Journal of Informetrics 7 (2013) 945–958

Contents lists available at ScienceDirect

Journal of Informetrics

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

The comparison of normalization procedures based on different classification systems

Yunrong Li, Javier Ruiz-Castillo*

Departamento de Economía, Universidad Carlos III, Spain

ARTICLE INFO

Article history:
Received 20 May 2013
Received in revised form
18 September 2013
Accepted 24 September 2013

Keywords:
Citation practices
Normalization procedures
Classification systems
Citation inequality

ABSTRACT

In this paper, we develop a novel methodology within the IDCP measuring framework for comparing normalization procedures based on different classification systems of articles into scientific disciplines. Firstly, we discuss the properties of two rankings, based on a graphical and a numerical approach, for the comparison of any pair of normalization procedures using a single classification system for evaluation purposes. Secondly, when the normalization procedures are based on two different classification systems, we introduce two new rankings following the graphical and the numerical approaches. Each ranking is based on a double test that assesses the two normalization procedures in terms of the two classification systems on which they depend. Thirdly, we also compare the two normalization procedures using a third, independent classification system for evaluation purposes. In the empirical part of the paper we use: (i) a classification system consisting of 219 sub-fields identified with the Web of Science subject-categories; an aggregate classification system consisting of 19 broad fields, as well as a systematic and a random assignment of articles to sub-fields with the aim of maximizing or minimizing differences across sub-fields; (ii) four normalization procedures that use the field or sub-field mean citations of the above four classification systems as normalization factors; and (iii) a large dataset, indexed by Thomson Reuters, in which 4.4 million articles published in 1998–2003 with a five-year citation window are assigned to sub-fields using a fractional approach. The substantive results concerning the comparison of the four normalization procedures indicate that the methodology can be useful in practice.

1. Introduction

Differences in publication and citation practices have been known for decades to create serious difficulties for the comparison of raw citation counts across different scientific disciplines. Since the early eighties various normalization proposals have been suggested (see the review by Schubert & Braun, 1996). Moreover, the normalization problem has recently attracted renewed interest.[1] Consequently, there is a need to develop methods for the comparison of the performance achieved by different normalization procedures in empirical situations.

Abbreviation: IDCP, citation inequality due to differences in citation practices.

* Corresponding author. Tel.: +34 91 624 95 88.
E-mail address: jrc@eco.uc3m.es (J. Ruiz-Castillo).

[1] Among the target, or cited-side variety of normalization procedures, see Glänzel (2011), Radicchi, Fortunato, and Castellano (2008), Radicchi and Castellano (2012), Crespo, Li, and Ruiz-Castillo (2013) and Crespo, Herranz, Li, and Ruiz-Castillo (2013), as well as the review of the percentile rank approach by Bornmann and Marx (2013). Among the source, or citing-side variety, see inter alia Zitt and Small (2008), Moed (2010), Leydesdorff and Opthof (2010), and Waltman and Van Eck (2013a).

1751-1577/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2013.09.005

Lacking information on the citing side, we focus on normalization procedures of the target or cited-side variety, where each procedure is based on a priori given classification system of publications in the periodical literature into a set of scientific disciplines. The paper studies the evaluation of alternative normalization procedures in two scenarios. In the first one, there is only a single classification system for the implementation as well as the evaluation of two (or more) normalization procedures. In this case, all that is needed is a method for assessing the performance of the contesting normalization procedures using the single classification system for evaluation purposes. In the second scenario, there are two (or more) classification systems for the implementation and the evaluation of two (or more) normalization procedures. As far as we know, this is the first paper that presents a complete discussion of this case (see, however, the contributions by Sirtes, 2012; Waltman & Van Eck, 2013b, that will be discussed below).

Given a classification system, we evaluate the performance of normalization procedures using the measurement framework recently introduced in Crespo, Li, et al. (2013), where the number of citations received by an article is a function of two variables: the article's underlying scientific influence, and the discipline to which it belongs. Consequently, the citation inequality of the distribution consisting of all articles in all disciplines – the *all-sciences case* – is the result of two forces: differences in scientific influence within each homogeneous discipline, and differences in citation practices across the set of heterogeneous disciplines. Essentially, as we will see below, the effect of the latter on citation inequality is captured by an *IDCP* term – where *IDCP* stands for citation inequality attributable to differences in citation practices.

A key aspect of this framework is that it serves to evaluate any set of normalization procedures in terms of any given classification system as required in the first scenario. The evaluation can take a graphical, or a numerical form.[2] In this paper, we establish that the graphical approach does not provide a complete ranking, i.e. we show that there are situations in which a pair of normalization procedures is non-comparable according to the graphical criterion. We also establish that the rankings according to the two approaches are logically independent, that is, we show that there exists at least one pair of normalization procedures that are ordered differently by the two rankings.

The canonical example of the second scenario arises in the presence of a number of classification systems at different aggregate levels. Assume for simplicity that there are only two hierarchically nested classification systems into what we call *sub-fields* and *fields*, so that every sub-field at the lower aggregation level belongs to only one field at the higher aggregate level. The question we study in this paper is how to compare one normalization procedure based at the sub-field level with another based at the field level. The problem is that we only know how to assess alternative normalization procedures using a single classification system for evaluation purposes. Therefore, the performance of the first procedure evaluated at the sub-field level cannot be directly compared with the performance of the second procedure evaluated at the field level. Our solution to this problem is the introduction of a new ranking based on a double test that assesses both normalization procedures in terms of the two classification systems on which they depend. For a procedure to dominate the other according to the double test, it should perform better than the other under both classification systems.

This idea is applicable to the comparison of any two normalization procedures based on different classification systems independently of the method followed for their evaluation. However, it should be remembered that in our measuring framework the evaluation of normalization procedures could take a graphical and a numerical approach. Therefore, in our case we must introduce *two* new rankings, each of them relying on a double test that compares the two normalization procedures using for evaluation purposes the two classification systems on which they depend. For a procedure to strongly dominate another according to the graphical (or the numerical) approach it should exhibit a better graphical (or numerical) performance under both classification systems. We establish that the two rankings are logically independent; therefore, strict dominance according to one ranking does not necessarily imply dominance according to the second.

This strategy deserves two closely related comments. Firstly, satisfying either of the two dominance criteria is a strong requirement. Consequently, we expect that neither of the two new rankings is complete. Secondly, Sirtes (2012) first suggested that the assessment of two classification-system-based normalization procedures would be generally biased in favor of the normalization procedure based on the system used for evaluation purposes. Waltman and Van Eck (2013b) concur with this idea, and provide further arguments about the possibility of this bias. In a double test, the presence of a bias of this type would favor the first (and the second) procedure under comparison when the first (and the second) classification system is used for evaluation purposes. Therefore, the bias would increase the probability that the two procedures are non-comparable. In any case, we confirm that neither of the two rankings is complete.

In order to avoid the bias, Waltman and Van Eck (2013b) compare source and target normalization procedures using an independent classification system for evaluation purposes. On our part, we believe that this is a recommendation worth pursuing. Thus, in the second scenario we suggest the comparison of any pair of classification-system-based normalization procedures using two strategies: the double tests that only involve the two classification systems on which the normalization procedures are based and, by analogy with Waltman and Van Eck's (2013b) procedure, the evaluation in terms of a third, independent classification system. Therefore, to illustrate this methodology in empirical situations we need to specify a

[2] Both forms have been previously used in two instances: (i) to compare the performance of different normalization procedures based on the same classification system (Crespo, Li, et al., 2013; Crespo, Herranz, et al., 2013; Li, Castellano, Radicchi, & Ruiz-Castillo, 2013), and (ii) to compare two types of normalization procedures, namely, those target procedures in which the disciplines' mean citations in different classification systems are used as normalization factors, and a variety of source normalization procedures independent of any classification system (Waltman & Van Eck, 2013b).