Jurriën Bakker[1] · Dennis Verhoeven[1] · Lin Zhang[1,2] · Bart Van Looy[1]

Received: 18 May 2015 / Published online: 2 November 2015

Abstract The number of citations that a patent receives is considered an important indicator of the quality and impact of the patent. However, a variety of methods and data sources can be used to calculate this measure. This paper evaluates similarities between citation indicators that differ in terms of (a) the patent office where the focal patent application is filed; (b) whether citations from offices other than that of the application office are considered; and (c) whether the presence of patent families is taken into account. We analyze the correlations between these different indicators and the overlap between patents identified as highly cited by the various measures. Our findings reveal that the citation indicators obtained differ substantially. Favoring one way of calculating a citation indicator over another has non-trivial consequences and, hence, should be given explicit consideration. Correcting for patent families, especially when using a broader definition (INPADOC), provides the most uniform results.

Keywords Patent citations · EPO · USPTO · PCT · Patent family · Multivariate analysis

JEL Classification O34

✉ Jurriën Bakker
Jurrien.Bakker@kuleuven.be

Dennis Verhoeven
Dennis.Verhoeven@kuleuven.be

Lin Zhang
Lin.Zhang@kuleuven.be

Bart Van Looy
Bart.Vanlooy@kuleuven.be

[1] Department of Managerial Economics, Strategy and Innovation, KU Leuven, Naamsestraat 69, 3000 Louvain, Belgium

[2] North China University of Water Conservancy and Electric Power, No. 36, Beihuan Road, Zhengzhou 450045, Henan, China

Mathematics Subject Classification 62H20 · 62H25 · 62H30

Introduction

The number of times that patents are cited by other patents[1] can be used to complement the mere *counting* of patented inventions in order to address the differences in value and impact between inventions. The idea of using patent citations as an indicator is relatively old and appears to have originated from Seidel in 1949 (Karki 1997). However, the first systematic empirical investigations only emerged in the 1980s, with Carpenter et al. (1981) showing that patents related to industry awards are cited more frequently.

A patent can be cited for various reasons: an inventive step, its industrial relevance, to qualify novelty, or to provide additional, relevant information to situate the claims advanced in the patent document. Patents that are cited (more often) are considered more important and valuable than patents that are not used (or used infrequently) to qualify subsequent technological activity. Therefore, one can approximate an individual patent's importance by the number of times it is cited. This argument is empirically supported by the work of Albert et al. (1991), Arts et al. (2012) and Gambardella et al. (2008) who show that patent citations correlate significantly with the value of the individual patent. Likewise, Hall et al. (2005), Narin et al. (1987), Neuhäusler et al. (2011) and Trajtenberg (1990) find a positive correlation between firm performance and the total number of forward citations that their patents receive, even after correcting for firm size. Lanjouw and Schankerman (2004) have determined that patent citations are correlated with other indicators of patent quality, which in turn are correlated with variations in firm value. Additionally, Neuhäusler and Frietsch (2012) and Frietsch et al. (2014) show that forward patent citation counts are strongly correlated with export volume.

While (front page) patent references are ultimately included by examiners, a number of researchers conceive citations as an approximation of knowledge flows: (Hall et al. 2005; Jaffe et al. 1993, 2000; MacGarvie 2006; Paci and Usai 2009). When this perspective is adopted, the number of patent citations received indicates the subsequent influence or impact of the knowledge implied in the patented invention.

A major advantage of using patent citations as an indicator of inventive quality, either conceived as value or impact, pertains to the relative simplicity of the measure: it merely requires counting the number of citations a patent receives. Since a large number of patents receive citations,[2] this measure allows for the construction of enriched indicators both on the patent level and on more aggregate levels (e.g. firm, industry, country). Currently, patent citations are considered an important indicator of the innovative output of companies (e.g. Hagedoorn and Cloodt 2003). They also enable statistics and rankings that can be used to determine the innovative performance of countries (e.g. Chakrabarti 1991; Criscuolo and Verspagen 2008; Neuhäusler and Frietsch 2012).

While these, and related studies, point to the relevance of counting the citations received by patent documents, the method of measuring this count is not singularly defined. Despite the simple conceptualization of the measure, calculating citation indicators involves a number of methodological decisions that, in turn, result in a variety of possible citation

[1] Often referred to as patent citations, forward citations or patent citation count. We will use these terms throughout this paper.

[2] Up to 88 % of applications score a non-zero citation count on at least one of the citation indicators we computed.

indicators. The first decision is to choose the data source from which to compile patent citations, given that patent systems are geographically bounded (e.g. US, EU, Japan, China). Since patent citations to one patent system can stem from different geographic areas, the second decision is to choose the source from which citations to the focal set of patents will be included. Finally, given the possible existence of multiple patent documents pertaining to a single invention, a viable option is to treat equivalent patent documents as one patent family, which will also affect citation counts. Currently, there are three different approaches to these decisions in the literature.

The National Bureau of Economic Research (NBER) set up a data platform that contains only patents filed at the United States Patent and Trademark Office (USPTO). This data has been available as early as 2001 (Hall et al. 2001). Additionally, the first analyses on patent citations relied on data from USPTO documents (e.g. Carpenter et al. 1981; Narin et al. 1987). The NBER database is still widely used as the high number of recent citations to the source paper from Hall et al. (2001)[3] attests.

A second set of studies has been conducted using European Patent Office (EPO) patent documents. European patent data is noticeably different from USPTO data: the EPO patents cover a different geographic area; they are heterogeneous in terms of the countries where they are filed; and finally, examiners tend to include fewer citations than their colleagues from the USPTO. Citation data from EPO patents have been compiled since 2003 (Webb et al. 2005) resulting in several EPO-based patent citations studies (e.g. Harhoff and Reitzig 2004; Neuhäusler et al. 2011; Schoenmakers and Duysters 2010).

Finally, some researchers have opted to go beyond the use of data stemming from a single source (patent office) and take into account the presence of patent families (hence, considering the equivalents of an invention that are present in multiple patent systems when calculating citations). This seems especially appropriate in correcting for 'home biases' (Criscuolo 2006) and in providing a more encompassing view of the impact of an invention. Examples of this approach can be found in the work of Gambardella et al. (2008), Graham and Harhoff (2006), Magerman et al. (2011), and Neuhäusler and Frietsch (2012).

When using a patent citation indicator, it is implicitly assumed that different calculation methods of this indicator will, in general, yield similar results. However, this may not necessarily be the case: patent citations from different offices may reflect 'national' impact rather than 'global' impact. Additionally, patent offices focus on their own geographical jurisdiction, which may result in a 'home bias' when looking at patent citations (Criscuolo 2006). Finally, offices and, hence, examiners' practices vary in terms of the average number of patent citations included: USPTO patent documents display (on average) more citations than EPO patent documents. This, in turn, can lead to a situation whereby citation indicators—derived from different computational choices—do not reflect the same information (Alcácer and Gittelman 2006). For these reasons, it makes sense to assess the effects of the methodological choices that researchers face when assessing patent quality through forward citations. To the best of our knowledge, no systematic analysis of this kind has been performed. This paper will assess the extent to which different methods yield (dis)similar results. Hence, we pose our research question as follows:

Do citation counts that are computed by different methods reveal similar information?

This question can be further refined by adopting the distinction between technological improvements of an incremental nature vis-à-vis inventions implying a more radical departure from what was previously possible (Baumol 2004; Dosi 1982). Accordingly, researchers have operationalized these 'breakthrough' inventions by identifying patents receiving exceptionally

[3] This paper needs to be cited when the NBER database is used.

high numbers of forward citations (e.g. Ahuja and Lampert 2001; Chakrabarti 1991; Schoenmakers and Duysters 2010). Since citation counts may depend on computational choices, it is of particular interest to compare different methods with respect to identifying highly cited patents. This leads to the following extension of the research question:

To what extent do different calculation methods affect the identification of highly cited patents?

In the remainder of this paper, we answer our first question using correlation and cluster analyses, which compare different methods to calculate citation counts of patent applications. To answer the second research question, we compute the degree of overlap observed between patents that are identified as highly cited by various methods. We start with a systematic discussion of the different computational choices, resulting in a set of indicators that this study then compares. We then present the empirical findings that we obtained and discuss their implications. Overall, our findings signal non-trivial differences among the variety of approaches envisaged.

Overview of the methodological choices when computing patent citation indicators

When counting patent citations, different choices need to be made. These choices pertain to the patent system (or protocol) in which the receiving and citing patent documents reside. The presence of patent families could also be taken into account. In this section, we discuss the general choices that are available when counting forward citations.

The patent office

The patent system in which the patent resides may affect the way in which the patent is cited. This is due to two reasons: the home bias and the inherent difference between the patent systems. A home bias, as discussed in the introduction, implies that patent examiners cite more prior art present in their own jurisdiction[4] (Michel and Bettels 2001). In addition, while patent systems are largely similar in terms of subject matter and application procedures, they nonetheless differ in several ways. Not only are there observable differences in terms of subject matter—between the USPTO and the EPO concerning the costs incurred (van Pottelsberghe de la Potterie and François 2009)—but practices such as the 'duty of candor'[5] in the US lead to an increase in references being included in patent documents, which may have an impact on citation-based indicators.

Selection of the citing patents

The second choice a researcher faces relates to selecting the patent documents that cite the focal patent. One can choose to count either the citations that an entity (application or patent family) receives from patents in the same patent office (e.g. EPO, USPTO) or to include citations from patents present in other patent systems. The reason this distinction is worth investigating is twofold.

[4] We show this later in Table 4.

[5] The 'duty of candor' rule requires that applicant and inventors involved in a patent application must disclose all known information which may adversely affect the probability of obtaining a granted patent.

First, we note that many researchers restrict themselves to a single source, which is often the EPO or the USPTO system, as noted in the introduction. This implies they only count citations that patent applications receive from documents residing in the chosen system. Therefore, it is interesting to examine the effects of this restriction: does restricting citations to the office of the focal application significantly alter the results?

Second, most documents tend, largely, to cite patent documents from within their chosen 'system', due to the examining process (Michel and Bettels 2001). This is not unexpected since patent examiners should have an overriding concern for the validity of the application within their own jurisdiction. At the same time, when specific procedures are in place, differences can become more pronounced. The case of USPTO is apposite in this respect. When applying to the USPTO, applicants have a so-called *duty of candor,* requiring them to disclose to the examiner any knowledge of prior art, even if this information could lead to the application being disqualified. Patent examiners then select from these references and/or add other references deemed relevant. However, USPTO examiners are most familiar with USPTO patents. In the case of foreign applicants, references stemming from prior art located outside the American patent system may be advanced relatively more frequently by such applicants. Indeed, Sampat (2004) observed that, in approximately 70 % of patents, references to foreign patents are initially advanced by the applicant (see Azagra-Caro et al. 2011 in this respect).

Correcting for patent families

Patents that represent and/or build on the same invention can also be grouped into so-called 'patent families'. It makes sense to correct citations for the presence of families since other patents can make reference to multiple family members besides the initial, focal application. If the researcher feels that such a citation is just as valuable as a direct citation of the initial patent application, then a correction based on the patent family seems appropriate. In general, this involves adding citations from family members to the citation count of the focal application itself. A case study by Nakamura et al. (2015) shows that accounting for patent families can improve analyses based on patent citations.

There are different definitions of the patent family: in this paper, we consider two. Martínez (2011) defines them as the extended patent family (INPADOC)[6] and the examiner's technology-based family (DOCDB).[7] The DOCDB definition centers on finding the closest equivalents of a patent document in other offices. These documents are usually characterized by having the same priority applications.[8] The INPADOC definition is less strict and is used to find documents protecting the same invention, including documents with a somewhat different priority profile (Albrecht et al. 2010). The members of INPADOC patent families share priority applications with at least one other member of the family. Therefore, patents that are members of the same DOCDB patent family should also be members of the same INPADOC patent family, since all DOCDB patent family

[6] INPADOC is an abbreviation for INternational PAtent DOCumentation, the patent data collected but not generated by the EPO (2014). It is also used to denote the extended patent family in the EPO PATSTAT databases.

[7] DOCDB is the EPO master documentation database (Martínez 2011). It is also used to denote the examiner's technology-based patent family in the EPO PATSTAT databases.

[8] Albrecht et al. (2010) define the DOCDB patent family as patent applications that have an equal 'priority picture': this can, under certain circumstances, include the priority application itself. Additionally, this family is corrected to include applications that have the same technical content but have been excluded due to a 'discrepancy in the priority picture' Albrecht et al. (2010: 283).

members have the same priority applications.[9] However, it is possible that two members of the same family share no priority applications. This can occur when they both share a priority application with a third member of the family (Lingua 2005). In this study, both family definitions will be adopted and assessed.

Data and methods

Data used

We used patent data from the October 2011 version of the EPO PATSTAT database. From this data, we extracted indicators for patent applications belonging to the EPO and the USPTO, as well as applications that were filed through the Patent Cooperation Treaty (PCT) route. We chose these applications for three reasons: First, most research that employs patent citation data uses patents from at least one of these three systems (or routes, in the case of PCT applications). Second, the data provided by these offices from the USPTO and the EPO is relatively complete in PATSTAT, compared to other offices (also included in PATSTAT). In the remainder of this paper, we shall refer to different origins by designating documents as EPO, USPTO and PCT patent applications.

The focal applications for which the indicators were calculated have been cleaned to remove—amongst others—duplicates caused by untraceable priorities and citations, incorrect conversions of patent numbers, and several issues caused by changes in the USPTO system in 2001.[10] In addition, we only considered USPTO applications that were granted. This is due to the observation that USPTO applications that did not lead to a granted patent are not completely covered by PATSTAT.

After the cleaning exercise, we were left with 8,658,272 focal applications from which 4,397,304 were applications filed at USPTO, 2,343,707 applications filed at EPO and 1,917,261 applications filed via the PCT route. The filing dates range from the 2nd of January, 1970 to the 6th of May 2011. However, it should be noted that the cleaning activity led to the removal of a large number of applications: 3,319,894 applications from the USPTO (mainly because no granted equivalent was yet present); 10,567 applications from the EPO, and 11,335 PCT applications.

With regard to the citing applications, we used all patent documents available in the 2011 October version of PATSTAT. We excluded only artificial applications.[11] Therefore, the cited applications involved more cleaning than the citing applications. This was carried out because we wanted to keep the citation indicators as close as possible to those obtained when using currently available databases (notably PATSTAT). Consequently, we did not correct all recently known issues that exist in patent citation indicators.[12]

[9] This statement holds for the vast majority of patent applications in the EPO PATSTAT database; there is a small minority of patents (0.09 % of DOCDB patent families) that do not fulfill this criterion due to discrepancies in their priority picture. However, these families do not affect the analyses presented later in this paper.

[10] These imply changes in publication types; patent duplicates that occur before and after 2001; and applications that are not available before 2001 but partly available thereafter.

[11] These are added to the database to maintain logical links and do not actually represent any patent applications.

[12] An example of this pertains to the well-known issue that EPO references other patents by referring to the references of their PCT equivalents via a non-patent reference in PATSTAT. This has been noted in Harhoff et al. (2003) and Neuhäusler et al. (2011).

Table 1 Simplified table of naming indicators

The origin of the prefix	Office of the focal patent	Application or patent family	If patent family correction only applied on the cited side	If patent family correction applied on both sides	If only citations from the office of the focal patent are used
Possible prefixes	EPO USPTO PCT	Application DOCDB INPADOC	Cited family	Full family	Within office

All indicator names consist of a number of prefixes and the word count. This table explains the origins of each prefix. Full definitions for each indicator can be found in Table 2

The patent citation indicators and their definitions

We performed four different permutations to calculate our indicators. These are based on patent origin, citation origin and a twofold family correction (see previous section). We have chosen these permutations in the belief that they represent virtually all possible permutations that researchers are likely to consider when working with patent citations. In this section, we explain how these permutations are used.

Starting with patent origin, we compare indicators resulting from three different data sources: EPO, USPTO, and applications filed through the PCT route. We use this data because the vast majority of publications dealing with patent citations use indicators drawn from these sources. Next, we distinguish two groups of indicators based on the source of the citation. This is done by comparing the number of citations received from applications in the office of the focal application, and the number of citations that were received irrespective of the patent office.[13] We will denote those indicators with a restricted source of citations by adding 'within office' to the indicator name.

A third permutation deals with applying a correction for citations received by family members of the focal application. Each family indicator is, therefore, replicated for each patent office. For the patent family definition, we compare both the INPADOC and DOCDB definitions. We denote patent citation indicators that correct for patent family on the cited side (i.e. an indicator that counts all applications that cite the family of the application) by including 'cited family count' in their name.

It is possible that a number of citations originate from applications that are part of the same patent family. It can be argued that these citations are mere duplicates since the patent is cited twice by the same invention. This could then create a bias towards citations received from larger patent families, since it is inherent that the size of the family increases the probability of two or more of its members citing the same patent. Therefore, as a final, fourth permutation, we correct for this bias by counting not the number of patent applications but rather the number of patent families that cite the focal family. We denote patent citation indicators that have this correction by replacing 'cited family count' with 'full family count' in their name. Table 1 provides an overview of the prefixes for the indicators used in this paper

This leads to a total of ten different indicators for each office: two indicators based on the application, four indicators based on the DOCDB family and four indicators based on the INPADOC family. To keep the list of indicators tractable, we provide names and definitions for each indicator in Table 2.

[13] In the case of applications filed through the PCT, other applications that followed this route were taken.

Table 2 Indicators and their definitions

Patent family	Patent citation indicator	Definition
N/A	Application count	Number of citations a patent application receives from all other patent applications, irrespective of their publication office
N/A	Application count within office	Number of citations a patent application receives from patent applications that were published in the same office as the focal application
DOCDB	Cited family count	Number of citations the DOCDB patent family of the focal application receives from all other patent applications, irrespective of publication office
DOCDB	Cited family count within office	Number of citations the DOCDB patent family of the focal application receives from patent applications that were published in the same office as the focal application
DOCDB	Full family count	Number of citations the DOCDB patent family of the focal patent receives from all other DOCDB patent families, irrespective of publication office
DOCDB	Full family count within office	Number of citations the DOCDB patent family of the focal patent receives from patent applications that were published in the same office as the focal application. This count is corrected for DOCDB patent family on the citing side
INPADOC	Cited family count	Number of citations the INPADOC patent family of the focal application receives from all other applications, irrespective of publication office
INPADOC	Cited family count within office	Number of citations the INPADOC patent family of the focal application receives from other patent applications that were published in the same office as the focal application
INPADOC	Full family count	Number of citations the INPADOC patent family of the focal patent receives from all other INPADOC patent families, irrespective of publication office
INPADOC	Full family count within office	Number of citations the INPADOC patent family of the focal patent receives from other patent applications that were published in the same office as the focal application. This count is corrected for INPADOC patent family on the citing side

These indicators are calculated for focal applications at the EPO, USPTO and PCT

We computed descriptive statistics for the indicators in Table 2; these are listed in Table 3. From these descriptive statistics, we can derive two main conclusions. The first is that a large number of patents receive at least one citation. However, the rate of patents with a non-zero citation count varies considerably, from 25 % (EPO application count within office) to 88 % (USPTO INPADOC full family count and USPTO INPADOC cited family count). Therefore, the distribution of the citation indicator varies from highly truncated to a more continuous spectrum. Second, we observe that the indicators vary greatly with respect to their averages and standard deviations. The average of the EPO application count within office is about 45 times smaller than the average of the USPTO INPADOC cited family count.

To perform the correlation analysis of the citation indicators, we use only applications that receive at least one citation for any of the indicators considered. In practice, this definition translates into selecting only those applications that receive at least one citation on the DOCDB level or the INPADOC family level. Consequently, other indicators can still have a score of 0. This was done in order to better assess the information contained in

Table 3 Descriptive statistics for the indicators that were computed for this paper

Focal patent source	Patent family	Patent citation indicator	Number of observations	Forward citation statistics			
				Average	SD	Median	Nonzero (%)
EPO	N/A	Application count	2,343,707	1.92	5.10	0	38
EPO	N/A	Application count within office	2,343,707	0.57	1.55	0	25
EPO	DOCDB	Cited family count	2,343,707	9.03	20.88	3	75
EPO	DOCDB	Cited family count within office	2,343,707	1.07	2.51	0	41
EPO	DOCDB	Full family count	2,343,707	7.28	16.21	3	75
EPO	DOCDB	Full family count within office	2,343,707	1.03	2.33	0	41
EPO	INPADOC	Cited family count	2,343,707	17.05	84.56	4	79
EPO	INPADOC	Cited family count within office	2,343,707	1.76	8.37	0	45
EPO	INPADOC	Full family count	2,343,707	11.21	47.774	3	79
EPO	INPADOC	Full family count within office	2,343,707	1.58	6.43	0	45
USPTO	N/A	Application count	4,397,304	9.91	18.22	5	82
USPTO	N/A	Application count within office	4,397,304	8.46	16.35	4	79
USPTO	DOCDB	Cited family count	4,397,304	13.05	24.66	6	86
USPTO	DOCDB	Cited family count within office	4,397,304	10.20	21.08	5	82
USPTO	DOCDB	Full family count	4,397,304	10.85	19.48	6	86
USPTO	DOCDB	Full family count within office	4,397,304	8.97	17.31	4	82
USPTO	INPADOC	Cited family count	4,397,304	25.95	129.50	8	88
USPTO	INPADOC	Cited family count within office	4,397,304	19.73	102.23	6	84
USPTO	INPADOC	Full family count	4,397,304	16.95	72.90	6	88
USPTO	INPADOC	Full family count within office	4,397,304	13.54	58.94	5	84
PCT	N/A	Application count	1,917,261	1.90	5.63	0	41
PCT	N/A	Application count within office	1,917,261	0.58	1.55	0	27
PCT	DOCDB	Cited family count	1,917,261	5.73	16.38	1	59
PCT	DOCDB	Cited family count within office	1,917,261	1.10	2.49	0	41
PCT	DOCDB	Full family count	1,917,261	4.63	12.73	1	59
PCT	DOCDB	Full family count within office	1,917,261	1.09	2.46	0	41
PCT	INPADOC	Cited family count	1,917,261	13.22	87.36	2	63
PCT	INPADOC	Cited family count within office	1,917,261	2.46	15.88	0	46
PCT	INPADOC	Full family count	1,917,261	8.63	50.28	1	63
PCT	INPADOC	Full family count within office	1,917,261	2.31	14.11	0	46

Table 4 Origin and destination of citations

Family correction	Focal office	EPO (%)	USPTO (%)	PCT (%)	EPO (national office)[a] (%)	Other (%)	Total (%)	Total citations received
None	EPO	31.35	36.14	19.84	12.31	0.36	100	4,501,136
None	USPTO	4.22	85.28	6.62	3.74	0.15	100	43,566,925
None	PCT	13.30	31.33	24.07	30.72	0.58	100	3,635,340
DOCDB family	EPO	12.03	64.16	14.58	8.85	0.39	100	21,160,972
DOCDB family	USPTO	6.45	78.18	8.54	6.61	0.22	100	57,379,697
DOCDB family	PCT	8.10	52.14	16.06	23.40	0.30	100	10,994,350
INPADOC family	EPO	10.33	66.25	15.17	7.94	0.31	100	39,950,651
INPADOC family	USPTO	6.99	76.09	10.69	6.01	0.22	100	114,120,819
INPADOC family	PCT	7.31	56.65	15.93	19.86	0.25	100	25,338,999

Citations are calculated as originating from applications from any office in the PATSTAT database to applications at the EPO, USPTO and PCT. Family correction implies that the citation is made to the patent family of applications at the EPO, USPTO and PCT. The citations are expressed in percentages of all citations to the (patent family of) applications at the focal office

[a] Patent offices that are located in the geographical area covered by the EPO

the citation counts. Its effects are quite substantial since—depending on the office[14]—a considerable share of patents in our sample have no citations, resulting in identical scores (0) for all indicators. The inclusion of applications that are never cited would have an inflating effect on the correlation and is, therefore, undesirable.

The distribution of citations

To better understand the behavior of the patent citation indicators, we compiled an overview of the origin and destination of citations, shown in Table 4. This table reveals that the USPTO is the main supplier of citations in the patent system. Not only does the vast majority of citations to USPTO entities come from the USPTO itself but the USPTO also supplies most citations to other documents. There are more USPTO citations to EPO documents than EPO citations to USPTO documents. A similar pattern emerges for PCT documents.

Correcting for patent family remedies this to some extent; at the same time, USPTO documents remain dominant since they account for the most citations overall. In the case of the EPO, INPADOC families with an EPO member receive 6.4 times more citations from USPTO documents than from EPO documents.

Note that the large majority of all citations stem from either USPTO, EPO or PCT documents; very few citations come from other offices such as the Japanese Patent Office (JPO) or the Chinese Patent Office (SIPO). It is interesting to observe that, from the remaining citations, the vast majority are from applications at the national level of the EPO. These citations may indeed represent a duplication of EPO patents, or they may be applications that were filed at only a single national office instead of the EPO, due to the costs of the EPO process [as noted by van Pottelsberghe de la Potterie and François (2009)].

[14] The exact figures are: 21 % for EPO applications, 12 % for USPTO applications and 37 % for PCT applications.

Table 5 Statistics of INPADOC and DOCDB families in our applications

Family	Number of families	% Singletons[a]	Average number of members	Overlap between both family definitions	% Overlap[b]	% Overlap[c]
INPADOC	5,309,452	21	2.64	4,179,052	79	73
DOCDB	6,017,825	35	2.01	4,179,052	69	63

[a] Families with only one member

[b] Including singletons

[c] Excluding singletons

Patent families

In this paper, we deploy two different family definitions: the DOCDB and the INPADOC definitions. We have compiled some descriptive statistics to understand the effects of correcting for patent family. These statistics are shown in Table 5. Here, we can see that a large number of patent families exist in the database. Note that, even though these families need at least one EPO, USPTO or PCT application, they may also have applications from other offices.

From these patent families, only between 21 and 35 % consist of a single patent application. Most patent families have at least two or more members. Finally, we see that a large number of patent families are equal for either family definition, even after excluding singleton families, which are equal by definition.

Results of the correlation analysis

The effects of expanding the sources of citing patents and correcting for patent family

We first determined the effect of correcting for family and citation origin for each office separately. For this purpose, we compared the 'application count within office' indicator with all other indicators in the office of the focal application. This was done for two reasons: First, the indicator is the most basic (i.e. it is uncorrected for family and only uses citations from its own office). Second, it is the indicator that is most widely used: the NBER citation indicator is the USPTO 'application count within office', while the aforementioned scholars who utilize EPO data often use the EPO 'application count within office'. The results of this exercise are presented in Table 6. The full correlation table can be found in "Appendix 1".

Table 6 shows that there is a substantial effect of citation origin (i.e. all citations vs. only those from within the office) on the patent citation indicators. This can be seen when inspecting the correlation of the 'application count within office' indicator with the 'application count indicator'. This effect is more pronounced for EPO and PCT indicators, with correlations of 0.77–0.79, than for their USPTO equivalent, which is less sensitive in this respect (see the correlation of 0.99). Given the citation information presented in Table 4, this should come as no surprise.

Correcting for patent family introduces considerable differences. The effects of this correction are more outspoken in the EPO and PCT systems than in the USPTO system: where the USPTO 'application count within office' has a correlation of 0.84 with the

Table 6 Correlation with the application count within office indicator for each office

Family	Compared indicator	EPO	USPTO	PCT
N/A	Application count	0.79	0.99	0.77
N/A	Application count within office	1	1	1
DOCDB	Cited family count	0.34	0.84	0.35
DOCDB	Cited family count within office	0.64	0.86	0.72
DOCDB	Full family count	0.33	0.84	0.34
DOCDB	Full family count within office	0.65	0.86	0.72
INPADOC	Cited family count	0.09	0.23	0.14
INPADOC	Cited family count within office	0.20	0.25	0.19
INPADOC	Full family count	0.12	0.25	0.16
INPADOC	Full family count within office	0.26	0.28	0.22

All correlations are significant at the 0.001 level

DOCDB family-corrected indicator, the equivalent correlations for EPO and PCT are situated around 0.33. Correcting for the INPADOC patent family has an even stronger effect than correcting for the DOCDB patent family. Finally, we see that correcting for patent family on the citing side has a relatively small effect. The values in Table 6 are almost equal for the cited family count and the full family count indicators. The tables in "Appendix 1" confirm this conclusion: the correlations between cited family count and full family count indicators are very close to 1 for both the DOCDB and the INPADOC family definitions.

The effect of using different sources (for patent documents present in all three systems)

For an inter-office comparison, we calculated the correlation for DOCDB patent families from which applications were filed at the EPO, the USPTO, and through the PCT route. This was done because the DOCDB family is based on the technical equivalence of the documents. Therefore, we can assume that the different elements in the DOCDB family are documents describing the exact same invention in different jurisdictions. Because of this equivalence, a direct comparison focusing on the source document is feasible.

Again, we considered only patents that had at least one citation in their largest (i.e. INPADOC) family. However, we found that all DOCDB patent families with applications in all three offices fulfilled this criterion. Therefore, this restriction did not change the analysis. These considerations led to the comparison of citation indicators for 388,512 DOCDB families. The full correlation matrix is presented in "Appendix 2". Here, we extracted the correlations that compare the different sources of patent data. These are listed in Table 7.

Table 7 shows that correlations for the basic indicators obtained for the same family but derived from relying on different offices are very low. The correlation between the EPO 'application count within office' and the USPTO 'application count within office' is only 0.09. Using citations from outside the office of the focal application ('application count') remedies this slightly by raising the correlation to levels ranging from 0.11 to 0.30.

Correlations observed when correcting for the DOCDB and INPADOC families are considerably higher. This is naturally the case for the DOCDB cited family count and the DOCDB full family count since the applications are all part of the same family. The

INPADOC cited family count and the INPADOC full family count indicators also have coefficients of 1, as shown in Table 7. This is due to the fact that applications that are members of the same DOCDB family are also members of the same INPADOC family. Interestingly, correcting for patent family increases compatibility, even when only citations from the office of the focal application are counted. Therefore, even when there is only application data from one patent office, correcting for the patent family of the focal applications is an interesting method for increasing compatibility with data from other patent offices.

Clustering the patent citation indicators

We performed a cluster analysis on the patent citation indicators by using the correlation table listed in "Appendix 2", i.e. pertaining to patent documents that have equivalents in all different systems under study. To define clusters, we performed a divisive cluster analysis, based on factor analysis (see "Appendix 3" for a technical description). Since the analysis compares patent applications with the counterparts of their DOCDB patent family, the indicators 'DOCDB cited family count' and the 'DOCDB full family count' give equal values regardless of the office of the focal application. Therefore, they are replaced by the general indicator. This is also carried out for the corresponding INPADOC family indicators since DOCDB family members are also part of the same INPADOC family: the INPADOC family is by definition larger. Including all INPADOC indicators would thus be redundant. The resulting indicators are denoted by the 'ALL' notation. The identified clusters are reported in Table 8.

We have created a graphical depiction of the variables and their relation to one another using multidimensional scaling. The result is shown in Fig. 1. The cluster analysis shows that citation indicators that are from different offices (the 'application count' indicators) are significantly different: the corresponding USPTO, EPO and PCT indicators are all grouped into different clusters. This indicates that, when using indicators from USPTO, EPO and PCT sources only, one is relying on different information.

Correcting for patent family substantially increases compatibility. The indicators that are based on the DOCDB family are grouped into only two clusters (clusters DOCDB A and DOCDB B) that appear close to each other (see Fig. 1). It is interesting to note that the USPTO DOCDB family indicators are clustered together with the overall family indicators. This is understandable given the large number of citations that originate from the USPTO system. Finally, we see that the INPADOC indicators are all grouped together in one cluster (cluster INPADOC). Therefore, we conclude that correcting for the INPADOC patent family results in more similar information across patent systems.

Robustness tests

We performed several robustness tests to verify the results of the correlation analysis under different assumptions and settings. These tests were performed both on the level of the individual sources of the applications (EPO, USPTO and PCT) and the combined set, unless otherwise indicated.

Table 7 Correlations between equal indicators derived from different sources

	Application count	Application count within office	DOCDB				INPADOC			
			Cited family count	Cited family count within office	Full family count	Full family count within office	Cited family count	Cited family count within office	Full family count	Full family count within office
EPO–USPTO	0.12	0.09	1	0.71	1	0.75	1	0.80	1	0.83
EPO–PCT	0.11	0.04	1	0.91	1	0.91	1	0.91	1	0.91
USPTO–PCT	0.30	0.20	1	0.78	1	0.81	1	0.93	1	0.95

These correlations were calculated on the basis of 388,512 DOCDB families and are significant at the 0.001 level

Table 8 Result of clustering the patent citation indicators

Source	Family	Indicator	Cluster	R^2 within cluster	R^2 closest cluster
ALL	INPADOC	Cited family count	INPADOC	0.9636	0.3586
ALL	INPADOC	Full family count	INPADOC	0.9758	0.3918
EPO	INPADOC	Cited family count within office	INPADOC	0.789	0.7103
EPO	INPADOC	Full family count within office	INPADOC	0.7857	0.7261
PCT	INPADOC	Cited family count within office	INPADOC	0.9923	0.4306
PCT	INPADOC	Full family count within office	INPADOC	0.9948	0.4448
USPTO	INPADOC	Cited family count within office	INPADOC	0.9352	0.3164
USPTO	INPADOC	Full family count within office	INPADOC	0.9545	0.3606
EPO	DOCDB	Cited family count within office	DOCDB B	0.9795	0.4549
EPO	DOCDB	Full family count within office	DOCDB B	0.9816	0.4747
PCT	DOCDB	Cited family count within office	DOCDB B	0.9808	0.599
PCT	DOCDB	Full family count within office	DOCDB B	0.9805	0.602
PCT	N/A	Application count	PCT	0.9486	0.203
PCT	N/A	Application count within office	PCT	0.9486	0.2062
USPTO	N/A	Application count	USPTO	0.9998	0.2108
USPTO	N/A	Application count within office	USPTO	0.9998	0.2041
EPO	N/A	Application count	EPO	0.9536	0.2817
EPO	N/A	Application count within office	EPO	0.9536	0.2909
ALL	DOCDB	Cited family count	DOCDB A	0.9891	0.6409
ALL	DOCDB	Full family count	DOCDB A	0.9804	0.6734
USPTO	DOCDB	Cited family count within office	DOCDB A	0.9737	0.5847
USPTO	DOCDB	Full family count within office	DOCDB A	0.993	0.5187

Using a full factor analysis

We performed a full factor analysis on the indicators. We used the principal component method and rotated the solution using the Quartimax algorithm, since this is the most capable method of assigning indicators to different factors. This led to five factors with an eigenvalue larger than 1. We grouped indicators that had loadings higher than 0.5. on the same factor. This analysis resulted in similar conclusions to the cluster analysis: all indicators that relate to patent applications are grouped according to office. However, the family indicators were grouped differently: there was one factor that had all family related indicators, with the exception of the EPO and the PCT DOCDB indicators, which were grouped separately. Thus, a factor analysis groups clusters 1 and 6. We can, therefore, derive the same conclusions as in the cluster analysis: patent citation indicators that relate to equal applications are different from each other, especially when they are related to applications from different patent offices. Family indicators are more similar, but the difference between DOCDB and INPADOC indicators remains present.

Inclusion of uncited applications

In our main analysis, we excluded patent applications that had zero citations on any indicator. This was carried out in order to improve the precision of the analysis. When we

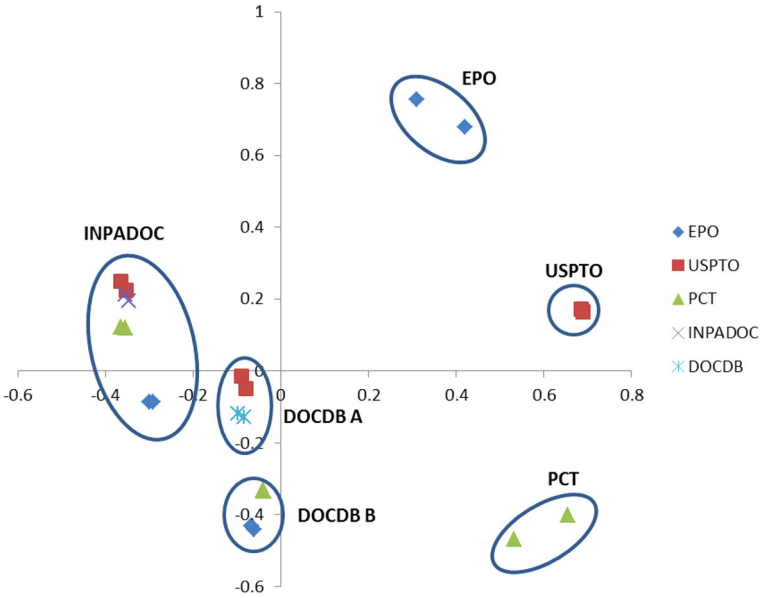

Fig. 1 Depiction of the differences between citation indicators on a 2D plane by multidimensional scaling. The dissimilarity between indicators, as defined by $1 - R^2$, is represented by the distance between them. Cluster names are related to clusters as described in Table 8

included the uncited applications, we found that the correlation of the different indicators increased slightly. However, this increase was small and equally distributed across the different correlation coefficients between the citation indicators. Consequently, we conclude that the inclusion of applications with zero citations does not substantially change the conclusions of the preceding section.

Using only granted applications

The main analysis of the paper pooled different kinds of patent application. It could be that the citation patterns of applications leading to a grant are different from those of other applications. Since granted patent applications are more valuable, researchers could opt to use only those in their analysis. Hence, it is important to determine if our results hold when only considering granted applications.

Patent applications that follow the PCT route cannot be granted (as PCT documents), since the WO is not a patent office with a territory over which it exercises patent grants. Since we only used granted patent applications from the USPTO, the USPTO indicators will not be affected by this step. Therefore, the analysis will only affect the EPO patent applications. For the overall analysis, we included the PCT and USPTO documents to derive a close comparison with the main analysis.

Using only granted applications from the EPO does not substantially change the correlation between the different indicators. Correlations between indicators on EPO and USPTO documents varied little with the main analysis. This then resulted in the same clusters being returned by the cluster analysis. Nor were the inter-office correlations substantially different. Thus, we conclude that our findings remain similar when including only granted applications.

Using log citations instead of normal citations

Many researchers include not the raw patent citation count but rather the logarithm of the citation count to account for the skewed distribution of patent citations. Therefore, we have also computed the indicators using the following transformation:

$$I^* = \ln(I + 1)$$

whereby I is any of our citation indicators and I^* is its transformed form. We have computed correlations between all transformed indicators.

This transformation yields indicators that are more similar to each other. This is because the difference between low and extremely high scores is diminished. Hence, all correlations are substantially improved. This leads the clustering algorithm to select fewer groups. In particular, all DOCDB indicators are now grouped together. All other groups are equal. So, we conclude that, even though the log transformation improves the correlations, this improvement is not sufficient to remove any significant differences that we found in the main analysis.

Using only patent data from before 2000

The main analysis was performed on patent data that cover the time period 1980–2011. Consequently, there are numerous patents that have not yet received (all of their) citations. Since different patent systems may well experience different time lags, this could create a difference in citation data that is due to these time lags, as opposed to an inherent difference in information. In order to control for a potential time lag effect, we repeated the correlation analysis using only patent applications that were filed before 2000. For our complete analysis, we only compared patent families from which at least one patent in each office had a filing date before 2000.

We find that indicators for patents filed before 2000 behave in a similar, albeit not identical, way to the main analysis. The major difference is that the correlations between family-based indicators, most notably those based on INPADOC, increase substantially. This was most pronounced when we computed the full correlation matrix over the three sources of patent data. Because of this, the cluster solution was altered with a reduced number of clusters: one large cluster with all family based indicators, thereby combining clusters INPADOC, DOCDB A and DOCDB B from the main analysis; and three small clusters with application counts from each office, equal to clusters EPO, USPTO and PCT from the main analysis. Consequently, we can conclude that family-based indicators are more similar in this sample, while non-family-based indicators remain very different from each other and from the family-based indicators.

Highly cited patents

Set-up of the analysis

We identified the groups of highly cited patents according to two different criteria: the top 100 patents in terms of citations received, and patents that score more than 5 standard deviations (SD) above the mean number of citations of all patents under study.[15] Highly

[15] The size of the groups of highly cited patents identified by the 5 SD outlier criterion varies between 765 and 35,145 depending on the source office and indicator specification.

cited patents were identified, reflecting the unit of analysis of the respective indicators (patent application, DOCDB patent family, INPADOC patent family).

The effects of expanding the sources of citing patents and correcting for patent family

The main observation from the analysis is that commonality between sets of highly cited patents, identified via different indicators, is rather low, whether one considers the top 100 cited patents or patents receiving 5 standard deviations more citation on average.

Table 9 reports the results obtained in calculating how many identical patent applications are identified when adopting different choices with respect to calculating citations. The reference group consists each time of the patent documents identified by applying the 'application count within office' indicator: citations to the focal document within the patent system of the focal document.

From Table 9, we can derive several conclusions: first, we observe that 5 standard deviation outliers of indicators are in general more similar than the top 100 scores. Second, the table resembles the pattern in Table 8: we observe low levels of overlap for EPO and PCT documents while, for USPTO documents, the overlap is consistently higher. Third, we again observe that both the correction for citation origin and the correction for family have a considerable effect on the indicators. In the case of the EPO and the PCT, we find that the patents identified in the top 100 of the 'application count within office' indicator and those identified by the family corrected indicators hardly overlap.

Even though the commonality improves for the 5 SD outlier and for the USPTO indicators, we conclude that the differences are non-trivial. Differences are larger for INPADOC than for DOCDB indicators.

Table 9 Qualified communalities between the 'application count within office' indicator and other indicators from the same office

Family	Indicator	EPO		USPTO		PCT	
		Top 100	5 SD	Top 100	5 SD	Top 100	5 SD
N/A	Application count	0.31	0.52	0.89	0.94	0.37	0.52
N/A	Application count within office	1	1	1	1	1	1
DOCDB	Cited family count	0.04	0.18	0.76	0.83	0.06	0.16
DOCDB	Cited family count within office	0.31	0.55	0.81	0.89	0.40	0.54
DOCDB	Full family count	0.05	0.18	0.75	0.82	0.06	0.15
DOCDB	Full family count within office	0.28	0.55	0.80	1.00	0.38	0.54
INPADOC	Cited family count	0.04	0.18	0.30	0.72	0.07	0.19
INPADOC	Cited family count within office	0.20	0.45	0.35	0.78	0.18	0.41
INPADOC	Full family count	0.04	0.18	0.28	0.66	0.06	0.17
INPADOC	Full family count within office	0.20	0.45	0.29	0.71	0.16	0.40

Fractions are computed as the amount of overlap divided by the maximum amount of possible overlap. Top 100 refers to the 100 most cited patents and 5 SD refers to patents present in the 5 standard deviation outlier of the distribution

Table 10 Comparison between indicators at different offices

Family	Indicator	USPTO–EPO		USPTO–PCT		EPO–PCT	
		Top 100	5 SD outlier	Top 100	5 SD outlier	Top 100	5 SD outlier
N/A	Application count	0.02	0.09	0.03	0.07	0.00	0.02
N/A	Application count within office	0.02	0.08	0.00	0.03	0.00	0.01
DOCDB	Cited family count	0.48	0.72	0.34	0.49	0.57	0.54
DOCDB	Cited family count within office	0.08	0.21	0.16	0.19	0.14	0.19
DOCDB	Full family	0.45	0.70	0.30	0.46	0.56	0.53
DOCDB	Full family count within office	0.09	0.24	0.16	0.24	0.14	0.19
INPADOC	Cited family count	0.88	0.99	0.78	0.92	0.84	0.74
INPADOC	Cited family count within office	0.40	0.37	0.43	0.41	0.45	0.33
INPADOC	Full family	0.85	0.99	0.76	0.87	0.85	0.73
INPADOC	Full family count within office	0.35	0.36	0.43	0.40	0.44	0.34

Commonality measures were computed by dividing the number of common members of highly cited groups by the maximum number of common members possible

The effect of using different sources of patent data

In this analysis, we focused on comparing similar indicators from each office with each other. Table 10 presents the result of this analysis. It is important to note that there are two mechanisms by which a highly cited patent does not appear in another patent system. It could be because its family members did not receive a sufficient number of citations, or because it did not have family members present in the other patent system.

In concordance with the results from the previous analysis, we see that using the top 100 rank criterion results in a similar overlap pattern as using the 5 SD outlier criterion. However, the qualified overlap scores are generally lower when using the top 100 rank criterion. Overlaps between indicators that score applications on the citations they receive from within their own offices are very low. This is only slightly improved when citations from other offices are included (moving from 'application count within office' to 'application count' yields, at best, an increase of 3 % for the top 100).

The use of citation indicators that correct for families drastically increases overlap scores between offices. While the use of DOCDB corrected indicators results in qualified overlaps of around 50 %, the highest overlap scores are obtained when INPADOC family corrected citation indicators, which use all citations, are used.

Conclusion

We set out to determine the (dis)similarity between different citation indicators. We achieved this by computing a set of commonly and less commonly used citation indicators and comparing them with one another. We relied on correlation and cluster analysis to assess (dis)similarities; in addition, we examined which highly cited patents were identified by different indicators. The results showed substantial dissimilarities between the various patent citation indicators.

The correlation and cluster analysis demonstrated that there are large differences in the information revealed by patent citations, depending on which indicator is used. First, a significant effect was present when comparing indicators that use citation information from all offices versus indicators that only use 'within office' citations. Second, indicators computed over different entities (patent application, DOCDB patent family, INPADOC patent family) display only modest levels of commonality. Finally, these effects are most pronounced for EPO and PCT patents. The USPTO indicators tend to be more similar, except when the INPADOC family is corrected for.

Cluster analysis revealed distinctive clusters for each office. Most family corrected indicators, whether they encompass all citations or not, were grouped in clusters reflecting the family definition. Only the indicators based on the DOCDB patent family definition were split into two clusters. Therefore, we conclude that patent citation indicators based on families are more comparable to each other, even when information from only one office is used. This conclusion remains robust under all tests that were performed.

The analysis of highly cited patents provides a similar picture. Correction for the family and the citation origin results in significant effects and leads to larger commonality between different indicators. Commonality is higher when adhering to the indicator reflecting '5 standard deviation' outliers compared to relying on the indicator consisting of the 100 most cited patents. The only indicator resulting in almost complete congruence pertains to the INPADOC corrected indicators.

Since this paper has established clear differences between different citation indicators, it may inspire additional research on the underlying drivers of these differences. Future efforts should be made to examine the origins of these differences. Are they fully explained by different practices in the different offices or do they indicate a separated impact from the regions over which these offices grant patents? A similar effort should be focused on the family indicators. While it appears that they give unbiased information of the global impact of an innovation, this may not be completely true: Family indicators correlate more with USPTO indicators than with their EPO or PCT counterparts. We suggest that this could be due to the higher number of citations that are present in the USPTO system, thus biasing the family indicators towards the greater importance of citation activity in the US. Therefore, efforts could be undertaken to examine the magnitude of this possible bias and, if necessary, derive an unbiased global patent citation indicator. Finally, the INPADOC patent family definition could be further investigated: while the DOCDB definition is clear and often used, this is not the case for the INPADOC patent family definition.

The observation that different indicators display low levels of commonality implies that choices with respect to citation indicators are non-trivial. As a result, we suggest researchers become more aware and explicit in deciding which citation indicator to use. This choice should be ultimately be guided by the underlying research question. At the same time, our results may also inspire further research into assessing the consistency of results obtained when deploying different citation indicators. If the intention is to strive for an indicator that is not sensitive to design choices, the INPADOC corrected indicator is clearly the prime candidate since it implies commonality approaching 100 %.

Acknowledgments Lin Zhang acknowledges the support of the National Natural Science Foundation of China Grant No. 71103064.

Appendix 1: Correlation between indicators from the same office

See Tables 11, 12 and 13.

Table 11 Correlation of indicators of patents filed at the EPO

	Family	Indicator	1	2	3	4	5	6	7	8	9	10
1	N/A	Application count	1.00									
2	N/A	Application count within office	0.79	1.00								
3	DOCDB	Cited family count	0.40	0.34	1.00							
4	DOCDB	Cited family count within office	0.51	0.64	0.66	1.00						
5	DOCDB	Full family count	0.39	0.33	0.99	0.65	1.00					
6	DOCDB	Full family count within office	0.52	0.65	0.66	0.99	0.65	1.00				
7	INPADOC	Cited family count	0.12	0.09	0.36	0.26	0.35	0.25	1.00			
8	INPADOC	Cited family count within office	0.17	0.20	0.28	0.39	0.28	0.38	0.88	1.00		
9	INPADOC	Full family count	0.14	0.12	0.40	0.30	0.40	0.30	0.91	0.77	1.00	
10	INPADOC	Full family count within office	0.23	0.26	0.35	0.47	0.35	0.48	0.81	0.89	0.87	1.00

Table 12 Correlation of indicators of patents filed at the USPTO

	Family	Indicator	1	2	3	4	5	6	7	8	9	10
1	N/A	Application count	1.00									
2	N/A	Application count within office	0.99	1.00								
3	DOCDB	Cited family count	0.85	0.84	1.00							
4	DOCDB	Cited family count within office	0.85	0.86	0.99	1.00						
5	DOCDB	Full family	0.84	0.84	0.99	0.98	1.00					
6	DOCDB	Full family count within office	0.85	0.86	0.98	0.99	0.99	1.00				
7	INPADOC	Cited family count	0.23	0.23	0.30	0.31	0.30	0.30	1.00			
8	INPADOC	Cited family count within office	0.25	0.25	0.31	0.32	0.31	0.32	0.99	1.00		
9	INPADOC	Full family count	0.25	0.25	0.33	0.33	0.33	0.33	0.95	0.94	1.00	
10	INPADOC	Full family count within office	0.27	0.28	0.34	0.35	0.35	0.35	0.94	0.95	0.99	1.00

Table 13 Correlation of indicators of patents filed at the PCT

	Family	Indicator	1	2	3	4	5	6	7	8	9	10
1	N/A	Application count	1.00									
2	N/A	Application count within office	0.77	1.00								
3	DOCDB	Cited family count	0.52	0.35	1.00							
4	DOCDB	Cited family count within office	0.61	0.72	0.69	1.00						
5	DOCDB	Full family count	0.49	0.34	0.99	0.68	1.00					
6	DOCDB	Full family count within office	0.61	0.72	0.69	1.00	0.68	1.00				
7	INPADOC	Cited family count	0.21	0.14	0.29	0.23	0.29	0.23	1.00			
8	INPADOC	Cited family count within office	0.20	0.19	0.20	0.28	0.20	0.28	0.88	1.00		
9	INPADOC	Full family count	0.22	0.16	0.32	0.26	0.32	0.26	0.93	0.82	1.00	
10	INPADOC	Full family count within office	0.22	0.22	0.23	0.31	0.23	0.31	0.84	0.94	0.88	1.00

Appendix 2: Correlation between indicators from different offices

See Tables 14, 15 and 16.

Table 14 Correlation coefficients of indicators pertaining to patents filed both at the EPO (columns) and the USPTO (rows)

Office	Indicator number	Family	Office Indicator	EPO 1	EPO 2	EPO 3	EPO 4	EPO 5	EPO 6
USPTO	1	N/A	Application count	0.12	0.09	0.27	0.28	0.17	0.18
USPTO	2	N/A	Application count within office	0.12	0.09	0.27	0.27	0.17	0.18
USPTO	3	DOCDB	Cited family count within office	0.11	0.06	0.71	0.72	0.81	0.82
USPTO	4	DOCDB	Full family count within office	0.12	0.07	0.74	0.75	0.82	0.83
USPTO	5	INPADOC	Cited family count within office	0.01	0.00	0.39	0.40	0.80	0.80
USPTO	6	INPADOC	Full family count within office	0.02	0.00	0.43	0.45	0.82	0.83

Table 15 Correlation coefficients of indicators pertaining to patents filed both at the EPO (columns) and the PCT (rows)

Office	Indicator number	Family	Office Indicator	EPO 1	EPO 2	EPO 3	EPO 4	EPO 5	EPO 6
PCT	1	N/A	Application count	0.11	0.07	0.47	0.47	0.33	0.33
PCT	2	N/A	Application count within office	0.07	0.04	0.34	0.33	0.22	0.22

Table 15 continued

Office	Indicator number	Family	Office Indicator	EPO 1	EPO 2	EPO 3	EPO 4	EPO 5	EPO 6
PCT	3	DOCDB	Cited family count within office	0.16	0.10	0.91	0.91	0.82	0.82
PCT	4	DOCDB	Full family count within office	0.16	0.10	0.91	0.91	0.82	0.82
PCT	5	INPADOC	Cited family count within office	0.03	0.01	0.54	0.55	0.91	0.91
PCT	6	INPADOC	Full family count within office	0.04	0.01	0.55	0.56	0.92	0.91

Table 16 Correlation coefficients of indicators pertaining to patents filed both at the USPTO (columns) and the PCT (rows)

Office	Indicator number	Family	Office Indicator	US PTO 1	US PTO 2	US PTO 3	US PTO 4	US PTO 5	US PTO 6
PCT	1	N/A	Application count	0.30	0.29	0.37	0.38	0.11	0.13
PCT	2	N/A	Application count within office	0.20	0.19	0.22	0.23	0.04	0.06
PCT	3	DOCDB	Cited family count within office	0.31	0.30	0.78	0.81	0.48	0.52
PCT	4	DOCDB	Full family count within office	0.31	0.30	0.78	0.81	0.48	0.52
PCT	5	INPADOC	Cited family count within office	0.10	0.10	0.78	0.76	0.93	0.94
PCT	6	INPADOC	Full family count within office	0.10	0.11	0.79	0.78	0.94	0.95

Appendix 3: Variable cluster method

This appendix explains the cluster algorithm that was used to cluster indicators. This method is an implementation of the VARCLUS procedure in the SAS® software package (SAS Institute 2008). What follows are excerpts from the SAS manual (SAS Institute 2008: 7461–7463) explaining the logic of the underlying procedure. Our specific settings are detailed in italics. Options not related to our analysis have been omitted.

'The VARCLUS procedure divides a set of numeric variables into disjoint or hierarchical clusters. Associated with each cluster is a linear combination of the variables in the cluster. The linear combination used here consists of the first principal component. (…) The first principal component is a weighted average of the variables that explains as much variance as possible.

(…)

The VARCLUS procedure tries to maximize the variance that is explained by the cluster components, summed over all the clusters. The cluster components are oblique, not orthogonal, even when the cluster components are first principal components. In an ordinary principal component analysis, all components are computed from the same

variables, and the first principal component is orthogonal to the second principal component and to every other principal component. In the VARCLUS procedure, each cluster component is computed from a different set of variables than all the other cluster components. The first principal component of one cluster might be correlated with the first principal component of another cluster. Hence, the VARCLUS algorithm is a type of oblique component analysis.

We use the correlation matrices as input for the principal component analysis used in the VARCLUS procedure (…)

The VARCLUS algorithm is both divisive and iterative. By default, the VARCLUS procedure begins with all variables in a single cluster. It then repeats the following steps:

1. A cluster is chosen for splitting. Depending on (…) the largest eigenvalue associated with the second principal component (…)
2. The chosen cluster is split into two clusters by finding the first two principal components, performing an orthoblique rotation (raw quartimax rotation on the eigenvectors; Harris and Kaiser 1964), and assigning each variable to the rotated component with which it has the higher squared correlation.
3. Variables are iteratively reassigned to clusters to try to maximize the variance accounted for by the cluster components.

(…)VARCLUS stops splitting when every cluster has only one eigenvalue greater than one, thus satisfying the most popular criterion for determining the sufficiency of a single underlying dimension.'

References

Ahuja, G., & Lampert, C. Morris. (2001). Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal, 22*(6–7), 521–543.

Albert, M. B., Avery, D., Narin, F., & McAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy, 20*(3), 251–259.

Albrecht, M. A., Bosma, R., van Dinter, T., Ernst, J. L., van Ginkel, K., & Versloot-Spoelstra, F. (2010). Quality assurance in the EPO patent information resource. *World Patent Information, 32*(4), 279–286.

Alcácer, J., & Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics, 88*(4), 774–779.

Arts, S., Appio, F., & Van Looy, B. (2012). Inventions shaping technological trajectories: Do existing patent indicators provide a comprehensive picture? *Scientometrics, 97*(2), 397–419.

Azagra-Caro, J. M., Mattsson, P., & Perruchas, F. (2011). Smoothing the lies: The distinctive effects of patent characteristics on examiner and applicant citations. *Journal of the American Society for Information Science and Technology, 62*(9), 1727–1740.

Baumol, W. J. (2004). *Education for innovation: entrepreneurial breakthroughs vs. corporate incremental improvements* (No. w10578). National Bureau of Economic Research.

Carpenter, M. P., Narin, F., & Woolf, P. (1981). Citation rates to technologically important patents. *World Patent Information, 3*(4), 160–163.

Chakrabarti, A. K. (1991). Competition in high technology: Analysis of patents of US, Japan, UK, France, West Germany, and Canada. *Engineering Management, IEEE Transactions on, 38*(1), 78–84.

Criscuolo, P. (2006). The 'home advantage' effect and patent families. A comparison of OECD triadic patents, the USPTO and the EPO. *Scientometrics, 66*(1), 23–41.

Criscuolo, P., & Verspagen, B. (2008). Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. *Research Policy, 37*(10), 1892–1908.

Dosi, G. (1982). Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change. *Research policy, 11*(3), 147–162.

EPO. (2014). What's in a name? *Patent Information News, 4*, 4.

Frietsch, R., Neuhäusler, P., Jung, T., & Van Looy, B. (2014). Patent indicators for macroeconomic growth—The value of patents estimated by export volume. *Technovation, 34*(9), 546–558.
Gambardella, A., Harhoff, D., & Verspagen, B. (2008). The value of European patents. *European Management Review, 5*(2), 69–84.
Graham, S., & Harhoff, D. (2006). Can post-grant reviews improve patent system design? A twin study of European and US patents. *CEPR Discussion Paper* No. 5680, CEPR London.
Hagedoorn, J., & Cloodt, M. (2003). Measuring innovative performance: Is there an advantage in using multiple indicators? *Research Policy, 32*(8), 1365–1379.
Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). *The NBER patent citation data file: Lessons, insights and methodological tools* (No. w8498). National Bureau of Economic Research.
Hall, B. H., Jaffe, A. B, & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics*,*36*(1), 16–38.
Harhoff, D., & Reitzig, M. (2004). Determinants of opposition against EPO patent grants—The case of biotechnology and pharmaceuticals. *International Journal of Industrial Organization, 22*(4), 443–480.
Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy,* 32(8), 1343-1363.
Harris, C. W., & Kaiser, H. F. (1964). Oblique factor analytic solutions by orthogonal transformations. *Psychometrika, 29*(4), 347–362.
Jaffe, A. B., Trajtenberg, M., & Fogarty, M. S. (2000). Knowledge spillovers and patent citations: Evidence from a survey of inventors. *The American Economic Review, 90*(2), 215–218.
Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics, 108*(3), 577–598.
Karki, M. M. S. (1997). Patent citation analysis: A policy analysis tool. *World Patent Information, 19*(4), 269–272.
Lanjouw, J. O., & Schankerman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal, 114*(495), 441–465.
Lingua, D. G. (2005). INPADOC: 30 years of endeavours yet unmapped territories remain! *World Patent Information, 27*(2), 105–111.
MacGarvie, M. (2006). Do firms learn from international trade? *Review of Economics and Statistics, 88*(1), 46–60.
Magerman, T., Van Looy, B., & Debackere, K. (2011). *In search of anticommons: Patent-paper pairs in biotechnology. An analysis of citation flows*. MSI FEB Working paper, KU Leuven.
Martínez, C. (2011). Patent families: When do different definitions really matter? *Scientometrics, 86*(1), 39–63.
Michel, J., & Bettels, B. (2001). Patent citation analysis. A closer look at the basic input data from patent search reports. *Scientometrics, 51*(1), 185–201.
Nakamura, H., Suzuki, S., Kajikawa, Y., & Osawa, M. (2015). The effect of patent family information in patent citation network analysis: A comparative case study in the drivetrain domain. *Scientometrics, 104*(2), 437–452.
Narin, F., Noma, E., & Perry, R. (1987). Patents as indicators of corporate technological strength. *Research Policy, 16*(2), 143–155.
Neuhäusler, P., & Frietsch, R. (2012). Patent families as macro level patent value indicators: Applying weights to account for market differences. *Scientometrics*, *96*(1), 1–23.
Neuhäusler, P., Frietsch, R., Schubert, T., & Blind, K. (2011). *Patents and the financial performance of firms—An analysis based on stock market data* (No. 28). Fraunhofer ISI discussion papers innovation systems and policy analysis.
Paci, R., & Usai, S. (2009). Knowledge flows across European regions. *The Annals of Regional Science, 43*(3), 669–690.
Sampat, B. N. (2004). *Examining patent examination: an analysis of examiner and applicant generated prior art* (Doctoral dissertation, University of Michigan).
SAS Institute Inc. (2008). *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.
Schoenmakers, W., & Duysters, G. (2010). The technological origins of radical inventions. *Research Policy, 39*(8), 1051–1059.
Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *The RAND Journal of Economics*,*12*(1), 172–187.
van Pottelsberghe de la Potterie, B., & François, D. (2009). The cost factor in patent systems. *Journal of Industry, Competition and Trade, 9*(4), 329–355.
Webb, C., Dernis, H., Harhoff, D., & Hoisl, K. (2005). *Analysing European and international patent citations: A set of EPO patent database building blocks*, OCDE Science. Technology and Industry Working Paper, 9.