Journal of Informetrics 11 (2017) 1069–1079

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

Confidence intervals for normalised citation counts: Can they delimit underlying research capability?

Mike Thelwall

Statistical Cybermetrics Research Group, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK

ARTICLE INFO

Article history:
Received 26 April 2017
Received in revised form 23 June 2017
Accepted 11 September 2017

Keywords:
Citation analysis
MNLCS
Confidence intervals
Field normalised indicators

ABSTRACT

Normalised citation counts are routinely used to assess the average impact of research groups or nations. There is controversy over whether confidence intervals for them are theoretically valid or practically useful. In response, this article introduces the concept of a group's underlying research capability to produce impactful research. It then investigates whether confidence intervals could delimit the underlying capability of a group in practice. From 123120 confidence interval comparisons for the average citation impact of the national outputs of ten countries within 36 individual large monodisciplinary journals, moderately fewer than 95% of subsequent indicator values fall within 95% confidence intervals from prior years, with the percentage declining over time. This is consistent with confidence intervals effectively delimiting the research capability of a group, although it does not prove that this is the cause of the results. The results are unaffected by whether internationally collaborative articles are included.

1. Introduction

The average citation impact of the publications of a group of researchers can be calculated with various field normalised indicators (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011a; Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011b). The results may inform funding decisions for departments or higher level policy-making (e.g., Science-Metrix, 2015). The underlying assumption is that the number of citations to a scholarly paper tends to reflect its contribution to science (Merton, 1973). Although this is frequently untrue (MacRoberts & MacRoberts, 1996) or an oversimplification for individual citations and citation counts (Borgman & Furner, 2002), it is reasonable when applied to sufficiently large groups of papers in many disciplines (van Raan, 1998).

Citation counts need to be normalised for the field and year of publication to avoid misleading results. Nevertheless, even normalised citation indicators do not accurately compute the impact of a group of publications because field and year normalisation is unavoidably imperfect. These imperfections include treating articles that are published on different days during the year as being published at the same time (giving earlier articles an advantage) and combining interdisciplinary research and related specialisms into "fields" for normalisation purposes or using another heuristic to define fields (giving articles from higher citation specialisms or multidisciplinary combinations an advantage). Moreover, citations do not reflect all types of contributions to scholarship or non-scholarly impacts.

This article investigates whether it is reasonable to use confidence intervals to delimit the likely underlying research capability of a group. It does this by first calculating 95% confidence intervals for indicators annually. It then checks how

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

http://dx.doi.org/10.1016/j.joi.2017.09.002
1751-1577/

often indicator values for a group fall within 95% confidence intervals from previous years. Finally, it heuristically assesses whether this frequency is broadly in line with the hypothesis that the confidence intervals delimit underlying research capability. This assessment necessarily relies upon subjective assumptions about the level of stability to be expected in different contexts.

2. Confidence intervals and random factors for citation-based indicators

A controversy exists over whether confidence intervals should be reported alongside field normalised indicators (Williams & Bornmann, 2016). For example, stability intervals have been suggested as a partial substitute, using bootstrapping to assess the effect of changes in underlying publication dataset on indicator values (Waltman et al., 2012). Citation-based evaluations usually incorporate all publications produced by the assessed group (e.g., department, institution, nation) within the given period or all those that meet a criterion (e.g., indexed in Scopus). They are therefore not a random sample of a larger population but a complete set or specified subset (Schneider, 2016). Thus, from a basic statistical perspective it does not make sense to calculate confidence intervals for them because confidence intervals are designed to help infer population parameters from a random sample. In other words, citation indicators calculated from complete publication sets are exact rather than estimates.

Despite the above logic, there are reasons why confidence intervals can be desirable. There are two main types of random factors that may influence the number of citations accrued by the publications of a research group: citation-related and article-related. Many citation-related factors are out of the control of the authors of a paper (Abramo, Costa, & D'Angelo, 2015; Waltman, 2016). Each individual citation to a paper in the Web of Science (WoS) or Scopus is a consequence of multiple factors, including someone's decision to write a new paper, the new authors finding the first paper and judging it worth citing, the new authors submitting their paper to a WoS/Scopus journal, and the journal's editors and referees agreeing to accept the paper. Thus, if a research group's outputs receive a total of 1000 citations then this outcome is the result of thousands of individual decisions. From this perspective, confidence intervals for normalised citation counts would be reasonable to assess the range of values that the research group *could* plausibly have achieved after publishing their papers. Here, the confidence interval is for factors external to the research group that affect their citations. The set of citation counts is treated as an apparent population (Berk, Western, & Weiss, 1995): a sample of the outcomes that might reasonably have been expected from the set of publications. Although randomness at the citation level is a relevant phenomenon, it is not the main focus of the current paper. Instead, the type of randomness primarily modelled in the current paper is at the publication level.

At the publication level, each output from a group is a consequence of the creative powers of its authors as well as their technical prowess and the availability of time and other resources to conduct the research. Thus, each group publication is partly due to creativity-related factors that are not fully within the authors' control (Lee, Walsh, & Wang, 2015; Simonton, 2004). Whilst a highly creative author may *tend* to produce high impact work, she is not able to *guarantee* that all her ideas have the same high impact. Instead, for reasons that she may not fully understand or control, some of her papers may remain uncited whereas others become citation classics. For example, the Google Scholar citation profile of Nobel Prize winner Michael Levitt includes articles with citation counts ranging from 0 to 3572. This level of variability seems too great to be explained by a citation-level random factors model.

In theory, creativity-related factors could be conceived and modelled as each person having an underlying research-related creative power but each of their publications randomly varying above or below this value. This explains why the same group of authors can produce works with substantially different research impacts (however assessed).

2.1. The research capability model

If the above publication-level random factors logic is accepted then a confidence interval for an impact indicator partly reflects the underlying ability of a group to produce impactful work in addition to external, citation-related factors. From this perspective, the set of publications produced by a group is not a definitive finite population (Nane, 2016) but is a sample from an infinite set of the publications that the group might have written in similar circumstances (Claveau, 2016; Williams & Bornmann, 2016). This publication-level randomness is primarily modelled in the current paper and the theoretical background introduced here will be called the *research capability model*. For clarity, it can be distinguished from two other perspectives.

- *The research capability model*: A group of researchers (of any size, including a single person) has a fixed underlying capability to produce research of a given quality or impact (or range of qualities) but the quality and impact of that research is affected by factors outside their control, including creativity variability, field changes, and mission changes. This "fixed underlying capability" may have distributional characteristics, such as 50% of the group being able to produce excellent research and 50% being able to produce average research. From a research capability perspective, citation analyses would aim to estimate the underlying research capability of a research group from the citations to its publications. This analysis would need to consider both publication-level and citation-level random factors.