Paper to be presented at

DRUID15, Rome, June 15-17, 2015

(Coorganized with LUISS)

A critical analysis of patent citations as a measure of patent value

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

Bart Van Looy
K.U. Leuven
Managerial Economics, Strategy and Innovation / INCENTIM-ECO
Bart.VanLooy@kuleuven.be

Abstract

The empirical community has been enamored with patent citations for a long time. In this paper we evaluate this relationship by examining the efficacy of the patent citation indicator to explain patent value as measured by patent renewal. We use this evaluation to test several qualities of the patent citation indicator.
First, we find that not all patent citation indicators are equally as effective in explaining patent renewal.

The indicator based on the DOCDB patent family appears as an alternative to citations solely based on the EPO and USPTO patent system. Second, we determine that patent citations display a logarithmic relation with patent value. Third, the analysis shows that the origin of the patent citation is of little relevance and that patent citations that directly involve the focal patent are more important than those to family members . We find that the best indicators are based on the DOCDB patent family and include citation information from multiple patent offices. Fourth, we show that self-citations explain patent renewal better than other citations. Fifth, we find that applicant generated citations provide more explanatory value, than examiner generated citations and sixth, we find that citation windows invariably lead to a substantial alteration of the citation indicator.

Finally and foremost, we observe that despite the infatuation of the econometric community with the indicator, patent citations are poor at best to explain patent renewal. We therefore conclude that better indicators of patent quality are needed

Jelcodes:O34,O32

A (re-)evaluation of the patent citation indicator

Jurriën Bakker[1] and Bart Van Looy[2]

Abstract

The empirical community has been enamored with patent citations for a long time. In this paper we evaluate this relationship by examining the efficacy of the patent citation indicator to explain patent value as measured by patent renewal. We use this evaluation to test several qualities of the patent citation indicator.

First, we find that not all patent citation indicators are equally as effective in explaining patent renewal. The indicator based on the DOCDB patent family appears as an alternative to citations solely based on the EPO and USPTO patent system. Second, we determine that patent citations display a logarithmic relation with patent value. Third, the analysis shows that the origin of the patent citation is of little relevance and that patent citations that directly involve the focal patent are more important than those to family members . We find that the best indicators are based on the DOCDB patent family and include citation information from multiple patent offices. Fourth, we show that self-citations explain patent renewal better than other citations. Fifth, we find that applicant generated citations provide more explanatory value, than examiner generated citations and sixth, we find that citation windows invariably lead to a substantial alteration of the citation indicator.

Finally and foremost, we observe that despite the infatuation of the econometric community with the indicator, patent citations are poor at best to explain patent renewal. We therefore conclude that better indicators of patent quality are needed.

[1] Corresponding author, contact at jurrien.bakker@kuleuven.be
[2] The authors wish to thank ECOOM and Tom Magerman for providing the EPO PATSTAT database in a readily accessible form. The authors also wish to thank Dietmar Harhoff for providing key insights that greatly helped this paper.

Introduction

Patent citation indicators are the most often used indicator to assess the 'value' of a patent. This value of a patent can refer to the private value (Gambardella et al. 2008), the social value (Trajtenberg 1991, Carpenter *et al.* 1981). as well as the knowledge value contained in the innovation (e.g. Hall *et al.*, 2005; Jaffe *et al.*, 1993,2000; Mcgarvie, 2006 and Paci & Usai, 2009) . Patent citation indicators are widely used because of the availability of the data and because it is the first quality indicator that has been validated as early as 1981 (Carpenter *et al.* 1981). In this paper we will examine the validity of this infatuation with patent citations.

Patent citation indicators came into play not only to assess the value of a single patent document, but to estimate the inventive performance of larger entities such as companies, universities and even countries. Patent citation indicators gradually replaced simple patent counts as the former are deemed to reflect mere inventive input(i.e. R&D expenditure) rather than inventive output(Grilliches,1998).

However, there has been growing concern that not all patent citations are equal. (Bakker et al. 2014) found that various widely used citation measures are substantially different from each other. Thus the question becomes, if patent citations are both useful and very different from each other, which indicator should one use to assess the (economic) value of a patent?

In this respect, it can be noted that few validation studies are available and the ones that are, rely on small scale samples. Moreover, the more advanced validation efforts found that citations only explained a marginal fraction of patent value. For example Trajtenberg(1991) used only 10 annual observations , Harhoff et al. (1999) used 192.Furthermore, the few large scale studies (e.g. Gambardella et al.2008, Hall et al. 2005,2007) found significant yet only small effects from the citation indicator. Thus even though patent citations are considered to be a consistent indicator of patent value, the low explanatory value of the indicator is leading to a growing unease at giving it such a prominent place in empirical research.

This paper is geared towards addressing this unease by attempting to validate patent citation indicators using a large scale study using patent renewal data. We will first validate the most commonly used patent citation indicators, second we will examine the exact relation between patent citation indicators and patent value. Finally, we will use the relation to deconstruct the patent citation indicators to understand more of their inner workings .

Patent citations, a critical review

Patent citations are often considered to be one of the more validated indicators. However, it is also known that this indicator tends to carry possible biases. In this section we will discuss both the validation efforts on this indictor as well as the potential problems it faces.

The validation of patent citations

Validation efforts of the patent citation indicator have been done in various ways ,using different dependent variables and by using different sources of patent citation data. Most notably patents are used either from the European Patent Office (EPO) or the United States Patent and Trademark Office (USPTO). Patent citations have been linked to social value, private value as well as inventive value. Table 1 provides an overview of the most important validation efforts, their methods and the results that were found.

Insert Table 1 about here

From this table we can deduce that patent citations are a significant indicator of patent value. In almost all studies, significant results were found. This is an interesting achievement considering that so many different dependent variables as well as estimation methods were used. However, most results were obtained with few other variables considered and then on small datasets and still find . We are not the first authors to note this(e.g. Gay and Le Bas (2005), Gittelman(2008)).

Later and larger validation efforts such as Gambardella et al.(2008) and Hall et al. (2005) still provide a bleak picture: patent citations are still significant but only explain a small part of the variance in the quality indicator, ranging between 2-4% percent. Finally, most validation studies only pick up significant effects for groups of highly cited patents, while patent citations do not appear to perform as well for patents with lower valuation. This raises further questions with regard to the use of the measure to identify value of more run of the mill patents.

The biases in patent citations

Unfortunately, the problems of the citation indicator are not confined to its low efficiency. There are also potential biases present when one uses the indicator. The first and most straightforward bias is that that patents of different years exhibit different citation patterns. This is not only due to a censoring problem, where one does not observe citations because they are outside the data set, but also due to the changing natures of citations, especially since citations are directly related to the number of patents applied after the focal patent has been applied for.

Patent citation indicators also tend to differ between industries and technological fields, as do most other patent indicators(e.g. Lanjouw et al. (1998)). Furthermore lengths of innovation cycles differ for different industries (Marchetti , 1980)., therefore the time in which a patent is eligible to be cited can vary.

Citation counts could also suffer from the phenomenon of preferential attachment in which a patent that is cited more gets an even higher chance of being cited (Hung and Wang, 2010). This brings then two problems: first, the noise in the indicator is related to relatively random occurrences in the time just after the application has been filed; second, this introduces a non-linearity in the relation between citations and quality which is problematic if we proxy quality using a linear indicator of patent citations.

Furthermore, citations are attributed for different reasons. In general there are three reasons for a patent to cite another patent document: to establish prior art, to establish a technological frontier (from which the patent needs to differ substantially) and to provide general back ground[3]. These different reasons to be cited might in turn not all be similarly related to the patent value of the cited patent and could interact with other characteristics such as the time and industry of the applications. In addition, citations might be added by different actors i.e. by applicants or examiners. These citations might behave differently as well as shown by Alcacer and Gittelman (2006) and Criscuolo and Verspagen (2008).

Another source for a potential bias is that not all citations come from the same source. As with academic citations analysis, one must always be vigilant of self-citations. In the case of patents, these are generally perceived as citations that come from patents with the same applicant. They are then not an external indication of the impact of the invention but rather an indication that the invention is being used in an ongoing project (Narin et al.,1987) or that subsequent innovation is built on it (Belenzon, 2012).

Finally, citations might be not only related to the value of the invented product, or even the value of the patent document, but also to the quality of the patent application in itself. After all, citations are established for legal reasons. Therefore better written and more detailed patent documents are preferential as they improve the legal understanding of the citation. However, these better written documents are not necessarily fully indicative of more important inventions.

Methods and data

In order to examine the effect of using different citation indicators we use the different patent citation indicators to explain patent renewal. Patent renewal is often seen as an indicator of value (e.g. Pakes and Schankerman, 1984; Lanjouw et al. 1998, Hegde and Sampat, 2009) as it implies an economic decision of the owner of the patent: Only those patents from which the expected value in the renewal period exceeds the cost of renewal will be renewed.

We choose renewal as a proxy of patent value because it is an indicator that covers virtually all patent applications and bears a close link to private value. Patent citations, which also are theorized to represent private economic value, should therefore be highly efficient at explaining the variance in this indicator. Moreover, if patent citations are indeed an accurate depiction of this value they should be able to explain the decision of renewal(i.e. highly cited patents are often renewed, others less so).

---

[3] However, the background references of a patent are not always included in patent citation databases such as EPO PATSTAT.

Therefore, allowing patent citations to classify patent renewal data allows for two observations: first we are able to perform a large scale validation study thus not only testing weather patent citations provide information on the private value of patents but also if this information is enough to substantially differentiate between valuable and less valuable patents. Second, a large scale study provides a means to compare the efficacy of different patent citation indicators and select those which prove to be superior.

Unfortunately there are limitations to this setup as it does not allow us to test other attributes of patent citations such as the knowledge flows that they could represent, or the social or innovative value contained in a patent. However, we assume that for many patents these quantities are closely related (e.g. a patent with lots of private economic value is also more likely to contain more socially valuable knowledge). Furthermore, it may be argued that renewal decisions are not always taken on the private value of a single patent but are taken on the scope of patent portfolios. However, we would argue that if citations would pertain to a substantial part of economic value then they should be providing a substantial increase in models that explain patent renewal.

There are still a number of data issues that need to be taken into consideration: First renewal data suffers from severe censoring as we cannot observe potential renewal after the maximum time (usually 20 year) of the patent. Also we observe censoring at the end of period of observation with patents that are still in force at this moment. In order to deal with this censoring we opted for binary analyses where we examine if the patent has been renewed up until its maximum lifetime. Thus we run analyses on the fact if a patent has been renewed for its maximum potential lifetime, which is 20 years in the case of EPO patents and 18/20 years (depending on the year of filing) for USPTO patents.

For this paper, we used indicators constructed in similar fashion to Bakker et al.(2014) but derived from the same data source: the October 2013 version of the EPO PATSTAT database. We used this data source not only for information on individual patents but also to extract data on patent families, most notably the DOCDB and INPADOC definitions[4]. We obtained renewal data from the INPADOC_PRS file provided by the 2014 April version of the EPO PATSTAT database. We observed patent renewal for all granted patents of both EPO and USPTO by examining renewal fees paid to the respective office[5]. In the case of EPO patents, we considered a patent renewed if it has been renewed at least one national office, which subscribes to the EPO regional system.

We used data on patents that were applied from 1980 onwards because we are unsure of the accuracy of renewal data of earlier patents[6]. The dataset is naturally limited by the time until a patent can be renewed. Since renewal for a EPO office is a yearly decision, we choose to include patents that are at least 20 years old, thus having application years up until 1993. USPTO renewal only occurs at three

---

[4] The DOCDB family is perceived as a family of technologically equivalent patent documents in different offices, whereas the INPADOC family is (often) larger and is thought to represent the underlying invention better. For exact definitions of these patent families consult Albrecht et al. 2010.

[5] i.e. we observed the last registered renewal payment of the patent at its respective office.

[6] We observed a larger number than expected of never renewed patents for earlier dates. This indicates that some renewal data might be missing, thus making an analysis of these years unreliable.

stages of the patent life time with full renewal decided at 12 years after application. Therefore USPTO patents up until 2001 are included, giving the USPTO analyses more observations than their EPO counterparts.

Because we want to compare indicators across different offices, we only used DOCDB patent families that had members in both EPO and USPTO offices. We further restricted the families to have granted patents in both EPO and USPTO, because we are working with renewal data and therefore only consider granted patent applications. We used basic data cleaning to avoid problems with the USPTO patents due to a rule change in 2001.

We performed the analyses using a basic set of control variables, reflecting any easy to obtain knowledge about the patent portfolio of any patent holder. Therefore we can assess in this analysis the added value of including the patent citation indicator. We included controls for both the year and the technological area (IPC3)[7] in which the patent was filed. Moreover, we also included controls reflecting attributes of the applicant of the patent[8]: the type of applicant and the size of the applicant. We included these dummies because we assume that different applicants likely write different kinds of patent documents, thus affecting their citation rates and also have different evaluation criteria of the renewal decision of a patent. The exact definitions of these controls and all other variables used in this paper are listed in table 2.

Insert table 2 about here

Which patent citation indicator is a good indicator of patent renewal?

The first analysis is centered around the efficacy of different patent citations at explaining patent renewal. As mentioned in the previous sections, there are many different patent citation indicators to be constructed. In this paper we will focus on the most widely used and applied indicators, these are: EPO, USPTO, DOCDB and INPADOC. Their definitions are provided in table 2. We have listed the descriptive statistics of all variables used in this paper in table 3.

Insert table 3 about here

For our main analysis we decided to present a log transform of the indicators for three reasons: first, the log transformation is very often used to counter the skewed distribution of citations and is therefore relatively prominent. Second, the log transformation proves to be better at explaining the variation in

[7] The third highest level of the International Patent Classification system, also known as the class level.
[8] For 7 patents no applicant was identified, these applications were not included in the analysis.

our data than the linear version of the indicators, as is shown in the analysis of the next section. Finally, we tested both the linear and log transformed specifications presented here and found the log transformed variables to be superior at explaining renewal. The results of this analysis are given in table 4 for EPO renewal and table 5 for USPTO renewal.

Insert table 4 about here

Insert Table 5 about here

From this analysis we can see that even though the indicators explain some of the variance, even the better ones do not explain but a few extra percent of recall and precision. This confirms the results found in other large scale studies: A significant indicator that does not explain much variance in the dependent variables.

The results from tables 4 and 5 also allow us to observe differences between the citation indicators. Looking at the log likelihood for each analysis and the pseudo R^2 that is derived from it, we see that for USPTO renewal the USPTO indicator is superior, while for EPO the INPADOC indicator is superior. However, we find that the DOCDB indicator in both cases performs very close to the superior indicator and thus appears to provide a compromise. Furthermore we observe that EPO renewal is less well explained than USPTO renewal with most particular a substantial lower recall for all analyses, which is probably due to the lower incidence of EPO renewal as shown in table 2.

The relation between citations and renewal

In this analysis we will further test the relationship between patents and renewal. We used as a primary citation indicator the DOCDB indicator as it is relatively comparable across offices (see Bakker et al. 2014), furthermore it has a decent correlation with patent renewal at both the USPTO and the EPO as is shown in the first analysis. Finally, we have repeated this analysis with other citation indicators and found no substantial difference in the results.

We performed an analysis where we created a dummy for each level of the DOCDB citation indicator. We took two additional steps: we combined patents that either 0 or 1 citations according to this indicator, since there are too few patents that are not cited. We then set this group as a referenceto compare patents with more citations. We also capped the indicator at a level of 100 citations because the amount of patents with each citation level above 100 was too low to make a meaningful analysis. Therefore we grouped all patents with more than 100 citations in the >100 citations group. The estimates from the logistic analysis are shown in figure 1 and figure 2 for EPO and USPTO. Note that the coefficients are as compared to a level of 0 citations, thus the starting level is set at 0 by definition.

Insert figure 1 about here

Insert figure 2 about here

First, there is a striking similarity in the figures, indicating that the relation between DOCDB citations and both USPTO and EPO renewal is follows the same relation. Second, we observe that the coefficients of the analysis are monotonously increasing until a relatively high number of citations. Thus we can deduct that the citation indicator possesses a substantial amount of information in this region. The relation is less obvious for higher citations. Finally, we seem to observe a logarithmic relation between the level of patent citations and the logistic coefficients. We test this relationship by taking the exponent of these coefficients, which incidentally are also the odds ratios for the patent levels. The results of this exercise are displayed in figures 3 and 4.

Insert figure 3 about here

Insert figure 4 about here

In these figures we observe that the correlation between the odds ratio and the citation score is very strong ($R^2 \approx 0.9$) for renewal at both the EPO and USPTO patent offices. Therefore we conclude that to determine the private value of a patent, the log transformation will yield results that fit very well the relation with patent value. We also observe that the value of citations behaves at low levels as a monotonous function and at high levels not very much so. However, this might also be a consequence of the declining power of the analysis due to there being fewer patents that are cited at an exceptional rate. The reduced power of the analysis can also be observed by the increasing size of the confidence intervals.

The relative importance of different citations

One of the issues plaguing citation based indicators is the weighing of citations. Either one uses citations belonging to a single source, thus setting the weight of all other sources at 0. Or one adds up all the citations indiscriminately, thus setting all weights equal. Unfortunately this does not match well with reality: citations outside the patent system of the focal patent are likely valuable in giving information, but are very likely to contain different information due to differences in patent systems.

Therefore we present an analysis in which we separate citations by citing patent office and cited patent office to determine the optimal weights of citation based indicators to explain patent renewal in our dataset. We can use this to test the hypotheses presented above: that either there is one perfect indicator and all other weights should be 0 or that all weights are equal. Furthermore this analysis allows us to identify the decay of value as citations are drawn further away from the focal document.

Starting from the equation which we used to estimate the previous effects of different citation indicators:

$$P(renewal) = L(C + \beta_{cit}\ln(1 + x_{cit}) + \vec{\beta}_{control}\vec{x}_{control})$$

Here P(renewal) denotes the chances of renewal, *L* represents the logistic transformation, C a constant, x_{cit} the citation indicator and $\vec{x}_{control}$ the vector with control variables. Finally β_{cit} and $\vec{\beta}_{control}$ denote the coefficients for the citation indicator term and the coefficient of the controls. We can write the citation indicator as a linear combination of separate sources of citations:

$$x_{cit} = w_1 Cit_1 + w_2 Cit_2 + \cdots = \vec{w}\overrightarrow{Cit}$$

Here *Cit* represents a count of citations from an identified source and *w* the weight of this source in the overall citation indicator. Thus we transform the previous equation into the following:

$$P(renewal) = L(C + \beta_{cit}\ln(1 + \vec{w}\overrightarrow{Cit}) + \vec{\beta}_{control}\vec{x}_{control})$$

In this equation, we are interested in estimating the weight vector $\vec{w}$as it will give is more insight into what citations matter and how they compare relative to each other. To estimate these weights we ran adapted logistic regressions[9] with a number of different configurations of the origin of citations. We describe these different sources in table 4. We included the same controls as in the analyses before.

It is to be noted that it is not possible to simultaneously estimate β_{cit} and $\vec{w}$. Therefore we ran a normal regression with $\vec{w}$ set as 1 (i.e. the normal regression) and used its estimate of β_{cit} in the analysis to estimate $\vec{w}$. We again used both EPO renewal and USPTO renewal to estimate these coefficients for different specifications. These can be found in table 5. Note that the denoted linear coefficients in this table are part of $C + \vec{\beta}_{control}\vec{x}_{control}$ while the denoted non-linear coefficients are part of $\vec{w}\overrightarrow{Cit}$ in the preceding equation

----

Insert table 6 about here

----

[9] For this we adapted the maximum likelihood function of the logistic regression to include the non-linear term.

Table 5 lacks the estimates for the log likelihood of all citation sources from the USPTO renewal data. The reason for this lack is that the coefficients related to the INPADOC additions were sufficiently negative to create a situation in which for at least some of the observations the non-linear term becomes imaginary. Since we assume that citations will not relate negatively with renewal, we see this as an indication that INPADOC added citations do not significantly add to the explanatory value of the citation indicator to explain patent renewal of USPTO patents.

From table 5 we can derive several conclusions: first, citations that directly involve the focal patent have very similar weights whereas citations that reference patent family members have much lower weights. It is interesting to note that there are few differences in weights attributed to citations derived from the office of the patent and citations from patent applications from other offices. Thus it appears that while it is important if the patent or its equivalent is referenced, it is less important from which office this reference comes.

The drop in relative weight between citations to the focal patent and to its equivalents is quite high, thus disproving the theory that all citations are equal. However, even citations from far away sources are still significant, thus they still give meaning full information. This leads to the conclusion that both hypothesis presented are dismissed: there is no single perfect indicator and not all citations are equal. Therefore we propose that researchers do include citations at least up until the DOCDB family level, but use a discounted rate for this inclusion.

We also observe that the coefficient of the DOCDB correction is negative for USPTO, which indicates that patents that are cited more often by the same DOCDB family have a higher chance of being renewed. This is a result that was unexpected as the DOCDB indicator usually removes these duplicate citations.

Self-citations: to include or not to include

Within bibliographic research the notion that citations from different sources have different value often comes up. The most ubiquitous origin of these citations is when the owner of the document cites the document in one of his other documents, this is then called a self-citation.

Considering the relative value of these self-citations in patents, there are two schools of thought. The first school is derived from the study of scientific citations and believes that self-citations should be corrected for (e.g. Van Raan(1996)). On the other side of the spectrum, it is argued that self-citations do provide a valuable indicator of innovative quality. This is due to the fact that self-citations indicate that the owner (i.e. the applicant) pursued further activity in the area. Therefore the area is apparently promising for the applicant which may be due to the quality of the work that was done before and is represented by the cited patent. Thus patent with many self-citations could indicate building stones for large projects within the firm (Narin et al. 1987, Belenzon, 2012).

To solve the paradox we performed an analysis similar in to the analysis on the different sources of self-citations in which we separate citations that are derived patents with the same applicant and those belonging to other applicants. Since it is difficult to correctly identify equal applicants with slightly different names, we again used the consolidated table of applicant names (Magerman et al. 2009, Peters et al. 2010). The results of this analysis are listed in table 6.

Insert table 7 about here

From table 6 we can draw the following two conclusions: first, we find that self-citations are a much higher indication of patent renewal than citations that come from patents from other applicants. Second, again both EPO renewal and USPTO renewal follow a similar pattern with respect to the citation indicator, thus the results appear robust. Finally, the difference in weights is quite high. Self-citations count for 7 times a normal citation. This might be because we use patent citations to explain a measure of private patent value. Therefore in certain applications, such as using patent citations as a proxy for private value, it would be prudent to give a higher weight to self-citations. However, this may not be necessarily true for all measures of patent value (e.g. social or innovative value) .

Citation origins

Patent citations are often viewed as representing knowledge flows (e.g. Jaffe et al. 1992, Jaffe and Trajtenberg 1999). Yet a substantial number of these citations are not added by the applicants of the invention but by patent examiners. This leads to a situation in which patent inventors are often unaware of the prior art that is cited in their application(e.g. Jaffe et al 1998,2000). Thus patent citations do not always represent citation flows. However, patent examiners may have a more unbiased view of the patent system and therefore might apply more relevant prior art. Therefore citations added by examiners could still represent a value yet different than those added by applicants

Citation practices vary across different patent offices. In particular the USPTO requires that applicants disclose all known prior art when they apply for a patent in what is known as the 'duty of candor'. Other large patent office's such as the EPO and JPO do not have such a requirement, even though they do accept citations to prior art from applicants. This 'duty of candor' has led to a relatively large number of applicant based citations in the USPTO patent system, compared to that of the EPO. Therefore the USPTO indicator is a good indicator to test the differences between examiner and applicant citations.

We again use the setup explained before to test the difference between examiner and applicant added citations. However, we note that the USPTO splits the citations in three: citations from the applicant, citations inserted during the search part of the examination procedure and citations to yet unpublished patents. Since the last two types of citations are inserted by examiners, we grouped them together. We also note that citations can, in principle be duplicated when examiners and applicants cite the same application. The results of this analysis are shown in table 8.

Insert table 8 about here

Another source of differences between the value of patent citations concerns the reason why patents are cited. Examiners usually state in their search report why certain patents are cited. The most frequent reasons are citations used to give a general back ground or to exclude novelty or inventive step. However, there a few other reasons why examiners might mention citations in their search report such as citations suggested by applicants or citations to potentially conflicting documents. Again these citations might be different as a value indicator.

Therefore we decided to use our analysis framework to test these differences. For this we used the EPO indicator as the EPO search reports are the best reported in our database. Moreover, the EPO does not have a tradition of applicants introducing citations and we feel therefore that the EPO citations are probably less biased by these inclusions.

For this analysis we only took into consideration citations that were made by EPO examiners in their search reports. We separated these between several categories that are found in the PATSTAT database, we describe these in table 2. It is again to be noted that some citations fall in several categories, some of which are even then omitted as is explained in the resources provided by the EPO(2013). Therefore patents can receive citations multiple times. The results of this analysis can be found in table 9.

Insert table 9 about here

In this result we find evidence that citations made for different reasons have a different value indication. Again citations made by applicants provide by far the largest indication of value. Citations made to prejudice inventive step appear to rank high but there are not enough of these citations to make a meaningful estimate (a problem several categories have). Finally citations to provide a general background, to prejudice novelty or that are part of a larger group score similar (i.e. are not significantly different). Therefore we conclude that the reasons examiners have to cite patent applications are not very indicative of differences in value and we find that citations introduced by applicants remain the most indicative of the private value of the cited patent.

Citation windows

In the empirical practice, citation windows are often employed to deal with data truncation issues. A citation window indicates that a researcher cuts the possible time that a patent can be cited. This is done to ensure that all patents have an equal time in which citations to them can be observed. By doing this, citation windows prevent truncation issues due to the fact that older patents have much more time for which citations to them can be observed than newer patents.

Citation windows can be justified if one of three conditions holds: First, the omitted number of citations after the citation window is sufficiently small that the citation indicator will not really be affected by their exclusion. Second, even if this number is not small enough, the citation indicator will still be relatively unaffected because the number of citations that is excluded correlates highly with the number of included citations. Third, even if the second condition is not met the citation indicator could be still of use if the excluded citations are relatively unimportant in explaining patent value.

In this section we will test these three conditions for different citation windows. We chose the citation windows of 5 and 10 years. These intervals were selected because the 5 year interval is often used in the literature, while the 10 year interval is the maximum citation window that we can test relatively accurately with our sample and is the largest window that is still somewhat commonly used.

First, the condition that citation frequencies drop of quickly after a citation window. For this we determined the average relative cumulative citation distribution for three indicators: EPO, USPTO and DOCDB[10]. This distribution is shown in figure 5. From this figure we can see determine that patents are still often cited after 5 years. Even after 10 years, patents only have received around 70% the total number of citations in the USPTO and DOCDB indicator. The corresponding number for the EPO indicator(85%) is higher. We believe that missing 15-30% of the total number of citations is still sufficient cause for concern. Therefore we reject the first condition for these indicators.

Insert figure 5 about here

To test the second condition, we calculate the correlations between the counts of the included and excluded citations. This is done for both the 5 year and the 10 year citation window. We do this again for the EPO,USPTO and DOCDB indicators. The results are listed in table 8. The correlations in this table are substantially lower than 1. This means that the second condition does not hold for any of the three indicators.

Insert table 8 about here

Finally we test the third condition: Do citations that are added later to the patent citation indicator still represent a meaningful indication of value? For this we again employed the same framework of analysis as in the analyses of the previous sections. We present the results of the DOCDB indicator as it is the most consistent indicator between the USPTO and EPO system. We also restricted the observations to those that had at least 18 years of citation information(i.e. the same restriction that we apply to EPO

[10] For this indicator the citation time was determined with reference to the application year of the oldest application in the patent family.

renewal observations).   The results of this exercise can be found in table 9. We first tested the contributions of each individual year to the indicator. We found that for both analyses of USPTO and EPO renewal that there is no substantial difference between the relative importance of each year. This indicates that the third condition also does not hold. We tested this also for citation windows of 5 and 10 years and again we did not found a substantial difference between the relative weights strong enough to accept condition three.

Insert table 9 about here

We have found that out of the three conditions which are sufficient to justify citation windows, none is confirmed by the data available. Therefore we conclude that citation windows will always result in a substantial loss of information in the patent citation indicator. Citation windows should, if they are used at all, include as much citation years as possible. Finally we suggest that citation years perhaps can be avoided by including dummies for the (first) application year of the focal patent or patent families in the analysis. This is a viable solution if it can be assumed that the truncation affects the citation counts of patents of the same application year equally.

Conclusion

In this paper we use the large dataset of patent data available through the EPO PATSTAT database to evaluate the relevance of the patent citation indicator. For this we used the available renewal data and evaluated patent citations by their ability to explain the occurrence of patent renewal. We note that we do not establish a causal link between patent citations and renewal in either direction. We then evaluate patent citations based on different patent systems, different citation origins and from different applicants. In addition, we also determined the nature of the relation between patents and renewal.

Our first and foremost conclusion is that patent citation indicators do not provide an adequate picture of the private value of a patent. In line with other studies, we find that even though patent citations correlate significantly with a measure of patent value (i.e. patent renewal) they do not explain most of its variation. In our study, the patent citation indicator, even those of the most effective indicators, only improved marginally the recall and precision of the logistic analyses. Therefore patent citations cannot be seen as an effective proxy much less a good indicator of economic value and by extension innovative value. However, at present there is no indicator of patent quality that can fulfill this role. We would therefore urge researchers who encounter this measure to critically assess its role in their research and look for alternatives.

Our other findings are directed at those researchers that despite this finding still wish to use this indicator, unfortunately often for lack of a better alternative. We find that different patent citation indicators perform substantially different in explaining patent renewal. It appears that patent indicators, solely based on USPTO data are better at explaining patent renewal, while EPO renewal is best explained by family based citation indicator. We would argue that the DOCDB indicator is the most intermediate indicator in that it performs relatively similar to the best indicators for both EPO and USPTO renewal, whereas the USPTO and the INPADOC indicator only perform well on one patent system. Based also in part on the work by Bakker et al. (2014)[11] and the argument of Webb et al. (2005)[12] we would suggest researchers to consider this indicator to observe patent value.

Next, we find that patent citations appear to have a logarithmic relation with patent value. Therefore creating a log transformation of the indicator is not only good to deal with distributional issues (e.g. extreme outliers), it is also good because it more accurately describes the relation with patent value. Furthermore we found that while the relation between value and citations is less clear as the number of citations increases, there is no intrinsic reason to create groups of highly cited patents as it is likely that patents in this category still adhere to this relation with patent value.

Using the findings from the previous analysis we performed an analysis on the relative importance of citation sources. We found that citations that are made directly to the patent are better at explaining patent renewal than citations made to a patent of the same family. Moreover we found that the origin of citations is of little relevance. As for patent references to other documents, we find that patent citations to INPADOC family members do not contribute substantially, while citations to DOCDB family members do contribute but at a severely discounted rate. Finally we observe that citation corrections on the citing side (i.e. discounting families that cite a focal patent or its family members more than once) do not have a significant effect on the explanatory power of the analysis. From this we can draw two conclusions: first: researchers using this measure should try and include references from multiple patent offices as they can provide more information on patent value. Second, the best indicator can be created by not only taking citations directly to the focal patent but also by adding citations to the DOCDB family members but at a discounted rate.

We also used the analysis to determine the efficacy of self-citations. We find that self-citations are better at explaining patent renewal than other citations. However, given that the arguments for self-citations often center around an increased value for the applicant and that patent renewal in theory is also fully governed by the value of the patent to the applicant, we cannot generalize this finding to other definitions of patent quality. However, we feel confident that this result indicates that self-citations should play an important role in determining the value of a patent.

We then investigated the relation between the nature of the citation and its efficacy at explaining patent renewal. Our results show that applicant based citations have a higher weight than examiner based citations, for both EPO and USPTO indicators. This strengthens the argument that applicant citations represent knowledge spillovers, since applicants are better at identifying relatively important patents. However, a part of this result might be explained also by the fact that applicants are more likely to cite their own patents and that we therefore find a part of the effect of self-citations that we

[11] Bakker et al. (2014) show that the DOCDB indicator creates a large conformity of patent citation indicators when considering different citation databases

[12] Webb et al. (2005) argue that citations to equivalent patents should also be counted.

described earlier. We also found that the reason for citing a patent in general appears to have little effect on the explanatory power.

Lastly, we focused our attention on the question of patent citation windows. We stated three sufficient conditions justifying their use. In our analysis we found that not one of these conditions was met. In addition we determined that there is little difference in the importance of patent citations with regard to their timing. Therefore we conclude that citation windows should be used as few as possible.

As a final note, we would like to return to our main finding: patent citations do a poor job in explaining patent value. Yet, patent citations are still the most used indicator to assess this very quality. It may very well be that this is due to a lack of better measures thus creating a necessity to find better measures of patent quality. However, at the moment we have to warn against any interpretation of patent citations as a well-founded proxy of innovative output. Next to being unable to explain adequately a decent measure of patent value, the indicator also suffers from biases as we noted in the review section. When using patent citations, it should always be taken into account how these biases interact with the rest of the research design.

Therefore, this paper would call upon the patent research community to: use a correct interpretation of patent citations(i.e. using a logarithmic functional form and using the correct patent citation indicator); include other indicators of patent value (e.g. patent family size, backward citations, see for an overview Squicciarini et al. (2013)); and finally give up the enamored relation with patent citations as the perfect indicator of patent quality and start a structured research effort to find a better indicator.

References

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.

Alcacer, 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. P., & Van Looy, B. (2013). Inventions shaping technological trajectories: do existing patent indicators provide a comprehensive picture?.*Scientometrics*, *97*(2), 397-419.

Bakker, J., Verhoeven, D., Zhang, L., & Van Looy, B. Patent Citation Indicators: One Size Fits All? *Paper presented at the Druid Society conference/AOM conference (June/August, 2014)*

Belenzon, S. (2012). Cumulative Innovation and Market Value: Evidence from Patent Citations*. *The Economic Journal*, *122*(559), 265-285.

Bessen, J. (2008). The value of US patents by owner and patent characteristics. *Research Policy*, *37*(5), 932-945.

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

Gambardella, A., Harhoff, D., & Verspagen, B. (2008). The value of European patents. *European Management Review*, *5*(2), 69-84.

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.

EPO (2013). ANNEX XIV - Categories of Cited Documents in: *EXCHANGE FORMAT EPO - Patent Information Resource.* distribution by : Exchange Partners

Gay, C., & Le Bas, C. (2005). Uses without too many abuses of patent citations or the simple economics of patent citations as a measure of value and flows of knowledge*. *Economics of Innovation and New Technology*, *14*(5), 333-338.

Griliches, Z. (1998). Patent statistics as economic indicators: a survey. In *R&D and productivity: the econometric evidence* (pp. 287-343). University of Chicago Press.

Gittelman, M. (2008). A note on the value of patents as indicators of innovation: Implications for management research. *The Academy of Management Perspectives*, *22*(3), 21-27.

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of economics*, 16-38.

Hall, B. H., Thoma, G., & Torrisi, S. (2007). The market value of patents and R&D: evidence from European firms. In *Academy of Management Proceedings* (Vol. 2007, No. 1, pp. 1-6). Academy of Management.

Harhoff, D., Narin, F., Scherer, F. M., & Vopel, K. (1999). Citation frequency and the value of patented inventions. *Review of Economics and statistics*,*81*(3), 511-515.

Hegde, D., & Sampat, B. (2009). Examiner citations, applicant citations, and the private value of patents. *Economics Letters*, *105*(3), 287-289.

Hung, S. W., & Wang, A. P. (2010). Examining the small world phenomenon in the patent citation network: a case study of the radio frequency identification (RFID) network. *Scientometrics*, *82*(1), 121-134.

Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1992). *Geographic localization of knowledge spillovers as evidenced by patent citations* (No. w3993). National Bureau of Economic Research.

Jaffe, A. B., Fogarty, M. S., & Banks, B. A. (1998). Evidence from patents and patent citations on the impact of NASA and other federal labs on commercial innovation. *The Journal of Industrial Economics*, *46*(2), 183-205.

Jaffe, A. B., & Trajtenberg, M. (1999). International knowledge flows: evidence from patent citations. *Economics of Innovation and New Technology*, *8*(1-2), 105-136.

Jaffe, A. B., Trajtenberg, M., & Fogarty, M. S. (2000). *The meaning of patent citations: report on the NBER/Case-Western Reserve survey of patentees* (No. w7631). National bureau of economic research.

Magerman, T., Van Looy, B., & Song, X. (2006). *Data production methods for harmonized patent indicators: Patentee Name Harmonization*. EUROSTAT Working Paper and Studies, Luxembourg.

Marchetti, C. (1980). Society as a learning system: discovery, invention, and innovation cycles revisited. *Technological forecasting and social change*, *18*(4), 267-282.

Narin, F., Noma, E., & Perry, R. (1987). Patents as indicators of corporate technological strength. *Research policy*, *16*(2), 143-155.

Peeters, B., Song, X., Callaert, J., Grouwels, J., & Van Looy, B. (2010). Harmonizing harmonized patentee names: an exploratory assessment of top patentees. *Eurostat Working Paper*.

Squicciarini, M., Dernis, H., & Criscuolo, C. (2013). *Measuring patent quality: Indicators of technological and economic value* (No. 2013/3). OECD Publishing.

Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. *The Rand Journal of Economics*, 172-187.

Van Raan, A. F. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, *36*(3), 397-420.

Webb, C., Dernis, H., Harhoff, D., & Hoisl, K. (2005). Analysing European and International Patent Citations. *OECD Science, Technology and Industry Working Papers*, 2005/09, OECD Publishing.

Tables

Table 1: An overview of selected studies that examine patent citations

Study	source	Independent Variable	Dependent Variable	Controls	Statistical Method	Number of observations	Result with respect to patent citations
Carpenter et al. (1981)	USPTO	Count	R&D award given	Year	2-way ANOVA	202	31.38 F score
Narin and Noma(1987)	USPTO	Count	Financial performance	N/A	Correlation	17	0.628 correlation
Albert et al.(1991)	USPTO	Groups	Relative importance in industry	Year/ company/ technology	ANOVA	77 in 8 groups	2.01 F score (only most cited group significantly different)
Trajtenberg(1991)	USPTO	Aggregated citations	Innovative value on industry level	N/A	Correlation	10	0.685 correlation
Harhoff et al. (1999)	USPTO	Log(count)	Replacement value of patent	none	OLS	192	6.3% of variation explained
Hall et al. (2005)	USPTO	Stock	Tobins Q	R&D, sales, year industry	Non-linear model	12188 (1983 with patents)	3.2% of variation explained
Hall et al. (2007)	EPO	Stock	Tobins Q	R&D, sales, year, industry	Market eq. estimation	1 779	Insignificant
Bessen (2008)	USPTO	Count	Renewal value	Applicant, backward citations. Generality Originality	OLS	48990	4-7% of variation explained
Gambardella et al. (2008)	EPO	Log(count)	Replacement value of patent	Year/country/ tech class[13]	OLS	8 217	1.4% of variation explained
Arts et al.(2013)	USPTO	Groups of highly cited	Importance of invention	Year/USPC class	Logistic analyses	74 072	Citation indicators are highly significant

[13] 30 classes of the ISI-INPI-OST

Table 2: Variables used in this paper and their definitions

Indicator	Definition
Patent citation indicators	
EPO	Count of citations that are received by EPO applications from other EPO applications
USPTO	Count of citations that are received by USPTO applications from other USPTO applications
DOCDB	Count of citations that are received by the application or members from its DOCDB family from other DOCDB families
INPADOC	Count of citations that are received by the application or members from its INPADOC family from other INPADOC families
Control variables	
Nr. Countries	Number of distinct patent offices in which the DOCDB family of the patent has at least 1 application present .
Ln(Applt. Size)	Logarithm of the total number of patents filed by applicant
IPC*	Code identifying the technology class of the patent at the *level of the International Patent Classification. Patents may have more than 1 class
Year	Year in which the application was applied for at the patent office
Applicant type	Type of applicant: Company, government, hospital, individual, university or unknown. Patents may have multiple types of applicant.
Exclusive patent citation indicators	
EPO out	Count of citations made to an EPO patent by non-EPO patents
USPTO out	Count of citations made to an USPTO patent by non-USPTO patents
DOCDB EPO	Count of citations made to a DOCDB family member of an EPO patent
DOCDB USPTO	Count of citations made to a DOCDB family member of an USPTO patent
DOCDB cor.	Correction made by deduplication of citations made by one DOCDB patent family to another.
INPADOC add.	Count of citations made to INPADOC family members of the patent patent minus the count of citations to DOCDB family members
INPADOC cor.	Correction made by deduplication of citations made by one INPADOC patent family to another.
Self-citation indicators	
DOCDB(self)	Count of citations that are received by the application or members from its DOCDB family from other DOCDB families which have the same applicant
DOCDB(not self)	Count of citations that are received by the application or members from its DOCDB family from other DOCDB families which do not have the same applicant
Citation by origin and reason	
USPTO(APP)	Count of citations that are given by applicants of USPTO patents to the USPTO application
USPTO(EXA)	Count of citations that are given by examiners of USPTO patents (in their search report) to the USPTO application
EPO type A	Count of awarded citations made to understand the state of the art in the field
EPO type D	Count of citations made by the applicant
EPO type E	Citations patents that have earlier priority dates but later publication dates
EPO type XI	Citations that are made to prejudice inventive step
EPO type XN	Citations that are made to prejudice novelty
EPO type Y	Citations to patents that need to be seen as part of a larger group with other patents that are cited under the same code
EPO type L	Citations for other reasons than those listed in the previous EPO indicators
Citation window analysis	
DOCDB year *	Count of citations added to the DOCDB indicator in the *th year after the first application was filed for the DOCDB family
DOCDB * window	Count of citations taking only into account the first * years after the first application was filed for the DOCDB family
DOCDB * window (out)	Count of citations that are excluded by taking a * citation window

Table 3: Descriptive statistics of variables used in this paper

Variable	N	Mean	Std. Dev.	Min	Max
Full term EPO	338405	0.21	0.40	0	1
Full term USPTO	667948	0.46	0.50	0	1
Nr. Countries	669186	7.25	4.31	2	51
Ln(Applt. Size)	669186	7.44	3.32	0	12.99
EPO	669186	1.71	3.77	0	311
USPTO	669186	15.24	28.71	0	2802
DOCDB	669186	26.12	58.61	0	3146
INPADOC	669186	33.75	12.04	0	4747
EPO out	669186	3.11	7.63	0	548
USPTO out	669186	3.09	10.24	0	2267
DOCDB EPO	669186	34.98	94.19	0	5217
DOCDB USPTO	669186	21.47	81.73	0	4121
DOCDB cor.	669186	-13.68	39.19	-2080	0
INPADOC add.	669186	14.27	148.82	-832	6991
INPADOC cor.	669186	-20.32	87.14	-3240	0
DOCDB(self)	669186	3.79	10.85	0	481
DOCDB(not self)	669186	22.33	53.20	0	3034
DOCDB 5 window	669186	8.09	16.76	0	588
DOCDB 5 window(out)	669186	31.70	83.95	0	5096
DOCDB 5 window	669186	31.01	74.68	0	2637
DOCDB 10 window(out)	669186	8.79	26.82	0	2589

Table 4: Logistic regressions with the dependent variable: patent being renewed to its maximum term at the EPO

	(1) Full term EPO	(2) Full term EPO	(3) Full term EPO	(4) Full term EPO	(5) Full term EPO
Ln(1+EPO)		0.433***			
		(0.00623)			
Ln(1+USPTO)			0.363***		
			(0.00443)		
Ln(1+DOCDB)				0.587***	
				(0.00594)	
Ln(1+INPADOC)					0.576***
					(0.00574)
Nr. countries	0.0569***	0.0572***	0.0608***	0.0445***	0.0446***
	(0.00129)	(0.00131)	(0.00113)	(0.00133)	(0.00134)
Ln(applt. size)	0.00983***	-0.00336	0.00841***	0.00672***	0.00726***
	(0.00171)	(0.00174)	(0.00158)	(0.00174)	(0.00175)
Constant	-2.493***	-2.714***	-3.301***	-3.915***	-3.876***
	(0.0603)	(0.0609)	(0.0564)	(0.0634)	(0.0632)
IPC3 Dummies	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes
sector dummies	Yes	Yes	Yes	Yes	Yes
N	292625	292625	330478	292625	292625
Pseudo R^2	0.022	0.039	0.054	0.059	0.060
Log likelihood	-139057	-136643	-159946	-133858	-133703
Recall(%)	0.583	1.864	5.030	3.952	4.352
Precision(%)	55.20	57.56	59.83	60.79	58.03

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Logistic regressions with the dependent variable: patent being renewed to its maximum term at the USPTO

	(1) Full term USPTO	(2) Full term USPTO	(3) Full term USPTO	(4) Full term USPTO	(5) Full term USPTO
Ln(1+EPO)		0.237^{***} (0.00367)			
Ln(1+USPTO)			0.465^{***} (0.00293)		
Ln(1+DOCDB)				0.486^{***} (0.00327)	
Ln(1+INPADOC)					0.460^{***} (0.00315)
Nr. countries	0.0241^{***} (0.000745)	0.0281^{***} (0.000708)	0.0207^{***} (0.000765)	0.0123^{***} (0.000766)	0.0133^{***} (0.000765)
Ln(applt. size)	0.0139^{***} (0.000969)	0.00695^{***} (0.000949)	0.0132^{***} (0.000993)	0.0120^{***} (0.000990)	0.0127^{***} (0.000989)
Constant	-3.999^{***} (0.0694)	-4.187^{***} (0.0687)	-5.068^{***} (0.0703)	-5.144^{***} (0.0703)	-5.056^{***} (0.0702)
IPC3 Dummies	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes
sector dummies	Yes	Yes	Yes	Yes	Yes
N	616883	658738	616883	616883	616883
Pseudo R^2	0.094	0.104	0.125	0.121	0.121
Log likelihood	-385521.4	-407735.5	-372078.9	-373809.1	-374083.3
Recall(%)	65.34	67.26	63.87	63.81	63.72
Precision(%)	57.73	59.18	62.04	61.72	61.78

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Adapted logistic regressions related to different sources of citations. The dependent variable: patent being renewed to its maximum term at the EPO or USPTO

	(1) Full term EPO	(2) Full term EPO	(3) Full term EPO	(4) Full term USPTO	(5) Full term USPTO
Linear					
Nr. Countries	0.0573***	0.0433***	0.0445***	0.0188***	0.0171***
	(0.00130)	(0.00134)	(0.00134)	(0.000766)	(0.000784)
Ln(applt. size)	-0.000867	0.00477**	0.00437*	0.0157***	0.0155***
	(0.00173)	(0.00176)	(0.00176)	(0.000995)	(0.000996)
Constant	-2.725***	-3.375***	-3.588***	-4.996***	-5.004***
	(0.0610)	(0.0683)	(0.0730)	(0.0708)	(0.0711)
Non-linear					
EPO	0.749***	0.722***	1.193***		
	(0.0560)	(0.0634)	(0.113)		
EPO out	0.599***	0.307***	0.631***		
	(0.0386)	(0.0369)	(0.0686)		
DOCDB EPO		0.178***	0.389***		
		(0.0162)	(0.0371)		
USPTO				0.678***	0.644***
				(0.0203)	(0.0219)
USPTO out				0.575***	0.569***
				(0.0271)	(0.0299)
DOCDB USPTO					0.0249***
					(0.00650)
DOCDB cor.		-0.270***			-0.00913
		(0.0247)			(0.0142)
INPADOC add.			0.154***		
			(0.0240)		
INPADOC cor.			0.0747*		
			(0.0364)		
IPC3 Dummies	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes
sector dummies	Yes	Yes	Yes	Yes	Yes
N	292626	292626	292626	616884	616884
Pseudo R^2	0.0399	0.0632	0.0633	0.127	0.127
Log Likelihood	-136518	-133201	-133192	-371529	-371421
Recall(%)	1.592	4.132	3.965	63.74	63.62
Precision(%)	57.63	61.91	61.42	62.15	62.22

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Adapted logistic regressions related to self-citations. The dependent variable: patent being renewed to its maximum term at the EPO or USPTO

	(1) Full term EPO	(1) Full term USPTO
Linear		
Nr. Countries	0.0379***	0.00746***
	(0.00135)	(0.000769)
Log(applt. size)	-0.0126***	-0.00296**
	(0.00182)	(0.00102)
Constant	-3.317***	-4.717***
	(0.0691)	(0.0716)
Non-linear		
DOCDB(self)	2.063***	2.374***
	(0.122)	(0.0837)
DOCDB(not self)	0.306***	0.344***
	(0.0217)	(0.0148)
IPC3 Dummies	Yes	Yes
Year dummies	Yes	Yes
Sector dummies	Yes	Yes
N	292626	616884
Pseudo R^2	0.065	0.126
Log likelihood	-132969	-371692
Recall(%)	4.33	63.69
Precision(%)	60.13	62.46

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Adapted logistic regressions related to citations from examiners and applicants . The dependent variable: patent being renewed to its maximum term at the or USPTO

	(1) Full term
Linear	
Nr. Countries	0.0202***
	(0.000765)
Log (applt. size)	0.0140***
	(0.000994)
Constant	-4.895***
	(0.0707)
Non-linear	
USPTO (APP)	0.819***
	(0.0234)
USPTO (EXA)	0.599***
	(0.0200)
IPC3 Dummies	Yes
Year dummies	Yes
Sector Dummies	Yes
N	616884
pseudo R^2	0.126
Log likelihood	-371964
recall(%)	63.19
precision(%)	62.22

Table 9: logit regression on reaching full term for granted EPO patents applications

	(1)
	Full term
Linear	
Nr. Countries	0.0563***
	(0.00131)
Log(applt. size)	0.00112
	(0.00174)
Constant	-2.649***
	(0.0608)
Non-linear	
EPO (type A)	0.528***
	(0.0334)
EPO (type D)	3.728***
	(0.205)
EPO (type T)	-0.371
	(0.236)
EPO (type E)	0.407*
	(0.183)
EPO (type L)	0.0973
	(0.929)
EPO (type XI)	1.262
	(0.827)
EPO (type XN)	0.642***
	(0.0571)
EPO (type Y)	0.723***
	(0.0795)
IPC3 Dummies	Yes
Year dummies	Yes
Sector dummies	Yes
N	292626
Pseudo R^2	0.0377
log likelihood	-136825.7
recall(%)	1.543
precision(%)	54.41

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Correlations between counts of citation that fall within a citation window and the count of citations that fall outside that citation window

Citation window	EPO	USPTO	DOCDB
5 year	0.34	0.44	0.53
10 year	0.36	0.44	0.60

Table 9: Adapted logistic regressions related to citation windows. The dependent variable: patent being renewed to its maximum term at the EPO or USPTO

	(1) Full term EPO	(2) Full term EPO	(3) Full term EPO	(4) Full term USPTO	(5) Full term USPTO	(6) Full term USPTO
Linear						
Nr. Countries	0.0450***	0.0450***	0.0450***	0.0250***	0.0251***	0.0251***
	(0.00133)	(0.00133)	(0.00133)	(0.00120)	(0.00120)	(0.00120)
Log(applt. size)	0.0104***	0.00966***	0.00990***	0.0265***	0.0268***	0.0264***
	(0.00176)	(0.00175)	(0.00175)	(0.00142)	(0.00142)	(0.00142)
Constant	-3.626***	-3.621***	-3.636***	-5.407***	-5.424***	-5.419***
	(0.0698)	(0.0702)	(0.0700)	(0.0825)	(0.0826)	(0.0826)
Non linear						
DOCDB						
Year 0	0.280**			0.978***		
	(0.0911)			(0.110)		
Year 1	0.227***			0.672***		
	(0.0611)			(0.0694)		
Year 2	0.250***			0.507***		
	(0.0558)			(0.0584)		
Year 3	0.252***			0.464***		
	(0.0552)			(0.0562)		
Year 4	0.376***			0.509***		
	(0.0612)			(0.0588)		
Year 5	0.248***			0.432***		
	(0.0597)			(0.0581)		
Year 6	0.461***			0.655***		
	(0.0689)			(0.0667)		
Year 7	0.502***			0.673***		
	(0.0730)			(0.0699)		
Year 8	0.452***			0.800***		
	(0.0738)			(0.0759)		
Year 9	0.618***			0.851***		
	(0.0835)			(0.0797)		
Year 10	0.697***			0.535***		
	(0.0472)			(0.0315)		
Year >10	0.865***			1.066***		
	(0.220)			(0.211)		
DOCDB 5 window		0.258***			0.594***	
		(0.0287)			(0.0385)	

DOCDB 5 window (out)		0.625***			0.591***	
		(0.0426)			(0.0325)	
DOCDB 10 window			0.360***			0.627***
			(0.0295)			(0.0360)
DOCDB 10 window (out)			0.731***			0.558***
			(0.0494)			(0.0328)
IPC3 Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes	Yes	Yes	Yes
N	292626	292626	292626	289258	289258	289258
Pseudo R^2	0.060	0.059	0.060	0.087	0.086	0.086
Log likelihood	-133695	-133740	-133722	-183008	-183052	-183048
Recall(%)	4.11	4.00	4.10	60.75	60.82	60.86
Precision(%)	61.06	61.03	61.55	62.14	62.15	62.15

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figures

Figure 1: Depiction of the coefficients of different citations scores of the DOCDB indicator in a logistic regression, using EPO renewal as a dependent variable. The coefficients are compared to a score of 0 citations. Error bars signal a 95% confidence interval.

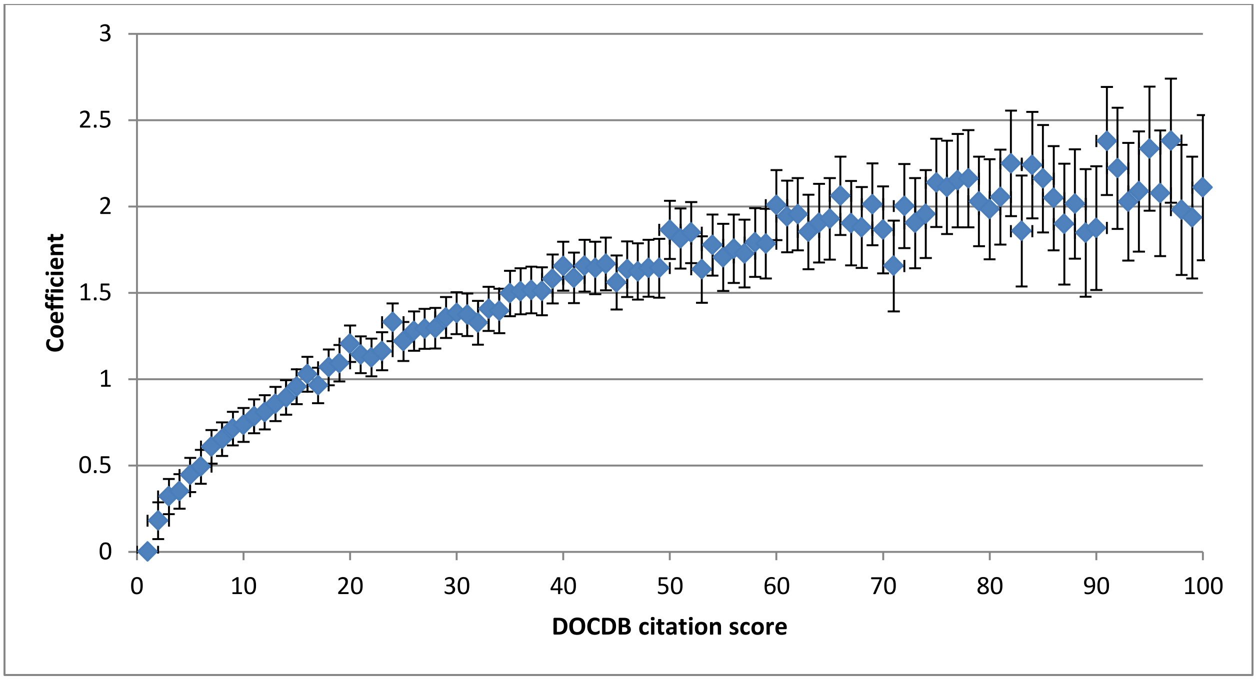

Figure 2: Depiction of the logistic coefficients of different citations scores of the DOCDB indicator in a logistic regression, using USPTO renewal as a dependent variable. The coefficients are compared to a score of 0 citations. Error bars signal a 95% confidence interval.

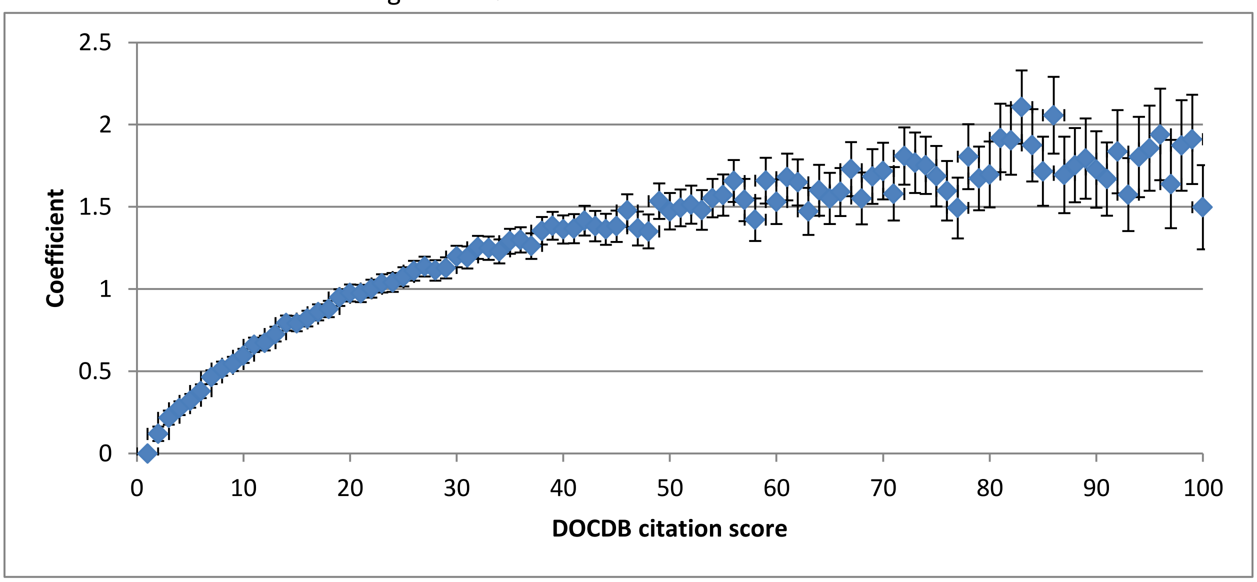

Figure 3: Depiction of the estimates of the odds ratios of different citations scores using the DOCDB indicator to explain EPO renewal, a 95% confidence interval is depicted. The odds ratios are relative to patents that have 0 citations. The fitted regression line has a slope of 0.0783, an intercept of 1.49 and an R^2 of 0.89.

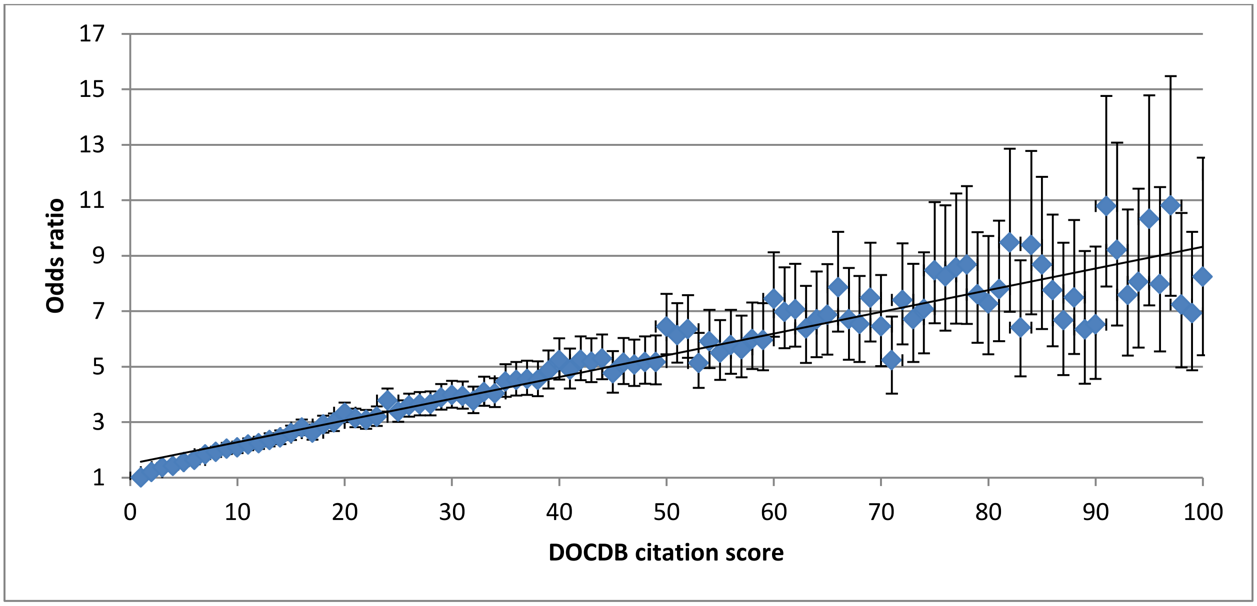

Figure 4: Depiction of the estimates of the odds ratios of different citations scores using the DOCDB indicator to explain USPTO renewal, a 95% confidence interval is depicted. The odds ratios are relative to patents that have 0 citations. The fitted regression line has a slope of 0.052, an intercept of 1.55 and an R^2 of 0.88.

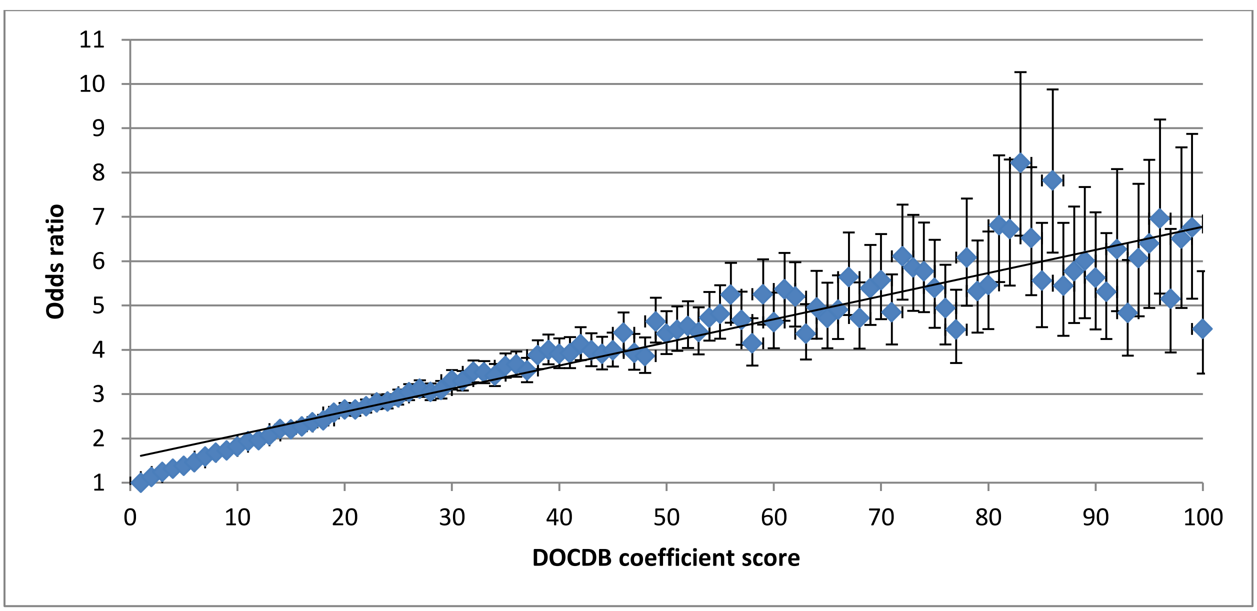

Figure 5: Depiction of the share of citations received on average in the years after an application has been filed for three citation indicators.