Journal of Informetrics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Informetrics

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

Normalization of zero-inflated data: An empirical analysis of a new indicator family and its use with altmetrics data☆

Lutz Bornmann[a,*], Robin Haunschild[b]

[a] Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany
[b] Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70569 Stuttgart, Germany

ARTICLE INFO

Article history:
Received 9 November 2017
Received in revised form
22 December 2017
Accepted 24 January 2018
Available online xxx

Keywords:
Zero-inflated data
Citation counts
Altmetrics
Equalized Mean-based Normalized Proportion Cited (EMNPC)
Mean-based Normalized Proportion Cited (MNPC)
Mantel-Haenszel quotient (MHq)

ABSTRACT

Recently, two new indicators (Equalized Mean-based Normalized Proportion Cited, EMNPC; Mean-based Normalized Proportion Cited, MNPC) were proposed which are intended for sparse scientometrics data, e.g., alternative metrics (altmetrics). The indicators compare the proportion of mentioned papers (e.g. on Facebook) of a unit (e.g., a researcher or institution) with the proportion of mentioned papers in the corresponding fields and publication years (the expected values). In this study, we propose a third indicator (Mantel-Haenszel quotient, MHq) belonging to the same indicator family. The MHq is based on the MH analysis – an established method in statistics for the comparison of proportions. We test (using citations and assessments by peers, i.e. F1000Prime recommendations) if the three indicators can distinguish between different quality levels as defined on the basis of the assessments by peers. Thus, we test their convergent validity. We find that the indicator MHq is able to distinguish between the quality levels in most cases while MNPC and EMNPC are not. Since the MHq is shown in this study to be a valid indicator, we apply it to six types of zero-inflated altmetrics data and test whether different altmetrics sources are related to quality. The results for the various altmetrics demonstrate that the relationship between altmetrics (Wikipedia, Facebook, blogs, and news data) and assessments by peers is not as strong as the relationship between citations and assessments by peers. Actually, the relationship between citations and peer assessments is about two to three times stronger than the association between altmetrics and assessments by peers.

1. Introduction

Alternative metrics (altmetrics) have been established as a new fast-moving and dynamic area in scientometrics (Galloway, Pease, & Rauh, 2013). Initially, altmetrics have been proposed as an alternative to traditional bibliometric indicators. Altmetrics are a collection of multiple digital indicators which measure activity related to research papers on social media platforms, in mainstream media, or in policy documents (National Information Standards Organization, 2016; Work, Haustein, Bowman, & Larivière, 2015). Haustein (2016) identified the following seven groups of platforms which are (currently) used for altmetrics: "(a) social networking (e.g., Facebook, ResearchGate), (b) social bookmarking and reference

☆ This paper is based on a presentation at the 16th International Conference on Scientometrics & Informetrics (ISSI) 2017.

* Corresponding author.
E-mail addresses: bornmann@gv.mpg.de (L. Bornmann), R.Haunschild@fkf.mpg.de (R. Haunschild).

https://doi.org/10.1016/j.joi.2018.01.010
1751-1577/

management (e.g., Mendeley, Zotero), (c) social data sharing including sharing of datasets, software code, presentations, figures and videos, etc. (e.g., Figshare, Github), (d) blogging (e.g., ResearchBlogging, Wordpress), (e) microblogging (e.g., Twitter, Weibo), (f) wikis (e.g., Wikipedia), as well as (g) social recommending, rating and reviewing (e.g., Reddit, F1000Prime)" (p. 417).

According to Adie (2014), there are three developments which foster the engagement in altmetrics. (1) Evaluators, funders, or national research assessments are not only interested in research impact inside but also outside of academia (Mohammadi, Thelwall, & Kousha, 2016; Thelwall & Kousha, 2015a). (2) There is a general shift from print to online. In an early study, Bollen, Van de Sompel, and Rodriguez (2008) demonstrated the richness of data from online activities. The data include web citations in digitized scholarly documents and from social media (Wilsdon et al., 2015). (3) The publication of the altmetrics manifesto by Priem, Taraborelli, Groth, and Neylon (2010) gave this new area in scientometrics a name and thus a focal point. Today, many publishers add altmetrics to papers in their collections (e.g., Wiley and Springer) (Thelwall & Kousha, 2015b). Altmetrics are also recommended by Snowball Metrics (Colledge, 2014) for research evaluation purposes – an initiative publishing global standards for institutional benchmarking in the academic sector (www.snowballmetrics.com).

In recent years, some altmetrics indicators have been proposed which are field- and time-normalized. These indicators were developed because evidences have been published that this data is – similar to bibliometric data – field- and time-dependent (see, e.g., Bornmann, 2014b). Obviously, some fields are more relevant to a broader audience or general public than others (Haustein, Larivière, Thelwall, Amyot, & Peters, 2014). Bornmann and Haunschild (2016b) and Haunschild and Bornmann (2016) introduced the mean discipline normalized reader score (MDNRS) and the mean normalized reader score (MNRS) based on Mendeley data (see also Fairclough & Thelwall, 2015). Bornmann and Haunschild (2016a) propose the Twitter Percentile (TP) – a field- and time-normalized indicator for Twitter data. This indicator was developed against the backdrop of a problem with altmetrics data which is also addressed in this study – the inflation of the data with zero counts. The overview of Work et al. (2015) on studies investigating the coverage of papers on social media platforms shows that many platforms have coverages of less than 5% (e.g., blogs or Wikipedia). This result is confirmed by the meta-analysis of Erdt, Nagarajan, Sin, and Theng (2016): their analyses across former empirical studies dealing with the coverage of altmetrics show that about half of the platforms are at or below 5%; except for three (out of eleven) the coverage is below 10%. Common normalization procedures based on averages and percentiles of individual papers are problematic for zero-inflated data sets (Haunschild, Schier, & Bornmann, 2016). Bornmann and Haunschild (2016a) circumvent the problem of zero-inflated Twitter data by including in the calculation of TP only journals with at least 80% of the papers with at least 1 tweet each. However, this procedure leads to the exclusion of many journals.

Recently, Thelwall (2017a, 2017b) proposed another family of field- and time normalized indicators which compare the proportion of mentioned papers (e.g. on Facebook or Wikipedia) of a unit (e.g., a researcher or institution) with the proportion of mentioned papers in the corresponding fields and publication years (the expected values). The family consists of the Equalized Mean-based Normalized Proportion Cited (EMNPC) and the Mean-based Normalized Proportion Cited (MNPC). In this study, we investigate the new indicator family empirically and add a further variant to this family. In statistics, the Mantel-Haenszel (MH) analysis is recommended for pooling the data from multiple 2×2 cross tables based on different subgroups (here: mentioned and not mentioned papers of a unit published in different subject categories and publication years compared with the corresponding reference sets) (Sheskin, 2007). We call the new indicator Mantel-Haenszel quotient (MHq).

In the first step of the empirical analysis, we analyze the convergent validity of the new indicator family by comparing the scores with ratings by peers. We investigate whether the indicators are able to discriminate between different quality levels assigned by peers to publications. Since the convergent validity can only be tested by using citations (which are related to quality), the first empirical part is based on citations. Good performance on the convergent validity test is an important condition for the use of the indicators in altmetrics. For altmetrics, the relationship to quality – as measured by peer assessments – is not clear. Since the first empirical part will show that the MHq is convergent valid, we test the ability of several altmetrics (e.g., Wikipedia and Facebook counts) to discriminate between quality levels. Thus, we investigate whether several altmetrics are related to the quality of publications – measured in terms of peers' assessments.

2. Indicators for zero-inflated count data

Whereas the EMNPC and MNPC proposed by Thelwall (2017a) are explained in Sections 2.1 and 2.2, the MHq is firstly introduced in Section 2.3. The next sections present not only the formulas for the calculation of the three metrics, but also the corresponding 95% confidence intervals (CIs). The CI is a range of possible indicator values: We can be 95% confident that the interval includes the "true" indicator value in the population. With the use of CIs, we assume that we analyze sample data and infer to a larger, inaccessible population (Williams & Bornmann, 2016). According to Claveau (2016), the general argument for using inferential statistics with scientometric data is "that these observations are realizations of an underlying data generating process . . . The goal is to learn properties of the data generating process. The set of observations to which we have access, although they are all the actual realizations of the process, do not constitute the set of all possible realizations. In consequence, we face the standard situation of having to infer from an accessible set of observations – what is normally called the sample – to a larger, inaccessible one – the population. Inferential statistics are thus pertinent" (p. 1233).

The relationship between 95% CIs and statistical significance (in case of independent proportions) is as follows:

Please cite this article in press as: Bornmann, L., & Haunschild, R. Normalization of zero-inflated data: An empirical analysis of a new indicator family and its use with altmetrics data. *Journal of Informetrics* (2017), https://doi.org/10.1016/j.joi.2018.01.010