Journal of Informetrics 8 (2014) 175–180

Contents lists available at ScienceDirect

Journal of Informetrics

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

How to improve the prediction based on citation impact percentiles for years shortly after the publication date?

Lutz Bornmann[a,*], Loet Leydesdorff[b], Jian Wang[c,d]

[a] Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany
[b] Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam, The Netherlands
[c] Institute for Research Information and Quality Assurance (iFQ), Schützenstraße 6a, 10117 Berlin, Germany
[d] Center for R&D Monitoring (ECOOM) and Department of Managerial Economics, Strategy and Innovation, Katholieke Universiteit Leuven, Waaistraat 6, 3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 13 August 2013
Received in revised form
18 November 2013
Accepted 19 November 2013
Available online 12 December 2013

Keywords:
Citation impact normalization
Percentile
Short citation window

ABSTRACT

The findings of Bornmann, Leydesdorff, and Wang (2013b) revealed that the consideration of journal impact improves the prediction of long-term citation impact. This paper further explores the possibility of improving citation impact measurements on the base of a short citation window by the consideration of journal impact and other variables, such as the number of authors, the number of cited references, and the number of pages. The dataset contains 475,391 journal papers published in 1980 and indexed in Web of Science (WoS, Thomson Reuters), and all annual citation counts (from 1980 to 2010) for these papers. As an indicator of citation impact, we used percentiles of citations calculated using the approach of Hazen (1914). Our results show that citation impact measurement can really be improved: If factors generally influencing citation impact are considered in the statistical analysis, the explained variance in the long-term citation impact can be much increased. However, this increase is only visible when using the years shortly after publication but not when using later years.

1. Introduction

Percentiles used in bibliometrics provide information about the citation impact of a focal paper compared with other comparable papers in a reference set; for example, all papers in the same research field and publication year. A percentile is the value below which a certain proportion of observations (here: papers) fall: the larger a paper's percentile, the higher citation impact it has – compared with papers in the same field and publication year. Since the percentile approach has been acknowledged in bibliometrics as a valuable alternative to the normalization of citation counts based on mean citation rates, some different percentile-based approaches have been developed (see an overview in Bornmann, Leydesdorff, and Mutz, 2013a). More recently, two of these approaches ($PP_{\text{top 10\%}}$ and the Excellence Rate, respectively) have been prominently used in the Leiden Ranking (Waltman et al., 2012) and the SCImago institutions ranking (Bornmann, de Moya Anegón, & Leydesdorff, 2012) as evaluation tools.

Using a publication set including all papers published in 1980 (nearly 500,000 papers), Bornmann et al. (2013b) investigated how the different percentile-based approaches are able to predict the long-term citation impact (in year 31, t_{31}) of

[*] Corresponding author. Tel.: +49 89 2108 1265.
E-mail addresses: bornmann@gv.mpg.de (L. Bornmann), loet@leydesdorff.net (L. Leydesdorff), Jian.Wang@kuleuven.be (J. Wang).

1751-1577/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2013.11.005

papers from citation impacts in previous years (years 1, t_1, to 30, t_{30}). In comparison to the other approaches, the SCImago approach demonstrated unexpected capabilities in accurately predicting the long-term citation impact on the basis of the citation impact in the first few years after publication. The consideration of the journal impact in this approach in solving the problem of tied citations seems to have generated this positive effect.

For the problem of ranks tying at the top 10% threshold level, SCImago introduces a secondary sort key in addition to citation counts: When citation counts are equal, the publication in a journal with the higher SCImago Journal Rank (SJR2) (Guerrero-Bote & de Moya-Anegon, 2012) obtains the higher percentile rank. Adding this journal metric takes into account not only the observed citations of the focal paper but also the prestige of the journal that a paper is published in.

Given the enduring tension between the practical needs for timely assessment of research outputs and the long time period it takes for research to reveal its full impact (Bornmann, 2013; Wang, 2013), we examine in the present study the added-value of considering the journal impact in predicting the long-term citation impact from citation impacts in previous years. For the statistical analyses, we use the same data set as Wang (2013) and Bornmann et al. (2013b). However, we consider not only the journal impact but also other factors (e.g., the number of authors) for better predicting long-term citation impact. Bibliometric studies have already pointed out several other factors – in addition to journal impact – with an (significant) effect on citation impacts (see an overview in Bornmann & Daniel, 2008). Thus, we examined, whether the prediction of long-term citation impact (based on years shortly after the publication date) can be improved by considering further factors. Since the approach of Hazen (1914) to calculate percentiles is widely used in statistical packages, we use it in this study. The results are also generalizable to other approaches (as we will exemplarily show).

2. Methods

2.1. The percentile approach of Hazen (1914)

Two steps are needed in order to calculate percentiles for a reference set based on the percentile-based approach of Hazen (1914):

First, all papers in the set are ranked in ascending order of their numbers of citations. Papers with equal citation counts are set equal by assigning the average rank. This is the default ranking method in the statistical package Stata (StataCorp., 2013). This method ensures that the sum of the ranks is fixed at $n^*(n+1)/2$, where n is the number of papers in the reference set.

Second, each paper is assigned a percentile based on its rank (percentile rank). Percentiles can be calculated in different ways (Bornmann, Leydesdorff, et al., 2013; Cox, 2005; Hyndman & Fan, 1996). In this study, we used the formula $(100^*(i-0.5)/n)$, derived by Hazen (1914). This formula is used very frequently nowadays for the calculation of percentiles and is wired into the official Stata command “quantile” (StataCorp., 2013). It ensures that the mean percentile is 50 and symmetrically handles the tails of the distributions.

2.2. Dataset used

In this study, we define a reference set for a paper under study as a set of papers with the same Web of Science (WoS, Thomson Reuters) subject category and document type. The reference sets were used to calculate the percentile-based approach developed by Hazen (1914). Each paper in WoS is classified into one unique document type but possibly into multiple subject categories. Therefore, for papers with multiple subject categories, the average percentile rank is used.

Furthermore, the citation percentiles could be too coarse if the size of the reference set is too small. Therefore, only reference sets with at least one hundred papers are included.[1] For example, if a paper belongs to two different reference sets: A and B, and A has more than 100 papers while B has less than 100 papers, then the percentiles based on B are discarded. If neither A nor B has more than 99 papers, then both results based on A and B are discarded, and this paper is excluded from the further analysis.

The dataset contains all journal papers published in 1980 and indexed in WoS, that is, 746,460 papers in total. Two restrictions are then imposed on the sample: (1) three document types – articles, reviews, and notes[2] – were kept while other document types were excluded, and (2) only papers having at least one reference set with hundred or more papers were included. As a result, we have 475,391 papers for analysis, and the annual citation counts (from 1980 to 2010) for these papers were retrieved from WoS.

2.3. Statistical procedures and variables (covariates)

We fitted 30 sets of regression models with the percentile of citations in year 31, t_{31}, as the dependent variable and the (short-) time-window citation percentiles (from year 1 to year 30) as one independent variable, correspondingly. For

[1] We decided to use 100 papers as a limit to produce reliable data. There is a high probability that the use of a limit of 50 or 200 would come to similar results as ours.

[2] Notes were removed from the database as a document type in 1997, but they were citable items in 1980.