Journal of Informetrics 10 (2016) 1143–1152

Contents lists available at ScienceDirect

Journal of Informetrics

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

Towards field-adjusted production: Estimating research productivity from a zero-truncated distribution

Timo Koski[a], Erik Sandström[b], Ulf Sandström[c,*]

[a] KTH Royal Institute of Technology, Dept Mathematics, SE-100 44 Stockholm, Sweden
[b] Doktorsvägen 43, SE-123 52, Stockholm, Sweden
[c] KTH Royal Institute of Technology, Dept Indek, SE-100 44 Stockholm, Sweden

ARTICLE INFO

Article history:
Received 7 April 2016
Received in revised form
15 September 2016
Accepted 15 September 2016

Keywords:
Research productivity
Waring distribution
Field adjusted production
Size-dependent indicators

ABSTRACT

Measures of research productivity (e.g. peer reviewed papers per researcher) is a fundamental part of bibliometric studies, but is often restricted by the properties of the data available. This paper addresses that fundamental issue and presents a detailed method for estimation of productivity (peer reviewed papers per researcher) based on data available in bibliographic databases (e.g. Web of Science and Scopus). The method can, for example, be used to estimate average productivity in different fields, and such field reference values can be used to produce field adjusted production values. Being able to produce such field adjusted production values could dramatically increase the relevance of bibliometric rankings and other bibliometric performance indicators. The results indicate that the estimations are reasonably stable given a sufficiently large data set.

1. Introduction

Whilst methods for citation analysis have developed significantly during the latest twenty years, the same cannot be said regarding methods for publication productivity analysis. “Research productivity”, “Scientific productivity” and “Publication productivity” are frequently used keywords in about one and a half thousand Web of Science-articles over the years, but a closer look reveals little of methodological development with regard to the measurement of “productivity” and few attempts to explicitly contribute to such a development (for an exception, see papers by Abramo & D’Angelo, 2014, 2016).

This paper will address a fundamental issue in the empirical study of scientific productivity, i.e. the calculation of the average number of peer reviewed papers[1] published by researchers in a given time period. This task, which at first sight seems quite simple, is often restricted by the properties of the data available. Publication databases, such as Web of Science and Scopus, only contain information on actual (publishing) authors within a given time period, not the full population of publishing and non-publishing “authors”. Hence, a paper frequency distribution based on such data will be zero-truncated; the zero-class (number of non-publishing, or potential, “authors”) will be missing.

* Corresponding author.
E-mail addresses: tjtkoski@kth.se (T. Koski), erik.sandstrom@zoho.com (E. Sandström), ulf.sandstrom@indek.kth.se (U. Sandström).

[1] In this paper, the Web of Science is used as to provide details of “peer reviewed papers”. It should be noted that all publications databases have coverage issues. Therefore, estimated in this paper are papers recorded in Web of Science (Arts & Humanities Citation Index, Social Science Citation Index, and Science Citation Index) as document types “Article”, “Letter”, or “Review” in the selected time period. For a discussion regarding coverage issues in Web of Science, see Mongeon and Paul-Hus (2016).

http://dx.doi.org/10.1016/j.joi.2016.09.002
1751-1577/

For example, if the productivity of two countries, e.g. China and Canada, are to be investigated, one might categorize the publishing authors by name (unique authors) and divide the total number of papers by the number of actual authors. However, this calculation will not produce a trustworthy measure of productivity since the proportion of non-publishing (potential) authors might differ between the two countries. A non-biased measure requires knowledge of the full population, including the number of non-publishing authors (i.e. the zero-class of the paper frequency distribution).

2. Actual authors and potential authors

"Potential authors" is a concept used by the Budapest group to ex ante denote the total population of researchers that could publish papers, or to ex post denote researchers that could have, but have not, been publishing within a given time period (Braun, Glänzel, & Schubert, 2001; Telcs & Schubert, 1986). The (ex post) potential authors include active researchers that have been publishing before or after the given time period but for different reasons have not been publishing during the given time period.

Productivity comparisons between different categories, e.g. field, gender or country requires knowledge about the total population of authors, potential as well as actual authors. In a field with low paper productivity, e.g. the social sciences, the proportion of zero-class (the potential authors) will be high in relation to the total population (actual and potential authors). In contrast, in fields with high productivity, such as the medical and natural sciences, the proportion of potential authors will be low. Comparisons based solely on actual authors will thus be misrepresentative.

Actual authors can be extracted from publication databases such as Web of Science and Scopus. Potential authors will, however, not be included in these. An estimate of the number of potential authors would either require detailed data concerning the entire researcher population or the use of statistical estimates, such as the one presented in this paper. A successful method for estimating the number of potential authors in a productivity distribution (i.e. the zero-class of the paper frequency distribution) would enable the creation of more advanced productivity measures.

The objective of this paper is to contribute to the discussion on productivity from a scientometric perspective. If it is possible to estimate the zero-class of a truncated paper frequency distribution, then it would, in principle, be possible to create advanced (e.g. field normalized) productivity indicators (Sandström & Sandström, 2008).

Hitherto, the most interesting discussion on publication productivity has been given within the framework of frequency distributions (Braun, Glänzel, & Schubert, 1990), which is a core element in bibliometric theory. The Waring distribution is a statistical distribution used for describing publication productivity processes. The distribution was originally introduced by Simon (1955) as a generalization of the Yule distribution and further analysed by Irwin in (1963), who gave the distribution its current name.

In this paper, we aim to further explore the Waring distribution as a method to estimate the zero-class of a truncated paper frequency distribution. We note that there are certainly other possible methods and that only the Waring method has been reviewed as part of this paper. We welcome further research focused on comparison with the results of alternative methods.

3. The waring distribution

The Waring distribution can be derived and justified by many means. We shall here present an argument related to a simple probabilistic picture of publishing. Let us suppose that an author during a certain period keeps on submitting new papers until rejected for the first time. Let us suppose that the probability of rejection of any paper is equal to p, $0 < p < 1$. Let us also suppose that the rejection or acceptance of a paper does not depend on the rejection or acceptance of any other paper.

Then it follows that the probability of publishing exactly k papers is the geometric distribution

$$Pr(k\, published\, papers|p) = (1-p)kp, (k) = 0, 1, \ldots \tag{1}$$

Let us next suppose that p itself is a random variable. This could be an expression of uncertainty about the actual value of p in the Bayesian sense, where the uncertainty is expressed as a prior probability density $f(p)$. Or, we could think that the author has been allotted p that has been drawn from probability density $f(p)$. Now we can compute a new probability of k published papers by averaging, i.e.,

$$Pr\,(k\ \ published\ \ papers) = \int_0^1 Pr\,(k\ \ published\ papers|\ \ p) f\,(p)\,dp = \int_0^1 (1-p)^k pf\,(p)\,dp. \tag{2}$$

Let us take $f(p)$ as the Beta density with parameters $\rho > 0$ and $\alpha > 0$, i.e,

$$fp\frac{\Gamma\alpha+\rho}{\Gamma\rho\Gamma\alpha}p^{\rho-1}1-p^{\alpha-1} \tag{3}$$