THE UNIVERSITY OF MELBOURNE

FACULTY OF BUSINESS & ECONOMICS

Melbourne Institute Working Paper Series

Working Paper No. 8/13

Selection Bias in Innovation Studies: A Simple Test

Gaétan de Rassenfosse, Anja Schoen and Annelies Wastyn

MELBOURNE INSTITUTE®
of Applied Economic and Social Research

Selection Bias in Innovation Studies: A Simple Test*

Gaétan de Rassenfosse[†], Anja Schoen[‡] and Annelies Wastyn[§]

[†] **Melbourne Institute of Applied Economic and Social Research and Intellectual Property Research Institute of Australia, The University of Melbourne**

[‡] **Technische Universität München**

[§] **Department of Managerial Economics, Strategy and Innovation, KU Leuven**

Melbourne Institute Working Paper No. 8/13

ISSN 1328-4991 (Print)

ISSN 1447-5863 (Online)

ISBN 978-0-7340-4297-2

March 2013

* A slightly modified version of this paper is forthcoming in *Technological Forecasting and Social Change*. Please consult the published version. The authors are grateful to Dirk Czarnitzki, Anne Leahy and John Micklewright as well as to the participants at various seminars and conferences for helpful comments. Gaétan gratefully acknowledges financial support from the ARC (grant number LP110100266). Corresponding author: <gaetand@unimelb.edu.au>.

Melbourne Institute of Applied Economic and Social Research
The University of Melbourne
Victoria 3010 Australia
***Telephone* (03) 8344 2100**
***Fax* (03) 8344 2111**
***Email* melb-inst@unimelb.edu.au**
***WWW Address* http://www.melbourneinstitute.com**

Abstract

The study of the innovative output of organizations often relies on a count of patents filed at one single office of reference such as the European Patent Office (EPO). Yet, not all organizations file their patents at the EPO, raising the specter of a selection bias. Using novel datasets of the whole population of patents by Belgian firms and German universities, we show that the single-office count results in a selection bias that affects econometric estimates of invention production functions. We propose a methodology to evaluate whether estimates that rely on the single-office count are affected by a selection bias.

JEL classification: O31, C18, C52, C81

Keywords: Knowledge production function, patent count, R&D, selection bias

1. Introduction

Economic and management research on innovation has greatly benefited from the increased availability of patent data, which provide a unique way of tracking the creation and the diffusion of innovation. Patent data are now being used in other fields including labor economics, development economics and economic geography. Yet, the measurement of innovation using patent data suffers from limitations. The two most severe of these limitations are that: (i) not all inventions are patentable and not all patentable inventions are patented; and (ii) the value of patents varies widely and the majority of patents is worthless. We refer the reader to [1], [2], [3] and [4] for in-depth discussions of these issues.[1]

This paper focuses on a third limitation, which is the selection bias that arises from the way patents are counted. It is common practice to count patents at a single patent office to assess organizations' inventive output (henceforth referred to as the 'single-office count'). A close look at a random sample of 20 scientific articles that use patent data, which were published recently in general economic and management journals as well as field journals, reveals that the overwhelming majority of studies rely on a single office count. However, this practice may result in selection bias since firms have the option of filing patents anywhere in the world. This is particularly true in Europe, where two overlapping patent offices coexist. Companies may file patents directly at their national patent office or they may take the more expensive 'European route' by filing patents at the European Patent Office (EPO). Companies that target an international market may file their patents at the World Intellectual Property Office (WIPO) in Geneva, at the US Patent and Trademark Office (USPTO), or in

[1] Researchers working on the economics of innovation are well aware of these limitations. A typical solution involves using additional innovation indicators such as the number of scientific publications or the share of sales from newly-introduced products. Alternative indicators are not always available and many studies rely exclusively on patent data.

any other jurisdiction. As long as filing decisions are random, the single-office count is a noisy proxy of the full patent count (*i.e.* the count that encompasses patents from all possible patent offices). However, as soon as systematic factors affect decisions to select a given filing route, the single-office count results in a selection bias.

Motivated by the tension between the popularity of the single office count and the threat of a selection bias, this paper proposes a way to test the existence of bias when the researcher observes patents at only one patent office. Using novel data on Belgian patenting firms and German universities, we show that the single-office count biases econometric estimates of invention production functions. Invention production functions relate an organization's inventive input to its output and are a key object of analysis in the innovation literature. We also show that our test, which uses information that is readily available to most researchers, successfully spots variables that are subject to a selection bias. It should be of interest to a wide audience given its ease of use and the popularity of the single-office count.

The paper is structured as follows. The next section surveys current practices in the way to count patents to estimate invention production functions. Section 3 explains the proposed methodology to detect a selection bias and section 4 presents the econometric framework. The test is put into practice using data on Belgian firms in section 5 and using data on German universities in section 6. Implications regarding data collection and the estimation methodology are presented in section 7, together with concluding remarks.

2. Measuring inventions with patent data: from theory to practice

Patent data are used in various ways and the appropriate patent indicator necessarily depends on the research objective. Here, our focus is on building a patent indicator to estimate invention production functions, a popular object of analysis in the innovation literature. Invention – or knowledge or patent – production functions relate organizations' research inputs such as R&D expenditures to their patented output. They have attracted considerable attention in the literature, dating back to [5], and have been used, amongst other things, to study the occurrence of innovation (e.g. [6], [7], [8]); to study the invention process and the effectiveness of innovation policies (e.g. [9], [10], [11]); or as an intermediate step to study the determinants of productivity (e.g. [12], [13]).

A patent provides protection only in the country in which it is filed. As a result, firms that want to protect their invention in different countries must file a patent in each relevant national patent office. The first patent describing the invention is called the 'priority filing', while the subsequent patents extending the protection in other jurisdictions are called 'second filings'. We use the terms 'priority filing' and 'priority patent application' interchangeably. The priority patent application is usually filed at the home patent office, although it could be filed at another patent office (the most popular being the USPTO, the EPO and the WIPO). Because applicants have a variety of patenting routes available to them, the patent count should theoretically include all *priority* patent applications filed anywhere in the world, regardless of the patent office of application. This global count of priority filings is explained in great detail in [14].

In practice, however, the operationalization frequently departs from this ideal situation. In particular, the count of patents is usually limited to a count at one reference office, usually the national patent office or the EPO for European firms. We studied a random sample of 20 papers that estimate patent production functions on European data and that were published in the recent past in general economic and management journals as well as in field journals (see Table A in Appendix).[2] We find that 75 per cent of the papers rely on the single office count, and the EPO is taken as the reference office in most of these instances. Surprisingly, very little information on the patent indicators is usually provided. In particular, the priority status of the patent documents (priority filings or second filings) is discussed in only two cases. Limiting the count to patents filed at one reference office is a simple and convenient way to count patents. It is, however, necessarily prone to measurement errors since only a fraction of the total patented output is observed. This measurement error is a random error if it results in an estimate of effect which is equally likely to be above or below the true value, and the single-office count is simply a noisy measure of the true count. However, non-randomness in the measurement error would lead to a selection that biases the estimates of the patent production function.

The question of whether the single office count results in a selection bias has not been studied explicitly, although some authors have reported evidence that systematic factors affect the decision of filing route. Seip [15] provides statistical evidence for Dutch patenting companies. He reports that 80 per cent of the Dutch companies that filed patents at the EPO or the WIPO in 2003–2007 were large companies (more than 200 employees). Yet, out of the 5,000 Dutch patent-filing companies, only 6 per cent have more than 200 employees,

[2] To build the sample we searched Google Scholar for journal articles that contain the keywords 'patent', 'production function', and the name of one European country (e.g. 'France'). We only kept articles that were published in A* or A journals according to the 2010 Excellence in Research for Australia (ERA) Ranked Journal List. Three exceptions are [42], [46] and [54], which are B journals.

suggesting a large selection bias in terms of firm size: large companies are more likely than SMEs to file their patents at the EPO or the WIPO. de Rassenfosse and van Pottelsberghe [16] show that the driving force of national and international patents differ. While nationally-filed patents are more reflective of the propensity to patent, international patents such as EPO patents are more reflective of the productivity of research (see also [17]). At the patent level, anecdotal evidence of a potential selection bias is provided in [18]. Using a large sample of patents granted by the EPO between 1990 and 1995, the authors find that firms adapt their filing strategies according to the expected value of the patent. Jensen *et al.* [19] come to a similar conclusion using Australian patents. They report evidence that patents filed by Australian inventors at the WIPO are more valuable on average than Australian patents filed at the Australian patent office.

In a nutshell, most authors count patents at one office, although this practice could induce a selection bias. The next section formalizes the selection bias in the framework of invention production functions and proposes a methodology to detect its presence.

3. Testing for a selection bias

Selection bias is a fundamental aspect of empirical research and many statistical remedies have been proposed. The most common forms of selection bias include the sample selection bias, data censoring and data truncation (see, for example, [20] and [21]). The selection effect of patent data is of a different nature, such that no standard method can be applied.

We study the selection bias in the framework of invention production functions first introduced by Pakes and Griliches [12]. Besides being a popular object of analysis in the

academic literature on innovation, invention production functions also allow formalizing the nature of the bias in an intuitive way, making them particularly suited for our purpose. It is useful to describe the nature of the selection bias using the log-linear specification of the patent production function. Let us write the total unobserved patented output for organization i (a firm or university), y_i^*, as:

$$\ln(y_i^*) = \boldsymbol{x'}_i \boldsymbol{\beta} + \varepsilon_i \qquad 1$$

where ε_i is an error term and bold letters denote matrices and vectors (variables in the right-hand side are taken to the logarithm). The single-office count implies that only a subset of the patents is observed. Let π_i be the organization-specific fraction of the total output that is observed at the reference office:

$$\ln(\pi_i) = \boldsymbol{x'}_i \boldsymbol{\alpha} + \upsilon_i$$

The output observed at the reference office is:

$$\ln(y_i) = \ln(\pi_i y_i^*) = \ln(\pi_i) + \ln(y_i^*) = \boldsymbol{x'}_i (\boldsymbol{\alpha} + \boldsymbol{\beta}) + \upsilon_i + \varepsilon_i \qquad 2$$

The observed output is an unbiased measure of the true output if $\boldsymbol{\alpha} = \boldsymbol{0}$, that is, if no systematic factor affects the choice of the filing route.

To detect the presence of selection bias, one should test whether π_i is random, *i.e.* to test whether the decision to file patents at the reference office is not affected by elements of $\boldsymbol{x}$. Unfortunately, since π_i is not observed, direct inference is not possible. The solution

proposed in this paper involves using information on the structural form of π_i to test for randomness. Patents observed at the reference office are of two types: priority filings, which are directly filed at the reference office; and second filings, which are filed at the reference office at a later stage. We can thus express π_i in a generic way as:

$$\pi_i = \pi_i^p + (1 - \pi_i^p)\pi_i^s$$

Where π_i^p is the proportion of priority patent applications among total priority patent applications that were filed at the reference office and π_i^s is the proportion of priority patent applications not filed at the reference office that are nevertheless observed at the reference office as second filings. We call π_i^p and π_i^s the 'components' of π_i. The variable π_i depends on $\boldsymbol{x}$ when at least one of the two components depends on $\boldsymbol{x}$. In this case, the following ratio:

$$\widetilde{\pi}_i = \frac{\pi_i^p}{\pi_i^p + (1 - \pi_i^p)\pi_i^s} \qquad 3$$

also depends on $\boldsymbol{x}$. The variable $\widetilde{\pi}_i$ is the proportion of priority filings at the reference office relative to total filings at the reference office (*i.e.* priority filings + second filings) and its exact value is usually known to the researcher.

Hence, one way of testing the presence of selection bias involves evaluating whether $\widetilde{\pi}_i$ is correlated with $\boldsymbol{x}$. If $\widetilde{\pi}_i$ is significantly correlated with $\boldsymbol{x}$, it is *likely* that there is a selection bias. Conversely, if $\widetilde{\pi}_i$ is not correlated with $\boldsymbol{x}$, it is *likely* that there is no selection bias. One can distinguish four general cases. First, if both components are independent of $\boldsymbol{x}$,

there is no selection bias and $\widetilde{\pi}_i$ is not correlated with $\boldsymbol{x}$. Second, if one component depends on $\boldsymbol{x}$ but not the other, there is a selection bias and $\widetilde{\pi}_i$ is unambiguously correlated with $\boldsymbol{x}$. Third, when both components increase (or decrease) with $\boldsymbol{x}$, there is a selection bias but the overall effect of $\boldsymbol{x}$ on $\widetilde{\pi}_i$ is ambiguous. Even though it is likely that $\widetilde{\pi}_i$ will be correlated with $\boldsymbol{x}$, there is a possibility that a change in the numerator is exactly offset by a similar change in the denominator. This occurs if:

$$\widetilde{\pi}_i = c \Leftrightarrow \frac{\pi_i^p}{\pi_i^p + (1-\pi_i^p)\pi_i^s} = c \Leftrightarrow \pi_i^s = \frac{\pi_i^p(1-c)}{(1-\pi_i^p)c} \qquad 4$$

In that particular scenario, $\widetilde{\pi}_i$ is not correlated with $\boldsymbol{x}$ but π_i depends on $\boldsymbol{x}$ and there is selection bias. Fourth, when one component increases with $\boldsymbol{x}$ and the other decreases with $\boldsymbol{x}$, there is not necessarily a selection bias but $\widetilde{\pi}_i$ is unambiguously correlated with $\boldsymbol{x}$. There is no selection bias if:

$$\pi_i^p + (1-\pi_i^p)\pi_i^s = c \Leftrightarrow \pi_i^s = \frac{\pi_i^p - c}{\pi_i^p - 1} \qquad 5$$

but $\widetilde{\pi}_i$ is correlated with $\boldsymbol{x}$. To sum up, if $\widetilde{\pi}_i$ is not correlated with $\boldsymbol{x}$, there is no selection bias unless Equation 4 is satisfied. If $\widetilde{\pi}_i$ is correlated with $\boldsymbol{x}$, there is a selection bias unless Equation 5 is satisfied. As a general rule, however, a significant effect of $\boldsymbol{x}$ on $\widetilde{\pi}_i$ would suggest the presence of a selection bias. Inversely, the selection bias can be ruled out if $\boldsymbol{x}$ is not correlated with $\widetilde{\pi}_i$.

Note that a second, similar, way of detecting the presence of selection bias involves comparing each coefficient of the patent production function estimated with priority filings at the reference office, with the corresponding coefficient estimated with total filings (priority filings + second filings) at the reference office. Equality of coefficients would suggest that there is no selection bias. We will report both the estimation of the determinants of $\widetilde{\pi}_i$ and the one-to-one Chow test of equality of coefficients in the empirical analysis.

Three additional comments are in order. First, the methodology detects the presence of a selection bias but is silent on the direction and the extent of the bias. As long as the output is observed at only one patent office, it is not possible to correct for the selection bias. Second, among the three patent counts available at one office (priority filings, second filings, and priority filings + second filings), the count of second filings is likely to be the least accurate. This is because it is prone to the two sources of errors: π_i^p and π_i^s. Second, the count of all patents at one office (priority filings + second filings) is likely to give more accurate estimates than the count of priority filings. The addition of second filings mitigates the potential bias induced by priority filings because the number of second filings that can be observed depends negatively on the number of priority filings already observed at the reference office. However, the count of all patents is not always more accurate than the count of priority filings since the possibility exists that the addition of second filings will reinforce the bias. As a result, it is good practice to report estimates of the patent production function with various counts (especially priority filings and total filings) to show the sensibility of the parameters, together with the estimation of the variable $\widetilde{\pi}_i$.

4. Econometric framework

The empirical analysis proceeds in two steps. First, patent production functions are estimated with multiple patent counts to explore the presence of a selection bias. Second, we study whether the variable $\tilde{\pi}_i$ and the Chow tests detect the selection bias.

Because patent number is a count variable, we use Poisson-based econometric models. The advantage of the Poisson regression is that it is consistent as long as the mean is correctly specified, see [22]. Taking the exponential form of equation 1 yields:

$$\begin{aligned} E[p_{it} \mid \boldsymbol{x_{it}}, \eta_i] &= \exp(\boldsymbol{x'}_{it}\,\boldsymbol{\beta} + \eta_i) \\ &= \mu_{it}\nu_i \text{ for } i = 1,...,N \text{ and } t = 1,...,T \end{aligned} \qquad 6$$

where p_{it} is the number of patents for firm (or university) i at time t, $\boldsymbol{x_{it}}$ is the vector of observable covariates, $\mu_{it} = \exp(\boldsymbol{x'}_{it}\,\boldsymbol{\beta})$, the term η_i is an unobservable individual firm-specific effect reflecting any permanent difference in the level of patents across firms. A popular estimation for count data models with fixed effects is the Poisson conditional maximum likelihood estimator proposed in [23]. However, consistency of the estimator relies on the strict exogeneity assumption of $\boldsymbol{x_{it}}$. This assumption is likely to be violated with patent production functions, because the patenting of an invention may call for further R&D. We adopt the 'pre-sample' estimator proposed in [24], which relaxes the strict exogeneity assumption (see also [25]). The fixed effect is approximated with the log of the pre-sample mean of the patent series, *i.e.* it reflects the patent practices and the entry-level knowledge stock of the firm. A dummy NO_PRE_PAT that takes the value of 1 if the firm had no

patents in the pre-sample period is added to capture the quasi-missing value in the log of patents. Recent studies that apply this estimation strategy include [26] and [27].

Three patent counts are used for the purpose of the analysis. The first, p^W, is the 'true' count of priority patent applications filed worldwide (the variable y_i^* in Equation 1). It is usually not observed by the researcher. Estimates with this benchmark count will be compared with estimates with single-office counts to study the effect of selection bias. The second, p^E, is the count of priority filings at the EPO. The third, $p^E + s^E$, is the count of priority and second filings at the EPO (the variable y_i in Equation 2). Although this count mixes patents of varying nature, it takes into account a broader set of patents than p^E, thereby potentially limiting the selection bias.

The measure for the single-office bias (variable $\tilde{\pi}_i$ in Equation 3) is estimated as a Bernoulli pseudo-maximum likelihood (PML) following [28] to account for the fact that the variable $\tilde{\pi}_i$ is bounded between 0 and 1:

$$E[\tilde{\pi}_{it} \mid \boldsymbol{x}_{it}] = h(\boldsymbol{x}'_{it}\,\boldsymbol{\gamma}) \qquad 7$$

where *h(z)* is a link function satisfying $0 \le h(z) \le 1 \;\forall\; z \in \mathrm{R}$ such as the logistic link function. An alternative approach, easier to implement but technically inaccurate, involves estimating the determinants of $\tilde{\pi}_i$ in a simple linear regression model using OLS.

We apply our methodology to data on Belgian firms and German universities. These two datasets are particularly appropriate for the purpose of illustrating the selection bias and

validating our test. Indeed, they provide us with two opposite institutional contexts: while Belgian firms rely on the EPO to a large extent, German universities mainly file their patents at home. Approximately 85 per cent of patents by Belgian firms in our sample are filed at the EPO, either as priority filings or as second filings. By contrast, almost 90 per cent of the patent applications by German universities are first filed at the German patent office and only half the patents eventually reach the EPO (mainly as second filings).[3] Since most of the patents by Belgian entities will end up at the EPO, one might expect little to no selection bias when the EPO is taken as the reference office. For the inverse reason, one might expect strong selection bias using German entities dataset given that a lower proportion of patents will end up at the EPO. It is thus interesting to study the presence of selection bias as well as how well our proposed test performs in these distinct institutional contexts.

5. Empirical test using data on Belgian firms

5.1 Data sources

Three databases were merged together for the purpose of the analysis. The first is the biannual R&D survey by the Government of the Flemish Community in Belgium. Three waves were used: 2004, 2006 and 2008, providing annual firm-level data on R&D-related variables for the period 2002–2008. The second is the Belfirst database by Bureau van Dijk, which provides yearly information on balance sheets and income statements. The databases are merged using the tax identification number. Finally, the Patstat database by the EPO (April 2009 version) was used to collect data on patents. Because patent applications are

[3] The patenting behavior of German universities seems to be similar to that of German firms in general: the 90 per cent figure matches the estimate presented in [14, Table 7] for the population of patents by German inventors.

published (hence observable) 18 months after the filing date, the data was collected up to 2007.

The worldwide count of patents used in this paper is novel and is a key aspect to evaluate the presence of selection bias. The construction of the patent indicator follows two logical steps. First, all the patent applications from inventions made in Belgium are identified. For the purpose of the analysis, it is particularly important to observe the *population* of priority filings invented in Belgium. It is done by identifying all the priority patent applications filed worldwide by inventors living in Belgium.[4] In practical terms, we look for patent applications by Belgian inventors in 52 patents offices.[5] Second, the name of the firms actually applying for the patents (the 'applicants' in jargon) is manually cleaned and harmonized. Patent applications are then matched with firm-level data using the harmonized name.

The sample is composed of all the companies that have at least one patent application in the period 2002–2007 and that are in at least two waves of the R&D survey. It contains 429 firm-year observations on 95 distinct firms and is thus slightly unbalanced.

5.2 The dependent variables

Table 1 presents descriptive statistics of the patent counts. Approximately 45 per cent of the priority filings by Belgian inventors over the period 2002–2007 were filed at the EPO (row (a)). This proportion is very large in comparison with the European average of less than 10

[4] The 'inventor' criterion reflects the origin of the inventive activity and ensures a good match with statistics on R&D, which specifically relate to the R&D expenditures within a country [29].
[5] These 52 offices account for 98.5 per cent of worldwide priority filings in 2005. See [14] for additional details.

per cent over roughly the same period [14]. Interestingly, the Belgian patent office receives less patent applications by Belgian inventors than the EPO: only 22 per cent of the priority patent applications by Belgian inventors are actually filed in Belgium (column (b) - column (a)). If both priority and second filings at the EPO are counted, the share of patents identified rises to around 85 per cent (row (e)).[6]

Table 1: Proportion of patents identified, Belgium firms

Number of priority patents:	507
Priority filings	
(a) EP [p^E]	0.45
(b) EP + BE	0.67
(c) EP + BE + US	0.73
(d) ALL [p^W]	1.00
Priority filings + second filings	
(e) EP [$p^E + s^E$]	0.85
(f) EP + BE	0.93
(g) EP + BE + US	0.96

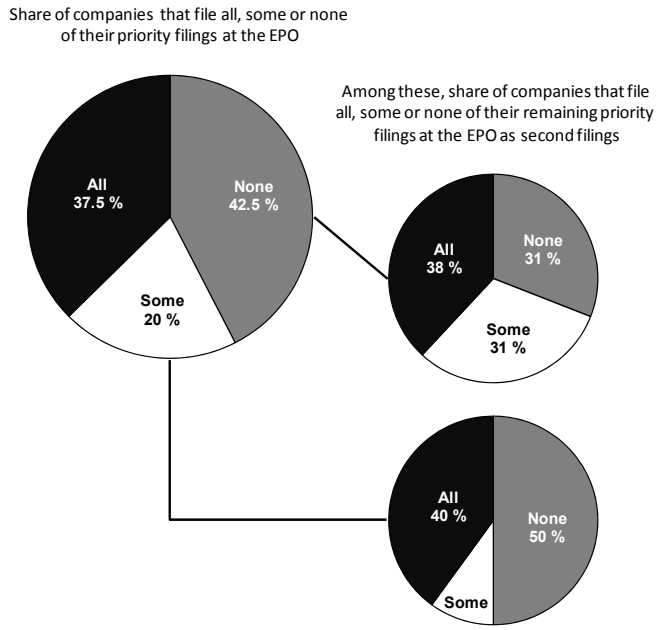
Notes: Figures computed at the patent level. 'BE' stands for the Belgian patent office, 'EP' for the EPO and 'US' for the USPTO.

Figures from Table 1 are computed using patent-level information and therefore hide any heterogeneity in the behavior of firms. Figure 1 reports firm-level statistics on the use of the EPO by Belgian patenting firms. As shown, 37.5 per cent of patenting firms in our sample file all of their priority patent applications at the EPO; 42.5 per cent of firms never file their priority applications at the EPO; and the remaining 20 per cent file some but not all their priority filings at the EPO. Among those that never file their priority filings at the EPO, 31 per cent do not file their second filings at the EPO. It follows that approximately 13 per cent of the Belgian patenting firms never file their patents at the EPO (31 per cent of 42.5 per cent) and are therefore excluded from the single office count. By contrast, the single office

[6] We consider a (priority) patent as identified when at least one member of its family is included in the sample.

count provides accurate information for about 61.5 per cent of firms if both priority filings and second filings are counted (37.5 per cent of firms that file all their priority patent applications at the EPO, plus 0.425*0.35 = 16 per cent of firms that have no priority filings at the EPO but all their second filings at the EPO, plus 0.20*0.40 = 8 per cent of firms whose patents reach the EPO through a mix of priority filings and second filings). Partial information is gleaned for the remaining 25.5 per cent of firms.

Figure 1: Share of Belgian companies that file their patent applications at the EPO



To sum up, 85 per cent of all patents by Belgian firms will eventually end up at the EPO (column (e) of Table 1), such that the single-office count seems a reasonable methodological choice. However, this high proportion masks important disparities across firms since partial or no information is collected for almost 40 per cent of firms (respectively 25.5 per cent and 13 per cent), as shown in Figure 1. The comparable figure is much higher in most other European countries, which rely on the EPO to a lesser extent. In this respect, the

Belgian case is a strong test of our claim. If a selection bias affects estimates for Belgian data, it is very likely that it will also affect estimates for data from other countries.

5.3 Covariates

We control for variables that are commonly found to affect the production of inventions such as firm size and research intensity [*e.g.,* 13, 27]. We include the number of full-time equivalent employees (EMP) as a measure of firm size. The ratio of tangible assets (CAPITAL) over the number of employees is a measure of the capital intensity of the firm. Similarly, the ratio of total R&D expenditures (RD) to the number of employees is an indication of the R&D intensity of the firm. We also include firm age (AGE), defined as the number of years the firm exists. Note that the logarithmic form of these variables is used in the empirical analysis, consistent with equation 1. The relationship between the degree of competition and innovation has been an ongoing topic of research for the last decades [30]. An attempt to capture the competitive environment of the firm is made by including a measure of international exposure (EXP). It is an ordinal variable that takes a value between 1 and 3 if the main competitors of the firm are located in Belgium (1), in Europe (2), or worldwide (3). Finally, the regression controls for 13 industry and five time dummies.

Table 2: Descriptive statistics

	Min	Mean	Max	Std. Dev.
EMP (FTE)	4	592	5,685	942
CAPITAL (in 000)	16	35,051	2,253,238	159,204
RD (in 000)	0	20,718	1,153,000	115,943
AGE	1	33	168	29
EXP (o)	1	2.59	3	0.60

Notes: N= 429. 'FTE' stands for full-time equivalent, 'in 000' for thousand Euros. '(o)' indicates an ordinal variable.

Table 2 provides the descriptive statistics. Firms in the sample are relatively large: the average firm has 592 employees and EUR 35 million in tangible assets, spends EUR 21 million in R&D, and is 33 years. The correlation structure of the variables used in the empirical analysis is presented in Table 3. Note that tangible assets and the research budget are divided by the number of employees to reduce multicollinearity.[7]

Table 3: Correlation coefficients

	ln(EMP)	ln(CAPITAL/EMP)	ln(RD/EMP)	ln(AGE)	EXP
ln(EMP)	-				
ln(CAPITAL/EMP)	0.14	-			
ln(RD/EMP)	0.05	-0.04	-		
ln(AGE)	0.49	0.05	0.16	-	
EXP	0.18	-0.04	0.42	0.05	-

5.4 Results

Table 4 presents estimates of the patent production function for the three dependent variables, as well as estimates of the tests for single-office bias.

[7] For instance, the correlation coefficient between ln(EMP) and ln(CAPITAL) is 0.84 (not reported) while the correlation coefficient between ln(EMP) and ln(CAPITAL/EMP) is only 0.14.

Table 4: Estimates of the patent production function and the selection bias

	(1)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
Equation:	6	6		6		7	
Dep. variable:	p^W	p^E		$p^E + s^E$		$\widetilde{\pi}$	
ln(EMP)	0.477***	0.473***	N	0.473***	N	0.0240	N
	(0.094)	*(0.096)*		*(0.102)*		*(0.209)*	
ln(CAPITAL/EMP)	-0.222**	-0.289**	N	-0.257**	N	0.369	N
	(0.102)	*(0.147)*		*(0.114)*		*(0.382)*	
ln(RD/EMP)	0.287***	0.597***	Y	0.240***	N	0.990**	Y
	(0.081)	*(0.127)*		*(0.093)*		*(0.448)*	
ln(AGE)	-0.055	-0.394**	Y	-0.053	N	-0.515*	Y
	(0.111)	*(0.158)*		*(0.135)*		*(0.273)*	
EXP	0.297	-0.124	N	0.632**	Y	-1.549*	Y
	(0.223)	*(0.300)*		*(0.299)*		*(0.938)*	
PRE_PAT	0.354**	-0.146		0.458***			
	(0.165)	*(0.256)*		*(0.172)*			
NO_PRE_PAT	0.219	-0.518		0.324			
	(0.173)	*(0.281)*		*(0.198)*			
NO_PATENT						-37.72***	
						(0.479)	
Constant	-4.878***	-3.759***		-5.862***		3.536	
	(0.920)	*(1.163)*		*(1.109)*		*(4.816)*	
Industry dummies	Y***	Y***		Y***		Y***	
Year dummies	Y***	Y***		Y***		Y	
R^2	0.58	0.56		0.55		0.81	

Notes: N = 429. The econometric method is a Poisson maximum likelihood in columns (1), (2a) and (3a), and a Bernoulli pseudo-maximum likelihood in column (4a). R^2 is computed as the square of the correlation coefficient between the dependent variable and its predicted value. Robust standard errors clustered at the firm level in parentheses. ***, **, * denote significance at the 1 per cent, 5 per cent and 10 per cent probability threshold respectively (test of joint significance for dummies).

The coefficients in column (1), obtained with the worldwide patent count p^W, should be compared with the coefficients estimated with the single-office count of priority filings in column (2a) and the coefficients estimated with the single-office count of total filings (priority filings + second filings) in column (3a). Column (2b) and (3b) report the results of the Chow tests for differences in coefficients. A value 'Y' indicates that the coefficient is different from the corresponding 'true' coefficient (column (1)) at the 10% probability threshold. For instance, ln(AGE) takes the value 'Y' in column (2b) because the coefficient

estimated with the count of priority filings at the EPO is statistically different from the coefficient in column (1).

Looking at column (1), firm age and the intensity of competition are not associated with differences in invention outcomes. The picture looks different if the count is limited to a single office, as shown in columns (2a) and (3a). Depending on the patent indicator, firm age and the exposure to larger markets are significant determinants of the patent count, seemingly suggesting that young firms and firms evolving in a global market are more 'innovative' than others. The true explanation, however, is different: these firms are simply more likely to file their patents at the EPO, reflecting a selection bias. A third bias occurs with respect to R&D intensity, which is significantly higher in column (2a) than in column (1).

Our proposed test, which uses only information available at the EPO, is twofold as explained in section 3. A first way of detecting bias is to estimate the determinants of the variable $\tilde{\pi}$, as shown in column (4a). The coefficients associated with the R&D intensity, age and exposure variables are significantly different from zero, suggesting that the methodology successfully identifies the presence of a selection bias. A one-to-one Chow test of difference in coefficients between column (2a) and column (3a) is a second way of detecting bias. The results, reported in column (4b), confirm the presence of bias for the same three variables.

In a nutshell, it seems that the variable $\tilde{\pi}$ contains information that allows detection of a selection bias. Two elements must be emphasized. First, both the count of priority filings and total filings at the EPO are biased, suggesting that researchers should estimate and report regression results for both counts. Second, the methodology has allowed successful detection

of bias and has not been affected by the risk of false negatives and false positives, as discussed in section 3.

5.5 Additional considerations

The patent production functions were estimated with a negative binomial regression model to account for possible over-dispersion of the dependent variables with no change to the results. The variable $\tilde{\pi}$ was also estimated with an OLS regression instead of the more sophisticated Bernoulli PML and the biases were successfully identified. We now explore two alternative approaches to control for selection bias. The first involves estimating zero-inflated Poisson regression models. The second involves weighting each patent observed by a measure of its value.

Zero-inflated Poisson regression models have been used to account for the fact that patenting is a rare event, particularly among small innovative firms. The zero-inflated Poisson distribution, introduced in [31], is a mixture between a degenerate distribution at zero with probability *p* and a Poisson distribution with probability 1-*p*. The aim is to increase the probability mass at zero to account for the greater occurrence of zero outcomes. Since the selection bias will exacerbate the occurrence of zero observations, one can wonder whether a zero-inflated Poisson regression model can be used to control for some of the effects of the selection bias. Estimates are presented in columns (1), (2a) and (3a) of Table 5. The upper panel presents estimates of the parameters of the Poisson distribution, while the lower panel models the probability of having a zero outcome (the inflation equation). Thus, a variable with a positive coefficient in the lower panel increases the probability of observing no patent. Looking at the results, it seems that the inflation equation does not eliminate the biases. This

is apparent in columns (2b) and (3b), which report the results of the Chow test for a difference in coefficients with column (1). As compared with the simple Poisson regression, the zero-inflated Poisson has made matters worse for priority filings.

Value-weighted counts are another possible way of removing the effect of selection bias, although their use in empirical studies remains the exception rather than the rule. For instance, only 25 per cent of the articles surveyed in section 2 use a value-weighted count. In theory, value-weighted patent indicators can mitigate the selection bias if more valuable patents are more likely to be filed at the reference office: since a low weight is given to low-value patents, which are also less likely to be observed at the reference office, the single office value-weighted count gets closer to the 'true' value-weighted count. There are three main measures of patent value: the number of years the patent has been maintained in force (*useful life*), the number of citations it has received (*citations*), and the number of countries in which it was filed (*family size*).[8] The first measure, useful life, is available only for patents that were filed 20 or more years ago, corresponding to the maximum number of years a patent can be held in force. For patents that are less than 20 years of age and still in force, the useful life will necessarily be truncated. Since our analysis uses recent data, this value variable would be severely truncated. The second measure, citations, raises practical hurdles in our context as it is not possible to build a proper benchmark. Patent citation practices vary greatly across patent offices and their interpretation is often office-specific [33]. As a result, it makes little sense to weight the exhaustive patent count p^W with the citations received across patent offices. In addition, the information on citations is not exhaustive in the Patstat database and is missing for some patent offices. The third measure, family size, is more appealing. It involves weighting each patent by the number of members in the patent family.

[8] We refer the reader to [32] for a recent review of patent value indicators.

Since the family may spread worldwide, this measure necessitates observing the whole population of patents. A researcher that is able to compute a family-weighted count is thus theoretically also able to build the exhaustive patent count. In other words, a proper value-weighted count cannot be computed if the researcher does not observe patents worldwide, precisely when our test should be used.

Nevertheless, we report family-weighted estimates in columns (4), (5a) and (6a) of Table 5 for the sake of completeness. The effect of a selection bias is still observed for both counts, as suggested by the Chow tests reported in columns (5b) and (6b). Interestingly, the bias seems smaller in size than the original, non-value-weighted, estimates, at least as far as total filings at the EPO are concerned. Three cautionary comments are in order. First, this methodology works only if the reference office attracts the most valuable patents. This is likely to be the case with the EPO, but not with national patent offices. If an office attracts the least valuable patents, then the use of a value indicator further distorts the count. Second, there are reasons for believing that the methodology will not work for many other countries. As described in section 3, Belgium has a very high share of patents that eventually ends up at the EPO (around 85 per cent). This situation is particularly favorable since the patents not observed at the EPO are likely to be of much lower value. Third, as already noted, a researcher that has enough data to compute a proper, value-weighted count has *a priori* also enough data to compute the exhaustive count.

Table 5: Patent production functions estimated with different specifications

	(1)	(2a)	(2b)	(3a)	(3b)	(4)	(5a)	(5b)	(6a)	(6b)
Dep. Variable:	p^W	p^E		$p^E + s^E$		$\tilde{p}^W$	$\tilde{p}^E$		$\tilde{p}^E + \tilde{s}^E$	
ln(EMP)	0.439***	0.713***	Y	0.470***	N	0.397***	0.413***	N	0.394***	N
	(0.080)	*(0.102)*		*(0.076)*		*(0.102)*	*(0.121)*		*(0.103)*	
ln(CAPITAL/EMP)	-0.305**	-0.542***	N	-0.477***	Y	-0.152	-0.340	N	-0.148	N
	(0.128)	*(0.159)*		*(0.010)*		*(0.133)*	*(0.242)*		*(0.144)*	
ln(RD/EMP)	0.152**	0.552***	Y	0.119	N	0.192	0.653***	Y	0.160	Y
	(0.069)	*(0.128)*		*(0.086)*		*(0.117)*	*(0.139)*		*(0.123)*	
ln(AGE)	-0.003	-0.355***	Y	-0.0323	N	0.200	-0.184	Y	0.250	N
	(0.092)	*(0.113)*		*(0.097)*		*(0.188)*	*(0.202)*		*(0.213)*	
EXP	0.527**	-0.270	Y	0.630**	N	0.589*	-0.368	Y	0.801**	Y
	(0.212)	*(0.271)*		*(0.252)*		*(0.309)*	*(0.392)*		*(0.348)*	
Inflation Equation:										
ln(EMP)	-0.125	0.853		-0.0746						
	(0.120)	*(0.595)*		*(0.118)*						
ln(CAPITAL/EMP)	-0.123	-0.779**		-0.191						
	(0.169)	*(0.389)*		*(0.165)*						
ln(RD/EMP)	-0.254**	-0.0295		-0.196						
	(0.102)	*(0.205)*		*(0.133)*						
ln(AGE)	0.397*	0.101		0.459*						
	(0.218)	*(0.280)*		*(0.268)*						
COMP	0.670*	-0.511		0.346						
	(0.378)	*(0.608)*		*(0.455)*						
R^2	0.57	0.58		0.54		0.63	0.45		0.61	

Notes: N = 429. Industry dummies, time dummies and pre-sample fixed effects included. '$\tilde{x}$' indicates that x is weighted by its family size. R^2 is computed as the square of the correlation coefficient between the dependent variable and its predicted value. Robust standard errors clustered at the firm level in parentheses. ***, **, * denote significance at the 1 per cent, 5 per cent and 10 per cent probability threshold respectively (test of joint significance for dummies).

6. Empirical test using data on German universities

6.1 Data sources

We apply our methodology to patents owned by German research universities. To identify the entire population of university-owned patents we first searched for the university names and variations thereof in Patstat (April 2010 version). Second, we manually cleaned and harmonized the applicant names. Third, we identified all priority applications filed worldwide

by these universities and matched patent applications to university-level characteristics. We obtain a final sample of 67 universities that have filed at least one patent application since the year 2000. We combined this dataset with data obtained from the German Federal Statistical Office (*Statistisches Bundesamt*), the German Research Foundation (*Deutsche Forschungsgemeinschaft*) and Schmude and Heumann [34]. Data is available for years 2006 and 2007.

6.2 The dependent variables

As in the previous empirical application, several patent counts are used: the worldwide count of priority patent applications (p^W); the count of priority filings at the EPO (p^E); and the count of all patents applied for at the EPO ($p^E + s^E$). We also use the count of all patents applied for at the German patent office (p^D). Results obtained with this latter count are interesting to look at given the predominance of the national filing route in Germany. The data is summarized in Table 6.

Table 6: Proportion of patents identified, German universities

Number of priority patents:	1,327
Priority filings	
(a) EP [p^E]	0.13
(b) EP + DE	0.98
(c) EP + DE + US	0.99
(d) ALL [p^W]	1.00
Priority filings + second filings	
(e) EP [$p^E + s^E$]	0.57
(f) EP + DE	0.99
(g) EP + DE + US	0.99

Notes: Figures computed at the patent level. 'DE' stands for the German patent office, 'EP' for the EPO and 'US' for the USPTO.

Around 13 per cent of patent applications by German universities in our sample were first filed at the EPO (row (a)), whereas approximately 85 per cent of the priority patent applications were filed at the German patent office (row (b) – row (a)). Counting both priority filings and second filings at the EPO (row (e)) allows for identification of 57 per cent of all patent applications by German universities. These figures contrast with those presented in Table 1 for Belgian firms. German universities rely on the EPO to a lesser extent and we expect a strong selection bias when the EPO is used as the reference office. The key question is whether our proposed test allows us to identify the sources of the bias given that less information (*i.e.* a lower number of patents) is available.

6.3 Covariates

Patent production functions have already been adapted to the university context. Much like firm-level studies, they typically control for size variables and measures of research capability [35, 36, 37]. Thus, we include the following university-level characteristics in the analysis: the number of professors (NB_PROF) to capture university size; the amount of third-party funding normalized by the number of professors (BUDGET/NB_PROF) to capture the financial situation of the university; the ratio of third-party funding obtained by private sources over total third-party funding (SHARE_PRIVATE) as a proxy for the intensity of the university–industry relationship; university age (AGE); a dummy variable indicating the presence of a medical school (UNI_HOSP); and a dummy variable that takes the value 1 if the university is a technical university (as opposed to traditional) university. Additionally, we include performance measures in other realms: a dummy variable indicating the outcome of the excellence initiative (*Exzellenzinitiative*) by the German federal government in respect of the second line of funding – excellence clusters – in 2006

(EXCELLENCE); universities' strength in basic research (RES_STRENGTH) measured by 2005–2007 rankings of the German Research Foundation; and a dummy variable for the top ten universities (ENTR_ORIENTATION) in the ranking of universities' entrepreneurial orientation available in Schmude and Heumann [34].[9,10]

Table 7 provides descriptive statistics and Table 8 provides the correlation matrix of all covariates. The average university in our sample has 283 professors and has attracted €41.5 million of third-party funding, 75 per cent of which came from industry. Around 20 per cent of universities are technical universities, and approximately half the universities have a hospital. Approximately one in four universities in our sample was awarded funds under the excellence initiative (13 universities in 2006 and 18 universities in 2007). Finally, note that two universities (TU Cottbus and U Hohenheim) rank 10 ex aequo in the ranking of most entrepreneurial universities, such that 16 per cent of universities in our sample are listed as being highly entrepreneurial.

Table 7: Descriptive statistics

	Min	Mean	Max	Std. Dev.
NB_PROF	24	283	699	141
BUDGET (in 000)	5,128	41,475	168,329	30,437
SHARE_PRIVATE	0.48	0.75	1.00	0.12
AGE	9	203	629	196
TU (d)	0	0.22	1	-
UNI_HOSP (d)	0	0.49	1	-
EXCELLENCE (d)	0	0.23	1	-
RES_STRENGTH (o)	0.01	0.07	1	0.14
ENTR_ORIENTATION (d)	0	0.16	1	-

Notes: N= 134.'in 000' for thousand Euros. '(d)' indicates a dummy variable, and '(o)' an ordinal variable.

[9] The first two rounds of the *Exzellenzinitiative*, started in 2005, competitively allocated €1.9 billion across three funding lines: graduate schools, centers of excellence, and university-level strategies over a period of six years. The majority of funds were spent on the second funding line.

[10] Regarding the variable RES_STRENGTH, we use the inverse of the rankings to facilitate interpretation in the empirical analysis.

Table 8: Correlation coefficients

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
(a) ln(NB_PROF)	-								
(b) ln(BUDGET/NB_PROF)	-0.09	-							
(c) ln(SHARE_PRIVATE)	0.18	0.05	-						
(d) ln(AGE)	0.34	0.06	0.07	-					
(e) TU	-0.20	0.57	-0.29	-0.03	-				
(f) UNI_HOSP	0.49	-0.09	0.14	0.35	-0.31	-			
(g) EXCELLENCE	0.43	0.20	0.17	0.22	0.16	0.28	-		
(h) RES_STRENGTH	0.34	0.32	0.13	0.21	0.17	0.26	0.32	-	
(i) ENTR_ORIENTATION	0.06	-0.02	0.25	-0.25	0.15	0.05	0.27	0.06	-

6.4 Results

Table 9 presents estimates of the patent production function for various dependent variables, as well as estimates of the single-office bias.

Five variables have a significant impact on the true number of patents produced as indicated in column (1). The patent count increases with university size (NB_PROF) and the research budget (BUDGET/NB_PROF). The higher the share of this budget that comes from private sources of funds, the lower the number of patents owned by universities (SHARE_PRIVATE). This result is probably explained by the fact that private providers of funds negotiate ownership of inventions produced during the period of collaboration. Finally, technical universities and universities with a research hospital file significantly more patent applications than other universities.

Table 9: Estimates of the patent production function and the selection bias

	(1)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)	(5a)	(5b)
Equation:	6	6		6		6		7	
Dep. variable:	p^W	p^E		$p^E + s^E$		p^D		$\widetilde{\pi}$	
ln(NB_PROF)	0.690***	1.988***	Y	0.835***	N	0.554***	Y	0.751	Y
	(0.178)	*(0.322)*		*(0.176)*		*(0.187)*		*(0.480)*	
ln(BUDGET/NB_PROF)	0.334*	1.399***	Y	0.209	N	0.317	N	-0.125	Y
	(0.176)	*(0.344)*		*(0.185)*		*(0.205)*		*(0.421)*	
ln(SHARE_PRIVATE)	-1.163**	-0.513	N	-0.105	Y	-1.341**	N	2.287*	N
	(0.468)	*(0.850)*		*(0.584)*		*(0.535)*		*(1.392)*	
ln(AGE)	0.108	0.159	N	0.0260	N	0.0825	N	0.200	N
	(0.071)	*(0.117)*		*(0.093)*		*(0.082)*		*(0.184)*	
TU	0.604***	-0.981***	Y	0.403***	N	0.737***	Y	-0.883**	Y
	(0.150)	*(0.364)*		*(0.147)*		*(0.178)*		*(0.408)*	
UNI_HOSP	0.377**	-0.372	Y	0.211	N	0.509***	Y	-0.490	Y
	(0.163)	*(0.285)*		*(0.184)*		*(0.178)*		*(0.476)*	
EXCELLENCE	0.0454	-0.271	N	-0.0644	N	0.0802	N	-0.408	N
	(0.094)	*(0.270)*		*(0.109)*		*(0.102)*		*(0.295)*	
RES_STRENGTH	-0.206	-0.300	N	0.629**	Y	-0.320	N	0.704	Y
	(0.349)	*(0.477)*		*(0.275)*		*(0.430)*		*(0.967)*	
ENTR_ORIENTATION	-0.160	0.294	Y	-0.341**	N	-0.239	N	0.696*	Y
	(0.168)	*(0.224)*		*(0.163)*		*(0.196)*		*(0.400)*	
PRE_PAT	0.148**	-2.667***		0.247**		0.172**			
	(0.072)	*(0.836)*		*(0.122)*		*(0.078)*			
NO_PRE_PAT	-0.221	-0.739		0.0543		-0.194			
	(0.204)	*(0.452)*		*(0.210)*		*(0.225)*			
NO_PATENT								-16.90***	
								(0.514)	
Constant	-4.676***	-17.80***		-4.677***		-4.088***		-4.492	
	(1.224)	*(2.544)*		*(1.379)*		*(1.380)*		*(3.515)*	
R^2	0.71	0.37		0.50		0.71		0.25	

Notes: N = 134. Year dummy included. The econometric method is a Poisson maximum likelihood in columns (1), (2a), (3a) and (4a), and a Bernoulli pseudo-maximum likelihood in column (5a). R^2 is computed as the square of the correlation coefficient between the dependent variable and its predicted value. Robust standard errors clustered at the university level in parentheses. ***, **, * denote significance at the 1 per cent, 5 per cent and 10 per cent probability threshold respectively.

Results obtained with the count of priority filings at the EPO in column (2a) provide a different picture. In particular, the size and budget variables have a significantly larger effect (see column (2b)), suggesting that larger and better-funded universities have a higher propensity to file priority patent applications at the EPO. A selection bias is also found for

technical universities, universities with a hospital, and universities with a high entrepreneurial orientation – although none of the coefficients in columns (1) and (2a) for the latter variable are actually significantly different from zero, they are statistically different from each other. Extending the patent count to all patent applications at the EPO (column (3a)) somewhat improves the accuracy of estimates, although significant biases remain or arise. For instance, the research strength (RES_STRENGTH) is associated with a significantly higher patent count, yet the true effect in column (1) is not significantly different from zero. Universities with strong research capability may produce inventions with high market potential, explaining the need to seek protection at the EPO. Overall, coefficients for seven out of the nine variables in our model are affected by selection bias.

Our proposed tests are reported in the last two columns of Table 9. The regression for the single-office bias in column (5a) rightly identifies three biased coefficients, while the Chow tests for differences in coefficients between columns (2a) and (3a) allows us to identify four additional biased coefficients. It follows that all the variables affected by selection bias are correctly identified by the procedure (no false negatives). There are also no false positives since the two coefficients that our tests suggest are not biased (AGE and EXCELLENCE) are, indeed, not affected by selection bias. Thus, this second analysis relating to German universities largely supports the findings of the first analysis relating to Belgium firms, despite the fact that the reference office attracts a lower proportion of total patents.

Much like the analysis of Belgian firms, estimates obtained with the total count of EPO patents are generally closer to the true value as compared with the count of EPO priority filings. However, the former indicator is not unambiguously better: using EPO total filings

instead of EPO priority filings induces a bias for the variables SHARE_PRIVATE and RES_STRENGTH.

We have also reported coefficients obtained with the count of patents filed at the German patent office (priority filings + second filings) in column (4a). Since the German patent office receives most patent applications, it is particularly interesting to investigate the presence of a selection bias. Our results suggest that the coefficients are usually closer to their true values than coefficients obtained with EPO patents. However, we also detect significant biases – in particular with the variables NB_PROF, TU and UNI_HOSP, as indicated in column (4b). Interestingly, these coefficients are only marginally closer to their true values than those obtained with EPO patents. For instance, the absolute value of the point estimate of the bias associated with the size variable is 0.136 for patents filed at the German patent office |0.690-0.544| compared with 0.145 for patents filed at the EPO |0.690-0.835|.[11]

7. Discussion and concluding remarks

This paper takes a close look at the widespread practice in innovation studies of using one single office of reference for counting patents. It uses novel datasets of the whole population of patents filed by Belgian firms and German universities to show that the single-office count of patents results in biased estimates of patent production functions.

The main contribution of the paper is a methodological one. It proposes a way to test for the existence of a selection bias. The methodology involves estimating the determinants

[11] Although we use the EPO as a reference office, we have also performed our test taking the German patent office as reference. The test led to inconclusive results, largely owing to the very small number of second filings at the German patent office. Among the 1,138 patent applications filed at the German patent offices, 1,131 were priority filings. However, there is no reason to suspect that our test does not work if the reference office attracts a large enough number of priority filings *and* second filings.

of the proportion $\tilde{\pi}$ of priority patent applications filed at the reference office among total patent applications at the reference office. Alternatively, the selection bias can also be detected by performing one-to-one comparisons of coefficients obtained with priority filings and total filings at the reference office. In fact, both tests should be used in conjunction to maximize the chance of detecting any bias. The empirical applications suggest that the tests successfully identify coefficients that are affected by selection bias. The issue of the single-office bias discussed in this paper extends beyond the framework of patent production functions and is relevant to other empirical settings. Generally speaking, the single-office count results in a selection bias when the choice of the filing route affects the accuracy of the patent indicator used (either as an input or as an output) in any empirical setting.

The study comes with one implication and one recommendation. A major implication is that estimates based on a single-office count of patents should be treated with caution. For instance, the empirical application of the test to Belgian firm data uses a variable that captures the market exposure of the firm. The effect of exposure on innovation is observed with international, high-value patents but not with total patents. Similarly, the empirical application of the test to German university data uses a variable that captures the research strength of the university. The effect of research strength is observed with EPO patents but not with total patents. These results provide clear evidence that the choice of the filing route is not random. In theory, many dimensions that affect the productivity of R&D also affect the choice of filing route. Hence, particular attention should be paid to the patent indicators that are used in future studies, and researchers should think carefully about and justify which patents they count.

The study also helps define best practices in patent statistics. It suggests that the count of patents should be global in order to avoid selection bias, and not limited to a single patent office. If it is not feasible to count patents globally, we would recommend using the patent office that attracts the larger proportion of patents, taking into account both priority filings and second filings. If the proportion of patents identified relative to the global count is large, as in Germany or the United States, then researchers can be reasonably confident of the accuracy of their results, although the threat of selection bias will always loom. Our empirical test using German university data suggests that coefficients obtained from patents filed at the German patent office are indeed generally closer to their true value than coefficients obtained with EPO patents. However, our results also demonstrate the presence of selection bias for patents filed at the German patent office. These biases are not necessarily smaller in magnitude (or in statistical significance) than when the EPO is taken as the reference office. Sometimes, however, for practical reasons often related to data quality or availability, the second or third most popular patent office is taken as the reference office (such as the EPO for European firms or the USPTO for Canadian and Israeli firms). If patents reach the reference office through a mix of priority filings and second filings, then good practice would involve reporting estimates of patent production functions for priority filings and total filings (*i.e.* priority filings + second filings) to show the sensibility of the parameters, together with estimates of the determinants of the variable $\widetilde{\pi}$. If the focal variable does not affect the variable $\widetilde{\pi}$, then our results suggest that the researcher can be confident that the coefficient associated with the focal variable is not biased by the patent indicator used. On the contrary, if the test detects the presence of a selection bias, then researchers should be aware that their findings cannot be generalized and may solely reflect a selection bias.

This study comes with a number of caveats and suggested possibilities for further research, which we discuss briefly here. First, it should be noted that the variable $\tilde{\pi}$ does not perfectly capture the selection bias. In particular there are well-defined, though very specific, conditions under which: i) the focal variable is not correlated with $\tilde{\pi}$ even though there is a selection bias; and ii) the focal variable is correlated with $\tilde{\pi}$ even though there is no selection bias. Although we did not come across such cases in the empirical analysis, the possibility of false negatives and false positives exists, at least in theory. The proposed methodology is a step in the right direction which we hope will contribute to improving empirical research in the field. Second, use of this test requires knowledge of the priority status of patent documents. This information is available in most databases either directly, or indirectly by looking at the priorities claimed by the patent document. By definition, a patent that does not claim any priority is itself a priority, thus the priority status of the document can be collected at a low additional cost.

References

[1] M. Jefferson, The geographic distribution of inventiveness, Geographical Review 19(4) (1929) 649–661.

[2] R. Merton, Fluctuations in the rate of industrial invention, Quarterly Journal of Economics 49(3) (1935) 454–474.

[3] K. Pavitt, Patent statistics as indicators of innovative activities: Possibilities and problems Scientometrics, 7(1–2) (1985) 77–99.

[4] Z. Griliches, Patent statistics as economic indicators: A survey, Journal of Economic Literature 28(4) (1990) 1661–1707.

[5] F. Scherer, Firms size, market structure, opportunity and the output of patent inventions, American Economic Review, 55(5) (1965) 1097–1125.

[6] P. Stoneman, Patenting activity: A re-evaluation of the influence of demand pressures, Journal of Industrial Economics 27(4) (1979) 385–401.

[7] A. Nielsen, Patenting activity and market structure – correcting for self-selection in a sample of Danish patent holders, Technological Forecasting & Social Change 66(1) (2001) 47–58.

[8] P. Aghion, N. Bloom, R. Blundell, R. Griffith, P. Howitt, Competition and innovation: An inverted-U relationship, Quarterly Journal of Economics 120(2) (2005) 701–728.

[9] A. Jaffe, Technological opportunity and spillovers from R&D: Evidence from firms' patents, profits, and market value, American Economic Review 76(5) (1986) 984–1001.

[10] M. Cincera, Patents, R&D, and technological spillovers at the firm level: Some evidence from econometric count models for panel data, Journal of Applied Econometrics 12(3) (1997) 265–280.

[11] D. Czarnitzki, B. Ebersberger, A. Fier, The relationship between R&D collaboration, subsidies and R&D performance: Empirical evidence from Finland and Germany, Journal of Applied Econometrics 22(7) (2007) 1347–1366.

[12] A. Pakes, Z. Griliches, Patents and R&D at the firm level: A first report, Economics Letters 5(7) (1980) 377–381.

[13] B. Crépon, E. Duguet, J. Mairesse, Research, innovation, and productivity: an econometric analysis at the firm level, Economics of Innovation and New Technology 7(2) (1998) 115–158.

[14] G. de Rassenfosse, H. Dernis, D, Guellec, L. Picci, B. van Pottelsberghe de la Potterie, The worldwide count of priority patents: A new indicator of inventive performance, Research Policy, accepted.

[15] M. Seip, Matching patent data in the Netherlands, Paper presented at the Patent Statistics for Decision Makers 2010. European Patent Office, Vienna, Austria (2010)

[16] G. de Rassenfosse, B. van Pottelsberghe de la Potterie, A policy insight into the R&D-patent relationship, Research Policy 38(5) (2009) 779–792.

[17] J. Azagra, A. Yegros, F. Archontakis, What do university patent routes indicate at regional level? Scientometrics 66(1) (2006) 219-230.

[18] N. van Zeebroeck, B. van Pottelsberghe de la Potterie, Filing strategies and patent value, Economics of Innovation and New Technology 20(6) (2011) 539–562.

[19] P. Jensen, R. Thomson, J. Yong, Estimating the patent premium: Evidence from the Australian Inventor Survey, Strategic Management Journal 32(10) (2011) 1128–1138.

[20] J. Tobin, Estimation of relationships for limited dependent variables, Econometrica 26(1) (1958) 24–36.

[21] J. Heckman, Sample selection bias as a specification error, Econometrica 47(1) (1979) 153–161.

[22] C. Gourieroux, A. Monfort, A. Trognon, Pseudo maximum likelihood methods: Applications to Poisson models. Econometrica 52(3) (1984) 701–720.

[23] J. Hausman, B. Hall, Z. Griliches, Econometric models for count data with an application to the patents-R&D relationship, Econometrica 52(4) (1984) 909–938.

[24] R. Blundell, R. Griffith, J. Van Reenen, Market share, market value and innovation in a panel of british manufacturing firms, Review of Economic Studies, 66(3) (1999) 529–554.

[25] R. Blundell, R. Griffith, F. Windmeijer, Individual effects and dynamics in count data models, Journal of Econometrics 108(1) (2002) 113–131.

[26] Y. Uchida, P. Cook., Innovation and market structure in the manufacturing sector: An application of linear feedback models, Oxford Bulletin of Economics and Statistics 69(4) (2007) 557–580.

[27] D. Czarnitzki, K. Kraft, S. Thorwarth, The knowledge production of ‘R’ and ‘D’, Economics Letters 105(1) (2009) 141–143.

[28] L. Papke, J. Wooldridge, Econometric methods for fractional response variables with an application to 401 (K) plan participation rates, Journal of Applied Econometrics 11(6) (1996) 619–632.

[29] Organisation for Economic Co-operation and Development (OECD), OECD Patent Statistics Manual, Paris, 2009, 158 p.

[30] R. Gilbert, Looking for Mr. Schumpeter : Where are we in the competition–innovation debate?, Innovation Policy and the Economy 6 (2006) 159–215.

[31] D. Lambert, Zero-inflated regression, with an application to defects in manufacturing, Technometrics 34(1) (1992) 1–14.

[32] N. van Zeebroeck, The puzzle of patent value indicators, Economics of Innovation and New Technology 20(1) (2011) 33–62.

[33] D. Harhoff, K. Hoisl, C. Webb, European Patent Citations – How to count and how to interpret them, Unpublished manuscript, University of Munich (2008)

[34] J. Schmude, S. Heumann, Vom Studenten Zum Unternehmer: Welche Universität bietet die besten Chancen? (2007) Handelsblatt Verlag, Düsseldorf.

[35] T. Coupé, Science is golden: Academic R&D and university patents, Journal of Technology Transfer 28(1) (2003) 31–46.

[36] A. Payne, A. Siow, Does federal research funding increase university research output?, Advances in Economic Analysis & Policy 3(1) (2003) Article 1.

[37] J. Azagra-Caro, N. Carayol, P. Llerena, Patent production at a European research university: exploratory evidence at the laboratory level, Journal of Technology Transfer 31(2) (2006) 257-268.

Appendix

INSERT TABLE A ABOUT HERE

References in Table A

[38] H. Ernst, Industrial research as a source of important patents, Research Policy 27 (1998) 1–15.

[39] E. Brouwer, A. Kleinknecht, Innovative output, and a firm's propensity to patent. An exploration of CIS micro data, Research Policy 28 (1998) 615–624.

[40] V. Meliciani, The relationship between R&D, investment and patents: a panel data analysis, Applied Economics 32(11) (2000) 1429–1437.

[41] J. Furman, M. Porter, S. Stern, The determinants of national innovative capacity, Research Policy 31(66) (2002) 899–933.

[42] M. Fritsch, Measuring the quality of regional innovation systems: A knowledge production function approach, International Regional Science Review 25(1) (2002) 86–101.

[43] L. Bottazzi, G. Peri, Innovation and spillovers in regions: Evidence from European patent data, European Economic Review 47 (2003) 687–710.

[44] R. Salomon, J. Shaver, Learning by exporting: New insights from examining firm innovation, Journal of Economics and Management Strategy 14(2) (2005) 431–460.

[45] H. Ulku, R&D, innovation, and growth: evidence from four manufacturing sectors in OECD countries, Oxford Economic Papers 59(3) (2007) 513–535.

[46] N. Carayol, Academic incentives, research organization and patenting at a large French university, Economics of Innovation and New Technology 16(2) (2007) 119–138.

[47] M. Mariani, M. Romanelli, "Stacking" and "picking" inventions: The patenting behavior of European inventors, Research Policy 36 (2007) 1128–1142.

[48] G. Tappeiner, C. Hauser, J. Walde, Regional knowledge spillovers: Fact or artifact?, Research Policy 37 (2008) 861–874.

[49] I. Akçomak, B. ter Weel, Social capital, innovation and growth: Evidence from Europe, European Economic Review 53 (2009) 544–567.

[50] J. Hoekman, K. Frenken, F. Oort, The geography of collaborative knowledge production in Europe, Annals of Regional Science 43(3) (2008) 721–738.

[51] M. Buesa, J. Heijs, T. Baumert, The determinants of regional innovation in Europe: a combined factorial and regression knowledge production function approach, Research Policy 39 (2010) 722–735.

[52] L. Picci, The internationalization of inventive activity: A gravity model using patent data, Research Policy 39 (2010) 1070–1081.

[53] D. Fornahl, T. Broekel, R. Boschma, What drives patent performance of German biotech firms? The impact of R&D subsidies, knowledge networks and their location, Papers in Regional Science, 90(2) (2011) 395–418.

[54] F. Rentocchini, Sources and characteristics of software patents in the European Union: Some empirical considerations, Information Economics and Policy 23(1) (2011) 141–157.

Non-listed references are in the main reference list.

Table A: Overview of patent indicators used in the literature

	Geographic area	Time period	Cross-section vs. panel	Sample size	Office(s)	PF vs. SF	Application vs. grant	Value
[10]	Worldwide	1983–1991	Panel	181 firms	EPO	Undisclosed	A	N
[38]	Europe and Japan	1990–1994	Cross-section	25 firms	EPO	Undisclosed	A	Y
[39]	The Netherlands	1992; 1998	Cross-section	148 firms	EPO	Undisclosed	A	N
[40]	OECD	1973–1999	Panel	180 country-sectors	USPTO	Undisclosed	G	N
[41]	OECD	1973–1996	Panel	17 countries	USPTO	Undisclosed	G	N
[42]	Europe	1995–1998	Cross-section	707 firms	Undisclosed	Undisclosed	A	N
[43]	Europe	1977–1995	Cross-section	86 regions	EPO	Undisclosed	G	N
[8]	U.K.	1973–1994	Panel	311 firms	USPTO	Undisclosed	G	Y
[44]	Spain	1990-1997	Panel	3,060 firms	Spanish PO, EPO	Undisclosed	A	N
[45]	OECD	1981–1997	Panel	68 country-sectors	USPTO	Undisclosed	G	N
[46]	France	1995–2000	Cross-section	941 scholars	French PO, EPO, PCT	PF & SF	A	Y
[47]	Europe	1988–1998	Cross-section	793 inventors	EPO	Undisclosed	A	Y
[48]	Europe	1999	Cross-section	51 regions	EPO	Undisclosed	A	N
[26]	Belgium	1993–2003	Panel	122 firms	EPO	Undisclosed	A	N
[49]	Europe	1990; 2000	Cross-section	102 regions	EPO	Undisclosed	A	N
[50]	Europe	1988–2001	Cross-section	1,316 regions	EPO	Undisclosed	Undisclosed	N
[51]	Europe	1995–2001	Panel	146 regions	EPO	Undisclosed	Undisclosed	N
[52]	Worldwide	1990–2005	Panel	42 countries	NPOs	PF	A	N
[53]	Germany	1997–2004	Panel	129 firms	EPO, PCT	Undisclosed	Undisclosed	N
[54]	Worldwide	2000–2003	Panel	979 firms	EPO	Undisclosed	A	Y

Notes: ‘PF’ stands for ‘priority filings’; ‘SF’ for ‘second filings’; ‘PO’ for ‘patent office’; and ‘NPO’ for ‘national patent office’.