KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

Selection bias in innovation studies: a simple test

Gaétan de Rassenfosse and Annelies Wastyn

DEPARTMENT OF MANAGERIAL ECONOMICS, STRATEGY AND INNOVATION (MSI)

OR 1201

Selection bias in innovation studies:
A simple test

Gaétan de Rassenfosse
The University of Melbourne. Melbourne Institute of Applied Economic and Social Research,
and Intellectual Property Research Institute of Australia.
Level 7, Alan Gilbert Building, Victoria 3010, Australia.
gaetand@unimelb.edu.au. Corresponding author

Annelies Wastyn
K.U.Leuven, Dept. of Managerial Economics, Strategy and Innovation,
Naamsestraat 69, 3000 Leuven, Belgium
annelies.wastyn@econ.kuleuven.be

December 06, 2011

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

JEL Codes: O31, C18, C52, C81
Keywords: innovation production function, patent, 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. Yet, the measurement of innovation using patent data suffers from limitations. The two most severe of these limitations, discussed extensively in the literature, 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 the original works by Jefferson (1929), Merton (1935), Pavitt (1985) and Griliches (1990) for in-depth discussions of these issues.

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 firms' innovation 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. In addition to filing patents at their national patent office, companies have the option of filing patents at the European Patent Office (EPO). Companies that target an international market may also file their patents at the World Intellectual Property Office (WIPO) in Geneva, at the US Patent and Trademark Office (USPTO), or in any other jurisdiction. As long as firms' 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, the objective of this paper is twofold. First, we exploit a novel dataset of patent applications to study whether the single-office count biases econometric estimates of innovation production functions. Innovation production functions, which relate a firm's inventive input to its output, are a key object of analysis in the innovation literature.

Using data on Belgian patenting firms, we find evidence of a selection bias. Second, we propose a simple way to test the existence of bias when the econometrician observes patents at only one patent office. We apply our methodology to the sample of Belgian firms to demonstrate its usefulness. The 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 among innovation scholars.

The paper is structured as follows. The next section surveys current practices in the way to count patents to estimate innovation production functions. Section 3 explains the proposed methodology to detect a selection bias and section 4 presents the econometric framework. The data is described in section 5 and the results in section 6. Implications regarding data collection and the estimation methodology are presented in section 7, together with concluding remarks.

2. Measuring innovation 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 innovation production functions, a popular object of analysis in the innovation literature. Innovation – or patent – production functions relate firms' research inputs such as R&D expenditures to their patented output. They have attracted considerable attention in the literature, dating back to Scherer (1965), and have been used, amongst other things, to study the occurrence of innovation (e.g. Stoneman, 1979; Aghion *et al.*, 2005; Correa, 2012); to study the innovation process and the effectiveness of innovation policies (e.g. Jaffe, 1986; Cincera, 1997; Czarnitzki *et al.*, 2007); or as an intermediate step to study the determinants of firm productivity (e.g. Pakes and Griliches, 1984; Crépon *et al.*, 1998).

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 companies 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 greater detail in de Rassenfosse *et al.* (2011).

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 in Table 5 in Appendix). 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 (2010) 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, 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 (2009) 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 Azagra *et al.*, 2006). At the patent level, anecdotal evidence of a potential selection bias is provided by van Zeebroeck and van Pottelsberghe (2011). 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.* (2011) 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 innovation 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, Tobin, 1958 and Heckman, 1979). The selection effect of patent data is of a different nature, such that no standard method can be applied.

It is useful to describe the nature of the selection bias using a log-linear specification. Let us write the total unobserved patented output for firm i (y_i^*) as:

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

where ε_i is an error term and bold letters denote matrices and vectors. The single-office count implies that only a subset of the patents is observed. Let π_i be the firm-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:

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

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 a selection bias, the econometrician 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 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:

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

also depends on $\boldsymbol{x}$. This ratio 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 econometrician.

Hence, a simple way of testing the presence of selection bias involves evaluating whether $\tilde{\pi}_i$ is correlated with $\boldsymbol{x}$. If $\tilde{\pi}_i$ is significantly correlated with $\boldsymbol{x}$, it is *likely* that there is a selection bias. Conversely, if $\tilde{\pi}_i$ 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 $\tilde{\pi}_i$ is not correlated with $\boldsymbol{x}$. Second, if one component depends on $\boldsymbol{x}$ but not the other, there is a selection bias and $\tilde{\pi}_i$ 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 $\tilde{\pi}_i$ is ambiguous. Even though it is likely that $\tilde{\pi}_i$ 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:

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

In that particular scenario, $\tilde{\pi}_i$ 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 $\tilde{\pi}_i$ is unambiguously correlated with $\boldsymbol{x}$. There is no selection bias if:

$$\pi_i^p + (1-\pi_i^p)\pi_i^s = c \Leftrightarrow \pi_i^s = \frac{\pi_i^p - c}{\pi_i^p - 1} \qquad 2$$

but $\tilde{\pi}_i$ is correlated with $\boldsymbol{x}$. To sum up, if $\tilde{\pi}_i$ is not correlated with $\boldsymbol{x}$, there is no selection bias unless Equation 1 is satisfied. If $\tilde{\pi}_i$ is correlated with $\boldsymbol{x}$, there is a selection bias unless Equation 2 is satisfied. As a general rule, a significant effect of $\boldsymbol{x}$ on $\tilde{\pi}_i$ would suggest the presence of a selection bias. Inversely, the selection bias can be ruled out if $\boldsymbol{x}$ is not correlated with $\tilde{\pi}_i$. However, since the individual components are not observed, this approach is not infallible.

Note that another, 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. It is useful to estimate the determinants of $\tilde{\pi}_i$ directly for two reasons. First, this approach makes the pitfalls more apparent, in particular regarding the fact that the two components can depend on $\boldsymbol{x}$ even though no correlation is observed (Equation 1). Second, this approach is also easier to implement since one statistical test is needed for all the variables in $\boldsymbol{x}$ instead of one statistical test for each variable in $\boldsymbol{x}$. The results of the one-to-one test of equality of coefficients will nevertheless also be reported in the empirical exercise.

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 (say priority filings and total filings) to show the sensibility of the parameters, together with the estimation of the variable $\tilde{\pi}_i$.

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 test whether the variable $\tilde{\pi}_i$ detects the selection bias.

Patent production functions are estimated as Poisson such that:

$$\begin{aligned} E[p_{it}|\boldsymbol{x}_{it}, \eta_i] &= \exp\left(\boldsymbol{x}_{it}^{\prime}\boldsymbol{\beta} + \eta_i\right) \\ &= \mu_{it}\nu_i \text{ for } i = 1, \dots, N \text{ and } t = 1, \dots, T \end{aligned} \qquad 3$$

where p_{it} is the number of patents for firm 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 by Hausman *et al.* (1984). 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 innovation may call for further R&D. We adopt the estimator proposed by Blundell *et al.* (1999), which relaxes the strict exogeneity assumption (see also Blundell *et al.*, 2002). 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 Uchida and Cook (2007), Lach and Schankerman (2008) and Czarnitzki *et al.* (2009).

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 section 3). It is usually not observed by the econometrician. 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 section 3). Although this count mixes patents of varying nature, it takes into account a broader set of patents, thereby potentially limiting the selection bias.

The measure for the single-office bias (variable $\tilde{\pi}_i$ in section 3) is estimated as a Bernoulli pseudo-maximum likelihood following Papke and Wooldridge (1996) to account for the fact that the variable $\tilde{\pi}$ is bounded between 0 and 1:

$$E[\tilde{\pi}_{it}|\boldsymbol{x}_{it}] = h(\boldsymbol{x}'_{it}\boldsymbol{\gamma}) \qquad 4$$

where $h(z)$ is a link function satisfying $0 \leq h(z) \leq 1 \; \forall \; z \in \mathbb{R}$ such as the logistic link function.

5. The Data

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 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. Finally, the Patstat database by the EPO (April 2009 version) was used to collect data on patents. Because patent applications are published (hence observable) 18 months after the filing date, the data was collected up to 2007.

The novelty of the patent data is a key aspect of our paper. The construction of the patent indicator follows two logical steps: all the patent applications from inventions made in Belgium are identified and are then merged with the R&D survey data. 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, using the counting methodology described in depth in de Rassenfosse *et al*. (2011).[1] The name of the firms actually applying for the patent (the 'applicant' in jargon) was then manually harmonized and matched with data from the R&D questionnaire.

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 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

[1] 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 (OECD, 2009).

(a)). This is a very large share in comparison with the European average of less than 10 per cent over roughly the same period (de Rassenfosse *et al.*, 2011). Interestingly, the Belgian patent office receives less patent applications 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)).

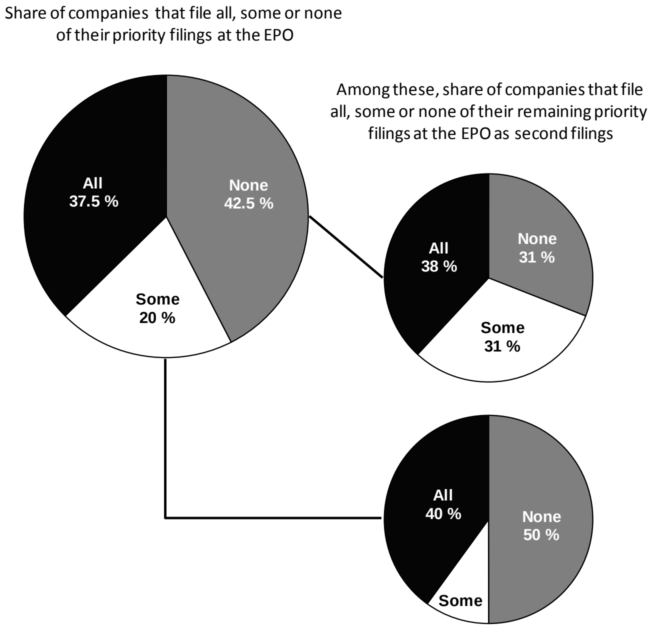
Table 1: Proportion of patents identified

Number of 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 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 certainly 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 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. Age (AGE) is defined as the number of years the firm exists. We also include a measure of the intensity of competition (COMP). 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 (‘000)	16	35,051	2,253,238	159,204
RD (‘000)	0	20,718	1,153,000	115,943
AGE	1	33	168	29
COMP (o)	1	2.59	3	0.60

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

Table 2 provides an overview of 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.

6. Results

Table 3 presents estimates of the patent production function for various dependent variables, as well as estimates of the single-office bias.

Table 3: Estimates of the patent production function and the selection bias

	(1)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
Equation:	3	3		3		4	
Dep. variable:	p^W	p^E		$p^E + s^E$		$\tilde{\pi}$	
ln(EMP)	0.477***	0.473***	N	0.473***	N	0.0240	N
	(0.094)	*(0.096)*		*(0.102)*		*(0.209)*	
ln(CAP/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)*	
COMP	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). Robust standard errors clustered at the firm level in parentheses. R^2 is computed as the square of the correlation coefficient between the dependent variable and its predicted value. 'Y' indicates that the coefficient is different from the corresponding 'true' coefficient at the 10% probability threshold. 'N' indicates that the coefficient is not significantly different from the corresponding 'true' coefficient.

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 test for difference in coefficients. A value 'Y' indicates that the coefficient is different from the corresponding 'true' coefficient 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 innovation 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 intensity of competition are significant determinants of the patent count, seemingly suggesting that young firms and firms evolving in a more competitive environment 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 reported in the last two columns. The estimation of the variable $\tilde{\pi}$ is presented in column (4a). The coefficients associated with the R&D intensity, age and competition variables are significantly different from zero, suggesting that the methodology successfully identifies the presence of a selection bias. The result of a one-to-one test of difference in coefficients between column (2b) and column (3b) is reported in column (4b) and confirms the presence of bias. Thus, our results suggest that the selection bias can be detected even though the econometrician observes patents at only one patent office.

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 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.

6.1 Additional considerations

The patent production functions were estimated with a negative binomial regression model with no change to the results. Similarly, the variable $\tilde{\pi}$ was estimated with a simple OLS regression 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 by Lambert (1992), 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 outcome. 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 4. 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*).[2] 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

[2] We refer the reader to van Zeebroeck (2011) for a recent review of patent value indicators.

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 (see e.g. Harhoff *et al.*, 2008). 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. 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 4 for the sake of completeness. The effect of a selection bias is still observed for both counts, as suggested by 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 4: Patent production functions estimated with different specifications

	(1)	(2a)	(2b)	(3a)	(3b)	(4)	(5a)	(5b)	(6a)	(6b)
Dep. Variable:	p^W	p^E		$p^E + s^E$		$\widetilde{p^W}$	$\widetilde{p^E}$		$\widetilde{p^E} + \widetilde{s^E}$	
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(CAP/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)*	
COMP	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(CAP/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: Estimates based on 429 observations. Industry dummies, time dummies and pre-sample fixed effects included. '$\tilde{\boldsymbol{x}}$' 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. A value 'Y' indicates that the coefficient is different from the corresponding 'true' coefficient at the 10% probability threshold. 'N' indicates that the coefficient is not significantly different from the corresponding 'true' coefficient.

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. Its contribution is twofold. First, it uses a novel dataset of the whole population of patents filed by Belgian firms to show that the single-office count of patents results in biased estimates of innovation production functions. Second, it proposes a simple way to test for the existence of a selection bias. The methodology involves estimating the determinants of the proportion $\tilde{\pi}$ of priority patent applications filed at the reference office among total patent applications at the reference office. The empirical application suggests that the test successfully spots coefficients that are affected by selection bias.

Two implications for research follow from the results presented in this paper. First, estimates based on a single-office count of patents should be treated with skepticism. For instance, the empirical application uses a variable that captures the competitive environment of the firm. While we are cautious not to interpret our results in any causal manner, we note that the type of patent indicator that is chosen affects the findings. In particular, the effect of competition on innovation is observed with international, high-value patents but not with total patents. Given that empirical studies have not generated clear conclusions about the relationship between innovation and competition (Gilbert, 2006), particular attention should be paid to the patent indicators that are used in future studies. This statement is true more generally for all variables that are likely to be affected by selection bias. Second, on a methodological level, the count of patents should be global to avoid a selection bias, and not limited to a single patent office. If the researcher is limited to a single office, then good practice would involve reporting estimates of patent production functions for both 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 $\tilde{\pi}$. If the focal variable does not affect the variable $\tilde{\pi}$, then our results suggest that the econometrician can be reasonably confident that the coefficient associated with the focal variable is not biased by the patent indicator used. Note that in countries such as the United Kingdom or the United States, the national patent office attracts more than 90 per cent of priority filings by national inventors (de Rassenfosse *et al.*, 2011). For these countries, restricting the count to patents filed at the home office is thus a sensible methodological choice. However, counting patents at the EPO would result in a severe restriction of patents. At the very least, our results suggest that researchers should think carefully about what patents they count.

This study also comes with a number of caveats and possibilities for further research, which we briefly discuss. First, it should be noted that the variable $\tilde{\pi}$ does not perfectly capture the selection bias. In particular, there are well-defined 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. Second, the test requires that the priority status of the patent document is available. 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. If the information on claimed priorities is not directly available in the patent database used by the researcher, it is still possible to identify priority filings by comparing the filing date with the priority date. If both dates are similar, the chance is high that the patent is a priority patent application. Thus, the priority status of the document can be collected at a low additional cost. Finally, even though a very high share of Belgian patents eventually ends up at the EPO, the effect of a selection bias was clearly visible in the data. In this respect, the Belgian case provides a very strong test of our claim: since a selection bias affects estimates for Belgian data, it is also very likely to affect estimates for data from other countries. Nevertheless, it would be useful to perform a similar exercise with data on firms from other countries to confirm the generality of our proposed methodology.

Acknowledgements

The authors are grateful to Dirk Czarnitzki and Anne Leahy 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).

References

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

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

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

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

Cincera, M., 1997. 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), 265–280.

Correa, J., 2012. Innovation and competition: An unstable relationship. *Journal of Applied Econometrics,* forthcoming.

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

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

Czarnitzki, D., Kraft, K., Thorwarth, S., 2009. The knowledge production of 'R' and 'D'. *Economics Letters,* 105(1), 141–143.

de Rassenfosse, G., Dernis, H., Guellec, D., Picci, L., van Pottelsberghe de la Potterie, B., 2011. The worldwide count of priority patents: A new indicator of inventive performance. Available at SSRN: http://ssrn.com/abstract=1883241

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

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

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

Hall, B., Ziedonis, R., 2001. The patent paradox revisited: An empirical study of patenting in the US semiconductor industry, 1979–1995. *The RAND Journal of Economics*, 32(1), 101–128.

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

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

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

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

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

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

Lach, S., Schankerman, M., 2008. Incentives and invention in universities. *The RAND Journal of Economics*, 39(2), 403–433.

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

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

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

Pakes, A., Griliches, Z., 1984. Patents and R&D at the firm level in French manufacturing: A first look. In: Z. Griliches ed., *Research and development, patents and productivity*, Chicago: The University Press of Chicago, 55–72.

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

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

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

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

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

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

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

van Zeebroeck, N., van Pottelsberghe de la Potterie, B., 2011. Filing strategies and patent value, *Economics of Innovation and New Technology*, forthcoming.

van Zeebroeck, N., 2011. The puzzle of patent value indicators. *Economics of Innovation and New Technology*, forthcoming.

Appendix

Table 5: 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
Cincera (1997)								
	Worldwide	1983–1991	Panel	181 firms	EPO	Undisclosed	A	N
Ernst (1998)								
	Europe and Japan	1990–1994	Cross-section	25 firms	EPO	Undisclosed	A	Y
Brouwer and Kleinknecht (1999)								
	The Netherlands	1992; 1998	Cross-section	148 firms	EPO	Undisclosed	A	N
Meliciani (2000)								
	OECD	1973–1999	Panel	180 country-sectors	USPTO	Undisclosed	G	N
Furman et al. (2002)								
	OECD	1973–1996	Panel	17 countries	USPTO	Undisclosed	G	N
Fritsch (2002)								
	Europe	1995–1998	Cross-section	707 firms	Undisclosed	Undisclosed	A	N
Bottazzi and Peri (2003)								
	Europe	1977–1995	Cross-section	86 regions	EPO	Undisclosed	G	N
Aghion et al. (2005)								
	U.K.	1973–1994	Panel	311 firms	USPTO	Undisclosed	G	Y
Salomon and Shaver (2005)								
	Spain	1990-1997	Panel	3,060 firms	Spanish PO, EPO	Undisclosed	A	N
Ulku (2007)								
	OECD	1981–1997	Panel	68 country-sectors	USPTO	Undisclosed	G	N
Carayol (2007)								
	France	1995–2000	Cross-section	941 scholars	French PO, EPO, PCT	PF & SF	A	Y
Mariani and Romanelli (2007)								
	Europe	1988–1998	Cross-section	793 inventors	EPO	Undisclosed	A	Y
Tappeiner et al. (2008)								
	Europe	1999	Cross-section	51 regions	EPO	Undisclosed	A	N
Czarnitzki et al. (2009)								
	Belgium	1993–2003	Panel	122 firms	EPO	Undisclosed	A	N
Akçomak and ter Weel (2009)								

Europe	1990; 2000	Cross-section	102 regions	EPO	Undisclosed	A	N
Hoekman et al. (2009)							
Europe	1988–2001	Cross-section	1,316 regions	EPO	Undisclosed	Undisclosed	N
Buesa et al. (2010)							
Europe	1995–2001	Panel	146 regions	EPO	Undisclosed	Undisclosed	N
Picci (2010)							
Worldwide	1990–2005	Panel	42 countries	NPOs	PF	A	N
Fornahl et al. (2011)							
Germany	1997–2004	Panel	129 firms	EPO, PCT	Undisclosed	Undisclosed	N
Rentocchini (2011)							
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'.

References in Table 5

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

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

Brouwer, E., Kleinknecht, A., 1999. Innovative output, and a firm's propensity to patent. An exploration of CIS micro data. *Research Policy*, 28, 615–624.

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

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

Ernst, H., 1998. Industrial research as a source of important patents. *Research Policy*, 27, 1–15.

Fornahl, D, Broekel, T, Boschma, R., 2011. 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), 395–418.

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

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

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

Mariani, M., Romanelli, M., 2007. “Stacking” and “picking” inventions: The patenting behavior of European inventors. *Research Policy,* 36, 1128–1142.

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

Picci, L., 2010. The internationalization of inventive activity: A gravity model using patent data. *Research Policy,* 39, 1070–1081.

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

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

Tappeiner, G., Hauser, C. and Walde, J., 2008. Regional knowledge spillovers: Fact or artifact? *Research Policy,* 37, 861–874

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

Non-listed references are in the main reference list.