Delineating the scientific footprint in technology: Identifying science within non-patent references

Callaert, Julie; Van Looy, Bart; Grouwels Joris
ECOOM & Research Division Incentim
Faculty of Business and Economics
K.U.Leuven

Abstract

Indicators based on non-patent references (NPRs) are increasingly being used for measuring and monitoring science-technology relations. But NPRs in patent documents contain noise, as not all of them can be considered 'scientific'. In this paper, we introduce the results of a machine-learning algorithm that allows to identify scientific references in an automated way. The obtained outcomes are used to analyze indicators based on non-patent references, with a focus on the difference between NPR- and SNPR-based indicators. Differences between both indicators are significant and dependent on the considered patent system, the applicant country and the technological field. These results indicate the relevance of delineating scientific references when developing and employing indicators of science-technology relations. Furthermore, the revealed sensitivity of these indicators to technological, national and patent system characteristics urges for a contextualized interpretation.

Introduction

In today's knowledge based systems of innovation, indicators of science-technology relations are highly relevant for monitoring innovative performance or potential at several levels of analysis. Indicators based on non-patent references are very popular in this respect. In spite of some discussion about their actual meaning (see Nelson, 2009), scientific references in patents are in any case indicative of relatedness or closeness between the developed technology and the cited science (Callaert, 2006; Meyer, 2000a; Tijssen et al. 2000; Van Looy et al., 2007). The presence of scientific research in the 'prior art' description of a patented invention should be considered an indicator of the relevance of scientific findings for assessing and contextualizing technology development. As such, indicators based on scientific references in patents references provide useful additional information on science-technology relatedness or vicinity, at least if their presence displays sufficient levels of occurrence (Callaert et al., 2006). Combine to this the widespread and consistent availability of reliable and comprehensive patent databases, and it is clear that there are many opportunities for systematic and objective quantitative analyses of patent-related issues, among which the linkage between science and technology as measures by NPRs. At the same time, NPRs in patent documents contain 'noise'. Some efforts have been made in the past to map types of non-patent references. Narin and Noma (1985) reported an average of 0.3 non-patent references per patent, 48% of which related to journals, 15% to books and 11% to abstracts. Van Vianen et al. (1990) found that 55.7% of non-patent references in Dutch patents were journal citations, the others were mostly books and abstract services. Harhoff et al. (2003) found approximately 40% of non-patent references referring to trade journals, firm publications or standard texts in technical fields. Callaert et al. (2006), in a sample of EPO and USPTO non-patent references, found that more than half of them were journal articles. The remainder include conference proceedings; industry-related documents; and reference books / databases. Even though some non-journal reference categories may still be considered scientific in a broader sense, it is clear from the above overview that not all non-patent references are scientific sources. Therefore, if one is interested in pinpointing those traces of prior art that refer to scientific research in a narrow sense: i.e. references to the serial scientific journal literature, large scale identification of the scientific character of non-patent references becomes highly relevant.

In this paper, a method for identification is developed and applied to study the occurrence of actual scientific references within the 'non-patent references'. Next we analyze to what extent NPR-based indicators change, depending on whether only scientific versus all non-patent references are taken into account. We consider two often-used NPR-based indicators of the science relatedness of patents. First, the proportion of patents with at least one scientific non-patent reference gives an idea of the size of the scientific footprint within technology (from now on, we refer to this indicator as 'size'). Second, the number of cited scientific non-patent references per patent – sometimes referred to as the 'science intensity' – gives an idea of the depth of the scientific footprint (from now on, we refer to this indicator as 'depth'). Both indicators are related, but capture different aspects of science-technology linkage. This implies that, although previous studies of ST relations consider mostly the depth component, a sound interpretation of the scientific footprint within technology requires that both the size and depth indicators are considered.

Methodology for characterizing non-patent references

In order to arrive at an algorithm that allows delineating scientific references, a 'learning set' has been created, consisting of 25.783 non patent references. These references have been selected randomly from 7 582 096 NPRS pertaining to all EP0, USPTO and PCT patent documents included in the Patstat Database (April 2008 version).

This sample was classified by a team of researchers (n=5) as either Journal, Proceedings or non-scientific (e.g. manuals, patent abstracts,…) We also included a category 'Maybe' to include inconclusive cases (e.g. N.N. (year)). In total 12.465 NPR's received the label 'Journal; 2.037 were classified as 'Proceedings'; 10.411 NPR's are Non-Scientific while 360 NPR's can be labeled as doubtful.

Within a next step, all these references have been parsed, indexed and stemmed in order to create a document by term-matrix (consisting of 25.783 rows (references) and 74.127 columns (unique stemmed terms); cell values equal the frequency of occurrence of each document by term combination)[14]. As in most text mining settings, this matrix is very sparse. Contrary to most text-mining applications, no weighting was applied. The main purpose of most standard weighing systems (e.g. TF-IDF), is to diminish the weight of the terms that occur very often in order to increase the discriminatory power. In this particular case (i.e. classification of a reference to be scientific or not) however, words that occur often can be very significant. For example, the two most occurring words in the sample index are "et" and "al". Their presence in a reference might be a strong indication of the reference being scientific.

In order to arrive at a robust algorithm, the learning set was partitioned in a random way by means of Monte Carlo simulations. Moreover, each time, a test set was created for verification purposes (consisting of 30% of the references). This test set was each time excluded when developing the classifier (which also implies that classifiers have been developed on smaller Document by Term matrices as the test documents (and the unique terms are left out)). This results in 10 different - but partially overlapping - training sets (consisting of about 17.500 documents and +/-56.600 terms) with their respective test sets of about 8.500 documents. Every one of these training sets will generate its respective classifier, which means that the

[14] See Salton, G., Wong, A. and Yang, C.S. (1975) on vector space models as well as Magerman, Van Looy & Song (2009) for a more elaborated account on vector space models for patent and publication documents.

simulation will yield 10 different classifiers. If the results of these classifiers on their test sets are similar, we can be confident that any classifier that is generated in this way will yield similar results.
For each of the 10 training sets, the terms (i.e. all terms that occur in the training set) are ordered by their frequency in this training set. Only terms that occur in 10 or more documents over the complete sample of 25000 are withheld as a certain term has to occur enough in order to have generalizing power over the whole dataset. This resulted in 4.148 terms withheld for the development of the qualifier. Next, a principal component analysis is performed on the training sets. The main purpose of this step is not dimensionality reduction, but de-correlating the variables. Highly correlated variables tend to be linearly dependent, which makes the covariance matrix positive semi-definite, while the discriminant analysis algorithms require positive definite covariance matrices. Retaining 99% of the variance removes the dimensions with an eigenvalue numerically equivalent to zero, solving the problem while retaining almost all information in the dataset. By applying a 99% threshold of withheld variance we obtain 3.500 components. Next, we multiplied all obtained eigenvalues with the outcome variable. The obtained scores for each component have been ranked; within different simulations the number of dimensions used to arrive at a classifier (based on discriminant analysis) equals 10, 50, 100, 500, 1000 and finally 3.500 components. The obtained findings reveal that the optimal amount of components for classification purposes amounts to 1000; When all dimensions are kept, over-fitting occurs (performance on training keeps on rising, but performance on test gets worse). The accuracy level obtained for training sets equals 94,1% (correctly classified/total number of references), for the test sets we obtained an accuracy level of 92%. The different simulations generated highly congruent outcomes, signaling the robustness of the overall approach.

Data

The above-described methodology was used to characterize all non-patent references (N= 11388123) in the PATSTAT database (version 04/2009). 58% of all NPRs were characterized as “scientific”: referring to the serial journal literature or to proceedings, implying that approximately 42% of all non-patent references are not scientific.
In what follows, we analyze indicators based on non-patent references, whereby we are primarily interested in the difference between NPR- and SNPR-based indicators. For these analyses, we consider EPO (applications and grants) and USPTO (grants) patents with application years 2000-2009, and with applicant countries in EU15, Switzerland, US, Canada, Japan and Korea. Indicators are broken down by patent system (EPO versus USPTO), application year, applicant country and technology domain (according to the FhG19 classification). Full counting schemes are used for patents that are assigned to different countries and technology fields. Starting from the number of patents, we calculate for each year/patent system/country/field combination: (1) the average number of patents with non-patent references, (2) the average number of references, (3) the size of the footprint as the proportion of patents with non-patent references and finally (4) the depth of the footprint as the average number of non-patent references per cited patent or. Each indicator is calculated twice: once using all NPRs and once using only scientific NPRs. Finally we include the traditional indicator denoting ‘Intensity’, i.e. the number of references divided by all patents.

Analyses and results

Difference between NPR-based and SNPR-based indicators

Scientific NPRs represent a fraction of all non-patent literature. In table 1, the differences between NPR- and SNPR-based indicators are shown, and their significance is evaluated

using a paired-sample t-test. Results are presented for EPO and USPTO patents separately, as well as for the aggregate set of EPO and USPTO patents.

Table 1. Paired-samples t-test: NPR- versus SNPR-based indicators (N=6540)

	EPO			*USPTO*			*AGGREGATE*		
	All NPRs	***Subset SNPRs***	***t-value (sign.)***	***All NPRs***	***Subset SNPRs***	***t-value (sign.)***	***All NPRs***	***Subset SNPRs***	***t-value (sign.)***
(1) Average # patents with references	124,77	67,43	*25,39***	201,83	113,67	*15,50***	159,33	88,17	*25,02***
(2) Average # of references	355,88	219,63	*24,26***	1888,80	1110,01	*10,77***	1043,35	618,94	*12,94***
(3) Average % of patents with references (as portion of total number of patents) → SIZE	29,78%	15,61%	*66,05***	43,80%	22,58%	*64,51***	36,07%	18,73%	*89,35***
(4) # references / # patents with references → DEPTH	2,65	2,75	*-4,44***	6,84	6,72	*1,55 (n.s.)*	4,52	4,52	*,03 (n.s.)*
# references / total # patents	0,85	0,53	*59,05***	3,31	2,05	*43,47***	1,95	1,21	*51,24***

Table 1 first reveals some differences between reference-based indicators for EPO and USPTO patents. USPTO patents have considerably higher volumes of references and of patents with references. Also in terms of impact, USPTO patents contain approximately 2,5 to 4 times more (S)NPRs per patent than EPO patents. These observations hold for NPR- as well as SNPR-based indicators and are in line with findings from previous studies. They largely result from different citation requirements and philosophies at the USPTO (duty of disclosure; documentary search) and at the EPO system (no duty of disclosure; patentability search) (Michel & Bettels, 2001).

As for the difference between NPR- and SNPR-based indicators, it can be seen that the fraction of NPRs referring to the scientific literature (journal articles and proceedings) is about 60%. This leads to considerable differences in derived indicators. The difference between NPR- and SNPR-based indicators is most prominent for the size indicator (indicator 3): the proportion of patents containing NPRs is 36% when all NPRs are considered, and is almost cut in half to 19% when only scientific NPRs are considered. The depth (intensity-indicator 4, measuring the number of references per citing patent) is 4,5 and appears to be independent of whether all NPRs versus only scientific NPRs are considered. This is because the denominators are adapted to the considered subset: whereas the volume of patents with NPRs is used as the denominator for the NPR intensity, the volume of patents with scientific NPRs is used as a denominator for the SNPR intensity indicator. If, alternatively, the same denominator is used for both (NPR and SNPR) intensities – namely the total number of patents – then both intensities differ significantly.

The observed differences between NPR- and SNPR-based indicators underline the relevance of singling out scientific NPRs when developing indicators of science-technology relatedness, based on references from patents to the non-patent literature. This relevance is especially pronounced for indicators that reflect the scientific size, i.e. the share of patents that contain references to the scientific literature. Not distinguishing scientific references leads to an approximate 200% overestimation of the scientific size within technology.

Further support for the relevance of singling out scientific NPRs is presented in table 2. This table relates the scientific or non-scientific character of NPRs to citation categories that are added in European search reports. When European search reports are drafted, categories are assigned to all documents cited in the search report. These categories reflect the relation or the

relevance of the reference to the patented technology. The most important - and most frequent - categories are the X, Y and A categories. The two former ones are assigned to references that question the novelty or inventive step of patent claims. More specifically, 'X' documents are considered highly relevant and related to the technology, as they can by itself and alone be prejudicial to the novelty. 'Y' documents from their part are also included for questioning the inventive step of a patent claim, but only if considered in combination with another document. 'A' references do not challenge novelty or inventive step, but rather document the technical background of the invention. These different citation categories hence have different degrees of linkage: whereas 'X' references have a very high degree of linkage, this degree may vary for 'Y' references. 'A' references usually represent lower proximity or less immediate relevance to the patented technology (see Meyer, 2000b). It can be seen in table 2 that the largest difference between scientific and non-scientific references lies in the proportion of A references (i.e. the ones with the lowest proximity to the invention). This proportion is higher for the non-scientific references, meaning that these references contain relatively more source material with a lower degree of linkage to the developed technology (significance of relations supported by chi square test).

Table 2. Breakdown of scientific and non-scientific references in citation categories (EPO patents only)

	non-scientific	% non-scientific	scientific	% scientific
A	130491	41,43%	189201	36,08%
X	114975	36,50%	183580	35,01%
Y	39063	12,40%	82352	15,71%
D	17625	5,60%	35935	6,85%
P	8718	2,77%	25066	4,78%
T	1707	0,54%	7325	1,40%
L	1672	0,53%	528	0,10%
E	608	0,19%	160	0,03%
O	132	0,04%	178	0,03%
Total	314991	100%	524325	100%

Finally, the in table 1 revealed differences between the EPO and USPTO patent system put forward the relevance of specificities in citation practices and characteristics. In the following section, we consider how several other potentially relevant factors influence the occurrence of scientific NPRs and hence the value of SNPR-based indicators.

Scientific size and depth: influencing factors

Many science-technology studies point out the importance of influencing factors when studying science-technology relations and related indicators (for an overview: see Van Looy et al., 2002). Patent system characteristics, national specificities and technological fields are among the most important factors.

The USPTO and EPO systems differ considerably in terms of search and examination procedures. It has been argued that the comprehensiveness and the quality of citation lists appearing in patent documents vary significantly as a function of the patent office (Meyer, 2000b; Michel and Bettels, 2001). At USPTO, patent applicants have a duty of disclosure, meaning that they must provide all information that is reasonably deemed necessary to properly examine the patent application and that they were aware of prior to the filing date of the application. When filing for a patent at EPO, applicants are under no such duty of disclosure. As Michel and Bettels (2001) point out, this leads to a situation where the average US search report has the characteristics of a documentary search, whereas the EPO search reports reflect patentability searches. The patentability search is not exhaustive in the same

sense as the documentary search in that it should be limited to what is directly relevant to patentability. In addition, USPTO and EPO patents differ in terms of the references that are published and hence made available in large-scale in patent databases. Whereas UPSTO patents list all examiner- and applicant-given references, the search reports for EPO patents contain examiner-given references. Applicant-given references are only included in EPO search reports if they are deemed relevant by the examiner in the patentability search. These differences imply several things. First, the volume of references in USPTO patents will be considerably higher than the volume of references in EPO patents, as already confirmed in our dataset of SNPR-based indicators (cf. supra table 1). Second, the proportion of applicant-given references – as registered in most available patent databases – will be higher for US patents than for EPO patents. These differences will influence the number and type of references cited in patents from both systems.

Moreover, the relations between science and technology and the indicators for measuring these are influenced by differences in national innovation systems (Harhoff et al., 2003; Van Looy et al., 2002). The scientific texture that characterizes a country and the extent of the national science-base will likely influence measures of science-technology relations, especially when indicators are based on the occurrence of references to the scientific literature. Such differences should be taken into account when comparing the scientific footprint within technologies across the boundaries of national innovation systems. It should be noted here that, if examiner-given references prevail, applicant countries may have less bearing on citation practices and characteristics. Examiners are located centrally and are not structurally related to the applicant countries of the patents that they examine. In line with the abovementioned argument on the higher share of applicant-given references in USPTO patents, an interaction effect may be assumed whereby the influence of applicant country on SNPR-based indicators would be more outspoken for USPTO patents than for EPO patents.

Finally, technology domains play an important role. Much in the same way that the propensity to publish and the propensity patent varies heavily between disciplines and fields, so will citation practices differ within and between science and technology spheres. Several studies point to field specific effects that need to be taken into account when using and interpreting indicators of science-technology relatedness (Callaert et al., 2006; Harhoff et al., 2003; Van Looy et al., 2002, 2003). Patents within technologies that are strongly linked to scientific progress and a knowledge base (e.g. pharmaceutics, chemistry,…) will show a higher proportion of scientific references than domains that are known to be less science-intensive (e.g. some machinery- and transport related subfields).

In what follows, we show to what extent these factors influence SNPR-based indicators. ANCOVA analyses were performed with scientific size and depth (measured by using only scientific NPRs) acting as dependent variables. They were logarithmically transformed to comply with normality assumptions. Independent variables are the applicant country, the technological field and the publication authority. To account for potential evolutions over time, a year covariate is included in the model. Interaction effects are included between patent system and applicant country, as well as between patent system and technological field. The results are shown in table 3.

Table 3. ANCOVA analyses: Influences on the size and depth of the scientific footprint

	DEP VAR = size of scientific footprint (ln)				DEP VAR = depth of footprint (ln)			
	Type III Sum of Squares	df	F	Sig.	Type III Sum of Squares	df	F	Sig.
Corrected Model	108,606	76	144,541	,000	1021,775	76	82,313	,000
Intercept	,357	1	36,072	,000	,483	1	2,957	,086
Application Year	,337	1	34,045	,000	,685	1	4,197	,041
Publication Authority	3,210	1	324,693	,000	258,690	1	1583,813	,000
FTC 19	97,832	18	549,741	,000	407,145	18	138,484	,000

Country of Applicant	3,235	19	17,219	,000	106,897	19	34,446	,000
Publication Authority * FTC 19	2,512	18	14,116	,000	34,456	18	11,720	,000
Publication Authority * Country	1,093	19	5,820	,000	59,983	19	19,328	,000
Error	63,898	6463			851,459	5213		
Total	334,867	6540			13561,553	5290		
Corrected Total	172,504	6539			1873,234	5289		
R-square	,630				,545			
Adj R-square	,625				,539			

Patent system, technology domain, and applicant country significantly influence both size and the depth of the scientific footprint (as measured by using only scientific NPRs). In addition, the interaction effects are significant, meaning that the influence of applicant countries differs between EPO and USPTO patents, as does the influence of technology domains. In what follows, we use simple descriptive statistics to shed further light on these influencing factors. Figures 1a and 1b show the influence of technological domains on scientific size and depth respectively. The figures distinguish between EPO and USPTO patents, to graphically present the interaction effects observed in table 3.

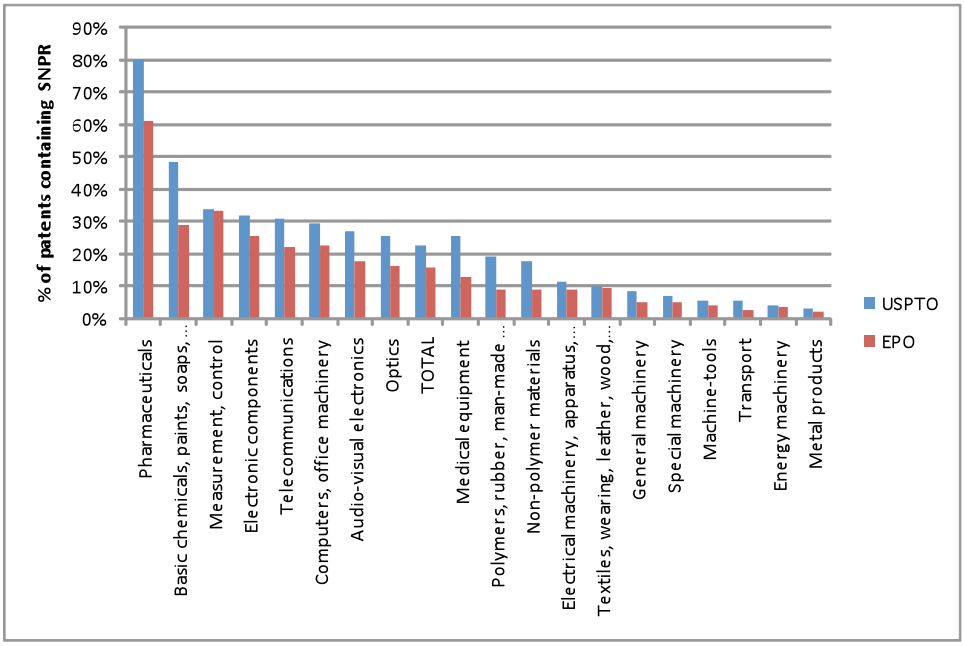

Figure 1a – Influence of technological domains on the size of the scientific footprint

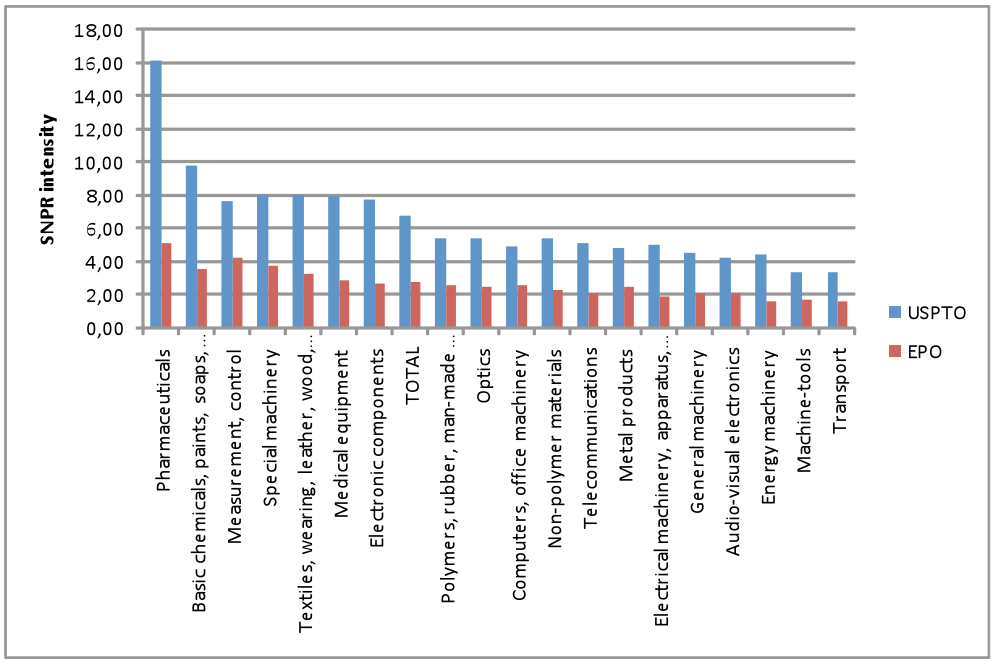

Figure 1b – Influence of technological domains on the depth of the scientific footprint

Overall, the observations confirm that a scientific footprint is most outspoken for fields like Pharmaceuticals and Chemicals. At the lower end are domains that are known to be less science-intensive, such as General machinery, Transport, Machine-tools, Energy machinery, Metal products. The technological specificity pattern of scientific size (Figure 1a) is quite similar for EPO and USPTO patents; although EPO patents show less variety. Also in terms of scientific depth (Figure 1b), the technological specificities are reflected more strongly in USPTO patents, with a notably 'deep' scientific footprint for Pharmaceuticals.

Figures 2a and 2b show the influence of applicant countries on scientific size and depth respectively. Only those countries are included with an annual average of at least 20 patents per technology domain.

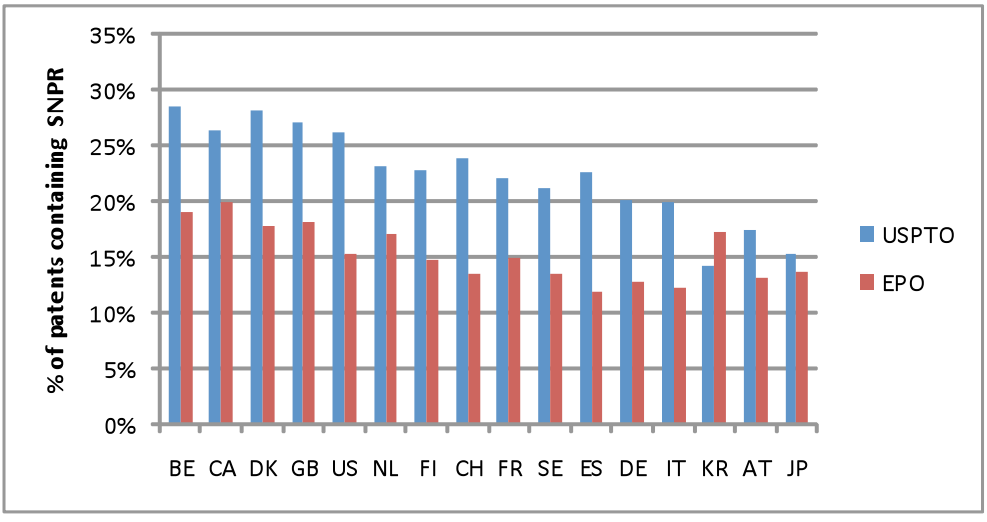

Figure 2a – Influence of applicant countries on the size of the scientific footprint

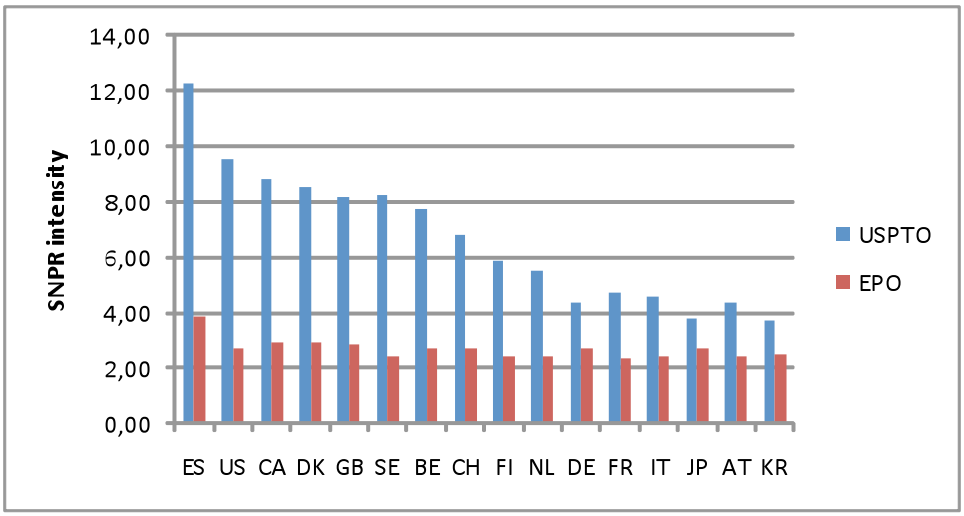

Figure 2b – Influence of applicant countries on the depth of the scientific footprint

Overall, these figures show that the scientific footprint is most prominent in patents from Canada, the United States, the United Kingdom and Denmark. For Japan, Austria and Korea, the scientific footprint is least apparent. It is no coincidence that English speaking countries have an advantage in terms of their ability to link up with the scientific literature, which consists mainly of English-written sources. Country-effects are at the same time shown to depend on which indicator is used. For instance, patents from Belgian applicants appear at the top in terms of scientific size (figure 2a), whereas they are more average in terms of scientific depth (figure 2b). The opposite holds for patents from Spanish applicants: although their scientific size is average, it can be seen that for those Spanish patents that cite scientific references, the average number of cited references is high in comparison to other countries. Finally, the influence of applicant countries on the scientific footprint in patents is much more outspoken for USPTO patents than for EPO patents. As suggested earlier, this likely results from the fact that USPTO patents include a higher proportion of applicant-given references, hence they will be more influenced by applicant-level characteristics.

3.3. Influences on differences between NPR- and SNPR-based indicators: altered ranks

We showed earlier that science-technology indicators differ, depending on whether only scientific NPRs or all NPRs are considered. Hence, reference-based indicators are sensitive to the specific subset of NPRs that is considered in building the indicators. The higher this sensitivity, the more appropriate it becomes to distinguish scientific from non-scientific NPRs.

In this section, we investigate to what extent such sensitivity – and hence the relevance of isolating scientific NPRs – is influenced by the factors identified in the previous section: patent system, technology domain, and applicant country. ANCOVA analyses are performed, with the ratios of SNPR- to NPR-based measures as dependent variables. Ratios were logarithmically transformed to comply with normality assumptions. Results are indicated in table 4.

Table 4. ANCOVA analyses: Influences on the ratio between SNPR- and NPR-based indicators

	DEP VAR = Ratio: size SNPR / size NPR (ln)				DEP VAR = Ratio: depth SNPR / depth NPR (ln)			
	Type III Sum of Squares	df	F	Sig.	Type III Sum of Squares	df	F	Sig.
Corrected Model	135,779	39	161,359	,000	19,620	39	19,641	,000
Intercept	2,754	1	127,633	,000	,139	1	5,429	,020
Application Year	2,866	1	132,852	,000	,097	1	3,789	,052

Publication Authority	,162	1	7,489	,006	,892	1	34,829	,000
FTC 19	124,221	18	319,851	,000	14,904	18	32,327	,000
Country	6,379	19	15,562	,000	3,598	19	7,393	,000
Error	127,559	5912			134,469	5250		
Total	1012,880	5952			2687,019	5290		
Corrected Total	263,338	5951			154,089	5289		
R-square	,516				,127			
Adj R-square	,512				,121			

The sensitivity of indicators to the considered subset of NPRs is shown to vary across patent systems, technological domains and applicant countries. The size ratio is somewhat higher for EPO than for USPTO patents. More notable however is the effect of the patent system on the depth ratio. For EPO patents, the scientific footprint becomes deeper if one considers only scientific NPRs (ratio depth SNPR/NPR > 1). For USPTO patents on the other hand, the scientific footprint becomes more shallow if only scientific references are considered.
Technology domains were shown to influence the size of the scientific footprint (see table 3 and figures 1a and 1b). In table 4, it is shown that they also affect the extent of the difference between NPR- versus SNPR-based indicators. An important implication for the practice of indicator development is that the ranking of technology fields by scientific size and depth will change, depending on which subset of NPRs is used for developing the indicator. This is demonstrated in table 5.

Table 5. NPR- versus SNPR-based indicators and ranks across technology domains

Technology Domain	size NPR	rank size NPR	size SNPR	rank size SNPR	depth NPR	rank depth NPR	depth SNPR	rank depth SNPR
Pharmaceuticals	76,6%	1	70,0%	1	11,64	1	10,30	1
Basic chemicals, paints, soaps, petroleum products	53,1%	2	37,8%	2	6,57	2	6,47	2
Measurement, control	47,4%	3	33,5%	3	5,54	3	5,79	3
Electronic components	45,6%	4	28,4%	4	4,60	5	4,94	7
Telecommunications	42,6%	6	26,0%	5	4,02	8	3,45	12
Computers, office machinery	44,8%	5	25,6%	6	4,15	6	3,65	10
Audio-visual electronics	41,8%	7	21,8%	7	3,43	13	3,02	16
Optics	40,0%	8	20,7%	8	3,63	10	3,81	9
Medical equipment	30,7%	11	18,5%	9	4,88	4	5,15	6
Polymers, rubber, man-made fibres	32,6%	10	13,5%	10	3,48	12	3,84	8
Non-polymer materials	35,4%	9	12,8%	11	3,76	9	3,64	11
Electrical machinery, apparatus, energy	29,6%	12	10,0%	12	3,20	14	3,26	14
Textiles, wearing, leather, wood, paper, domestic appliances, furniture, food	25,4%	16	9,6%	13	4,09	7	5,28	5
General machinery	26,3%	13	6,4%	14	2,79	15	3,17	15
Special machinery	23,4%	18	5,9%	15	3,51	11	5,64	4
Machine-tools	25,7%	14	4,6%	16	2,37	17	2,42	18
Transport	25,7%	15	4,1%	17	2,46	16	2,35	19
Energy machinery	24,2%	17	3,8%	18	2,36	18	2,85	17
Metal products	14,9%	19	2,7%	19	2,29	19	3,39	13
	Kendall's tau rank order correlation: 0,833**				***Kendall's tau rank order correlation: 0,661*****			

It can be seen that the top 3 technology domains (Pharmaceuticals, Chemicals and Measurement) are ranked robustly across the different indicators. For countries outside the top 3 however, there are some non-trivial shifts in ranks, depending on whether NPRs or SNPRs

are considered. This is especially pronounced for the depth indicator (see also its lower rank order correlation): the focus on only scientific references causes the domains of 'Telecommunications' and 'Computers, Office Machinery' to each drop 4 places in the ranking, whereas the domains of 'Metal Products' and 'Special Machinery' climb up 6 and 7 places respectively in terms of scientific depth. Just like technology domains, applicant countries were shown to influence not only the scientific footprint in technology (see table 3 and figures 2a and 2b), but also the extent of the difference between NPR- versus SNPR-based indicators (see table 4). Table 6 demonstrates what this implies in terms of shifts in national rankings. Only those countries are included with an annual average of at least 20 patents per technology domain.

Table 6. NPR- versus SNPR-based indicators and ranks across applicant countries

Applicant country	size NPR	rank size NPR	size SNPR	rank size SNPR	depth NPR	rank depth NPR	depth SNPR	rank depth SNPR
BE	41,1%	2	23,3%	1	4,37	8	4,85	7
CA	38,5%	4	23,0%	2	5,92	2	5,87	3
DK	35,4%	10	22,5%	3	4,85	4	5,37	4
GB	41,1%	1	22,3%	4	4,70	5	5,32	5
US	36,5%	8	20,6%	5	6,29	1	6,01	2
NL	38,8%	3	20,0%	6	3,75	10	3,84	10
FI	36,7%	7	18,6%	7	3,92	9	3,97	9
CH	37,2%	6	18,4%	8	4,48	6	4,65	8
FR	37,5%	5	18,3%	9	3,29	13	3,44	12
SE	35,8%	9	17,1%	10	4,44	7	5,06	6
ES	29,0%	16	16,6%	11	5,24	3	7,04	1
DE	35,0%	11	16,2%	12	3,38	11	3,48	11
IT	32,3%	14	15,8%	13	2,96	15	3,39	13
KR	31,5%	15	15,8%	14	2,95	16	3,06	16
AT	32,9%	13	15,0%	15	3,03	14	3,19	15
JP	34,0%	12	14,4%	16	3,33	12	3,25	14
	Kendall's tau rank order correlation: 0,550**				***Kendall's tau rank order correlation: 0,883*****			

It can be seen in table 6 that there are considerable shifts already at the top, depending on whether NPRs or SNPRs are used to calculate scientific footprint indicators. Shifts are most outspoken for the scientific size indicator (as reflected also in the lower rank order correlation). Especially Denmark and Spain benefit when SNPRs are considered instead of all NPRs: they climb up 7 and 5 positions respectively. Denmark hence enters the top 3, whereas it was ranked only 10th when all NPRs were considered. Spain makes a sizeable jump from the 16th to the 11th rank. Japan and France each drop with 4 positions in the size ranking. For the depth indicator, the shifts in national rankings are more modest. It should be noted that even small shifts in ranks can be non-trivial, especially at the top. For example, Belgium becomes ranked first rather than second in terms of scientific size. The US drops from the first to the second rank in terms of scientific depth. Due to the apparent sensitivity of these rankings to the NPR subset on which the indicators were built (i.e. NPRs versus scientific NPRs), and due to the non-triviality of even minor shifts in the rankings, one should be cautious when interpreting these indicators.

4. DISCUSSION AND CONCLUSION

Significant differences were revealed between NPR- and SNPR-based indicators. This indicates the relevance of distinguishing only scientific references, at least if one is concerned with content validity of the reference-based measures for capturing relations between

scientific and technological activities. It is known from the literature and from previous studies that NPR-based indicators are noisy, and not as straightforwardly interpretable as references between scientific articles. This is not only due to the specificities in citation practices (e.g. examiner versus applicant-given references) but it is also due to the heterogeneity present in non-patent references cited in patent documents. By means of text mining techniques and machine learning algorithms, we are able to discern scientific from non-scientific references with considerable levels of accuracy (92%). Identifying 'real' scientific references (i.e. references to the serial literature of journal articles and proceedings) is instrumental for reducing the noise in NPR-based indicators considerably.

In addition, the observed differences between the scientific size and depth indicators imply that they should be used as complementary indicators, thereby providing additional support to pleas for compound indicator systems. Third, our results underline the importance of taking into account influencing factors when interpreting SNPR-based indicators of science technology interaction. Differences between EPO and USPTO are at least partly a reflection of different underlying practices in the search for prior art. At the same time, the results show specificities relating to technology domains and applicant countries. Moreover, the patterns of these specificities change when only scientific NPRs are considered for the indicator development. Due to the consequent sensitivity of technological and national rankings to the NPR subset on which the indicators were built (i.e. NPRs versus scientific NPRs), one should be very cautious when interpreting indicators, even more so when these indicators are used to inform policy making on science-technology interactions.

REFERENCES

Callaert, J., Van Looy, B., Verbeek, A., Debackere, K., Thijs, B. (2006). "Traces of Prior Art: An analysis of non-patent references found in patent documents." Scientometrics, 69 (1), 3-20.

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

Magerman, T., Van Looy, B., Song, X. 2010, "Exploring the feasability and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications". Scientometrics, 82 (2), 289-306.

Meyer, M. (2000a), "Does science push technology? Patents citing scientific literature," Research Policy, 29, 409-434.

Meyer, M. (2000b), "What is Special about Patent Citations? Differences between Scientific and Patent Citations" Scientometrics, 49(1), 93-123.

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., and Noma E. (1985). "Is technology becoming science?" Scientometrics, 7, 369-381.

Nelson, A.J. (2009), "Measuring knowledge spillovers: What patents, licenses and publications reveal about innovation diffusion," Research Policy, 38, 994-1005.

Salton, G., Wong, A. and Yang, C.S. 1975 "A Vector Space Model for Automatic Indexing" Communications of the ACM, 18:11, 613-620.

Tijssen, R.J.W., R.K. Buter and Th.N. Van Leeuwen (2000), "Technological relevance of science: validation and analysis of citation linkages between patents and research papers", Scientometrics, 47, pp. 389-412

Van Looy, B., Callaert, J., Debackere, K., & Verbeek, A. (2002). "Patent-related indicators for assessing knowledge-generating institutions: Towards a contextualised approach," The Journal of Technology Transfer, vol. 28, no. 1, pp. 53-61

Van Looy, B., Magerman, T., & Debackere, K. (2007), "Developing technology in the vicinity of science: An examination of the relationship between science intensity (of patents) and technological productivity within the field of biotechnology," Scientometrics, 70(2), 441-458.

Van Looy, B., Zimmermann, E., Veugelers, R., Verbeek, A., Mello, J., and Debackere, K. (2003). "Do science-technology interactions pay off when developing technology?" Scientometrics, 57 (3), 355-367.

Van Vianen, B., Moed H. and Van Raan A. (1990). "An exploration of the science base of recent technology", Research Policy, 19, 61-81.
Verbeek, A., Debackere, K., Luwel, M. & Zimmermann, E. (2002). "Measuring progress and evolution in science and technology – I: The multiple uses of bibliometric indicators." International Journal of Management Reviews, 4(2), 179-211.
Verbeek, A., E. Zimmermann, P. Andries, K. Debackere, M. Luwel, R. Veugelers (2001), 'Linking Science to Technology: Bibliographic References in patents – State of the Art', Report to the European Commission, Part A, Brussels
Verspagen, B. (2008) "Knowledge Flows, Patent Citations and the Impact of Science on Technology," Economic Systems Research, 20(4), 339-266.