OECD Science, Technology and Industry Working Papers
2015/05

The Use of Patent Statistics for International Comparisons and Analysis of Narrow Technological Fields

Ivan Haščič,

Jérôme Silva,

Nick Johnstone

https://dx.doi.org/10.1787/5js03z98mvr7-en

OECD

OECD SCIENCE, TECHNOLOGY AND INDUSTRY WORKING PAPERS

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcomed, and may be sent to OECD Directorate for Science, Technology and Innovation, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France; e-mail: sti.contact@oecd.org.

The release of this working paper has been authorised by Andrew Wyckoff, OECD Director for Science, Technology and Innovation.

This paper has been authored by Ivan Haščič, Jérôme Silva (OECD Environment Directorate) and Nick Johnstone (OECD Directorate for Science, Technology and Innovation). The paper draws on past work of the Environment Directorate and its project on "Environmental policy and technological innovation" (http://www.oecd.org/environment/innovation.htm).

A version of this paper has been reviewed by the OECD Working Party on Environmental Information (WPEI) at its meeting in November 2014 and has benefited from the comments received. The authors are grateful to Dominique Guellec and Hélène Dernis for helpful comments on a previous version of this paper, as well as the participants of the *Conference on Patent Statistics for Decision Makers* (October 2013 in Rio de Janeiro, Brazil) and the *Workshop on Evaluating Green Innovation Policies* (December 2012 at Bruegel Institute in Brussels, Belgium) where previous versions of this paper were presented.

THE USE OF PATENT STATISTICS FOR INTERNATIONAL COMPARISONS AND ANALYSIS OF NARROW TECHNOLOGICAL FIELDS

by Ivan Haščič, Jérôme Silva and Nick Johnstone (OECD)

ABSTRACT

Patent data provide an increasingly used means to analyse innovation performance worldwide including in countries with incomplete data coverage, such as some developing countries. This paper discusses the specific issues associated with using patent data for measuring and analysing innovation in narrow technological fields, such as many environment-related technologies. To improve cross-country comparability of patent statistics, the paper advocates the use of indicators based on patent family size because they are more flexible and can be adapted to various applications. The paper also examines certain idiosyncratic characteristics of patent databases and proposes approaches to mitigate potential biases in empirical cross-country analyses. While doing so is particularly important for analyses of narrow technological fields such as many environment- and climate-related technologies, some of these issues are relevant for patent analysis more broadly.

Keywords: innovation, indicators, environmental technologies

JEL classification: O3; O31; O34; Q2; Q4; Q5

EXECUTIVE SUMMARY

Patent data provide an increasingly used means to analyse innovation performance worldwide including in countries with incomplete data coverage, such as some developing countries. This paper focuses on the construction of indicators and analytical methodologies for measuring innovation in climate- and other environment-related technologies. A companion paper discusses the use of such indicators for policy purposes (see Haščič and Migotto 2015).

Various approaches have previously been developed to improve cross-country comparability of patent statistics. However, they were generally limited either because they require additional data that might not be available in all countries or because they impose restrictions that are less appropriate for analysis of narrow technological fields. The simultaneity of two conditions – narrow technological scope and comparability across a wide range of countries – often renders existing indicators inapt for such analyses. This paper advocates the use of indicators based on patent family size because they are more flexible and can be adapted to such applications.

This paper also examines certain idiosyncratic characteristics of patent databases. The sources of such idiosyncrasies include unequal country and temporal coverage of patent databases, asymmetric information on patent protection in member states through regional applications, missing descriptive information in patent databases, and non-homogeneous assignment of patent classes and the consequent difficulty of identifying relevant patent documents. This suggests that all patent databases should be exploited with care. The paper proposes approaches to mitigate potential biases in empirical cross-country analyses. While doing so is particularly important for analyses of narrow technological fields such as environment and climate related technologies, many of these issues are relevant for patent analysis more broadly.

TABLE OF CONTENTS

1. INTRODUCTION

Patent data provide an increasingly used means to analyse innovation performance worldwide including in countries with incomplete data coverage, such as some developing countries. This paper focuses on the construction of indicators and analytical methodologies for measuring innovation in climate- and other environment-related technologies. A companion paper discusses the use of such indicators for policy purposes (see Haščič and Migotto 2015).

More generally, there is a growing demand for analysis of patenting and innovation performance worldwide including in countries with incomplete data coverage, such as some developing countries (see e.g., Collier and Venables 2012; Belward et al. 2011; Barton 2007). However, past analyses have not properly taken into account the varying geographic coverage of patent databases (including those by the authors of this paper). This paper builds on the first attempts to reflect upon this issue in Haščič et al. (2012). Although the sources of idiosyncrasy are discussed in the context of the EPO's PATSTAT database, similar approaches could be used to mitigate biases arising out of other patent databases.

In addition, patent data are increasingly used for international comparisons and analysis in relatively 'narrow' fields of technology, such as many environment-related technologies (see e.g. OECD 2011, 2012; Dechezleprêtre et al. 2011). However, achieving both comparability across countries and representativity within a country (statistical robustness) creates additional challenges. Although several approaches have been developed to improve cross-country comparability of patent statistics, these are not always suitable for such analyses – either because they require additional data that might not be available in all countries or because they impose restrictions that are less appropriate for analysis of narrow technological fields. The simultaneity of these two conditions – narrow technological scope and comparability across a wide range of countries – often renders many existing indicators inapt for such analyses.

This paper first discusses why different patent indicators might be more or less suitable to achieve the two objectives simultaneously. It then advocates the use of indicators based on patent family size that are more flexible and can be adapted for such applications, provided that certain idiosyncratic characteristics of patent databases are addressed. The following section discusses in detail these idiosyncratic characteristics of patent databases that are of general relevance (unequal country and temporal coverage of patent databases, asymmetric information on designation of member states in regional and PCT applications) as well as those that are of relevance particularly for narrow technological fields (missing information, identification of relevant patent documents). The paper proposes approaches to mitigating potential biases arising out of such idiosyncrasies – both in construction of patent statistics (treatment of underlying data) as well as in empirical analyses that use such statistics. In particular, it advocates the use of a robust control variable (e.g. TOTPAT) for cross-country analyses of narrow technological fields. All methodological issues discussed here and examples provided are based on the PATSTAT database, but could be applied to patent databases more generally.

2. ADAPTING PATENT INDICATORS TO DIFFERENT CONTEXTS

Many of the existing patent indicators, including counts of patent applications registered at a given national or regional patent office, Patent Cooperation Treaty (PCT) applications, or the triadic patent family (TPF) indicator, have one or both of the following shortcomings: i) either they are not suitable for international comparisons; and/or b) they are less suitable for analysis of narrow technological fields. The joint existence of two requirements on the indicators – comparability across countries and statistical robustness – often renders these indicators inappropriate for many applications. The reasons range from a home bias in national and regional filings, differences in propensities to use the PCT route (a special type of a home bias)[1], to an overly restrictive quality threshold applied precisely to address cross-country differences in patent quality (such as the TPF). Moreover, much of the complementary data that is commonly used to assess patent value (e.g. citation data) are currently available only in a very limited number of countries.

To illustrate the differences, Figures 1-3 give examples of alternative patent indicators for technological fields of different 'thickness' and for different geographical scales. Four alternative indicators are shown – PCT applications, TPF applications, "claimed priorities" (equivalent to patent family size>=2, or PF2) and all priorities (equivalent to patent family size>=1, or PF1).[2]

Given our reliance upon data obtained from a source which does not include the full population of patented inventions within a given field (see below) we do not have any external measure against which we can assess which statistic is optimal. However, in order to assess the reliability of the use of alternative statistics as indicators of inventive activity, and their suitability for empirical analysis it is interesting to compare: a) the degree of correlation between alternative measures; b) the extent of variation within fields over time and across countries; and, c) the presence of zero values.

To measure inventive activity worldwide in relatively 'thick' fields – several of the indicators seem satisfactory (for instance, see renewable energy and water pollution abatement in Figure 1). The different measures are highly correlated, and at the country-level the variation is 'plausible" for most countries. The situation changes with a focus on narrower technological fields (e.g. wind power, geothermal energy) which compromises performance of the most restrictive versions of these indicators. For example, in the case of geothermal energy, the frequency of zero counts is so high that it limits the information value of the TPF indicator (Figure 1).

These problems are accentuated when the geographic scope is reduced. Focusing on individual countries results in high frequency of zero counts for all but the largest inventor countries (Figures 2-3) Indeed, even for thick fields some of the more restrictive counts (e.g. TPF and PF2) give relatively low variation at the country level (Figures 2 and 3).

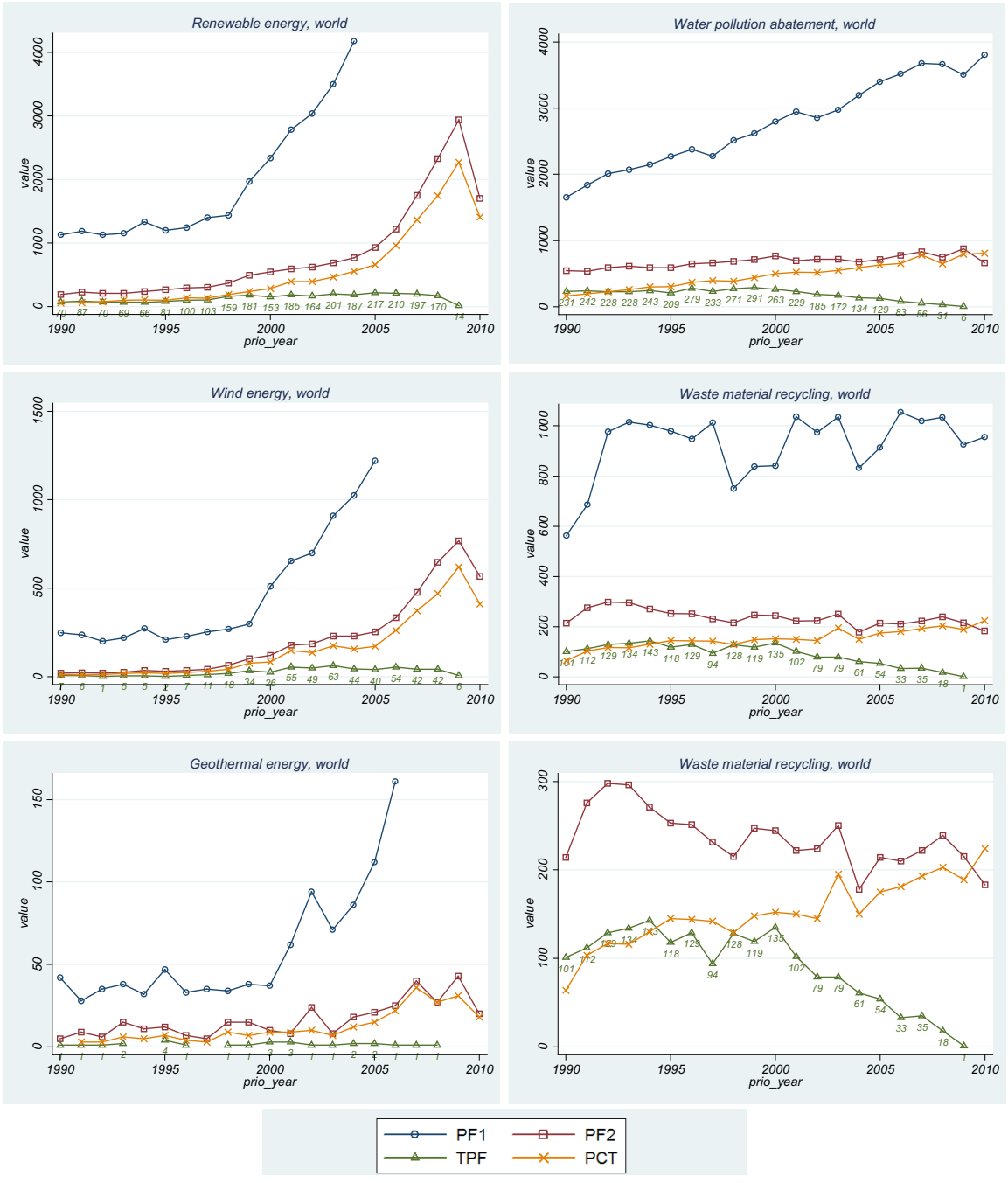
In general, PF2 seems as a suitable compromise between comparability (quality threshold) and practicality (variance), except for very narrow technological fields or small inventor countries – in which case PF1 might be the only option.

In sum, while restricting patent family size is a useful way to impose a certain quality threshold across countries (Dernis et al. 2001), the TPF condition of a specific EP-US-JP triad (Dernis and Khan 2004) is overly restrictive and potentially biased for the contexts discussed here. In such cases, it might be better to relax this restriction and instead adjust it depending on the application. While for some applications no restriction will be the only option (PF1), for others the dyadic (any office, or PF2) restriction might be feasible, and cases of particularly "thick" technological fields or broad geographic scopes will permit imposing the triadic (any office, or PF3) restriction, or higher. For these

reasons, this paper advocates the use of a range of patent family indicators, as a generalization of the TPF concept.

It must be noted that PF3 differs from the TPF indicator in two respects – (i) the TPF requires a specific triad (EP-US-JP) while PF3 can include any triad; (ii) the TPF is based on an extended patent family while PF3 is based on simple patent family. It would be a natural extension of the TPF concept to define a "dyadic patent family" (DPF) indicator as a dyad, perhaps any dyad of two offices. This would however require addressing the problem of regional and international patent filings (see Section 3 for a discussion).

Figure 1. Patent indicators in different contexts: Technological scope



Note: Figures in the left-hand panel are dominated by PF1; therefore the y-axis is reduced cutting off PF1 values for later years in the time series to allow better visibility of the remaining series.

Figure 2. Patent indicators in different contexts: Inventor country

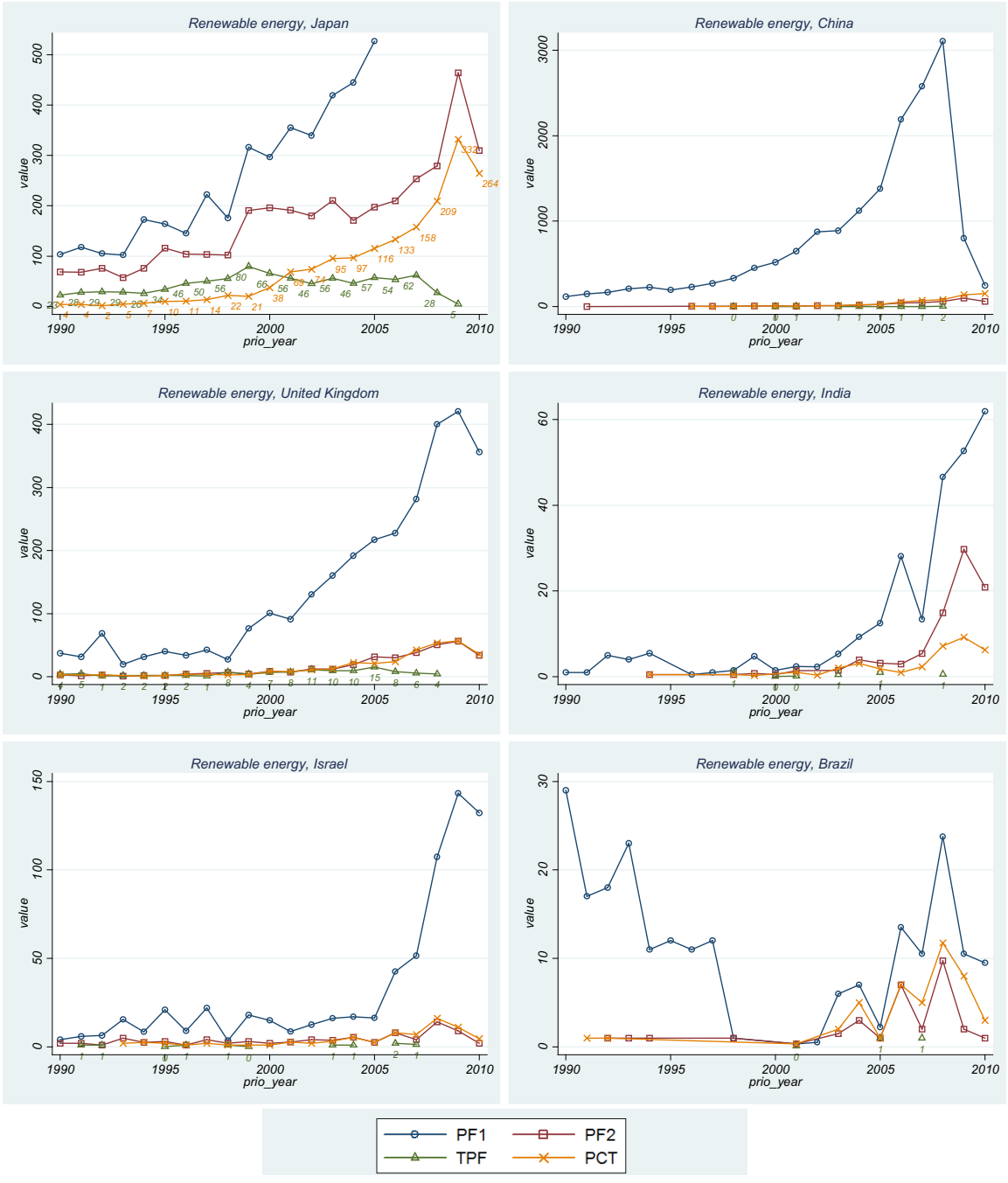


Figure 3. Patent indicators in different contexts: Inventor country

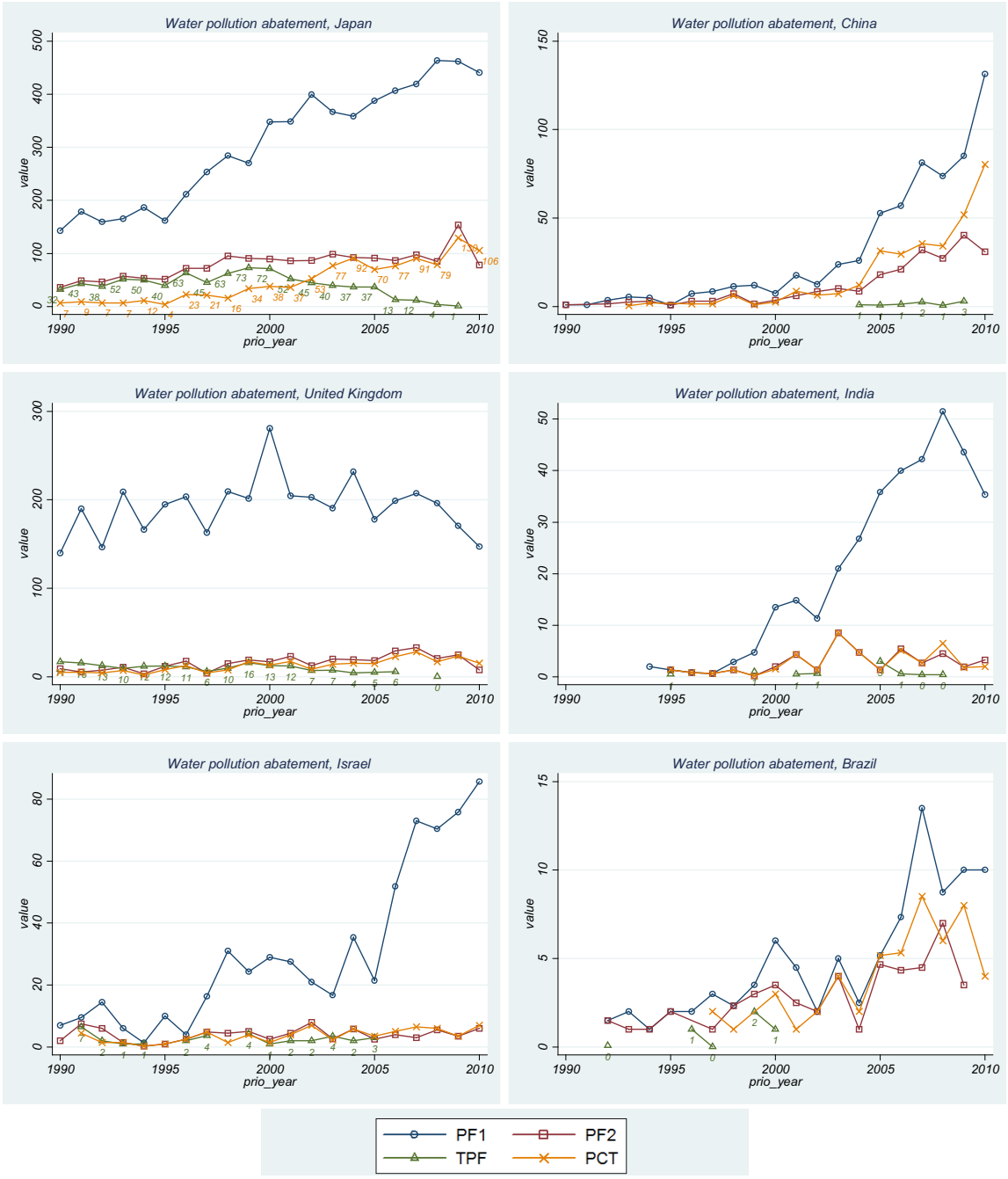


Table 1 summarizes the proposed range of indicators and places them in the context of the existing indicators.

Table 1. Patent indicators

Indicator	Definition	Main strengths and weakness	Purpose
National applications	*Singletons, claimed priorities and duplicates*	*Commensurate in terms of patent breadth and quality; subject to a home bias.*	*Measures of patenting activity (based on counting applications or grants)*
Regional applications	*Singletons, claimed priorities and duplicates*	*As above; difficulty with designation of member states.*	
PCT applications	*Singletons, claimed priorities and duplicates*	*Subject to its own kind of 'home bias'; propensity to use the PCT route varies across countries.*	
Triadic patent family (TPF)	*Triad (incl. EP, JP, US grants); based on extended patent family.*	*Only high-quality inventions; too restrictive for narrow tech fields.*	*Measures of inventive activity (based on counting patent families)*
PF1 – All priorities	*Family size >=1, any office, based on simple patent family (singletons and claimed priorities).*	*A complete picture of inventive activity worldwide because the entire stock of patent priorities is considered; too lax because it gives equal weight to low- and high-quality inventions; does not account for differences in patent breadth across offices.*	
PF2 – Claimed priorities	*Family size >=2, any office, based on simple patent family.*[3]	*A limited quality threshold suitable for many narrow tech fields.*	
PF3	*Family size >=3, any office, based on simple patent family.*	*Only high-quality inventions, but less restrictive than the TPF.*	

For the purpose of international comparisons, patent statistics based on counting distinct patent families (patent priorities) are preferable for several reasons: (i) considering only priority applications (and not their duplicates) avoids double-counting – which would occur if data from multiple patent offices were pooled. The data is thus better suited for cross-country analysis; (ii) the data are truly world-wide in their coverage – and thus less subject to bias – because the entire stock of patent priorities is considered (see also de Rassenfosse et al. 2013).[4]

In addition, a subset of these statistics – those based on counting only the 'claimed priorities' (i.e. patent priorities that have been 'claimed' as a priority by another patenting filing abroad) provides a quality threshold because priority applications which have never been claimed (singletons) are excluded. This helps contain concerns over low-value and strategic patenting. Counting the claimed priorities gives rise to the PF2 statistic (i.e. patent families with size 2 or greater).

It turns out that for many applications related to environmental technologies, the PF2 restriction (claimed priorities) is particularly suitable in order to achieve the dual purpose of comparability and robustness. The count of claimed priorities (i.e. patent applications deposited at any office world-wide, that have been claimed as priority elsewhere in the world) is a suitable statistic for the purpose of international comparisons because only the priority applications protecting "high-value" invention are

counted. The reason that claimed priorities can be viewed as representing inventions of higher value is that patenting is costly (e.g. translation and maintenance fees). As such, a firm will only go abroad to protect its intellectual property if it expects that the commercial value of its invention justifies it. Previous research has shown that the number of additional patent applications (other than the priority application) is a good indicator of patent value (Guellec and van Pottelsberghe 2000; Harhoff et al. 2003). The results in Guellec & van Pottelsberghe (2000) suggest that patent value increases up to family size of 4 (or 5), and decreases thereafter. However, few patent applications have family size greater than five.

The use of claimed priorities based on an economic threshold criterion was advocated already by Faust and Schedl 1983 and Faust 1990. For example, Faust (1990) argued that by excluding priority applications which have never been claimed abroad (singletons) this approach will exclude the large number of exclusively domestic Japanese patent applications with usually only one claim.[5]

Moreover, counts of claimed priorities (PF2 statistics) have the advantage that they do not suffer from biases due to changing publication practices at the USPTO. Prior to 2001, only grants (not applications) were published by the USPTO. Even after 2001, the rules allow applicants to opt-out from having their patent applications made public before grant (i.e. the non-publication request). These distinctions between grants and applications create biases in cross country comparisons when using the PF1 statistic. This is a limitation of using PF1 statistics compared with PF2, PF3, etc. that do not suffer from such bias.

The downside of counting claimed priorities is the duplication lag that weighs more on indicators with higher minimum family size. Another potential problem for determining family size is the question of how to treat applications filed through the regional (EP, AP) and international (PCT) routes. For example, should a regional application have an equal weight as a national application (e.g. treating EPO as a single jurisdiction), or should it rather be taken to represent individual member states? In the latter case how should the individual weights be determined?[6] This is similar to the problem of 'nowcasting' PCT applications in international phase towards national/regional phase. And again, should PCT applications in national/regional phase have an equal weight as a national application, or should a weight be estimated? We propose a solution in Section 3.2 below.

While the PF2 is indeed often a good compromise between the too restrictive TPF and the too lax count of all priorities (PF1), in some circumstances PF1 can be the only practical approach – e.g. for very narrow technological fields or for analysis requiring a high degree of spatial disaggregation. This approach can be generalized by allowing the analyst to choose the degree of restriction – the number of unique countries (patent offices) within a simple patent family – the minimum 'quality' threshold that still allows for sufficient variation in the measured statistic (avoids too many zeros). The choice of the restriction is then a function of 'scope' (narrow vs thick) of the technological field analysed, geographic coverage, as well as country context (propensity to protect inventions abroad). However, construction of a reliable indicator also requires addressing a certain number of potential pitfalls associated with patent databases. They are discussed next.

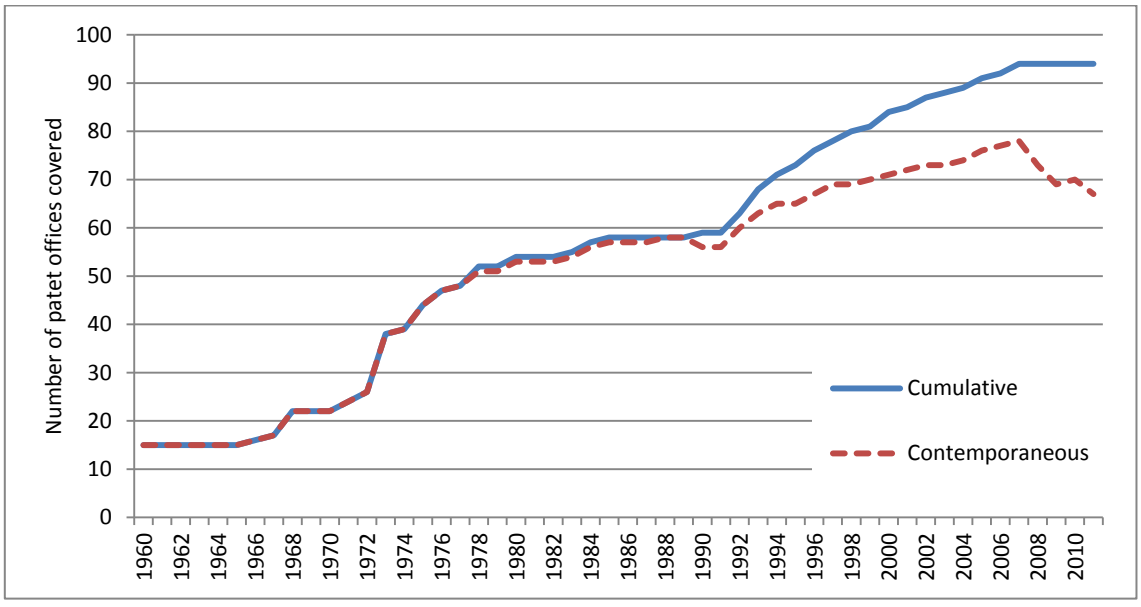
3. IDIOSYNCRATIC ISSUES IN CONSTRUCTION AND ANALYSIS OF PATENT STATISTICS

3.1 Country and temporal coverage of patent databases

Analyses using patent data have typically focused on OECD countries for which data availability is generally satisfactory. However, the recent growing interest in developing countries raises the importance of idiosyncrasies arising out of unequal database coverage which often varies highly among countries worldwide. Accounting for such idiosyncrasies is all the more important for analyses that study narrow technological fields because the high level of disaggregation leads to high frequency of zero (or very low) counts.

Ideally, one would need information on the representativeness of PATSTAT with respect to national registers. In the absence of such information, we use the EPO's weekly updates of the "contents and coverage of the DOCDB bibliographic file" – the DOCDB is the source database from which a 'snapshot' is drawn biannually and released as PATSTAT. The April 2012 release of PATSTAT includes batches of data provided from as many as 94 different patent offices, including all major patent offices worldwide. This is shown as cumulative coverage in Figure 4. However, for some offices the time series are incomplete and thus the contemporaneous coverage (for a given year) is typically somewhat lower. The apparent drop in coverage in recent years is most likely only a temporary phenomenon because some batches of new data are included with a lag. On-going efforts of the EPO to acquire historical batches of data would shift the curves upward, and bring them closer together.

Figure 4. Contemporaneous and cumulative data coverage in PATSTAT APR12



Note: Excluding data batches for utility models or petty patents.

The status of coverage in PATSTAT can be classified in four categories: (i) For 1980-2009 there are 39 offices with **complete coverage** during the entire 30-year period (e.g. JP, US, EP, IB); (ii) Further 16 offices have a complete coverage although defined for less than 30 years (e.g. DD, CS, YU)[7]; (iii) Other 40 offices have only **partial coverage** of patent applications deposited in any given year, for instance only certain months in a year or only certain types of patent applications (e.g. AR,

IN, MA); (iv) Finally, the remaining offices or countries have **no coverage**, theoretically, although for various reasons PATSTAT includes data for additional 19 offices (e.g. inclusion of data on prior art), meaning that there is data on an additional 129 inventor countries (Table 2).

Coverage of OECD countries is generally complete, although some gaps exist, notably for Chile (1980-2004 and 2009 are not covered at all, 2005 and 2008 only partly). Among non-OECD countries it is noteworthy that countries such as Brazil, South Africa, Egypt and Russia have complete coverage. However, gaps in coverage exist for some other countries, such as Argentina (1995-96 are only partly covered), India (2007 is only partly covered, 2008-09 not at all), China (1980-84 are not covered at all, 1985 only partly), Indonesia (1980-95 and 2002-09 are not covered at all, 1996-97 only partly) and Philippines (1999 only partly covered, 2000-09 not at all).

Table 2. Theoretical and empirical data coverage in PATSTAT APR12

Country/office		Theoretical coverage, data batches for 1980-2009		Empirical coverage, 1980-2009	
				as Application Authority(1)	as Inventor Country(2)
JP	Japan	30	complete	10,658,985	3,908,024
US	United States	30	complete	5,449,125	5,933,195
EP	European Patent Office [EPO]	30	complete	2,707,381	-
DE	Germany	30	complete	2,574,630	3,090,843
KR	Korea (South)	30	complete	1,921,667	1,633,271
AU	Australia	30	complete	982,172	144,914
CA	Canada	30	complete	963,367	366,256
SU	Soviet Union (former)	30	complete	766,954	544,381
AT	Austria	30	complete	663,518	183,190
GB	United Kingdom	30	complete	517,375	905,490
ES	Spain	30	complete	516,678	155,283
FR	France	30	complete	484,821	1,113,411
BR	Brazil	30	complete	408,144	74,490
DK	Denmark	30	complete	192,652	126,214
MX	Mexico	30	complete	176,070	17,390
IT	Italy	30	complete	165,372	509,374
PL	Poland	30	complete	157,305	69,593
NO	Norway	30	complete	150,368	70,045
IL	Israel[8]	30	complete	132,758	127,959
ZA	South Africa	30	complete	132,370	23,184
SE	Sweden	30	complete	116,434	332,107
FI	Finland	30	complete	109,235	189,493
NZ	New Zealand	30	complete	99,370	24,980
HU	Hungary	30	complete	93,013	49,931
NL	Netherlands	30	complete	89,231	383,733
PT	Portugal	30	complete	87,305	8,962
IB	International Bureau of the WIPO	30	complete	77,984	-
HK	Hong Kong, China	30	complete	69,446	23,408
CH	Switzerland	30	complete	66,335	380,553
IE	Ireland	30	complete	46,544	34,043
TR	Turkey	30	complete	43,447	27,571
RO	Romania	30	complete	42,023	31,733
BE	Belgium	30	complete	41,252	158,641
BG	Bulgaria	30	complete	32,092	23,625
EG	Egypt	30	complete	8,474	2,285
LU	Luxembourg	30	complete	7,300	12,507
CY	Cyprus[9]	30	complete	1,370	1,396
MC	Monaco	30	complete	957	2,722
LI	Liechtenstein		complete	-	5,737
AP	African Regional IPO [ARIPO][3]	25.2	complete	5,524	-
YU	Yugoslavia – Serbia/Montenegro	23.3	complete	14,576	8,939
DD	Eastern Germany (former)	19.5	complete	105,734	85,185
LT	Lithuania	17.2	complete	3,745	2,484
SI	Slovenia	17.1	complete	29,838	10,079
RU	Russian Federation	16.9	complete	479,655	243,024
CZ	Czech Republic	16.8	complete	67,030	42,134
SK	Slovak Republic	16.2	complete	23,904	7,731
LV	Latvia	15.8	complete	4,259	2,854
MD	Republic of Moldova	15.6	complete	4,486	4,430
HR	Croatia	15.4	complete	11,924	6,393
EE	Estonia	14.0	complete	6,164	4,089
CS	Czechoslovakia (former)	14.0	complete	78,257	49,765

EA	Eurasian Patent Org. [EAPO][4]	13.5	complete	23,976	-
GE	Georgia	10.0	complete	3,693	2,993
RS	Republic of Serbia	3.2	complete	4,708	1,923
GR	Greece	29.4	partial	50,567	11,333
AR	Argentina	28.7	partial	59,471	8,662
GT	Guatemala	27.6	partial	1,010	212
IN	India	27.4	partial	41,207	73,617
SG	Singapore	26.6	partial	54,160	30,356
CN	China	24.3	partial	3,587,728	2,141,368
EC	Ecuador	20.0	partial	7,098	678
CU	Cuba	19.9	partial	1,688	2,763
PH	Philippines	19.2	partial	14,701	3,050
PE	Peru	17.3	partial	10,752	1,116
MA	Morocco	16.2	partial	11,621	1,728
IS	Iceland	16.2	partial	5,481	3,562
CO	Colombia	14.9	partial	13,458	3,040
MW	Malawi	14.8	partial	428	18
OA	African IP Organisation [OAPI][5]	14.7	partial	6,819	-
ZM	Zambia	14.4	partial	788	42
ZW	Zimbabwe	14.4	partial	2,094	268
PA	Panama	13.6	partial	2,411	1,002
VN	Viet Nam	12.8	partial	187	775
MT	Malta	12.4	partial	126	516
TW	Chinese Taipei	12.2	partial	583,999	566,918
SM	San Marino	10.0	partial	190	210
SV	El Salvador	9.8	partial	1,329	320
KE	Kenya	9.7	partial	557	438
MN	Mongolia	9.5	partial	108	210
MY	Malaysia	9.2	partial	6,321	9,516
DO	Dominican Republic	8.8	partial	982	230
TJ	Tajikistan	8.7	partial	386	1,111
UA	Ukraine	8.6	partial	48,114	54,103
HN	Honduras	5.0	partial	235	117
ID	Indonesia	5.0	partial	14,326	2,146
GC	Patent Office of the Gulf Coop. Council[6]	4.4	partial	400	-
NI	Nicaragua	4.4	partial	199	36
KZ	Kazakhstan	4.2	partial	123	1,790
CL	Chile	3.8	partial	3,445	2,739
BA	Bosnia and Herzegovina	3.5	partial	264	341
DZ	Algeria	3.3	partial	1,391	586
CR	Costa Rica	3.0	partial	3,443	814
UY	Uruguay	2.2	partial	5,948	1,177
BY	Belarus	0.003	partial	166	5,180
AM	Armenia	-	none	55	548
AZ	Azerbaijan	-	none	51	1,252
MK	Macedonia (FYROM)	-	none	44	188
SD	Sudan	-	none	31	162
SY	Syrian Arab Republic	-	none	30	276
KP	DPR of Korea (North)	-	none	22	751
TN	Tunisia	-	none	22	1,128
UZ	Uzbekistan	-	none	15	1,036
LK	Sri Lanka	-	none	12	811
TT	Trinidad and Tobago	-	none	9	356
GH	Ghana	-	none	6	230
KG	Kyrgyzstan	-	none	6	362
TH	Thailand	-	none	5	3,561

BB	Barbados	-	none	2	685
LR	Liberia	-	none	2	28
AL	Albania	-	none	1	147
BZ	Belize	-	none	1	109
LS	Lesotho	-	none	1	10
LY	Libya	-	none	1	34
VE	Venezuela	-	none	-	2,682
SA	Saudi Arabia	-	none	-	2,285
IR	Iran	-	none	-	2,255
TM	Turkmenistan	-	none	-	1,650
AE	United Arab Emirates	-	none	-	1,021
LB	Lebanon	-	none	-	933
BD	Bangladesh	-	none	-	652
NG	Nigeria	-	none	-	460
BO	Bolivia	-	none	-	220
TZ	Tanzania	-	none	-	94
	…and other >100 countries and territories	-	none	-	

Notes:

1 Based on TOTPAT #7 counts (singletons + claimed priorities + duplicates)

2 Single-priority patent families, fractional counts

3 The African Regional Intellectual Property Organisation (ARIPO) has 17 member states: Botswana, Gambia, Ghana, Kenya, Lesotho, Malawi, Mozambique, Namibia, Rwanda, Sierra Leone, Somalia, Sudan, Swaziland, Uganda, Tanzania, Zambia and Zimbabwe.

4 The Eurasian Patent Organization (EAPO) has 8 member states: Turkmenistan, Belarus, Tajikistan, Russia, Azerbaijan, Armenia, Kazakhstan and Kyrgyzstan.

5 The African Intellectual Property Organisation (OAPI) has 15 member states: Benin, Burkina Faso, Cameroon, Central African Republic, Chad, Congo, Côte d'Ivoire, Gabon, Guinea, Guinea-Bissau, Mali, Mauritania, Niger, Senegal and Togo.

6 The Patent Office of the Cooperation Council for the Arab States of the Gulf (GCC) has 6 member states: United Arab Emirates, Bahrain, Saudi Arabia, Oman, Qatar and Kuwait.

This is a wealth of information from a large number of patent offices. However, given the unequal coverage across countries/offices and over time, the question is how to use this body of data for statistical analysis. Clearly, 'raw' patent counts that do not account for the unequal coverage should be interpreted with caution, particularly those with emphasis on countries and time periods for which the coverage is incomplete. Moreover, while country coverage is an important piece of information to assess patenting trends in general, it is particularly so with respect to developing countries for whom the coverage varies most.

The bias might be particularly important with respect to statistics of market protection (number of patent applications deposited at a national office). Implications for invention and co-invention indicators are less acute since some of this information can be obtained indirectly through other IP offices.

Using coverage weights to generate more reliable patent statistics

In this paper we take a step towards developing a methodology to account for differences in coverage. The detailed metadata described above can be used for this purpose and improve the reliability and representativeness of patent statistics. In particular, our intention is twofold:

i) Identify cases when one can reliably distinguish 'true' zero counts from **'missing' values**. As explained above, benefits of doing so are particularly high for analyses of narrow technological fields and of countries with low levels of patenting activity (e.g. many developing countries);

ii) In cases when values are not missing, the intention is to provide an indication of **reliability** of the observed frequency counts. Again, this is important in data-poor environments,

including those of narrow technological fields or countries with relatively low levels of patenting activity.

We use the PATSTAT metadata to develop coverage weights to accompany standard patent data extractions. In some sense, the weights will give the probability that the frequency counts obtained from PATSTAT reflect the 'true' count. For example, one can reasonably expect that if the coverage metadata indicate than no data batches were included in PATSTAT from patent office in country X, then the probability to have records of duplicates deposited at this country's office will be zero or very low (for singleton priorities it will indeed be zero) and for claimed priorities it will be low but non-zero. This approach will allow choosing the minimum level of coverage that is desired or appropriate for a given application (i.e. it allows sample selection).

First, we summarize the information provided by the EPO in the "contents and coverage of DOCDB bibliographic files" and construct a statistic that represents the fraction of a given year for which PATSTAT has batches of data from a given patent office, ranging between "0" (office *i* in year *t* not covered at all) and "1" (full coverage) (e.g. if data is available only for 3 months of a year ten weight=0.25). Table 3 gives the weights for offices with at least a partial coverage.

Table 3. Coverage weights of patent offices in PATSTAT APR12, 1980-2009

	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
AP	0	0	0	0	0.167	1	1	1	1	1	1	1	1	1	1	1
AR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.630
AT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
AU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BG	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CH	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CN	0	0	0	0	0	0.310	1	1	0.992	0.992	1	1	1	1	1	1
CO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.882
CR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CS	1	1	1	1	1	1	1	1	1	1	1	1	1	0.956	0	0
CU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.945
CY	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CZ	0	0	0	0	0	0	0	0	0	0	0	0	0	0.795	1	1
DD	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DK	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DZ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EC	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
EE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.047
EG	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
EP	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ES	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
FI	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
FR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GB	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GR	1	1	1	1	1	1	1	1	0.352	1	1	1	1	1	1	1
GT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
HK	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
HN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.392	1
HU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ID	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
IL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
IN	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
IS	0	0	0	0	0	0	0	0	0	0	0	0	0	0.189	1	1
IT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 3. (cont.)

	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
JP	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
KE	1	1	1	1	1	1	1	1	1	0.668	0	0	0	0	0	0
KR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
KZ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LT	0	0	0	0	0	0	0	0	0	0	0	0	0.213	1	1	1
LU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
LV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.814	1
MA	0	0	0	0	0	0	0	0	0	0	0	0	0	0.197	1	1
MC	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.589	1
MN	1	1	1	1	1	1	1	1	1	0.455	0	0	0	0	0	0
MT	1	1	1	1	1	1	1	1	1	1	1	1	0.352	0	0	0
MW	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.781	0
MX	0.997	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MY	1	1	1	1	1	1	1	1	1	0.205	0	0	0	0	0	0
NI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NO	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NZ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
OA	0	0	0	0	0	0	0	0	0	0	0	0	0.380	1	1	1
PA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PE	0	0	0	0	0	0	0	0	0	0	0	0	0.254	1	1	1
PH	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
RO	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
RS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RU	0	0	0	0	0	0	0	0	0	0	0	0	0	0.877	1	1
SE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SG	0	0	0	0.849	1	1	1	1	1	1	1	1	1	1	1	0.762
SI	0	0	0	0	0	0	0	0	0	0	0	0	0.096	1	1	1
SK	0	0	0	0	0	0	0	0	0	0	0	0	0	0.167	1	1
SM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
UA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
US	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
UY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VN	0	0	0	0	0.489	1	1	1	1	1	1	1	1	1	1	1
WO	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
YU	1	1	1	1	1	1	1	1	1	1	1	1	0.407	0	0	0
ZA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ZM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.397	0
ZW	0.328	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.068

Table 3. (cont.)

	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	1980-2009
AP	1	1	1	1	1	1	1	1	1	1	1	1	1	1	25.167
AR	0.077	1	1	1	1	1	1	1	1	1	1	1	1	1	28.707
AT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
AU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
BA	0	0	0.825	1	1	0.704	0	0	0	0	0	0	0	0	3.529
BE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
BG	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
BR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
BY	0	0.003	0	0	0	0	0	0	0	0	0	0	0	0	0.003
CA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
CH	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
CL	0	0	0	0	0	0	0	0	0	0.984	1	1	0.814	0	3.798
CN	1	1	1	1	1	1	1	1	1	1	1	1	1	1	24.293
CO	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14.882
CR	0	0	0	0	0	0	0	0	0	0	0	0.992	1	1	2.992
CS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13.956
CU	0	0	0	0	0	0	0	0	0	0	0.948	1	1	1	19.893
CY	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
CZ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	16.795
DD	1	1	1	0.537	0	0	0	0	0	0	0	0	0	0	19.537
DE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
DK	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
DO	0	0	0	0	0	0.756	1	1	1	1	1	1	1	1	8.756
DZ	0	0	0	0	0	0	0.934	1	1	0.408	0	0	0	0	3.342
EA	0.503	1	1	1	1	1	1	1	1	1	1	1	1	1	13.503
EC	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20.000
EE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14.047
EG	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
EP	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
ES	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
FI	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
FR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
GB	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
GC	0	0	0	0	0	0	0.173	1	1	1	1	0.247	0	0	4.419
GE	0	0	0	0	0.975	1	1	1	1	1	1	1	1	1	9.975
GR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	29.352
GT	1	1	1	1	1	1	1	1	1	1	1	0.636	0	0	27.636
HK	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
HN	0	0	0	0	0	0	0	0	0	0.967	1	1	1	1	4.967
HR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15.392
HU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
ID	0.478	0.479	1	1	1	1	0.008	0	0	0	0	0	0	0	4.966
IE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
IL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
IN	1	1	1	1	1	1	1	1	1	1	1	0.359	0	0	27.359
IS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	16.189
IT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000

Table 3. (cont.)

	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	1980-2009
JP	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
KE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9.668
KR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
KZ	0	0	0	0	0	0	0	0	0.874	1	1	1	0.372	0	4.246
LT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17.213
LU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
LV	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15.814
MA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	16.197
MC	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
MD	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15.589
MN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9.455
MT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12.352
MW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14.781
MX	1	1	1	1	1	1	1	1	1	1	1	1	1	1	29.997
MY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9.205
NI	0	0	0	0	0	0	0	0.156	1	1	1	1	0.230	0	4.386
NL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
NO	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
NZ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
OA	1	1	1	1	1	1	1	1	1	1	1	0.282	0	0	14.662
PA	0.617	1	1	1	1	1	1	1	1	1	1	1	1	1	13.617
PE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17.254
PH	1	1	1	0.167	0	0	0	0	0	0	0	0	0	0	19.167
PL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
PT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
RO	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
RS	0	0	0	0	0	0	0	0	0	0	0.181	1	1	1	3.181
RU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	16.877
SE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
SG	1	1	1	1	1	1	1	1	1	1	1	1	1	1	26.611
SI	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17.096
SK	1	1	1	1	1	1	1	1	1	1	1	1	1	1	16.167
SM	0	0	0	0	1	1	1	1	1	1	1	1	1	1	10.000
SU	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
SV	0	0	0	0	0.801	1	1	1	1	1	1	1	1	1	9.801
TJ	0	0	0.132	1	1	1	1	1	1	1	1	0.540	0	0	8.671
TR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
TW	0	0.167	1	1	1	1	1	1	1	1	1	1	1	1	12.167
UA	0	0	0	0.225	1	1	1	1	1	1	1	1	0.402	0	8.626
US	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
UY	0	0	0	0	0	0	0	0	0	0	0	0.170	1	1	2.170
VN	1	0.315	0	0	0	0	0	0	0	0	0	0	0	0	12.804
WO	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
YU	0.981	1	1	1	1	1	1	1	1	1	0.956	0	0	0	23.344
ZA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30.000
ZM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14.397
ZW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14.396

Second, we define a set of rules to assign these "coverage weights" to different types of patent documents extracted from PATSTAT (Table 4). On the one hand, any information derived from a singleton patent application will be assigned a weight corresponding to the coverage of its priority application office. The weight is equal to the theoretical coverage of the priority application office in the application year. On the other hand, information derived from a family of patent applications (claimed priorities) is assigned a weight corresponding to maximum weight associated with the priority and/or duplicate application offices of corresponding family members. And finally, information derived from duplicate patent applications due to absence of data from the priority office is assigned a weight corresponding to the duplicate application office.

Table 4. Assignment rules for coverage weights

Document type	Coverage weight
Singleton priority	w^{PRIO}
Claimed priority	$max\ (w^{PRIO}, w^{DUPL})$
Duplicate application	w^{DUPL}

The unequal country coverage has only a limited effect on our ability to identify inventions that have sought protection elsewhere ('claimed priorities'), because we can impute inventor information from other members of the same patent family whose coverage might be superior. However, it does potentially compromise our ability to identify inventions that have sought protection only in one jurisdiction (singletons) with incomplete coverage, or duplicate applications of foreign patents (duplicates) registered in jurisdictions with incomplete coverage.

It must be noted that the approach presented here has certain limitations: First, the coverage weights as derived here represent only the direct coverage of countries in PATSTAT. Indirect coverage, derived by exploiting information on patent families, is not reflected (see e.g. the lower part of Table 2). Second, the coverage weights are only assigned to patent offices, not inventor countries (this would help with interpretation of the last column in Table 2). In both cases, addressing these issues would require estimating such propensities econometrically. However, this would require access to data external to PATSTAT.

Third, we generate frequency counts along selected vectors – by office, inventor country, priority or application year, technology, as well as the coverage weight. This allows making judgements about the reliability of observed patent counts, notably to:

- Identify 'true zeros' – i.e. when patent count=0 and coverage weight=1 (or close to 1). Such distinction between 'true zeros' and 'missing values is otherwise not possible.

- Identify nonzero counts with 'low reliability' – i.e. when patent count>0 and weight<1 (or far from 1). In some cases, one might observe a positive count and a zero coverage weight at the same time. Such counts arise through 'indirect' channels discussed above, and should be considered as significant underestimates. Among other things, this allows selecting a data sample based on a minimum threshold value of coverage. For example, consider only those counts derived from information that is 95% complete. However, applying such a 'filter' might involve a trade-off between geographical scope and data quality.

3.2 Protection in national jurisdictions through regional (international) patent filings

Regional IP offices play a more-or-less important role in the world. For example, patenting at the European Patent Office (EPO) is an increasingly important route of protecting IP rights in Europe. Similarly, patenting at the two African regional IP offices (ARIPO, OAPI) also plays an important role.[10] However, based on a regional application it is difficult to infer the true protection strategy of a patentee. As a consequence, when constructing patent indicators researchers frequently either treat EPO and other regional filings as if they were equivalent to national applications or simply set them aside as if they never existed.[11]

To address the potential biases, in this paper we propose a method to account for filings at the EPO by estimating the "protection propensities". We then use the propensities to apportion EPO counts onto national jurisdictions. And while here we focus on the EPO, this approach could potentially be extended to other regional offices subject to data availability.

Apportionment of EPO patent filings using protection propensities

There are several potential sources of data that are potentially useful for inferring patentee's market protection strategies, including (a) list of designated states using the PAT_EP database – a straightforward but not satisfactory solution because designation rules have changed over time and certain countries are designated 'by default' reducing the value of such data; (b) publication kind codes in PATSTAT that identify grants of patents – while recent additions to PATSTAT have made this approach more convenient[12], it is not clear whether a validation of an EP patent in a member state is always associated with a corresponding record in PATSTAT; and (c) payment of validation and maintenance fees – the major conceptual upside is that the data should be a reliable indicator of patentees' preferences because they are associated with financial cost. The latter approach is discussed next.

In this paper, we use the European Patent Office's Inpadoc Legal Status database, also called Patent Register Service (PRS) Legal Status database, to construct "protection propensities" over time. The database contains data from 41 patent granting authorities going back to 1970. It contains over 120 million records, each referring to a distinct "legal status event", identified by a PRS code. The great advantage is that the PRS records are linked to PATSTAT.

There are several PRS codes that are potentially useful as evidence that patent applicants have designated, validated and maintained their patent rights in certain EPO member states against payment of fees.

In this paper we use data on post-grant fee payment (PGFP) to calculate the protection propensity (other alternatives that were also considered are summarized in the Annex). The major advantage of PGFP is that the data should be complete because patent offices have an interest in maintaining reliable records of fee payments; however, we identify several years and offices for which data in the PRS database are missing or appear incomplete.[13] Another potential downside is the time lag due to granting of a patent and fee payment which causes problems for generating timely statistics. Therefore, we cannot calculate the protection propensities for the specific subset of relevant patents; instead we do so for all patents (PATSTAT total). More formally, **the protection propensity is calculated as the ratio of a frequency count of EP applications that enter the national phase in a given member state as evidenced by payment of post-grant fees, and the frequency count of all (distinct) EP applications with PGFP evidence from any state.** The protection propensity thus shows how often a typical EP application truly sought protection in a given member state (the sum of the propensities is thus more than 100%). We construct such propensities for all member states for the

time period starting with the year when they joined the EPO (Table 5). Propensities for selected offices are also shown in Figure 5.

The propensities implied by this statistic seem to be of reasonable magnitudes. Overall, four broad groups of member states can be distinguished: (i) the most frequently protected countries with propensities close to 50%, including Germany, the United Kingdom and France, and followed by Italy (about 30%); (ii) mid-size markets with propensities between 10% and 20%, including Netherlands, Switzerland, Belgium, Sweden, Austria, and Spain; (iii) smaller markets with propensities between 1% and 10%, including markets with stable propensities over time (e.g. Greece, Ireland, Portugal) as well as more dynamic markets whose propensities are growing quickly (e.g. Turkey, Finland, Poland); and finally (iv) markets with propensities of less than 1%.

Note that over time there has been a progressive move from single country designation to automatic designation of eventually all EPO members in EPO applications after 2004. The legalistic interpretation would thus lead to apportioning all EPO member countries to every EPO application after 2004. However, such automatic (or default) list of designated countries inflates the "true" intentions of the applicants. Instead, in this paper we adopt the economic interpretation and seek to approximate what markets applicants really care about. Our objective is to apportion an EPO filing only to those member states where the applicant truly seeks to protect the invention. Evidence of post-grant fee payment is thus a suitable option.

Table 5. Protection propensities in EPO member states (1990-2012, 3-year moving average)

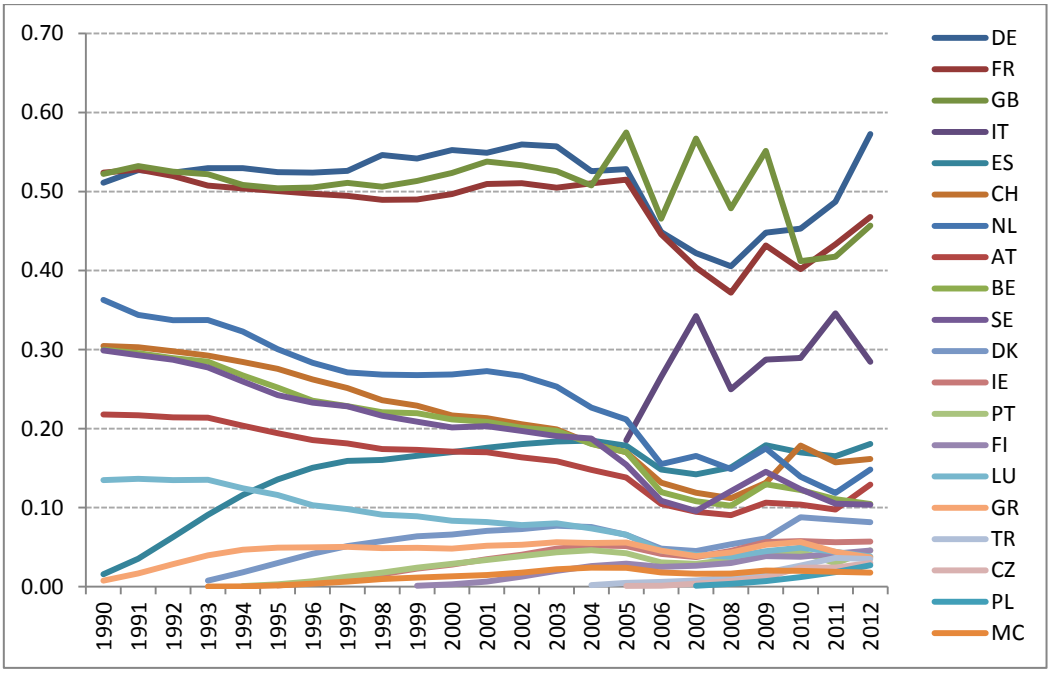
	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
DE	0.511	0.527	0.524	0.530	0.530	0.524	0.524	0.526	0.546	0.542	0.553	0.549
GB	0.522	0.532	0.525	0.522	0.508	0.504	0.505	0.511	0.506	0.513	0.524	0.538
FR	0.524	0.528	0.520	0.507	0.504	0.501	0.497	0.495	0.489	0.490	0.497	0.510
NL	0.363	0.344	0.337	0.337	0.323	0.301	0.283	0.271	0.268	0.268	0.269	0.273
CH	0.305	0.303	0.298	0.292	0.284	0.276	0.262	0.251	0.236	0.229	0.217	0.213
BE	0.300	0.295	0.289	0.285	0.267	0.252	0.235	0.228	0.221	0.220	0.212	0.209
SE	0.299	0.293	0.287	0.278	0.260	0.242	0.233	0.228	0.216	0.209	0.201	0.203
AT	0.218	0.217	0.214	0.214	0.204	0.194	0.185	0.181	0.174	0.173	0.171	0.170
ES	0.016	0.035	0.063	0.091	0.116	0.136	0.150	0.159	0.160	0.166	0.170	0.176
LU	0.135	0.136	0.135	0.135	0.125	0.116	0.103	0.098	0.091	0.089	0.083	0.082
DK				0.008	0.018	0.030	0.042	0.051	0.058	0.064	0.066	0.071
GR	0.008	0.017	0.028	0.040	0.047	0.049	0.050	0.050	0.049	0.049	0.048	0.052
IE						0.002	0.005	0.010	0.016	0.023	0.028	0.035
PT					0.001	0.003	0.007	0.013	0.018	0.024	0.029	0.034
FI										0.001	0.003	0.006
MC				0.000	0.000	0.001	0.004	0.006	0.010	0.012	0.013	0.015
Total	3.200	3.228	3.221	3.238	3.186	3.130	3.086	3.080	3.058	3.070	3.082	3.134

Table 5. (cont.)

	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	1990-2012 average
DE	0.560	0.557	0.526	0.528	0.449	0.422	0.405	0.448	0.453	0.487	0.573	0.513
GB	0.533	0.526	0.508	0.575	0.466	0.567	0.479	0.551	0.412	0.418	0.457	0.509
FR	0.511	0.505	0.510	0.515	0.446	0.404	0.372	0.432	0.402	0.433	0.468	0.481
IT				0.185	0.265	0.342	0.250	0.287	0.289	0.346	0.285	0.281
NL	0.267	0.253	0.227	0.212	0.155	0.165	0.149	0.175	0.139	0.119	0.148	0.245
CH	0.206	0.199	0.181	0.170	0.132	0.119	0.112	0.131	0.179	0.157	0.162	0.214
BE	0.201	0.197	0.181	0.170	0.120	0.108	0.102	0.130	0.122	0.111	0.105	0.198
SE	0.197	0.191	0.188	0.154	0.109	0.096	0.121	0.145	0.123	0.105	0.104	0.195
AT	0.164	0.159	0.148	0.138	0.105	0.095	0.090	0.106	0.104	0.098	0.129	0.159
ES	0.180	0.184	0.185	0.178	0.148	0.142	0.151	0.179	0.170	0.165	0.181	0.144
LU	0.078	0.080	0.074	0.066	0.046	0.040	0.038	0.045	0.049	0.044	0.038	0.084
DK	0.073	0.077	0.076	0.066	0.048	0.045	0.054	0.061	0.088	0.085	0.082	0.058
GR	0.053	0.056	0.055	0.056	0.046	0.039	0.043	0.053	0.057	0.044	0.037	0.045
IE	0.041	0.048	0.052	0.051	0.042	0.037	0.045	0.056	0.058	0.056	0.057	0.037
PT	0.039	0.043	0.046	0.042	0.031	0.029	0.035	0.043	0.044	0.030	0.046	0.029
FI	0.013	0.020	0.026	0.029	0.025	0.026	0.030	0.039	0.038	0.042	0.046	0.025
TR			0.002	0.005	0.006	0.008	0.012	0.018	0.027	0.036	0.035	0.016
MC	0.018	0.022	0.024	0.024	0.018	0.016	0.017	0.020	0.020	0.019	0.018	0.014
CZ				0.001	0.001	0.004	0.007	0.012	0.024	0.024	0.030	0.013
PL						0.001	0.004	0.007	0.012	0.018	0.027	0.012
CY	0.001	0.004	0.006	0.008	0.006	0.007	0.009	0.013	0.015	0.013	0.012	0.008
RO						0.001	0.003	0.005	0.010	0.013	0.013	0.008
HU				0.000	0.001	0.002	0.004	0.007	0.010	0.013	0.015	0.006
SK				0.000	0.001	0.002	0.004	0.006	0.009	0.011	0.013	0.006

SI				0.000	0.000	0.001	0.003	0.005	0.006	0.008	0.010	0.004
BG				0.000	0.000	0.001	0.002	0.005	0.006	0.008	0.009	0.004
EE				0.000	0.000	0.001	0.003	0.004	0.005	0.006	0.007	0.003
IS							0.000	0.001	0.002	0.003	0.004	0.002
NO										0.001	0.002	0.002
LV								0.001	0.001	0.002	0.003	0.002
LT						0.000	0.000	0.001	0.002	0.003	0.004	0.002
MT											0.001	0.001
HR										0.001	0.001	0.001
AL												
LI												
MK												
RS												
SM												
Total	3.132	3.121	3.013	3.174	2.664	2.722	2.543	2.988	2.876	2.916	3.118	3.317

Figure 5. Propensity to protect in an EPO member state (PGFP, 3-year MA)

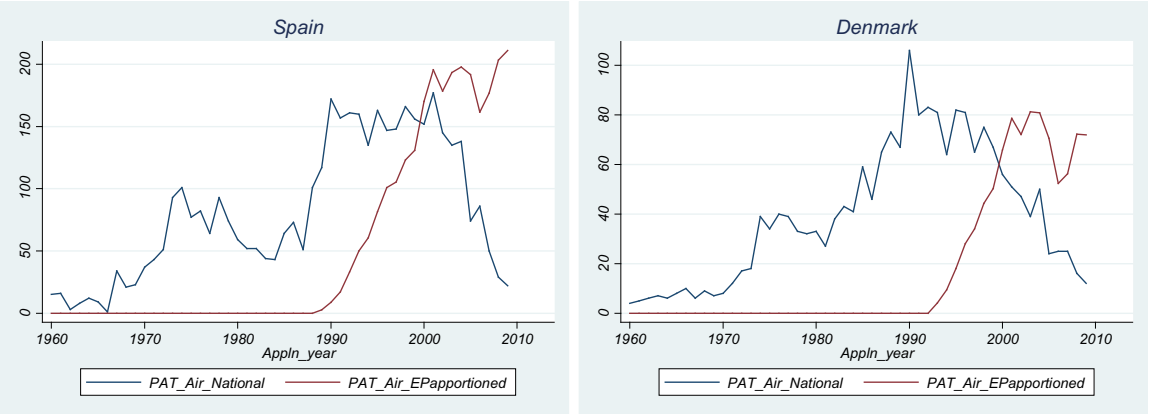


In this paper, the protection propensities are generated only over time, assuming that all patent applicants in a given year are homogeneous with respect to protection propensities. This could be refined further by allowing the propensities to vary not only over time but also across priority office as well as duplicate office, inventor country or applicant country (although such information is sometimes missing, reducing the sample size). EPO applications could then be apportioned along all these vectors.

The PRS data necessarily involve certain time lags. Therefore, in this paper we construct summary statistics of protection propensities for TOTPAT and apply these on EPAT. This is based on the assumption that protection propensities do not vary systematically across technological fields. We

thus use the full sample of PRS data (all patents, not only environmental) to calculate the propensities. We then apply the propensities on the "environmental" counts and apportion EPO applications onto national patent offices of the member countries (as unit counts, not fractionally). Figure 6 shows the evolution over time of national and EP-apportioned patenting activity in air pollution abatement technologies for selected countries. An alternative approach would be to estimate the protection propensities econometrically as a function of characteristics of patent applicants, inventors, priority office, and technological field, and then nowcast the recent years.

Figure 6. Patenting activity at the national office versus EP-apportioned filings



In sum, construction of 'protection propensities' allows apportioning of regional patent filings onto national jurisdictions, and thus allows to correct for the drop in national patenting caused by growing tendency to protect inventions at the regional rather than national authorities. Subject to availability of protection data, this methodology could be extended to other regional offices.

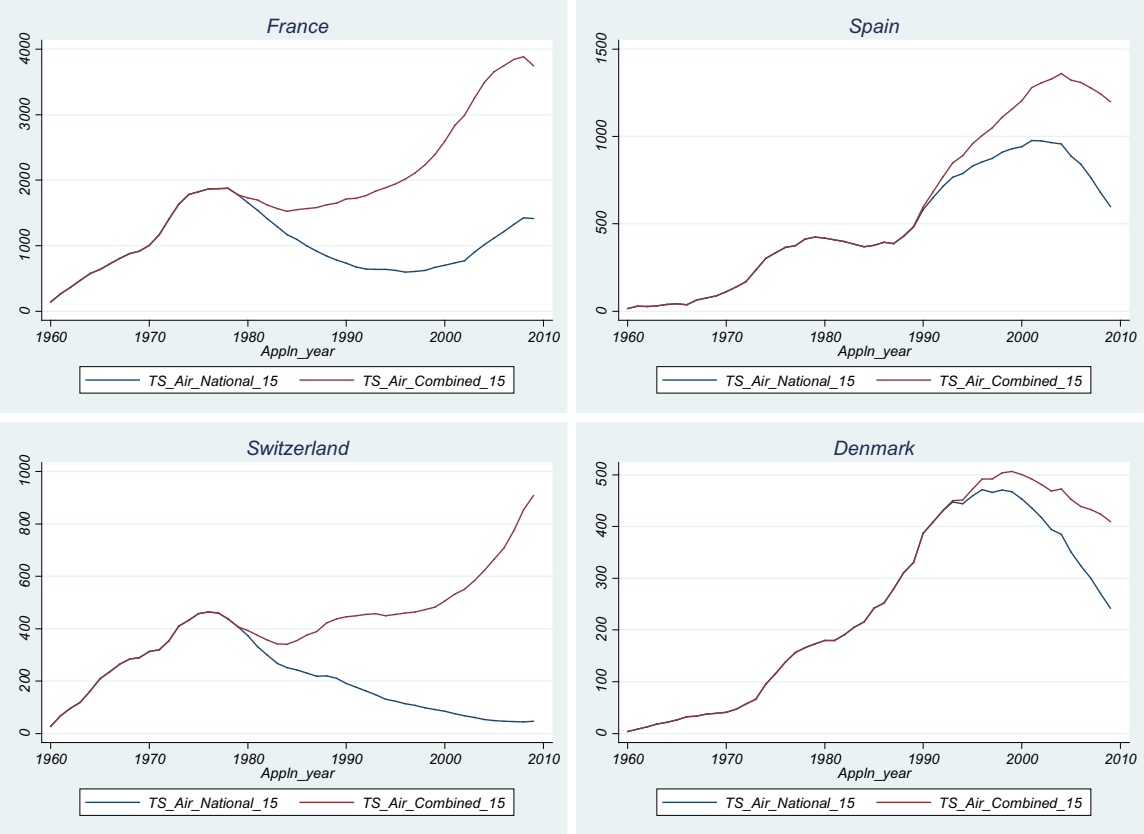
Such EPO protection propensities can be useful for at least two types of applications. First, the propensities can be used to **estimate patent family size** when constructing patent family-based indicators. Consider the following example using data from Table 5: Summing across protection propensities of all EPO member states in a given year yields the "aggregate propensity" – that is, the average size of an EP patent family. For most of the 1990-2012 period the aggregate propensity varies between 2.5 and 3.2 (last row in Table 5). For example, in year 2000 this is estimated to be 3.082, so a singleton application deposited in that year at the EPO corresponds, on average, to a patent family with an approximate family size of 3. However, if PATSTAT also contains a trace of a German family member then the German application must contribute its full weight (unity) rather than its estimated weight in that year (1 rather than 0.553). Therefore, the equivalent size of such patent family is approximately 3.5 (Table 6 summarises this example).

Table 6. Estimation of patent family size using EP protection propensities

Observed family	*Estimated family size in year 2000*
EP singleton	*3.082*
EP + DE	*3.082 - 0.553 + 1 = 3.530*
EP + US	*3.082 + 1 = 4.082*

Second, the protection propensities can be used to **apportion patenting activity onto national jurisdictions** of EPO member states, for example to construct patent stocks to be used in empirical analyses (Figure 7).

Figure 7. Technology stock based on national and EP-apportioned patents (15% discount rate)



3.3 Missing information

For narrow technological fields, small inventor countries, or countries with low levels of patenting activity – that is, situations with frequent zeros and generally low patent counts – increasing the 'yield' of patent data extractions is valuable. However, this should be done in a manner that minimizes a potential bias in the analysis (this will be discussed in Section 4 below).

Imputing missing information on inventors

One way to increase the magnitude and variation of counts is to impute missing inventor information. While basic bibliographic information such as application authority and application date is always available in patent databases, additional descriptive information such as name and address of the inventor(s) and applicant(s) are often missing or incomplete. To mitigate this problem we retrieve inventor data for all patent family members and impute from duplicate applications information on inventor countries that is not listed in the priority document. The imputed inventor information is then stored at the level of the priority because it is common to all members of the single-priority patent family. This typically allows increasing the volume of data with known inventor information by 3 to 4 percentage points (Table 7).[14]

While technically the same imputation procedure could be conducted to mitigate missing applicant information, conceptually it might be preferable to retain all sets of applicant data within a family because applicants might be different for each patent application, i.e. due to change of ownership.

Table 7. Benefit of imputing inventor information from duplicate filings

	Priorities with known inventor country	*Priorities with known inventor country* ***retrieved within PATSTAT***
Renewable energy (Y02E10) [1]	42.3%	46.2% (+3.9)
Geothermal energy (Y02E10:1) [1]	54.0%	58.1% (+4.1)
Wind power (Y02E10:7) [1]	52.6%	55.7% (+3.1)
Wind power (F03D) [2]	53.2%	56.3% (+3.1)
(Waste)water treatment (C02F) [2]	35.3%	38.8% (+3.5)

Notes:

[1] Based on searches in APPLN_ECLA table
[2] Based on searches in both APPLN_ECLA and APPLN_IPC table

Imputing missing information on patent classifications

Another source of missing information are missing classification symbols. For search strategies defined in terms of IPC symbols, the yield can be increased by searching also in the APPLN_ECLA (more recently APPLN_CPC) table in addition to APPLN_IPC table. This is because EPO examiners often assign ECLA symbols even to foreign filings if they consider that the existing classifications are not sufficient or appropriate. For example, in the case of wastewater treatment technologies this allows increasing the volume of data by 3 percentage points compared with a search conducted only in the APPLN_IPC table (Table 8).

Table 8. Benefit of imputing IPC symbols using the APPLN_ECLA table

	Nb. of documents identified (appln_id's)		
	search in APPLN_IPC only	*search in both APPLN_IPC & APPLN_ECLA*	*search in APPLN_ECLA only*
(Waste)water treatment (C02F)	*433,698*	*448,427 (+3%)*	*199,435*
Wind energy (Y02E10/7)	-	-	*62,702*
Wind motors (F03D)	*64,339*	*69,476 (+8%)*	*43,151*
Climate mitigation in transport (Y02T)	-	-	*293,670*
Electric & hybrid cars (IPC-based)	*113,038*	*128,991 (+14%)*	*71,357*
Climate mitigation in buildings (Y02B)	-	-	*209,898*
Energy efficiency in buildings (IPC-based)	*260,855*	*289,178 (+11%)*	*134,912*

Expansion of patent families

To construct patent families, we proceed in three steps. First, the patent database is searched for the relevant (e.g. "green" or "environment-related") patents using a given definition – a search strategy. Second, we search for potential family members outside of this subset of documents using the DOCDBFAM table that gives patent applications protecting the same set of claims. Third, we identify priority relationships within this enlarged subset of patent documents.

Figure 8. Expansion of patent families

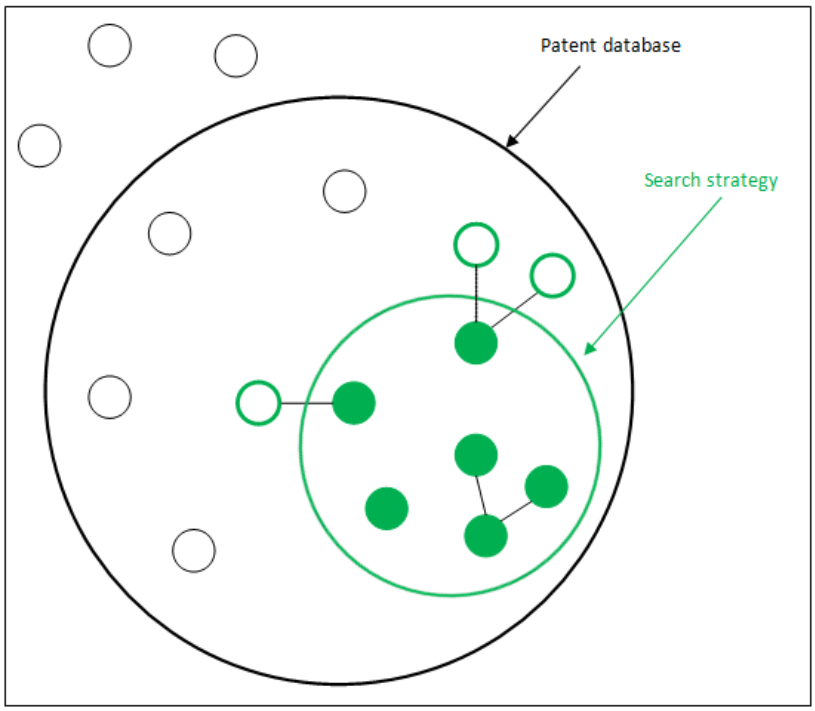


Table 9. Benefit of expansion: Additional patent family members identified

	Within	*Outside*
	Appln_id's	*Full expansion (additional Appln_id's)*
Wind energy (Y02E10/7)	*62,702*	
Wind motors (F03D)	*69,476*	*3,879 (+5%)*

This approach also helps to mitigate the bias historically created by the distinction between core- and advanced-level IPC classes and the differentiated obligations of patent offices to classify in the latter (only patent offices in the "PCT-minimum requirements" are mandated to classify at the advanced level). Hence, when a search strategy is based on advanced-level IPC symbols (which is almost always the case), small patent offices that only classify at the core level would be under-represented in the searches. The next Section discusses this issue in greater detail.

Constructing variables to normalize (or control for) missing information

While strategies to impute missing information are valuable because they increase the "yield" of patent extractions, they too might introduce potential biases in analysis. So both, the presence of missing information, as well as imputation thereof, might give rise to biases. One way to mitigate them is to generate a corresponding count of patent "totals" (so-called TOTPAT). By *corresponding* we mean a statistic that is constructed in a manner that is identical to the "environmental" count (EPAT) with the one exception that all documents (not only "environmental") are considered. For example, if missing information on inventors is imputed, this should be done identically for EPAT as for TOTPAT. Ways in which such TOTPAT variable can be used for the purposes of normalization in descriptive analyses or as a control variable in econometric analyses will be discussed in Section 4 below.

3.4 Non-systematic classification of patent documents

Assignment of patent classification symbols to patented inventions is not homogeneous across patent offices and this is another potential source of idiosyncratic biases in patent analysis. Questions arise concerning the best ways to apply a given search strategy to identify the relevant patents in a patent database; and, how to identify the relevant population, or a relevant subset in an unbiased manner?[15]

The use of the international patent classification system (IPC) varies across offices to a certain extent. For example, while patent offices listed in the "PCT minimum requirements" are obliged to assign IPC symbols at the 'advanced' level, other patent offices must do so only at the 'core' level although in practice they might assign advanced level symbols in areas of particular interest. This is important because search strategies for narrow technological fields are often based on symbols at the advanced level of IPC. Such heterogeneous practices in assignment of IPC symbols thus create a potential source of a bias.

Even more importantly, many national patent offices have developed their own patent classification systems (ECLA, USPC, F-terms, etc.) to classify patent applications, the related prior art, and perhaps patents in selected domains that are of specific interest to their respective countries. In any case, the national patent classification symbols are assigned only to a certain subset of the world's "population" of patent applications. For example, the EPO uses the ECLA symbols[16] to systematically classify all patent documents from the "PCT minimum requirements" in one of the EPO official languages (incl. German, English and French), with documents filed in other languages (e.g. Japanese, Russian, Spanish) are excluded. Patent documents in Dutch (e.g. those filed in the Netherlands, Belgium and Luxembourg) are also classified as well as those with first filings residents from Austria, Australia and Canada. In addition, Korean, Chinese and Indian documents are classified if they are in one of the four EPO languages. Additional documents might be machine-translated if they are in selected priority areas (determined based on IPC symbols assigned by home offices). Finally, some Japanese documents were classified in the past based on abstracts, and they might be re-classified if a subsequent equivalent is published in one of the four EPO languages.[17] The situation is in many respects similar, if not "worse", at other patent offices worldwide because they too face resource constraints.

In sum, while the practice of selective classification in national systems might be economically justified, from a researcher's point of view it creates another potential source of bias. For example, while a large majority of patent documents in PATSTAT have an ECLA symbol assigned, it is not true of all the patent documents. Therefore, a search strategy that uses an ECLA symbol might yield biased results insofar it cannot identify relevant documents among those without an ECLA symbol.

The situation is even more complicated with search strategies that use the recently introduced Y02 class (so-called Y-tags) developed by the EPO. While this new tagging scheme greatly facilitates identification of many climate-related inventions by non-specialists, users should be aware of its implications.[18] To be clear, the new tagging scheme does not substitute for the standard ECLA (or CPC) classification symbols, rather it is complementary. Technically, the tagging does not involve examination of each individual patent application, rather it is constructed as a set of search algorithms that exploit the full stock of descriptive attributes in DOCDB[19], including the IPC and ECLA symbols, as well as EPO's internal schemes of in-computer-only (ICO) classification symbols and keywords (KW). In addition, for some technologies keyword searches on English titles and abstracts are conducted (either as a single keyword or in combination). Finally, where available, Derwent WPI index might be used as well.[20] Therefore, the population of patents from which such searches draw is

unknown (or imperfectly defined, at least). Next we discuss strategies that help mitigate any potential biases.

Constructing variables to normalize (or control for) non-systematic classification

Ideally, every "environmental" patent (EPAT) search strategy needs its own corresponding TOTPAT control variable. A researcher using an IPC-based (or ECLA-based) search strategy can easily generate a *corresponding* count of patent "totals" (e.g. all patents with an/any IPC (ECLA) symbol assigned). Similarly, if a search strategy is based on English keywords in titles and abstracts, then the corresponding TOTPAT would be constructed as a count of all patent families with an English-language title or abstract. In these cases, construction of such variables is straightforward because the set of patents from which such searches draw is known – PATSTAT allows identifying these relevant sets of documents (see Table 10). At the aggregate level TOTPAT_IPC and TOTPAT_ECLA are highly correlated because as many as 80% of all documents (appln_id's) in PATSTAT have both IPC and ECLA symbols assigned. However, there are important differences at the more disaggregated level.

However, in the case of a Y-based search strategy the set of "potentially identifiable" documents is only imperfectly defined, and therefore it is more difficult to generate a corresponding "total". As described above, the set of documents that could have theoretically been searched and assigned a given Y-tag is only imperfectly defined. Therefore, constructing TOTPAT_YTAG correctly is an impossible task, unless a separate TOTPAT count is generated for every Y-symbol which is in most practical uses of patent data too costly.[21] In most practical applications, the researcher might use an approximation of TOTPAT_YTAG (#4) based on some combination of the corresponding totals – for example, the union of the IPC and ECLA TOTPATs (#5 in Table 10).

Table 10. Construction of corresponding totals

	If EPAT search strategy is based on:	*…then TOTPAT should be constructed as:*	*Share of appln_id's in PATSTAT*
(1)	*IPC symbols*	*All documents (single-priority patent families) that are 'identifiable' using IPC symbols (i.e. all appln_id's listed in the APPLN_IPC table).*	*79%*
(2)	*ECLA symbols*	*All documents (single priority patent families) that are 'identifiable' using ECLA symbols (i.e. all appln_id's listed in the APPLN_ECLA table with EC as classification scheme).*	*49%*
(3)	*Keyword searches on titles and/or abstracts*	*All documents (single priority patent families) that are 'identifiable' using keyword searches (i.e. all documents with title/abstract in the corresponding language).*	*58% (English)*
(4)	*Y02 tags*	*All documents (single-priority patent families) that could have potentially been tagged by the EPO.*	*? (this will vary by individual Y-symbol)*
(5)	*IPC and ECLA symbols*	*The union of (1) and (2) above.*	*80%*
(6)	*IPC, ECLA, ICO, or English title/abstract*	*The union of the respective counts.*	*84%*

For these reasons, and when there is a choice between an IPC-based and Y02-based search strategy, it might be preferable to use IPC because it allows correctly specifying the "counterfactual" (all patents with and assigned IPC symbol, not only the 'environmental' IPC symbol), a task that is more difficult (if not impossible) to do for Y-tagging because the individuals to be tagged are not drawn randomly from this population (we do not know the 'taggable' population). This might be particularly important for regions of the world that are not well covered by the Y02 scheme.

This problem will diminish over time because many of the elements (ECLA, ICO, KW) have now become part of the new CPC system. As patent classification systems become increasingly harmonised, and the CPC is used more widely, the precision of TOTPAT_YTAG will approach that of TOTPAT_CPC. Until then, for research hypotheses where it is of key importance to have a precise control (counterfactual) IPC might indeed be the better option.

In some cases it might be possible to define a "sectoral total" instead of a "PATSTAT total" as discussed above. For example, analysis of electric and hybrid cars patenting can be complemented with a "total" count for the entire sector of motor vehicles or road vehicles which it is possible to define using IPC symbols (Haščič and Johnstone 2011; see also Aghion et al. 2012). Similarly, analysis of fuel-efficient energy generation technologies can be complemented with a "total" count for the entire sector of energy generation from fossil fuel combustion (Lanzi et al. 2011; Noailly and Smeets 2013).

4. IMPLICATIONS FOR EMPIRICAL ANALYSIS

The need for a normalization count (control variable)

Dealing with a narrow technological field is both a boon and a bane. On the one hand, this poses certain difficulties in terms of indicator construction (Section 2), but on the other hand it allows generating a "total" variable that is an appropriate answer to these, and other, difficulties. The conceptual justification for the use of such variable is twofold: First, there are **conceptual (economic) reasons** for using a TOTPAT arising out of the need to control for many generic factors that affect patenting in a narrow field, including:

- Differences in inventive capacity
- Differences in propensity to patent
- Differences in patent breadth and patent 'quality'
- Other factors that might affect patenting in general

Obviously, there might be suitable substitutes for TOTPAT to control for these differences across countries, sectors, firms, and over time (while such substitutes may be even better, the data is not always readily available).[22]

Second, there is another set of reasons that might be even more important and for which it is impossible to substitute with other data. These are the **methodological (idiosyncratic) reasons** calling for the use of a TOTPAT arising out of:

- Incomplete information due to differences in coverage of patent databases (Section 3.1)
- Imperfect information on jurisdictions where patent protection is sought through regional procedures (Section 3.2)
- Extent of missing information on inventors, applicants, patent classes (Section 3.3)
- Non-systematic classification of patent documents (Section 3.4)

Using a TOTPAT variable deals perfectly with the latter (idiosyncratic) sources of biases (except for those discussed in 3.4) and imperfectly with the former (conceptual) sources of biases. How exactly should such a TOTPAT variable be constructed will depend on three factors:

i) Indicator selected (e.g. PF2 for invention, PF1 for co-invention)

ii) Search strategy used (e.g. based on IPC, ECLA, keyword searches, or other)

iii) Algorithm used to construct the EPAT count (i.e. treatment of idiosyncrasies, other programming details that might affect the final outcome).

The corresponding TOTPAT is identical to EPAT in all the three aspects above. It is measured using the same indicator (e.g. PF2, TPF, count of co-inventions), based on the same type of search strategy (see Table 10), and constructed using an otherwise identical algorithm.

In descriptive analyses, inclusion of a TOTPAT variable provides suitable context and a basis for interpretation of trends observed in EPAT. Another option is to 'normalize' the EPAT count, for example, as a measure of specialisation expressed as a ratio of a country's share of EPAT on TOTPAT

over the share of EPAT over TOTPAT worldwide (relative propensity to invent, relative propensity to patent) although such measures might be difficult for a layperson to interpret.

However, this approach yields the greatest benefits in econometric analyses. Indeed, many of the potential biases discussed above can be successfully mitigated by inclusion of a TOTPAT control variable (of the same dimensionality as EPAT). Such variable has the potential to partially control for changes in patenting propensities over time and across countries and offices, and to correct measurement error due to the idiosyncrasies detailed above. In empirical research one might be tempted to disregard such potential errors; while they might indeed be negligible in the EU-US space, shifting the analysis outside of the OECD countries increases the potential for erroneous conclusions, and hence the need to control for potential idiosyncratic biases.

5. CONCLUDING REMARKS

This paper draws attention to several issues that are specific to analysis of narrow technological fields in cross-country international comparisons. Moreover, there is a rising interest in patent analyses in developing countries and this raises new types of issues. First, the paper advocates the use of a range of patent indicators based on international patent family size. While it concludes that determining the "optimal" family size for a given application is largely an empirical question, PF2 or higher-order restrictions are preferable than PF1, unless the technological field studied is so narrow that this restricts the variation in the data 'too much' (too many zeros and low counts). In such cases one has no other option but to use PF1. There is indeed a trade-off between patent quality and breadth of technological fields studied.

Second, the paper reviews four types of idiosyncratic problems in the underlying data, including how to tell zeros from missings; how to address the problem of regional patent filings; how to increase yield of database extractions; and how to face patent identification problems. The paper advocates using a normalization (or control) variable constructed in an analogous manner as the indicator of primary interest. All patent databases come with a warning – do not blindly estimate on the contents of the database!

NOTES

[1] PCT count is not always a suitable indicator for inter-country comparisons because the propensity to use the PCT route varies across countries and over time (even after year 2000). Inventors from different countries vary in their likelihood of pursuing the PCT through to the national/regional phase (e.g. 80% for NL, 30% RU and KR), and as such PCT counts suffer from their own specific type of 'home bias'. It is equivalent to a call option rather than a true patent application, and because the initial costs are rather low (esp. in the international phase) PCT applications often covers inventions of little value (OECD 2009).

[2] Working with patent families is essential to avoid double-counting of inventions when patent data from multiple patent offices are pooled together. We adopt the notion of single-priority patent family (Martinez 2010) because it facilitates many of the steps discussed herein. We calculate the "international patent family size" – that is, the number of distinct patent authorities within a family. As such, we distinguish between 3 types of equivalents – singleton priorities, claimed priorities and duplicates.

[3] Claimed priorities with regional weights would yield a "dyadic patent family" indicator.

[4] In the same vein, de Rassenfosse et al. (2013) advocate a metric based on "a worldwide count of priority patents" as a measure of inventive activity (equivalent to PF1 presented here). They highlight the advantages of such a measure relative to PCT counts.

[5] Claimed priorities account for a relatively small proportion of the stock of patent applications. For example, in a study focusing on innovation in climate change mitigation technologies found that only about 11% of the relevant stock of patent applications included in PATSTAT were CPs, with 34% being their duplicates, and 55% being singletons (Haščič et al. 2010). In other words, in climate change mitigation technologies CPs represent only 16% of the stock of patented inventions (simple patent families), while the large majority (84%) of these inventions were only protected at a single patent office (singletons). It must be noted that there is variation in these proportions across patent offices.

[6] Note that account has to be taken of changing membership through time.

[7] For example, former Eastern Germany (DD) complete data until 1999; former Czechoslovakia (CS) complete data until 1993, thereafter complete data separately for SK and CZ; former Yugoslavia (YU) data until 1992.

[8] The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

[9] a) Note by Turkey:

The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

b) Note by all the European Union Member States of the OECD and the European Union:

The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

[10] ARIPO and OAPI together accounted for 34% of patent applications deposited in Africa during 1980-2009, excluding the South African office (Haščič et al. 2012).

[11] The ignorance of the "do nothing" approach is not necessarily bad as long as one is comfortable with its implications, such as (i) wrongly classifying an EP patent application as a singleton when in fact it is a claimed

priority, or (ii) wrongly classifying a patent family as a claimed priority when in fact it is only a singleton (e.g. consider a patent family consisting of EP and DE applications with DE as the only designated country).

[12] Previously, one needed to determine whether a patent has been granted at each of the 38 member states patent authorities, hence making sense of the 38 different publication kind code systems. Fortunately, this has been much facilitated in recent editions of PATSTAT.

[13] Specifically, data for years prior to 1990 seem rather incomplete; therefore we only construct the propensities for the period after 1990 (for Italy only after 2004). Even after 1990, data for some years and offices are missing, with two different situations occurring: (1) A year count is missing (GR 2007, LT 2007) and years around seem to be correct (the same trend before and after the missing year). In this case we simply impute a value for the missing year corresponding to the mean of the year before and the year after. (2) A year count is missing (FR 2007, IT 2007) and the years before and after show implausibly high numbers above the long-term trend. In this case we assume that for an unknown reason the patents have been distributed among the years before and after the missing year.). In this case we re-allocate a third of the count proceeding the missing year, and a third of count in the year following the missing year. Moreover, the total on which the propensities are calculated also has to be corrected (the total for 2007 is very low compared with other years). The total is the count of distinct applications among all countries (no double-counting for applications with protection in multiple states). We use the "duplicate protection" coefficient observed in 2005 and 2009, take the mean of those two values, and divide through the values re-allocated from 2006 to 2007 and from 2008 to 2007 to inflate the total for 2007.

[14] A similar approach was developed by De Rassenfosse et al. (2013).

[15] Development of a suitable search strategy – that is, a mapping between patent classification systems and a definition of "environmental" innovation - is a distinct problem and is not addressed here. A companion paper is devoted to this issue.

[16] Starting 2013 the EPO and USPTO have started using a common classification system (CPC) and consequently the use of ECLA and USPC is being discontinued.

[17] Personal communication from Mr. Pierre Held of the European Patent Office, November 2012 in Hamburg, Germany.

[18] The benefit of the new Y02 scheme is twofold. First, it allows identifying patent documents in technological fields that would otherwise not be possible by a non-specialist! The tags were developed by patent examiners at the EPO who are specialised in the different fields covered. Second, it provides a much greater level of detail than what would be possible using the IPC. This is a significant advantage over previous searches based on the IPC. See Veefkind et al. (2012) for further details.

[19] DOCDB, also referred to as EPODOC, is the EPO's master documentation database. This is the resource used by patent examiners for searches of prior art.

[20] Personal communication by Mr. Victor Veefkind of the European Patent Office, 2012.

[21] In the ideal case, one would construct a different TOTPAT for each of the individual Y02 symbols – depending on type of attributes used in the algorithm to assign a given symbol (e.g. some algorithms use keywords, others do not). While such information could be made available, it is of limited practical use because (i) analysis is rarely conducted at the level of an individual symbol (e.g. Y02E10:43 Fresnel lenses); rather, it is the hierarchically more aggregated level (e.g. Y02E10:4 Solar thermal energy) that is of interest to decision-makers. In such case it is not clear how can one generate a TOTPAT at the aggregate level when individual tags use different sets of attributes; (ii) Even if the same set of attributes were used within a given "field" of technology, analyses often compare several of such "fields" implying the use of multiple "totals" (e.g. a total for solar energy, a total for wind energy, etc.).

[22] See Johnstone et al. (2012) for a discussion.

REFERENCES

Aghion P., Dechezleprêtre, A., Hemous, D., Martin, R. and J. Van Reenen (2012) "Carbon Taxes, Path Dependency and Directed Technical Change: Evidence from the Auto Industry," NBER Working Papers 18596, National Bureau of Economic Research.

Barton J.H. (2007), "Intellectual Property and Access to Clean Energy Technologies in Developing Countries", International Centre for Trade and Sustainable Development (ICTSD).

Belward A., B. Bisselink, K. Bódis, A. Brink, J.-F. Dallemand, A. de Roo, T. Huld, F. Kayitakire, P. Mayaux, M. Moner-Girona, H. Ossenbrink, I. Pinedo, H. Sint, J. Thielen, S. Szabó, U. Tromboni, L. Willemen (2011), "Renewable energies in Africa", European Commission Joint Research Centre.

Collier, P., Venables, A.J. (2012), "Greening Africa? Technologies, endowments and the latecomer effect", *Energy Economics*, Elsevier, vol. 34(S1), pages S75-S84.

de Rassenfosse, G., Dernis, H., Guellec, D., Picci, L. and B. van Pottelsberghe de la Potterie (2013) "The worldwide count of priority patents: A new indicator of inventive activity", *Research Policy*, Volume 42, Issue 3, Pages 720–737.

Dechezleprêtre, A., M. Glachant, I. Haščič, N. Johnstone, and Y. Ménière (2011), "Invention and transfer of climate change mitigation technologies on a global scale: A study drawing on patent data," in *Review of Environmental Economics and Policy*, Volume 5, Issue 1, Pages 109-130.

Dernis, H. and M. Khan (2004), "Triadic patent families methodology", *STI Working Paper 2004/2*. OECD Publishing.

Dernis, H., Guellec, D. and B. van Pottelsberghe (2001), "Using patent counts for cross-country comparisons of technology output", STI Review No. 27, OECD Publishing.

Faust, K. (1990), "Early identIfication of Technological Advances on the Basis of Patent Data", *Scientometrics*, Vol. 19(5-6), pp. 473-480.

Faust, K. and H. Schedl (1983), "International Patent Data: Their Utilisation for the Analysis of Technological Developments", *World Patent Information*, Vol. 5(3), pp. 144-157.

Guellec, D. and B. van Pottelsberghe de la Potterie (2000), "Applications, Grants and the Value of a Patent", *Economics Letters*, Vol. 69, pp. 109-114.

Harhoff, D., F.M. Scherer and K. Vopel (2003), "Citations, family size, opposition and the value of Patent Rights", *Research Policy*, Vol. 32, pp. 1343-63.

Haščič, I. and N. Johnstone (2011), "Innovation in Electric and Hybrid Vehicle Technologies: The Role of Prices, Standards and R&D", in OECD, *Invention and Transfer of Environmental Technologies*, OECD Publishing. http://dx.doi.org/10.1787/9789264115620-5-en

Haščič, I. and M. Migotto (2015) "Measuring Environmental Innovation Using Patent Data: Policy Relevance", *OECD Environment Working Papers*, OECD Publishing (*forthcoming*).

Haščič, I., J. Silva and N. Johnstone (2012), "Climate Mitigation and Adaptation in Africa: Evidence from Patent Data", *OECD Environment Working Papers*, No. 50, OECD Publishing, Paris. http://dx.doi.org/10.1787/5k8zng5smxjg-en

Haščič, I., Johnstone, N., Watson, F., Kaminker, C. (2010), "Climate Policy and Technological Innovation and Transfer: An Overview of Trends and Recent Empirical Results", *OECD Environment Working Papers*, No. 30, OECD Publishing, Paris. http://dx.doi.org/10.1787/5km33bnggcd0-en

Johnstone, N., I. Haščič, J. Poirier, M. Hemar and C. Michel (2012) "Environmental policy stringency and technological innovation: evidence from survey data and patent counts", *Applied Economics*, Volume 44, Issue 17, 2012, pages 2157-2170.

Lanzi, E., Verdolini, E. and I. Haščič (2011) "Efficiency-improving fossil fuel technologies for electricity generation: Data selection and trends," in *Energy Policy*, Volume 39, Issue 11, Pages 7000-7014.

Martinez, C. (2010), "Insight into Different Types of Patent Families", *OECD Science, Technology and Industry Working Papers*, No. 2010/02, OECD Publishing, Paris. http://dx.doi.org/10.1787/5kml97dr6ptl-en

Noailly J. and R. Smeets (2013) "Directing Technical Change from Fossil-Fuel to Renewable Energy Innovation: An Empirical Application Using Firm-Level Patent Data," FEEM Working Papers 2013.34, Fondazione Eni Enrico Mattei.

OECD (2012), *Energy and Climate Policy: Bending the Technological Trajectory*, OECD Studies on Environmental Innovation, OECD Publishing. http://dx.doi.org/10.1787/9789264174573-en

OECD (2011), *Invention and Transfer of Environmental Technologies*, OECD Studies on Environmental Innovation, OECD Publishing. http://dx.doi.org/10.1787/9789264115620-en

OECD (2009), *OECD Patent Statistics Manual*, OECD Publishing. http://dx.doi.org/10.1787/9789264056442-en

Veefkind, V., J. Hurtado-Albir, S. Angelucci, K. Karachalios, and N. Thumm (2012), "A new EPO classification scheme for climate change mitigation technologies", *World Patent Information* 34 (2): 106-111.

ANNEX

We consider four alternative (candidate) statistics based on different combinations of PRS codes and publication kind codes as extracted from the PRS dataset (summarized in Table A1). The order is chronological and most likely it correlates with reliability of the statistic – data obtained close to grant reflect patentee's intentions better than data gathered at the beginning of the patenting process; moreover, patentee's preferences may evolve over time.

Table A1. Candidate statistics and legal status search strategy

Candidate statistic		*Search strategy*
1. Propensity to designate states at application	*AK-A*	*prs_code=AK and publ_kind_code=A%*
2. Propensity to designate states at payment of fee	*AKX-RBV*	*prs_code=AKX or prs_code=RBV*
3. Propensity to designate states at grant	*AK-B*	*prs_code=AK and publ_kind_code=B%*
4. Propensity to pay post-grant fees (annual maintenance fees)	*PGFP*	*prs_code=PGFP*

While the data underlying statistic #3 seem as a suitable option, there seems to be a problem of scale. Three countries have propensities that approach (or exceed) 90% (Germany, France, UK). While one would expect that these countries rank among the most often designated countries, the propensities suggested here are surprisingly high. Moreover, several rather minor economies show rather high propensities (e.g. Estonia has propensity almost as high as the Czech Republic, and Cyprus is not far from Denmark). This indicates that patentees might designate many countries at grant but then do not translate the application in the local language (validation) neither pay maintenance fees. Given this, statistic #4 is the best option.