Journal of Informetrics 11 (2017) 128–151

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

Three practical field normalised alternative indicator formulae for research evaluation

Mike Thelwall

Statistical Cybermetrics Research Group, University of Wolverhampton, UK

ARTICLE INFO

Article history:
Received 27 July 2016
Accepted 3 December 2016
Available online 4 January 2017

ABSTRACT

Although altmetrics and other web-based alternative indicators are now commonplace in publishers' websites, they can be difficult for research evaluators to use because of the time or expense of the data, the need to benchmark in order to assess their values, the high proportion of zeros in some alternative indicators, and the time taken to calculate multiple complex indicators. These problems are addressed here by (a) a field normalisation formula, the Mean Normalised Log-transformed Citation Score (MNLCS) that allows simple confidence limits to be calculated and is similar to a proposal of Lundberg, (b) field normalisation formulae for the proportion of cited articles in a set, the Equalised Mean-based Normalised Proportion Cited (EMNPC) and the Mean-based Normalised Proportion Cited (MNPC), to deal with mostly uncited data sets, (c) a sampling strategy to minimise data collection costs, and (d) free unified software to gather the raw data, implement the sampling strategy, and calculate the indicator formulae and confidence limits. The approach is demonstrated (but not fully tested) by comparing the Scopus citations, Mendeley readers and Wikipedia mentions of research funded by Wellcome, NIH, and MRC in three large fields for 2013–2016. Within the results, statistically significant differences in both citation counts and Mendeley reader counts were found even for sets of articles that were less than six months old. Mendeley reader counts were more precise than Scopus citations for the most recent articles and all three funders could be demonstrated to have an impact in Wikipedia that was significantly above the world average.

1. Introduction

Citation analysis is now a standard part of the research evaluation toolkit. Citation-based indicators are relatively straightforward to calculate and are inexpensive compared to peer review. Cost is a key issue for evaluations designed to inform policy decisions because these tend to cover large numbers of publications but may have a restricted budget. For example, reports on government research policy or national research performance can include citation indicators (e.g., Elsevier, 2013; Science-Metrix, 2015), as can programme evaluations by research funders (Dinsmore, Allen, & Dolby, 2014). Although funding programme evaluations can be conducted by aggregating end-of-project reviewer scores (Hamilton, 2011), this does not allow benchmarking against research funded by other sources in the way that citation counts do. The increasing need for such evaluations is driven by a recognition that public research funding must be accountable (Jaffe, 2002) and for charitable organisations to monitor their effectiveness (Hwang & Powell, 2009).

E-mail address: m.thelwall@wlv.ac.uk

http://dx.doi.org/10.1016/j.joi.2016.12.002
1751-1577/

The use of citation-based indicators has many limitations. Some well discussed issues, such as the existence of negative citations, systematic failures to cite important influences and field differences (MacRoberts & MacRoberts, 1996; Seglen, 1998; MacRoberts & MacRoberts, 2010), can be expected to average out when using appropriate indicators and comparing large enough collections of articles (van Raan, 1998). Other problems are more difficult to deal with, such as language biases within the citation databases used for the raw data (Archambault, Vignola-Gagne, Côté, Larivière, & Gingras, 2006; Li, Qiao, Li, & Jin, 2014). More fundamentally, the ultimate purpose of research, at least from the perspective of many funders, is not to understand the world but to help shape it (Gibbons et al., 1994). An important limitation of citations is therefore that they do not directly measure the commercial, cultural, social or health impacts of research. This has led to the creation and testing of many alternative types of indicators, such as patent citation counts (Jaffe, Trajtenberg, & Henderson, 1993; Narin, 1994), webometrics/web metrics (Thelwall & Kousha, 2015a) and altmetrics/social media metrics (Priem, Taraborelli, Groth, & Neylon, 2010; Thelwall & Kousha, 2015b). These indicators can exploit information created by non-scholars, such as industrial inventors' patents, and may therefore reflect non-academic types of impacts, such as commercial value.

A practical problem with many alternative indicators (i.e., those not based on citation counts) is that there is no simple cheap source for them. It can therefore be time-consuming or expensive for organisations to obtain, say, a complete list of the patent citation counts for all of their articles. This problem is exacerbated if an organisation needs to collect the same indicators for other articles so that they can benchmark their performance against the world average or against other similar organisations. Even if the cost is the same as for citation counts, alternative indicators need to be calculated in addition to, rather than instead of, citation counts (e.g., Dinsmore et al., 2014; Thelwall, Kousha, Dinsmore, & Dolby, 2016) and so their costs can outweigh their value. This can make it impractical to calculate a range of alternative indicators to reflect different types of impacts, despite this seeming to be a theoretically desirable strategy. The problem is exacerbated by alternative indicator data usually being much sparser than citation counts (Kousha & Thelwall, 2008; Thelwall, Haustein, Larivière, & Sugimoto, 2013; Thelwall & Kousha, 2008). For example, in almost all Scopus categories, over 90% of articles have no patent citations (Kousha & Thelwall, in press-b). These low values involved make it more important to use statistical methods to detect whether differences between groups of articles are significant. Finally, the highly skewed nature of citation counts and most alternative indicator data causes problems with simple methods of averaging to create indicators, such as the arithmetic mean, and complicate the task of identifying the statistical significance of differences between groups of articles.

This article addresses the above problems and introduces a relatively simple and practical strategy to calculate a set of alternative indicators for a collection of articles in an informative way. The first component of the strategy is the introduction of a new field normalisation formula, the Mean Normalised Log-transformed Citation Score (MNLCS) for benchmarking against the world average. As argued below, this is simpler and more coherent than a previous similar field normalisation approach to deal with skewed indicator data. The second component is the introduction of a second new field normalisation formula, the Equalised Mean-based Normalised Proportion Cited (EMNPC), that targets sparse data, and an alternative, the Mean-based Normalised Proportion Cited (MNPC). The third component is a simple sampling strategy to reduce the amount of data needed for effective field normalisation. The final component is a single, integrated software environment for collecting and analysing the data so that evaluators can create their own alternative indicator reports for a range of indicators with relative ease. The methods are illustrated with a comparative evaluation of the impact of the research of three large medical funders using three types of data: Scopus citation counts; Mendeley reader counts; and Wikipedia citations.

2. Mean normalised log-transformed citation score

The citation count of an article must be compared to the citation counts of other articles in order to be assessed. The same is true for collections of articles and a simple solution would be to calculate the average number of citations per article for two or more collections so that the values can be compared. This is a flawed approach for the following reasons that have led to the creation of improved methods.

Older articles tend to be more cited than younger articles (Wallace, Larivière, & Gingras, 2009) and so it is not fair to compare averages between sets of articles of different ages. Similarly, different fields attract citations at different rates and so comparing averages between sets of articles from different mixes of fields would also be unfair (Schubert & Braun, 1986). One solution would be to segment each collection of articles into separate sets, one for each field and year, and only compare corresponding sets between collections. Although this may give useful fine grained information, it is often impractical because each set may contain too few articles to reveal informative or statistically significant differences.

The standard solution for field differences in citation counts is to use a field (and year) normalised indicator. The Mean Normalised Citation Score (MNCS), for example, adjusts each citation count by dividing it by the average for the world in its field and year. After this, the arithmetic mean of the normalised citation counts is the MNCS value (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011a, 2011b). This can reasonably be compared between different collections of articles or against the world average, which is always exactly 1, as long as all articles are classified in a single field. If some articles are in multiple fields then weighting articles and citations with the reciprocal of the number of fields containing the article ensures that the world average is 1 (Waltman et al., 2011a).

A limitation of the MNCS is that the arithmetic mean is inappropriate for citation counts and most alternative indicators because they are highly skewed (de Solla Price, 1976; Thelwall & Wilson, 2016). In practice, this means that confidence limits must be calculated with bootstrapping and large sample sizes are needed for accurate results. An alternative approach that solves both of these problems is to switch from the arithmetic mean to the geometric mean because this is suitable for skewed