Paper to be presented at the

DRUID Society Conference 2014, CBS, Copenhagen, June 16-18

Patent Citation Indicators: One Size Fits All?

Jurrien Bakker
KU LEUVEN
Department of Managerial Economics, Strategy and Innovation
Jurrien.Bakker@kuleuven.be

Dennis Verhoeven
KU Leuven
Department of Managerial Economics, Strategy and Innovation
dennis.verhoeven@kuleuven.be

Lin Zhang
KU Leuven
Department of Managerial Economics, Strategy and Innovation
Lin.Zhang@kuleuven.be

Bart Van Looy
KU Leuven
Department of Managerial Economics, Strategy and Innovation
bart.vanlooy@kuleuven.be

Abstract

The number of citations that a patent receives is considered as an important indicator of the quality and impact of the patent. However, different methods and data sources are used to calculate this measure. This paper evaluates the similarities between citation indicators obtained when counting within different patent offices (USPTO, EPO and patent applications filed through the PCT). Additionally we discuss the effects of using indicators that correct for patent family and the effects of including citations from more than one source. Our findings reveal that the obtained citation indicators differ substantially. This is confirmed when comparing sets of highly cited patents identified by these different indicators. Correcting for patent families remedies this situation considerably, at least if a broad definition (INPADOC) of families is being adopted. Our findings suggest that favoring one way of calculating a citation indicator over another, has non-trivial consequences and hence should be considered explicitly. If one aims to discard differences arising from source choices (both in terms of cited and citing documents), patent citation indicators based on INPADOC families present themselves

as the preferred option.

Jelcodes:O34,-

Patent Citation Indicators: One Size Fits All?

ABSTRACT

The number of citations that a patent receives is considered as an important indicator of the quality and impact of the patent. However, different methods and data sources are used to calculate this measure. This paper evaluates the similarities between citation indicators obtained when counting within different patent offices (USPTO, EPO and patent applications filed through the PCT). Additionally we discuss the effects of using indicators that correct for patent family and the effects of including citations from more than one source. Our findings reveal that the obtained citation indicators differ substantially. This is confirmed when comparing sets of highly cited patents identified by these different indicators. Correcting for patent families remedies this situation considerably, at least if a broad definition (INPADOC) of families is being adopted. Our findings suggest that favoring one way of calculating a citation indicator over another, has non-trivial consequences and hence should be considered explicitly. If one aims to discard differences arising from source choices (both in terms of cited and citing documents), patent citation indicators based on INPADOC families present themselves as the preferred option.

INTRODUCTION

The number of times that patents are cited by other patents[1] can be used to complement the mere count of patented inventions. The idea of using patent citations as an indicator is relatively old and seems to originate from Seidel in 1949 (Karki, 1997). However, the first systematic empirical investigations only emerged in the 1980's, with Carpenter et al. (1981) revealing that patents related to industry awards are cited more frequently.

A patent can be cited for different reasons: to qualify novelty, inventive step, industrial relevance or to provide additional – relevant - information to situate the claims advanced in the patent document. Patents that become cited (more often) are considered to be more important and valuable than patents which are used not (or less) to qualify subsequent technological activity. Therefore one can approximate the importance of an individual patent by the number of times it is cited. This argument is empirically supported by the works of and Gambardella et al.. (2008) and Trajtenberg (1990) who show that patent citations correlate significantly with the value of the individual patent. Likewise, Hall et al. (2005), Narin and Noma (1987) and Neuhausler et al. (2011) 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, Neuhausler and Fritsch (2012) show that forward patent citation counts are well correlated with the export volume.

While (front page) patent references are ultimately included by examiners, a number of researchers conceived citations as an approximation of knowledge flows: Hall et al., 2005; Jaffe et al., 1993,2000; Mcgarvie, 2006 and Paci & Usai, 2009). When adopting this perspective, the

[1] Often referred to as patent citations, forward citations or patent citation count

number of received patent citations then indicates the future influence or impact of the knowledge implied within the patented invention.

A major advantage of using patent citations as an indicator of inventive quality pertains to the availability and relative simplicity of the measure: one simply counts 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 innovative performance of countries (e.g. Chakrabarti, 1991; Criscuolo and Verspagen, 2008; Neuhausler & Fritsch, 2012).

While these, and related studies, point at the relevance of counting the citations to 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, which in turn result in a variety of possible citation indicators. The first decision is to choose the data source from which to compile patent citations as patent systems are geographically bounded (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 to include citations to the focal set of patents. Finally, given the possible existence of multiple patent documents pertaining to a single invention, one can opt for treating equivalent patent documents as one patent family which will also affect citation counts. Currently, one observes three different approaches to these decisions in the literature.

The National Bureau of Economic Research (NBER) initiated a data platform that entails only patents filed at the United States Patent and Trademark Office (USPTO). This data has been

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

available in good detail and reliability as early as 2001 (Hall et al. 2001). Additionally, the first analyses on patent citations also rely on data from USPTO documents (e.g. Carpenter et al. 1981, Narin and Noma, 1987). The NBER database is still in large use as the high number of recent citations to the source paper from Hall et al. (2001)[3] suggests.

A second set of studies have been conducted by using EPO patent documents. European patent data is remarkably different from USPTO data: the EPO patents cover a different geographic area; they are heterogeneous in terms of countries they actually are being 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). This lead to researchers using EPO based patent citations (e.g. Harhoff and Reitzig, 2004; Neuhausler et al. 2011; Schoenmakers and Duysters, 2010).

Finally, some researchers have opted to use data stemming not just from a single source (patent office), but rather take into account the presence of patent families (and hence consider the equivalents of an invention that is present in multiple patent systems when calculating citations). This seems especially appropriate to correct for a home bias (Criscuolo, 2006) and to provide a more encompassing view on 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 Neuhausler and Fritsch (2012).

The general assumption when calculating a patent citation count is that counts from different methods will yield, in general, the same information. However, this might not necessarily be the case: citations from patents from different offices might reveal 'national' impact, rather than providing a 'global' quality indication. Additionally, patent offices have more patents from inventors and applicants from their own geographical location, resulting in a 'home

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

bias' (Criscuolo, 2006). Finally, offices, and hence examiner's 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 the situation whereby citation indicators – resulting from different computational choices - do not reflect the same information (Alcácer and Gittelman, 2006). We would thus argue that if one would assess the inventive performance of a nation or a firm, the results of that analysis could depend on the method used for computing citations.

Therefore it would be warranted to assess the effects of 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 to what extent different methods yield (dis)similar results. Therefore we pose our research question as follows:

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

Research in innovation has since long distinguished between technological improvements which include smaller, incremental, steps forward on the one hand, and inventions entailing large shifts in what is technologically possible (Baumol, 2004; Dosi, 1982). Accordingly, researchers have operationalized these 'breakthrough' inventions by identifying patents receiving exceptionally high numbers of forward citations (e.g. Ahuja and Lampert, 2001, Chakrabarti, 1991; Schoenmakers and Duysters, 2010). Since citation counts might depend on computational choices, it becomes especially interesting 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 will answer our first question with correlation and cluster analyses of the citation counts of patent applications. To answer the second research question we compute the degree of overlap observed between patents that are identified as breakthrough by various methods.

Within the next section, we discuss systematically the different computational choice resulting in a set of indicators which will be compared in this study. Next we will present the obtained empirical findings and discuss their implications. Overall, our findings signal non-trivial differences among the variety of approaches that can be envisaged.

OVERVIEW OF METHODOLOGICAL CHOICES WHEN COMPUTING PATENT CITATION INDICATORS

Patent citation counts are subject to choices made by the researcher. Researchers will need to weigh different options as to which indicator they use to approximate patent quality by its citations. In this section we will discuss the general choices that need to be made when counting forward citations.

The patent office

The patent system in which the patent resides might 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. The home bias, as discussed in the introduction, implies that patent examiners cite more prior art, already present within the own jurisdiction (Criscuolo, 2006). In addition, patent systems - while to a large extent similar in terms of subject matter and procedures – still differ in a number of ways. In terms of subject matter, the USPTO allows the patenting of business methods, software and biological advances such as living organisms while this is not the case

within EPO. The cost of patenting is also different, resulting in fewer EPO patents (van Pottelsberghe de la Potterie and François, 2009). Additionally, practices such as the 'duty of candor'[4] in the US, lead to an increase in references that are being included in patent documents which might impact 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 within the same patent office (e.g. EPO, USPTO, PCT) or to include citations from patents present within other patent systems. The reason that this distinction is worthwhile to investigate is twofold.

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

Second, most documents tend to cite primarily patent documents from within their 'system', due to the examining process (Michel and Bettels, 2001). This is not unexpected as patent examiners should be mainly concerned with the validity of the application in their own system. At the same time, when additional procedures (between patent systems) are in place, differences can be become more outspoken. The case of USPTO is very interesting in this respect. When applying to the USPTO, applicants have a so-called duty of candor where they are required to disclose any knowledge of prior art to the examiner even if this information would

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

contribute to a disqualification of the application. Patent examiners then select among these references and/or add other references deemed relevant. However, USPTO examiners are most familiar with USPTO patents. In case foreign applicants apply, references stemming from prior art situated outside the American patent system might be advanced relatively more by such applicants. Sampat (2004) indeed observed that in about 70% of the patents references to foreign patents are initially advanced by the applicant (see also 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 in so-called patent families. It makes sense to correct citations for the presence of families as other patents can reference towards multiple family members and necessarily the initial, focal application. If the researcher feels that such a citation is equally as valuable as a direct citation of the initial patent application, then a correction by introducing the patent family seems appropriate. In general this involves adding citations to family members to the citation count of the focal application itself.

There are different definitions of the patent family, in this paper we will consider two of them. Martinez (2011) defines these as an extended patent family (INPADOC) and an examiners technology based family (DOCDB). The DOCDB definition centers on finding as close as possible equivalents of a patent document in other offices. These documents are usually characterized by having the same priority applications.[5] The INPADOC definition is, in general, less strict and is used to find documents protecting the same invention, thus also including documents with different priority applications (Albrecht et al., 2010). The members of

[5] 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 were excluded due to a "discrepancy in the priority picture" Albrecht et al (2010: 283)

INPADOC patent families share priority applications with at least one other member of the family. However, it is possible that two members of the same family do not share any priority applications. This can happen if they both share a priority application with a third member of the family. Because of this definition the INPADOC family members are not necessarily equivalents, but are still connected through similar claims in their applications.

We decided to include these two definitions of the patent family because of two reasons: First, they cover two important aspects of the patent family. These are the identification of documents with similar technical content and the identification of the same invention in different documents. Second, they are the most easily available: they are provided in the EPO PATSTAT database and are therefore the most frequently used by researchers.

Indicators

We chose to perform four different permutations to calculate our indicators. These will be based on patent origin, citation origin and a twofold family correction (see above). We have chosen for these permutations because we believe they are representative for virtually all possible permutations that researchers will encounter when working with patent citations. We will shortly explain them in this section.

Starting from the patent origin, we will compare indicators resulting from three different data sources: EPO, USPTO and applications filed through the PCT-route. We use these data because the vast majority of publications dealing with patent citations use indicators based on these sources. Next, we will distinguish two groups of indicators based on the source of the citation: This will be done by comparing the number of citations that were received from

applications within the office of the focal application, and the number of citations that were received irrespective of the patent office[6].

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. Again this measure of patent citations is also calculated within the office of the application and outside its office. As for the patent family definition: we will compare both the INPADOC and DOCDB definitions.

It is possible that a number of citations originate from applications that are part of the same patent family. One could argue that these citations are in fact duplicates as the patent is cited twice by the same invention. This could then create a bias towards citations received from larger patent families, since the size of the family inherently increases the probability of two or more of its members citing the same patent. Therefore, as a final, fourth, permutation, we correct for this by counting not the number of patent applications, but rather the number of patent families that cite the focal family.

This leads to a total of 10 different indicators for each office: 2 indicators based on the application, 4 indicators based on the DOCDB family and 4 indicators based on the INPADOC family. To keep the indicators tractable we have provided names for each indicator, these are listed in table 1.

Insert Table 1 about here
---

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

DATA AND METHODS

Data used

We used patent data from the October 2011 version of PATSTAT. From this data we extracted indicators for patent applications belonging to the EPO and USPTO, as well as applications that were filed through the PCT route. We chose these applications for three reasons: First, most research that uses patent citation data uses patents from at least one of these three systems (or routes in case of PCT applications). Second, the data provided by these offices from the USPTO and EPO are relatively complete within PATSTAT, as compared to other offices (also included in PATSTAT). In the remainder of the paper we shall refer to different origins by denoting 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, wrong conversions of patent numbers and several issues caused by the changes in the USPTO system in 2001[7]. In addition we have only taken USPTO applications that were granted. This is due to the observation that USPTO applications are not completely covered by PATSTAT.

After the cleaning 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. Due to the aforementioned cleaning efforts a large number of applications were removed consisting of 3,319,894 (mainly because no granted equivalent was (yet) present) applications from the USPTO; 10,567 applications from the EPO and 11,335 PCT applications.

[7] 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.

The citation indicators where calculated using all citation data available in the 2011 October version of PATSTAT. We only excluded artificial applications[8]. Therefore the cited applications implied more cleaning than the citing applications. This was done because we wanted to keep the citation indicators as close as possible to those that are obtained when using currently available databases (notably PATSTAT). Therefore we did not correct all recently known issues that exist in patent citation indicators.[9]

Descriptive statistics

Insert Table 2 about here

From the descriptive statistics as listed in table 2, we can derive two main conclusions. The first is that a large number of patents receive at least 1 citation. However, the rate of patents with a nonzero citation count varies considerably. The rate varies from 20% (EPO simple in count) to 88% (USPTO INPADOC family count). Therefore the distribution of the citation indicator varies from highly truncated, i.e. having a zero or non-zero citation count, to a more continuous spectrum.

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 as shown in table 3. This table shows the following: 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, the USPTO also

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

[9] 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 (2006) and Neuhausler et al.(2011)

supplies most citations to other documents. There are more USPTO citations to EPO documents than EPO citations to these documents. A similar pattern emerges for PCT documents.

This changes when patent families, instead of single applications, are introduced. Correction for patent family makes the indicators become more similar. While this implies that the home bias is reduced substantially for all offices, one observes an overall dominance of the USPTO documents as they count for the most citations overall (and these USPTO citations are now becoming included in the family corrected indicators). In the case of EPO, INPADOC families with an EPO member receive 6.4 times more citations from USPTO documents than from EPO documents.

Notice that the large majority of all citations stem from either USPTO, EPO or PCT documents (80 and 97% for patent documents depending on the source; 76-93% for DOCDB families and 80-94% for INPADOC families respectively). It is interesting to note 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 are applications that were filed at only a single national office instead of the EPO due to costs of the EPO process (as noted by van Pottelsberghe de la Potterie and François (2009)).

Insert Table 3 about here

Patent families

In this paper we deploy two different family definitions: the DOCDB and the INPADOC definition. We have compiled some descriptive statistics to understand the effects of correcting for patent family. These statistics are shown in table 4. In this table we see that there exists a large number of patent families in the database. Note that even though these families need to

consist of 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 percent are consisting of only one patent application. Most patent families have at least 2 or more members. We can also see that these effects are larger for families based on the INPADOC definition than for those that are based on the DOCDB definition. Finally we see that a large number of patent families are equal for either family definition, even after excluding the singleton families (which are equal by definition).

Insert Table 4 about here

RESULTS OF THE CORRELATION ANALYSIS

This section describes how patent citations vary between indicators. We are specifically interested in the effects of citation origin, correction for patent family and the comparison between patent offices.

To construct the citation indicators we only use 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 level of the DOCDB or the INPADOC family level. This still allows for other indicators to have a score of 0. This was done in order to assess better the information that was contained in the citation counts. The effects of this are quite substantial as - depending on the office[10] - a considerable share of patents in our sample have no citations, resulting in identical scores (0) for all indicators. We feel that the

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

inclusion of applications that are never cited will have an inflating effect on the correlation and is therefore undesirable.

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 simple in count indicator with all other indicators within 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 'simple in count', while the EPO studies use the EPO 'simple in count'. The results of this exercise are presented in table 5. The full correlation table can be found in appendix A.

Insert Table 5 about here

Table 5 reveals that there is a substantial effect of citation origin (EPO/USPTO/PCT) on the patent citation indicators. This can be seen when inspecting the correlation with the simple count indicator. This effect is more outspoken for EPO and PCT indicators than for their USPTO equivalent, which appears almost not affected. Given the citation information that was presented in table 3, it is likely that most of the citations from outside the EPO and PCT systems are of USPTO origin. This effect is less pronounced in the USPTO system, since that has proportionately much fewer citations from outside.

We also find that correcting for patent family introduces considerable differences. The effects of this correction are stronger in the EPO and PCT systems than in the USPTO system: where the USPTO simple in count has a correlation of 0.84 with the DOCDB family corrected indicator, the equivalent correlations for EPO and PCT are only 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.

Therefore we can conclude that both the decision of using citations of all offices and the decision of correcting for patent family have substantial effects on the patent citation indicators. Even though these effects are more pronounced in the EPO and PCT systems they are also to a large extend present in the USPTO system.

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

For an interoffice comparison, we calculated the correlation for DOCDB patent families from which applications were filed at EPO, USPTO and through the PCT route. This was done because the DOCDB family is based on 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 becomes feasible.

Again we only considered patents which had at least one citation in their family. Even though we linked applications solely on the DOCDB family definition, the citation criterion was applied to the larger of INPADOC and DOCDB definitions. However, we found that all DOCDB patent families with applications in all three offices had at least 1 citation. Therefore this restriction did not change the analysis. These considerations lead to the comparison of citation indicators for 388.512 DOCDB families. The full correlation matrix is listed in appendix B. Here we have extracted the correlations that compare the different sources of patent data. These are listed in table 6.

Insert Table 6 about here

Table 6 shows that correlations for the basic indicators between different offices are very low. The correlation between the EPO 'simple in count' and the USPTO 'simple in count' is only 0.09. Using citations from outside the office of the focal application ('simple count') remedies this slightly by raising the correlation towards levels ranging between 0.11 and 0.30.

Correlations observed when correcting for the DOCDB or INPADOC family are considerably higher. This is naturally the case for the DOCDB family cited count since the applications are all part of the same family. INPADOC family indicators also have coefficients near 1. This is due to the fact that applications in the same DOCDB family are often in the same INPADOC family, since the INPADOC family is the larger family. Interestingly correcting for patent family increases compatibility, even when only citations from within the office of the focal application are counted. Therefore, when there is only application data from one patent office, correcting for the patent family of the focal applications is an interesting method to increase 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 B, i.e. pertaining to patent documents which have equivalents in all different systems under study. To define clusters, we performed a divisive cluster analysis, based on factor analysis (see appendix C for a technical description). Since the correlation table is linked by the DOCDB family definition, the indicators DOCDB family cited and the DOCDB full family count are equal for all three sources. Therefore they are replaced by the general indicator. This is also done for the corresponding INPADOC family indicators as the majority of DOCDB family members are part of the same INPADOC family, which was observed in table 6. Including all INPADOC indicators would thus be redundant. The resulting indicators are denoted

by the 'ALL DOCDB/INPADOC CITED/FULL' notation. The identified clusters are reported in table 7.

Insert Table 7 about here

We have created a graphical depiction of the variables and their relation to one another using multidimensional scaling. The result is shown in figure 1. The cluster analysis shows that citation indicators that are from different offices (the simple 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 increases the compatibility substantially. The indicators that are based on the DOCDB family are clustered into only two clusters (clusters 2 & 6) that appear to close to each other (see figure 1). It is interesting to note that the USPTO DOCDB 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 1). Therefore we conclude that correcting for the INPADOC patent family results in similar information across patent systems.

Insert Figure 1 about here

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 applications (EPO, USPTO and PCT) and the combined set unless otherwise indicated. The obtained results will be discussed shortly here.

Using a full factor analysis. We also performed a full factor analysis on the indicators. We used the principal component method and rotated the solution using the Quartimax algorithm, since that is most capable of grouping indicators into different factors. This lead to 5 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 as the cluster analysis: all indicators that relate to patent applications are grouped according to office. However, the family indicators were grouped differently: there was a factor that had all family related indicators, with the exception of the EPO and PCT DOCDB indicators, which were grouped separately. Thus a factor analysis groups clusters 1 and 6. Therefore we can derive the same conclusions as from 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 zero citations. In our main analysis, we excluded patent applications that had zero citations on any indicator. This was done in order to improve the precision of the analysis. When we included the zero applications, we found that the correlation of the different indicators increased slightly. However there was not a substantial increase in correlations across the different indicators. Therefore 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 applications. 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 only use them in their analysis. Therefore it is important to determine if our results hold on the set of granted applications.

Patent applications that follow the PCT route cannot be granted (as PCT document), since the WO is not a patent office with a territory to grant patents over. Since we only used granted patent applications from the USPTO, the USPTO indicators will also 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 EPO does not change the correlation between the different indicators substantially. 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. Also the inter office correlations were not substantially different. Therefore we conclude that our findings remain when examining 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 after using the following transformation:

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

Whereby I is any of our citation indicators and I^* is the transformed form of it. We have computed correlations between all transformed indicators.

This transformation does yield indicators that are more similar to each other. This is because the difference between low and extremely high scores is diminished. Therefore all correlations were significantly improved. This leads the clustering algorithm to select fewer

groups. In particular, all DOCDB indicators are now grouped together. All other groups are equal. Therefore 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. As such there are numerous patents that have not yet received (all of their) citations. Since different patent systems might experience different time lags, this could create a difference in citation data that is due to these time lags and not due 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 that were 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 compute the full correlation matrix over all three sources of patent data. Because of this the cluster solution was altered with a reduced number of clusters: 1 large cluster with all family based indicators, thereby combining clusters 1,2 and 6 from the main analysis; and 3 small clusters with simple counts from each office, equal to clusters 3,4 and 5 from the main analysis. Therefore 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: 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[11]. Highly cited patents were identified reflecting the unit of analysis of the respective indicators (patent document, 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 communality between sets of highly cited patents, identified via different indicators, is rather low, whether one considers top 100 cited patents or patents receiving 5 standard deviations more citation as average.

Table 8 reports the results obtained when calculating how many identical patent documents 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 'simple in count' indicator: citations towards the focal document within the patent system of the focal document.

Insert Table 8 about here

From table 8 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 of table 7: we observe low levels of overlap for EPO and PCT, while for USPTO documents, the overlap is consistently higher. Third, we again observe that both the

[11] The size of the groups of highly cited patents identified by the 5 sd. outlier criterion varies between 765 and 35145 depending on the source office and indicator specification.

correction for citation origin and the correction for family have a considerable effect on the indicators. In the case of EPO and PCT we find that the patents identified in top 100 of the simple in count indicator and those identified by the family corrected indicators are very different.

Even though the communality improves for the 5 sd. outlier and for the USPTO indicators, we can conclude that the indicators remain substantially different. Consistent with our findings at the correlation analysis, this effect is larger for INPADOC than for DOCDB indicators. This increased effect of the INPADOC correction is also explicitly present in USPTO indicators.

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 depicts the result of this analysis. It is important to note that there are two mechanisms by which a highly cited patent does not appear within 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.

Insert Table 9 about here

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 also

included (moving from Simple in Count to Simple Count yields at best an increase of 3% for the top100).

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 50%, the highest overlap scores are obtained when INPADOC family corrected citation indicators, that use all citations, are used.

CONCLUSION

We set out to determine the (dis)similarity between different citation indicators. We did 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 family, INPADOC family) display only modest levels of communality. Finally, these effects are most outspoken for EPO and PCT patents. The USPTO indicators tend to be more similar, except when INPADOC family corrected indicators are being introduced.

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 DOCDB definition was split into two clusters. Therefore we conclude

that patent citation indicators based on families are more comparable to each other, even when only information from 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 communality between different indicators. Communality 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.

As this paper has established that there are clear differences between different citation indicators, it might inspire additional research on the underlying drivers for these differences. Future efforts should be made to examine the origins of these differences. Are they fully explained by different practices at 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 might 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 a higher importance of citation activity within the US. Therefore efforts should 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 more 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 communality, implies that choices with respect to citation indicators become non-trivial. Therefore we suggest researchers to become more conscious and explicit when deciding which citation indicator to use. This choice should be ultimately guided by the underlying research question. At the same time, our results might also inspire further research which assesses the consistency of results obtained when deploying different citation indicators. If one would strive for an indicator which is not sensitive with respect to design choices, INPADOC corrected indicators present themselves as the prime candidate as they imply communality approaching 100%.

REFERENCES

Ahuja, G., & Morris Lampert, C. 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.

Campbell, R. S., & Nieves, A. L. 1979. Technology Indicators Based on Patent Data: The Case of Catalytic Converters: Phase I Report, Design and Demonstration. **Battelle Pacific Northwest Laboratories**.

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.
Gambardella, A., Harhoff, D., & Verspagen, B. 2008. The value of European patents. **European Management Review**, 5(2), 69-84.
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**, 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. 2006. Citations, Family Size, Opposition and the Value of Patent Rights'. **International Library of Critical Writings in Economics**, 197(2), 256.
Harris, C. W., & Kaiser, H. F. 1964. Oblique factor analytic solutions by orthogonal transformations. **Psychometrika**, 29(4), 347-362.
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.
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.
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.
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.
Narin, F., Noma, E., & Perry, R. 1987. Patents as indicators of corporate technological strength. **Research Policy**, 16(2), 143-155.
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.
Neuhäusler, P., & Frietsch, R. 2012. Patent families as macro level patent value indicators: applying weights to account for market differences. **Scientometrics**, 1-23
Paci, R., & Usai, S. 2009. Knowledge flows across European regions. **The Annals of Regional Science**, 43(3), 669-690.

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.
Sampat, B. N. 2004. Examining patent examination: an analysis of examiner and applicant generated prior art (Doctoral dissertation, University of Michigan).
SAS institute. 2009. **SAS/STAT(R) 9.2 User's Guide, Second Edition**
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**, 172-187.
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.

APPENDIX A: CORRELATION BETWEEN INDICATORS OF THE SAME OFFICE

Insert Table A1 about here

Insert Table A2 about here

Insert Table A3 about here

APPENDIX B: CORRELATION BETWEEN INDICATORS OF DIFFERENT OFFICES

Insert Table B1 about here

Insert Table B2 about here

Insert Table B3 about here

APPENDIX C: 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, 2009). What follows are excerpts from the SAS manual (SAS Institute, 2009: 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 (…)

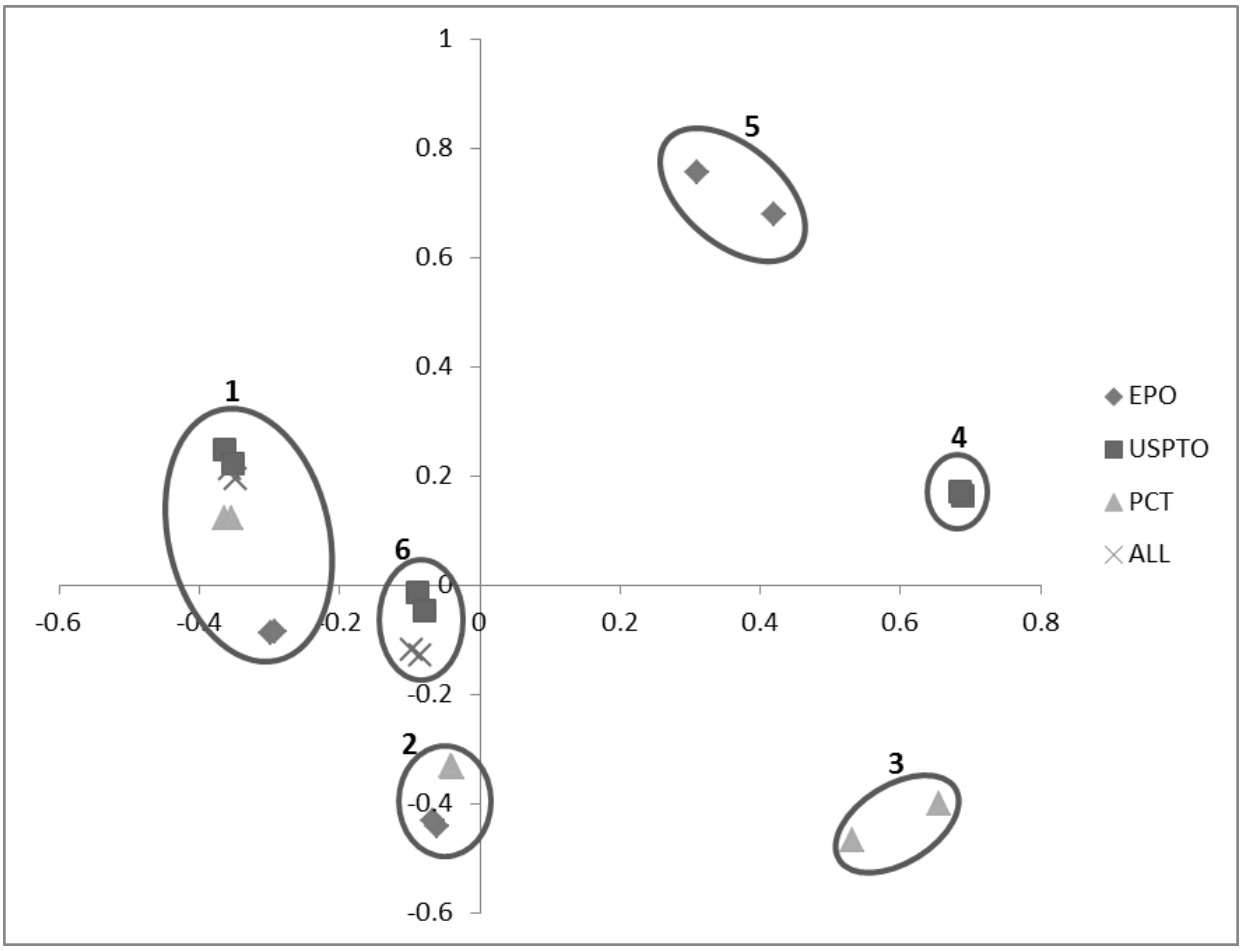
The VARCLUS algorithm is both divisive and iterative. By default, the VARCLUS procedure begins with all variables in a single cluster. It then repeats the following steps:

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. ”

FIGURES

Figure 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. Clusters numbers are related to clusters as described in table 7.



TABLES

Table 1: Indicators and their definitions. These indicators are calculated for focal applications at the EPO, USPTO and PCT.

Patent Family	Patent citation indicator	Definition
N/A	Simple count	Number of citations a patent application receives from all other patent applications, irrespective of publication office.
N/A	Simple in count	Number of citations a patent application receives from all other patent applications within its own publication office.
DOCDB	Family cited	Number of citations the DOCDB patent family of the focal application receives from all other patent applications, irrespective of publication office.
DOCDB	Family in cited	Number of citations the DOCDB patent family of the focal application receives from other patents in the publication office of 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 in count	Number of citations the DOCDB patent family of the focal patent receives from other applications, corrected for DOCDB patent family, in the publication office of the focal application
INPADOC	Family cited	Number of citations the INPADOC patent family of the focal application receives from all other applications, irrespective of publication office.
INPADOC	Family in cited	Number of citations the INPADOC patent family of the focal application receives from other patents in the publication office of 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 in count	Number of citations the INPADOC patent family of the focal patent receives from other applications, corrected for INPADOC patent family, in the publication office of the focal application

Table 2: 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	Standard deviation	Median	Nonzero
EPO	N/A	Simple count	2.343.707	1.92	5.10	0	38%
EPO	N/A	Simple in count	2.343.707	0.57	1.55	0	25%
EPO	DOCDB	Family cited	2.343.707	9.03	20.88	3	75%
EPO	DOCDB	Family in cited	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 in count	2.343.707	1.03	2.33	0	41%
EPO	INPADOC	Family cited	2.343.707	17.05	84.56	4	79%
EPO	INPADOC	Family in cited	2.343.707	1.76	8.37	0	45%
EPO	INPADOC	Full Family count	2.343.707	11.21	47.77	3	79%
EPO	INPADOC	Full Family in count	2.343.707	1.58	6.43	0	45%
USPTO	N/A	Simple count	4.397.304	9.91	18.22	5	82%
USPTO	N/A	Simple in count	4.397.304	8.46	16.35	4	79%
USPTO	DOCDB	Family cited	4.397.304	13.05	24.66	6	86%
USPTO	DOCDB	Family in cited	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 in count	4.397.304	8.97	17.31	4	82%
USPTO	INPADOC	Family cited	4.397.304	25.95	129.50	8	88%
USPTO	INPADOC	Family in cited	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 in count	4.397.304	13.54	58.94	5	84%
PCT	N/A	Simple count	1.917.261	1.90	5.63	0	41%
PCT	N/A	Simple in count	1.917.261	0.58	1.55	0	27%
PCT	DOCDB	Family cited	1.917.261	5.73	16.38	1	59%
PCT	DOCDB	Family in cited	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 in count	1.917.261	1.09	2.46	0	41%
PCT	INPADOC	Family cited	1.917.261	13.22	87.36	2	63%
PCT	INPADOC	Family in cited	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 in count	1.917.261	2.31	14.11	0	46%

Table 3: Origin and destination of citations. 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.

Family correction	Focal office	EPO	USPTO	PCT	EPO (National office)[12]	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

Table 4: Statistics of INPADOC and DOCDB families in our applications

Family	Number of families	% Singletons[13]	Average number of members	Overlap between both family definitions	% Overlap[14]	% Overlap[15]
INPADOC	5,309,452	21%	2.64	4,179,052	79%	73%
DOCDB	6,017,825	35%	2.01	4,179,052	69%	63%

[12] Patent offices which are located on the geographical area that is covered by the EPO.
[13] Families with only one member
[14] Including singletons
[15] Excluding singletons

Table 5: Correlation with the simple in count indicator for each office.
All correlations are significant at the 0.001 level.

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

Table 6: Correlations between equal indicators derived of different sources. These correlations were calculated on the basis of 388512 DOCDB families and are significant at the 0.001 level.

	Simple count	Simple in count	DOCDB				INPADOC			
			Family Cited	Family In cited	Full Family	Full Family in	Family Cited	Family In cited	Full Family	Full Family in
EPO-USPTO	0.12	0.09	1	0.71	1	0.75	1.00	0.80	1.00	0.83
EPO-PCT	0.11	0.04	1	0.91	1	0.91	1.00	0.91	1.00	0.91
USPTO-PCT	0.30	0.20	1	0.78	1	0.81	1.00	0.93	1.00	0.95

Table 7: Result of clustering the patent citation indicators.

source	Family	Indicator	cluster	R^2 within cluster	R^2 closest Cluster
ALL	INPADOC	CITED	1	0.9636	0.3586
ALL	INPADOC	FULL	1	0.9758	0.3918
EPO	INPADOC	Family in cited	1	0.789	0.7103
EPO	INPADOC	Full Family in count	1	0.7857	0.7261
PCT	INPADOC	Family in cited	1	0.9923	0.4306
PCT	INPADOC	Full Family in count	1	0.9948	0.4448
USPTO	INPADOC	Family in cited	1	0.9352	0.3164
USPTO	INPADOC	Full Family in count	1	0.9545	0.3606
EPO	DOCDB	Family in cited	2	0.9795	0.4549
EPO	DOCDB	Full Family in count	2	0.9816	0.4747
PCT	DOCDB	Family in cited	2	0.9808	0.599
PCT	DOCDB	Full Family in count	2	0.9805	0.602
PCT	N/A	Simple count	3	0.9486	0.203
PCT	N/A	Simple in count	3	0.9486	0.2062
USPTO	N/A	Simple count	4	0.9998	0.2108
USPTO	N/A	Simple in count	4	0.9998	0.2041
EPO	N/A	Simple count	5	0.9536	0.2817
EPO	N/A	Simple in count	5	0.9536	0.2909
ALL	DOCDB	CITED	6	0.9891	0.6409
ALL	DOCDB	FULL	6	0.9804	0.6734
USPTO	DOCDB	Family in cited	6	0.9737	0.5847
USPTO	DOCDB	Full Family in count	6	0.993	0.5187

Table 8: Qualified communalities between the simple in count indicator and other indicators from the same office. 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.

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

Table 9: Comparison between indicators at different offices. Communality measures were computed by dividing the number of common members of highly cited groups by the maximum number of common members possible.

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	Simple count	0.02	0.09	0.03	0.07	0.00	0.02
N/A	Simple in count	0.02	0.08	0.00	0.03	0.00	0.01
DOCDB	Family cited	0.48	0.72	0.34	0.49	0.57	0.54
DOCDB	Family in cited	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 in	0.09	0.24	0.16	0.24	0.14	0.19
INPADOC	Family cited	0.88	0.99	0.78	0.92	0.84	0.74
INPADOC	Family in cited	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 in	0.35	0.36	0.43	0.40	0.44	0.34

Table A1: Correlation of indicators of patents filed at the EPO.

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

Table A2: Correlation of indicators of patents filed at the USPTO

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

Table A3: Correlation of indicators of patents filed at the PCT

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

Table B1: Correlation of indicators of patents filed at the PCT

	Family	Indicator	1	2	3	4	5	6
1	N/A	Simple count	0.12	0.09	0.27	0.28	0.17	0.18
2	N/A	Simple in count	0.12	0.09	0.27	0.27	0.17	0.18
3	DOCDB	Family in cited	0.11	0.06	0.71	0.72	0.81	0.82
4	DOCDB	Full Family in count	0.12	0.07	0.74	0.75	0.82	0.83
5	INPADOC	Family in cited	0.01	0.00	0.39	0.40	0.80	0.80
6	INPADOC	Full Family in count	0.02	0.00	0.43	0.45	0.82	0.83

Table B2: Correlation of indicators of patents filed at the PCT

	Family	Indicator	1	2	3	4	5	6
1	N/A	Simple count	0.11	0.07	0.47	0.47	0.33	0.33
2	N/A	Simple in count	0.07	0.04	0.34	0.33	0.22	0.22
3	DOCDB	Family in cited	0.16	0.10	0.91	0.91	0.82	0.82
4	DOCDB	Full Family in count	0.16	0.10	0.91	0.91	0.82	0.82
5	INPADOC	Family in cited	0.03	0.01	0.54	0.55	0.91	0.91
6	INPADOC	Full Family in count	0.04	0.01	0.55	0.56	0.92	0.91

Table B3: Correlation of indicators of patents filed at the PCT

	Family	Indicator	1	2	3	4	5	6
1	N/A	Simple count	0.30	0.29	0.37	0.38	0.11	0.13
2	N/A	Simple in count	0.20	0.19	0.22	0.23	0.04	0.06
3	DOCDB	Family in cited	0.31	0.30	0.78	0.81	0.48	0.52
4	DOCDB	Full Family in count	0.31	0.30	0.78	0.81	0.48	0.52
5	INPADOC	Family in cited	0.10	0.10	0.78	0.76	0.93	0.94
6	INPADOC	Full Family in count	0.10	0.11	0.79	0.78	0.94	0.95