Journal of Informetrics 9 (2015) 102–117

Contents lists available at ScienceDirect

Journal of Informetrics

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

Field-normalized citation impact indicators using algorithmically constructed classification systems of science

Javier Ruiz-Castillo[a,*], Ludo Waltman[b]

[a] *Departamento de Economía, Universidad Carlos III of Madrid, Spain*
[b] *Centre for Science and Technology Studies, Leiden University, The Netherlands*

ARTICLE INFO

Article history:
Received 17 July 2014
Received in revised form 10 October 2014
Accepted 27 November 2014
Available online 18 December 2014

Keywords:
Field normalization
Classification systems
Clustering methodology
Citation impact indicators
University rankings

ABSTRACT

We study the problem of normalizing citation impact indicators for differences in citation practices across scientific fields. Normalization of citation impact indicators is usually done based on a field classification system. In practice, the Web of Science journal subject categories are often used for this purpose. However, many of these subject categories have a quite broad scope and are not sufficiently homogeneous in terms of citation practices. As an alternative, we propose to work with algorithmically constructed classification systems. We construct these classification systems by performing a large-scale clustering of publications based on their citation relations. In our analysis, 12 classification systems are constructed, each at a different granularity level. The number of fields in these systems ranges from 390 to 73,205 in granularity levels 1–12. This contrasts with the 236 subject categories in the WoS classification system. Based on an investigation of some key characteristics of the 12 classification systems, we argue that working with a few thousand fields may be an optimal choice. We then study the effect of the choice of a classification system on the citation impact of the 500 universities included in the 2013 edition of the CWTS Leiden Ranking. We consider both the MNCS and the $PP_{top\,10\%}$ indicator. Globally, for all the universities taken together citation impact indicators generally turn out to be relatively insensitive to the choice of a classification system. Nevertheless, for individual universities, we sometimes observe substantial differences between indicators normalized based on the journal subject categories and indicators normalized based on an appropriately chosen algorithmically constructed classification system.

1. Introduction

In this paper, we deal with the problem of normalizing citation impact indicators based on a classification system of science. As we know, the choice of a *classification system*, that is, the assignment of individual scientific publications (or journals) to research areas, remains an open question in Scientometrics. Together with the well-known classification systems included in Thomson Reuters' Web of Science (WoS hereafter) and Elsevier's Scopus databases, there are a number of interesting proposals suggested by individual researchers (see inter alia the references in Waltman & Van Eck, 2012).

In practice, the choice of the WoS classification system is often made because it is the only classification system that is readily available. However, a number of studies question the appropriateness of the WoS classification system for the purpose

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

http://dx.doi.org/10.1016/j.joi.2014.11.010
1751-1577/

of normalizing citation impact indicators. Neuhaus and Daniel (2009) contrast the assignment of individual publications to WoS subject categories based on the journal where they have appeared with a novel methodology for Chemistry and related fields where each publication is directly assigned to one of the 80 sections of the *Chemical Abstracts* database. Taking the journal *Angewandte Chemie* as an example, they illustrate the limitations of the WoS journal classification scheme in the case of general journals. On the other hand, using the 20 sections under the Biochemistry heading they clearly illustrate that citation habits vary extensively not only *between* fields but also *within* fields. Similarly, Van Eck, Waltman, Van Raan, Klautz, and Peul (2013) establish the existence of heterogeneous sub-groups (corresponding to clinical and basic medical research) with different citation practices within WoS subject categories. In this case, the mean citation of an entire category is simply the weighted average of different, and hence non-comparable sub-group mean citations. This is exactly the same problem found by Van Leeuwen and Calero-Medina (2012) inside the Thomson Reuters broad field of Economics and Business. Within the dominant WoS journal subject category in that field, denoted *Economics*, these authors find strong differences across 19 specialties defined in the EconLit electronic bibliography produced by the American Economic Association. Finally, Leydesdorff and Bornmann (2014) point out that the WoS subject categories were developed decades ago for the purpose of information retrieval and evolved incrementally with the database; the classification is machine-based and partially manually corrected. This contribution shows the potential problems for research evaluation in one discipline that is attributed a WoS category – *Information Science and Library Science* – and one specialty which is not – *Science and Technology Studies*.

Clearly, the comparison of the WoS system with some relevant alternatives is an important research problem. In this paper, we search for alternatives within the publication-level algorithmic methodology introduced by Waltman and Van Eck (2012) (the WVE methodology hereafter). This methodology is able to handle very large datasets, and uses a transparent clustering technique that classifies publications into clusters solely based on direct citations between them. Contrary to the WoS system, each publication is assigned to a single cluster. Moreover, the WVE methodology can be used to construct classification systems that, unlike for instance the *Chemical Abstracts* and EconLit systems, cover all scientific fields.

In the first large-scale application of the WVE methodology (Waltman & Van Eck, 2012), three types of parameter values needed to be chosen: the number of what we call *granularity levels* and, at each level, the *minimum number of publications per cluster*, and the *resolution parameter* that determines the level of detail of the clustering (i.e., a small number of large clusters vs. a large number of small clusters). In this paper, we consider a set of twelve granularity levels that are not restricted to be hierarchically linked. Thus, by fixing the resolution parameter at twelve different values, we build a sequence of independent classification systems in each of which the same set of publications is assigned to an increasing number of clusters. Furthermore, no minimum number of publications per cluster is imposed at any granularity level. Thus, at every step, the WVE algorithm freely determines a cluster size distribution.

We apply this scheme to a WoS dataset consisting of 3.6 million articles published in 2005–2008 in academic journals – excluding trade journals, national journals, etc. – and the citations they receive during a five-year citation window for each year in that period. The number of clusters in the WVE sequence ranges from 390 to 73,205 in granularity levels 1–12. This contrasts with the 236 clusters (i.e., journal subject categories) in the WoS classification system.

Which granularity level is used in practice in the calculation of normalized citation impact indicators is a very important issue. As clearly argued by Zitt, Ramana-Rahari, and Bassecoulard (2005), "*An article may exhibit very different citation scores or rankings when compared within a narrow specialty or a large academic discipline.*"(op. cit., p. 391). If we choose a granularity level dominated by a relatively small number of broad fields, the danger is that they are too heterogeneous, in which case comparisons between publications within the same cluster may be biased. For instance, this may affect the *Essential Science Indicators* of Thomson Reuters that provide reference standards solely for 22 broad fields of research. However, when we go in the opposite direction and choose a classification system including too many clusters, we face difficulties of a different nature. Firstly, some clusters may mostly include the output of a subset of closely connected authors citing each other, and isolated from *bona fide* scientific communities whose output is classified in other clusters. Secondly, some clusters may be so small as to jeopardize their statistical properties. Thirdly, some clusters may have artificially low mean citations, so that standard normalization procedures that use cluster mean citations as normalization factors will tend to over-estimate these clusters' publications against those in high impact clusters characterized by a high mean citation. It may very well be the case that classification systems characterized by high granularity levels are plagued with clusters that present the above three difficulties together.

As a consequence of the above issues, the evaluation of research units based on citation impact is likely to be dependent on the granularity level at which the evaluation takes place. As Zitt et al. (2005) conclude, "*The fact that citation indicators are not stable from a cross-scale perspective is a serious worry for bibliometric benchmarking. What can appear technically as a* '*lack of robustness*' *raises deeper questions about the legitimacy of particular scales of observation*." (op. cit., p. 392). Adams, Gurney, and Jackson (2008) reach a similar conclusion: "*the fact that more than one view and hence more than one interpretation of performance might exist would need to be taken into account in any evaluation methodology*" (op. cit., p. 94). For other studies on this problem, we refer to Colliander and Ahlgren (2011) and Glänzel, Thijs, Schubert, and Debackere (2009).

In this paper, we investigate two questions. Firstly, what are the main characteristics of the twelve WVE classification systems, and how do they compare with those of the WoS alternative? Secondly, what are the consequences of using the WoS classification system or an appropriately selected member of the WVE sequence for the evaluation of the citation impact of universities?

For the first purpose, we study how the following characteristics evolve as the granularity level increases: the cluster size and the cluster mean citation distributions, the degree of skewness and the similarity of this characteristic across cluster