Economics Letters 147 (2016) 72–77

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

On the distribution of patent citations and its fundamentals[☆]

Peter H. Egger[a,*], Florian Seliger[b], Martin Woerter[b]

[a] ETH Zurich, Switzerland
[b] KOF at ETH Zurich, Switzerland

HIGHLIGHTS

- We estimate power-law coefficients on 19.2mn patent-family citations in PATSTAT.
- We explain these coefficients by economic fundamentals in quantile regressions.
- Citations are more dispersed where R&D levels and outward FDI are higher.

ARTICLE INFO

Article history:
Received 29 September 2015
Received in revised form
8 August 2016
Accepted 10 August 2016
Available online 16 August 2016

JEL classification:
O34
C21

Keywords:
Patent citations
Power law
Pareto regression
Fixed-effects quantiles regression

ABSTRACT

This paper analyzes features of the distribution of 3.7 mn. patent family applications from PatStat in 34 countries, 17 industries, and 11 years. Power-law regressions suggest that higher levels of R&D intensity and outward foreign direct investment in a country and sector are associated with a reduced concentration of patent citations.

1. Introduction

Patents and their citations are particularly relevant for two reasons. First, patents reflect technological change as a key driver of economic growth (see Jaffe and Trajtenberg, 2002). Second, patent citations reflect the quality and value of innovations (see, e.g., Trajtenberg, 1990; Lanjouw and Schankerman, 2004; Hall et al., 2005). Understanding how they are linked to complementary factors of an economy is of key importance to policy makers.

Patent-activity distributions can be relatively well described by power laws. E.g., O'Neale and Hendy (2012) document this for patent filings across countries, Silverberg and Verspagen (2007) provide evidence for patent citations, and Pakes (1986), Scherer (1998), and Scherer and Harhoff (2000) illustrate it for patent values. The power-law form of patent citations and values suggests that only a small fraction of patented inventions yields high technical and economic value. We do not fully understand yet economic fundamentals behind citation power laws as such. The saliency of certain features of patent-citation distributions suggests that interesting latent processes might generate them (see Scherer, 1998), but it is not yet well understood how fundamentals such as R&D expenditures, market size, or openness contribute to the concentration or dispersion of patent citations across countries and sectors.

This paper characterizes the citations for all 3.7 mn. patent-family applications over the period 1995–2005 across 17 industries and 34 countries by power-law regressions and then determines to which extent the power-law characteristics relate to economic and institutional characteristics. This is achieved by combining estimates from power-law regressions with a subsequent fixed-effects quantiles regression analysis on the power-law coefficients. Key findings are that a greater R&D intensity and outward foreign direct investment in a country and sector suggest a greater dispersion of patent citations and, hence, patent values.

☆ Acknowledgements: We gratefully acknowledge numerous helpful comments by an anonymous reviewer. No funding was received for this research project.

* Correspondence to: Department of Management, Technology, and Economics at ETH Zurich, Leonhardstrasse 21, LEE G101, Switzerland. Fax: +41 0 44 6321218.
E-mail addresses: egger@kof.ethz.ch (P.H. Egger), seliger@kof.ethz.ch (F. Seliger), woerter@kof.ethz.ch (M. Woerter).

http://dx.doi.org/10.1016/j.econlet.2016.08.011

2. Estimating power-law regressions for patent citations

Let us denote the shape parameter of the Pareto distribution of patent citations in country i, sector s, and time t by k_{ist}. Moreover, let $1-\Pr_{pist}$ denote the probability for a patented invention of having at least as many citations as patented inventions in the pth percentile of the distribution, and let $rank_p$ denote the corresponding rank. Hence, $1-\Pr_{99ist}$ is associated with $rank_1$. Let us refer to the average number of citations of all patented inventions up until the pth percentile of the distribution in country i, sector s, and time t by $\bar{c}_{pist}$, which is computed cumulatively in the first percentile, in the first-plus-second percentile, etc. Then, k_{ist} can be estimated per $\{ist\}$ by a generalized-linear exponential-family model of the form

$$\bar{c}_{pist} = \exp(\underbrace{\alpha_{ist}}_{=-(1/k_{ist})} \ln rank_{pist} + \mu_{ist} + error_{pist}), \quad (1)$$

where μ_{ist} is a fixed country-sector-time effect, and $rank_{pist} \in \{1, \ldots, 99\}$. There are 99 degrees of freedom to estimate each one of the parameters k_{ist}. A parameter α_{ist} that is larger (smaller) in absolute value suggests that citations drop off more (less) severely as we move down the citation ranks. Hence, the concentration of citations rises with the absolute value of α_{ist}. We consider two implementations of (1), one where α_{ist} is estimated in a pooled fashion for all centiles p per $\{ist\}$, and one where we relax pooling to acknowledge a deviation from the single-parameter Pareto power law, estimating slope parameters per decile d among the centiles p, α^d_{ist}. The latter allows for a deviation from linearity in the otherwise linear index in (1).

We estimate (1) based on the universe of all 19.2 mn. patent family citations reported in the European Patent Office (EPO) PatStat data for all 34 OECD countries[1] and 17 sectors[2] for patent-family applications between 1995–2005 per country, sector, and year that occur within 5 years after publication of the cited patent family application. The sector assignment is based on the concordance tables from Lybbert and Zolas (2014). The number of forward citations per patent and industry is weighted based on the technological fields (International Patent Classification on subclass level) a patent family belongs to, yielding 6358 triplets $\{ist\}$. Industry and country data on possible determinants are taken from the OECD STAN database, the World Bank World Development Indicators, and the United Nations Conference on Trade and Development (UNCTADSTAT).

The upper four panels in Fig. 1 summarize the relationship postulated in (1) for two exemplary pairs of countries (Germany and the Netherlands) and sectors (Electrical/Optical Equipment and Wood Products) for $\hat{\alpha}_{ist}$ in the average year in 1995–2005. The lower two panels portray all estimates $\hat{\alpha}_{ist}$ as well as their demeaned value, $\breve{\hat{\alpha}}_{ist}$ (subtracting the average per is over time) by way of kernel-density plots. We observe some deviation from the (log-linear) power law which varies across sectors, supporting an estimation of decile-specific α^d_{ist} for better approximation. Moreover, the distribution of $\hat{\alpha}_{ist}$ looks twin-peaked and asymmetric due to time-invariant, country-sector-specific factors.

[1] The OECD comprises the following countries: Australia, Austria, Belgium, Canada, Chile, Canada, Chile, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, United Kingdom, United States.

[2] Agriculture, Hunting, Forestry and Fishing; Mining and Quarrying; Food, Beverages and Tobacco; Textiles and Textile Products, Leather and Footwear; Wood and Products of Wood and Cork; Pulp, Paper, Printing and Publishing; Coke, Refined Petroleum and Nuclear Fuel; Chemicals and Chemical Products; Rubber and Plastics; Other Non-Metallic Mineral; Basic Metals and Fabricated Metal; Machinery; Electrical and Optical Equipment; Transport Equipment; Other Manufacturing and Recycling; Electricity, Gas and Water Supply; Construction.

3. Explaining power-law regression coefficients in fixed-effects quantile regressions

We are particularly interested in determining the concentration of patent citations by relating $\hat{\alpha}_{ist}$ and $\hat{\alpha}^d_{ist}$ to economic fundamentals. We treat them as estimated, heteroskedastic dependent variables (see Saxonhouse, 1977), explaining them by a number of lagged time-variant fundamentals, X_{ist-1}, in linear (see Baltagi, 2008) and quantiles fixed-effects regressions (see Canay, 2011). These are three measures capturing the knowledge environment in a country (log patent stock per country and sector, the R&D intensity, the share of people with tertiary school enrolment per country and sector); three variables measuring sector size (log employment), tangible-investment intensity (investment over sales), and profitability (log operating profits from sales) per country and sector; and four measures of openness (export intensity and import intensity per country and sector, outward and inward FDI stocks relative to GDP per country). As we allow for some endogeneity of all determinants of $\hat{\alpha}_{ist}$, we generally include residuals from first-step regressions in X_{ist-1} which are obtained from regressions of fundamentals on lagged differences of those determinants (see the footnote to Table 1; notice that such a control function approach is consistent with the idea in Chernozhukov and Hansen, 2008). The proposed models are

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta,\tilde{\lambda}} \mathbb{E}[(\breve{\hat{\alpha}}_{ist} - \breve{X}_{ist-1}\beta - \breve{\lambda}_{is})^2],\\ \hat{\beta}_q &= \arg\min_{\beta_q} \mathbb{E}[\rho_q(\breve{\hat{\alpha}}_{ist} - \breve{X}_{ist-1}\beta_q - \breve{\hat{\lambda}}_{is})],\end{aligned} \quad (2)$$

where "$\breve{\ }$" denotes quantities that are adjusted for (power-law-regression-)imprecision, $\breve{\lambda}_{is}$ are is-specific fixed effects, q indicates a quantile, ρ_q is the *check* or *asymmetric absolute loss* function (see Wooldridge, 2010; Canay, 2011), and $\breve{\hat{\lambda}}_{is}$ are step-1 estimates of fixed effects which are not estimated in Step 2 of the quantiles model, where $\hat{\beta}_q$ is estimated. We conduct all estimations for both pooled $\breve{\hat{\alpha}}_{ist}$ and decile-specific $\breve{\hat{\alpha}}^d_{ist}$. A negative (positive) parameter on one of these variables suggests that a higher value of that variable is associated with a bigger (smaller) concentration of citations in a sector, country, and subsequent year.

Table 1 summarizes the results and is organized in four blocs. These blocs refer to all estimates of $\breve{\hat{\alpha}}_{ist}$ (upper left) and, using super-script d to denote deciles of $\breve{\hat{\alpha}}_{ist}$, the subsamples of first-decile (upper right), fifth-decile (lower left), and ninth-decile estimates (lower right) of $\breve{\hat{\alpha}}^d_{ist}$, respectively. Each of the four blocs in Table 1 contains six columns for the linear model and the 10th, 25th, 50th, 75th, and 90th percentiles for the quantiles models.[3]

The results suggest that the association between the considered fundamentals and the power-law of patent citations should rely on nonlinear estimation techniques such as quantiles models rather than linear models: some of the parameter estimates differ between these types of models even in a qualitative dimension, and the quantitative differences are often large relative to the quantiles point estimates. Specifically, bigger patent stocks tend to be associated with a more dispersed pattern of citations (and values) of patents, but this effect is estimated to be weaker by the quantiles model than the linear one (see, e.g., the top-left panel in Table 1). A higher R&D intensity of a country and sector also tends to reduce the concentration of patent citations (and values) on average and for two quantiles of the parameter distribution in the

[3] Table 2 provides the same set of results for the pooled parameter estimated with a Generalized Linear Model instead of OLS as a robustness test.