Journal of Informetrics 11 (2017) 164–175

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

# Skewness of citation impact data and covariates of citation distributions: A large-scale empirical analysis based on Web of Science data

Lutz Bornmann[a,*], Loet Leydesdorff[b]

[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, PO Box 15793, 1001 NG Amsterdam, The Netherlands

ARTICLE INFO

*Article history:*
Received 2 November 2016
Received in revised form
28 November 2016
Accepted 1 December 2016
Available online 23 December 2016

*Keywords:*
Citation impact
Factors influencing citations
Percentile shares
Impact factors
Normalization

ABSTRACT

Using percentile shares, one can visualize and analyze the skewness in bibliometric data across disciplines and over time. The resulting figures can be intuitively interpreted and are more suitable for detailed analysis of the effects of independent and control variables on distributions than regression analysis. We show this by using percentile shares to analyze so-called "factors influencing citation impact" (FICs; e.g., the impact factor of the publishing journal) across years and disciplines. All articles (n = 2,961,789) covered by WoS in 1990 (n = 637,301), 2000 (n = 919,485), and 2010 (n = 1,405,003) are used. In 2010, nearly half of the citation impact is accounted for by the 10% most-frequently cited papers; the skewness is largest in the humanities (68.5% in the top-10% layer) and lowest in agricultural sciences (40.6%). The comparison of the effects of the different FICs (the number of cited references, number of authors, number of pages, and JIF) on citation impact shows that the JIF has indeed the strongest correlations with the citation scores. However, the correlation between FICs and citation impact is lower, if citations are normalized instead of using raw citation counts.



## 1. Introduction

van Raan (2014) listed the skewness of citation data as one of several methodological problems in citation analysis. The skewness of bibliometric data has been a topic in this field since its beginnings in the 1920s. The issue is associated with the "laws" of Alfred Lotka, Samuel Bradford, and George Zipf: "the concentration of items on a relatively small stratum of sources" (de Bellis, 2009, p. xxiv). Since then a large number of papers have appeared demonstrating the skewness of citation data. Seglen (1992), for example, argued that "50% of the citations and the most cited half of the articles account for nearly 90% of the citations" (p. 628). He concluded that citation distributions follow approximately an inverse power-law distribution (the number of citations larger than $x$ is proportional to $-\log(x)$) (Katz, 2000). Albarrán and Ruiz-Castillo (2011) showed empirically that the "existence of a power law cannot be rejected in ALL SCIENCES taken together as well as in 17 of 22 fields whose articles represent 74.7% of the total" (p. 48). Using a replication and scale invariant technique–the Characteristic Scores and Scales (CSS) (Glänzel, 2011)–the results of Albarrán, Crespo, Ortuño, and Ruiz-Castillo (2011) show

* Corresponding author.
*E-mail addresses:* bornmann@gv.mpg.de (L. Bornmann), loet@leydesdorff.net (L. Leydesdorff).

http://dx.doi.org/10.1016/j.joi.2016.12.001
1751-1577/

that citation distributions are highly skewed: "the mean is 20 points above the median, while 9–10% of all articles in the upper tail account for about 44% of all citations" (p. 385).

In this study, we analyze the skewness of citation impact data in six major disciplines (natural sciences, engineering and technology, medical and health sciences, agricultural sciences, social sciences, and humanities) based on all articles in Web of Science (WoS) published in 1990, 2000, and 2010. First, we use percentile shares–a recently introduced visualization and analysis technique–to quantify the proportions of total citation impact that go into different groups (e.g., the 10% most-frequently-cited papers). Percentile shares can be intuitively and appealingly interpreted and are especially suitable "for the detailed analysis of distributional changes" (Jann, 2016, p. 3).

In a next step, we use percentile shares to analyze covariates of the citation distributions. Journal Impact Factors (JIF) are often used as proxies for the citation impact of papers published in the respective journals. Are JIFs indeed a factor influencing citation impact? We show the advantages of using percentile shares in the case of a number of co-variates of citation scores indicated in the literature as "factors influencing citation impact" (FICs, e.g., the number of authors, see Bornmann & Daniel, 2008). We compare the association of JIFs as co-variates with citation scores at the level of individual papers with other FICs mentioned in the literature, such as the number of co-authors, the numbers of pages, and the number of cited references. How much does each covariate enhance the likelihood of being cited in the top-10% layer of citation scores? Finally, we address the question of whether normalization of the citation scores increases or reduces this chance.

Let us note that the analysis is correlational. One cannot conclude that FICs influence citation impact since a third factor such as the quality of the paper may be involved (Bornmann & Leydesdorff, 2015). If a paper is of high quality, e.g., it may attract large numbers of citations and be published in a journal with a high impact. The relation between the JIF and the citation impact is then spurious. From this perspective, the terminology "factors influencing citation" is unfortunate.

## 2. Factors influencing citation counts: a short review of the literature

In the following, only a short review focusing on recent and overview studies is provided. Recently, Tahamtan, Safipour Afshar, and Ahamdzadeh (2016) published a comprehensive review of FICs. The results of many studies question the usefulness of citation counts for measuring research impact or of using citation counts as a proxy for research quality. Overviews of FICs listed in tabular forms can be found in Onodera and Yoshikane (2014) and Didegah and Thelwall (2013). Both tables emphasize the Journal Impact Factor (JIF, Garfield, 2006) as an important factor in receiving citations.

In a recent study, Didegah and Thelwall (2014) investigated a range of factors which may be associated with the citation counts of social-science papers. The authors conclude that "journal and [cited] reference characteristics, and particularly journal and reference impact, are the main extrinsic properties of articles that associate with their future citation impact in the social sciences. Journal and reference internationality can also help with the prediction of future citation impact. Research collaboration, and particularly individual and institutional collaboration, can help to predict citation counts for articles but international collaboration alone is not important, unless it is with a high impact nation. Paper length, abstract length and abstract readability are also significant determinants of citation counts, but not all make a substantial difference. In the world top institutions, articles with more readable abstracts (i.e., easier to read) were less cited but in the social sciences more readable abstracts are more cited" (Didegah & Thelwall, 2014, pp. 173–174).

Robson and Mousquès (2016) focused on papers in environmental modelling published since 2005 and studied a range of FICs which were quantified or classified. The results of the study reveal that "papers with no differential equations received more citations. The topic of the paper, number of authors and publication venue were also significant. Ten other factors, some of which have been found significant in other studies, were also considered, but most added little to the predictive power of the models. Collectively, all factors predicted 16–29% of the variation in citation counts, with the remaining variance (the majority) presumably attributable to important subjective factors such as paper quality, clarity and timeliness" (Robson & Mousquès, 2016, p. 94). Onodera and Yoshikane (2014) studied samples of papers in six selected fields (condensed matter physics, inorganic and nuclear chemistry, electric and electronic engineering, biochemistry and molecular biology, physiology, and gastroenterology) and tried to reveal some general patterns. "Some generality across the fields was found with regard to the selected predicting factors and the degree of significance of these predictors. The Price index [the proportion of references within 3 and 5 years (de Solla Price, 1963)] was the strongest predictor of citations, and number of references was the next. The effects of number of authors and authors' achievement measures were rather weak" (Onodera & Yoshikane, 2014, p. 739).

For the field of "information science & library science", Yu, Li, and Wang (2014) used stepwise regression to produce a model predicting citation counts. The authors included a range of possible FICs and claimed that they can predict–with relative good accuracy–citation impact using a citation window of five years.

## 3. Methods

### 3.1. Percentile shares and Gini coefficients

We build on our argument for using percentiles in bibliometric evaluations (Bornmann, Leydesdorff, & Mutz, 2013; Leydesdorff, Bornmann, Mutz, & Opthof, 2011). Here, percentiles are used to field- (and time-) normalize citation counts; a percentile is the percentage of papers with lower or higher citation impact – depending on the percentile formula. Hicks,