Journal of Informetrics 8 (2014) 25–28

Contents lists available at ScienceDirect

Journal of Informetrics

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

Letter to the Editor

The comparison of classification-system-based normalization procedures with source normalization alternatives in Waltman and Van Eck (2013)

Along with other co-authors, Ludo Waltman and Nees Jan van Eck have recently contributed to significantly increasing the fairness of bibliometric research assessments, in particular that of multidisciplinary assessments involving comparisons of citation impact between different fields of science (Van Eck, Waltman, Van Raan, Klautz, & Peul, 2013; Waltman & Van Eck, 2013a; Waltman, Van Eck, Van Leeuwen, & Visser, 2013). This note addresses some pending issues in their latest contribution – Waltman and Van Eck (2013b) for this journal, referred to as WVE hereafter – concerning a systematic large-scale empirical comparison of classification-system-based *versus* source normalization procedures.[1] Among the former, they focus on what we will call classification system *WoS*, namely, the system where publications are classified into fields based on the journal subject categories in the Web of Science bibliographic database. WVE study the normalization procedure based on this system that uses field mean citations as normalization factors.[2] To differentiate this procedure from others of the same type, we denote it by *NWoS* (rather than *NSC* as do WVE). On the other hand, according to WVE $SNCS^{(3)}$ exhibits the best performance among the source normalization procedures. Therefore, for our purposes, the issue put forward by WVE is the comparison of *NWoS versus* $SNCS^{(3)}$.

A key methodological feature of WVE's contribution is the distinction between the use of a classification system in the implementation and the evaluation of a normalization procedure. Sirtes (2012) first suggested that using a certain classification system for evaluation purposes would be generally biased in favor of normalization procedures based on that particular system. WVE concur with this idea, and provide further arguments about the possibility of this bias (see footnote 6 and Appendix C). Therefore, they recommend that the comparison between *NWoS* and $SNCS^{(3)}$ should be done using a second, independent classification system for evaluation purposes. Following this recommendation, WVE use three systems algorithmically constructed according to the methods in Waltman and Van Eck (2012), systems *A*, *B*, and *C*, consisting of 21, 161, and 1334 scientific fields at different granularity or aggregation levels.

Given a classification system, the degree to which differences in citation practices between fields have been corrected is indicated by the degree to which the field-normalized citation distributions coincide with each other. In particular, WVE use the measurement framework introduced in Crespo, Li, and Ruiz-Castillo (2013) where, given a classification system, the effect on citation inequality of differences in citation practices is captured by an *IDCP* between-group inequality term in a certain partition of the overall citation distribution by field and quantile, where *IDCP* stands for citation *I*nequality attributable to *D*ifferences in *C*itation *P*ractices. The evaluation of any set of normalization procedures in terms of a given classification system can take a graphical or a numerical form. Following the graphical approach, WVE reach the following conclusion:

> The $SNCS^{(3)}$ procedure generally performs better than the NWoS procedure, specifically at higher levels of granularity.

Li and Ruiz-Castillo (2013) establish that the graphical and the numerical approaches are logically independent. Therefore, they can be used in a complementary fashion. To save space, in most of this note I will follow the numerical approach where, given a classification system, each normalization procedure is assessed in terms of the reduction it generates in the *IDCP* term. To understand the evaluation results obtained with this approach, some notation is needed. Recall that we have four classification systems, which will be indexed by $K = WoS, A, B$, and *C*. Given system *K*, denote by *IDCP*(*K*) the *IDCP* term that captures the effect on citation inequality of differences in citation practices across fields in *K*. Similarly, given system *K*, denote by *NK* the associated normalization procedure. Finally, given a classification system *G* for the evaluation of procedure

[1] As Li, Castellano, Radicchi, and Ruiz-Castillo (2013) and WVE indicate, the normalization need arises at the cardinal level, that is, in situations where the actual number of citation counts of individual publications – and not only their location in a percentile distribution (or a percentile class) – is needed. At the ordinal level, the percentile rank approach provides a sort of perfect normalization where, for any classification system, all citation distributions become equally distributed. For the percentile rank approach, see *inter alia* Bornmann and Marx (2013), Bornmann (in press), and Bornmann et al. (in press).

[2] This well-known, traditional, and inexpensive normalization procedure has been favorably evaluated from a number of different perspectives in recent contributions (Radicchi et al., 2008; Radicchi and Castellano, 2012; Crespo, Li, et al., 2013; Crespo, Herranz, Li, & Ruiz-Castillo (2013); Li, Castellano, Radicchi, & Ruiz-Castillo, 2013). Therefore, in this note we will always use it as a convenient representative of classification-system-based normalization procedures.

1751-1577/$ – see front matter
http://dx.doi.org/10.1016/j.joi.2013.10.002

Table 1
The impact of normalization under the four classification systems.

Change in the value of the *IDCP* term after normalization by the different procedures, in %				
Normalization	Classification system used for evaluation purposes			
Procedures	*A*	*B*	*C*	*WoS*
WoS	87.0	82.5	72.2	86.8
$SNCS^{(3)}$	85.4	83.7	73.4	80.1
NA	88.8	75.3	61.0	68.7
NB	88.8	87.8	71.4	75.6
NC	89.2	88.6	86.9	80.2

NK, with *G* not necessarily equal to *K*, denote by $IDCP^{NK}(G)$ the *IDCP* term within system *G* after normalization with *NK*. To rank any pair of procedures *NK* and *NL* under classification system *G*, we compare $IDCP^{NK}(G)$ with $IDCP^{NL}(G)$. We find it more useful to express the result as the percentage that the differences in the *IDCP* terms before and after normalization, $[IDCP(G) - IDCP^{NK}(G)]$ and $[IDCP(G) - IDCP^{NL}(G)]$, represent relative to the initial situation, $IDCP(G)$. Fortunately, WVE provide the values for $IDCP(G)$ and $IDCP^{NK}(G)$ for $K = WoS, SNCS^{(3)}$ when $G = A, B, C$ in Tables D1, D2, and D3 in Appendix D. The remaining values – for $K = A$, *B*, and *C*, and when *WoS* is used for evaluation purposes – have been kindly provided by Ludo Waltman (to save space, this information is available on request). Thus, we have constructed Table 1 presenting the change in the *IDCP* term before and after each of the five normalization procedures using the four classification systems for evaluation purposes.

Consider, for example, the case where normalization procedure *NWoS* is applied to the data organized according to system *A*. The consequences are captured by $IDCP^{NWoS}(A) = 0.0237$ (row 2 and column *IDCP* in Table D1 in WVE). In turn, $IDCP(A) = 0.1818$ (row 1 and column 3 in Table D1 in WVE). We are interested in the percentage change in the *IDCP* term before and after applying *NWoS* in *A*, that is, in the expression

$$\frac{100[IDCP(A) - IDCP^{NWoS}(A)]}{IDCP(A)} = \frac{100(0.1818 - 0.0237)}{0.1818} = 87.0.$$

The value of this expression appears in row *NWoS* and column *A* in Table 1, indicating that the effect of differences in citation practices across fields in system *A* has been reduced by 87% as a consequence of normalization by *NWoS*. This figure can be compared, for example, with the 88.8% reduction caused by normalization with *NA* using *A* itself for evaluation purposes (row *NA* and column *A* in Table 1). On the other hand, the figures in columns *B*, *C*, and *WoS* in row *NWoS*, for example, are the values in expression 100 $[IDCP(K) - IDCP^{NWoS}(K)]/IDCP(K)$ when the evaluation system is $K = B, C$, and *WoS* rather than *A*.

Our first observation is that when we evaluate *NWoS* and $SNCS^{(3)}$ using the independent classification systems *A*, *B*, and *C* in Table 1, we obtain the same results as WVE with the graphical approach: $SNCS^{(3)}$ performs better than *NWoS* (albeit by a small margin) using systems *B* and *C*, but the opposite is the case when we use system *A*. Since system *A* is less discriminating than *B* and *C*, we conclude that $SNCS^{(3)}$ performs generally better than *NWoS* using the numerical approach.[3] Given the independence between the two approaches, this is an important result pointing toward a certain superiority of source over classification-system-based normalization procedures. However, the next two observations go against this provisional conclusion.

1. As WVE point out in their concluding Section 5, any classification system can be expected to introduce certain biases in normalization, simply because any organization of the scientific literature into a number of perfectly separated fields of science is artificial. The obvious advantage of source normalization approaches is that they are independent of any classification system. However, in most practical situations researchers are limited to working within a single classification system, and do not have access to the (active) references needed to implement any citing-side normalization procedure. In this situation, only normalization procedures of the cited-side variety are available. Assume, for example, that we are limited to working with only a single classification system *K*, where *K* could be equal to $K = WoS$, *A*, *B*, or *C*. According to Table 1, normalization by *NK* reduces the $IDCP(K)$ term in the range 87–89%.[4] Furthermore, one could use other cited-side normalization procedures that perform even better. For example, judging from the results obtained in Li, Castellano, Radicchi, and Ruiz-Castillo (2013), one could use the two-parameter scheme originally suggested by Radicchi and Castellano (2012). There might be better alternatives, but this large reduction in the effect on citation inequality of

[3] Li and Ruiz-Castillo (2013) also find that the evaluation using a less discriminating classification system that assigns publications to fields in a random manner, so as to make the differences between them as small as possible, may lead to results that contradict the conclusions obtained under other, preferable approaches.

[4] This reduction in the *IDCP* term is of the same order of magnitude as the reduction generated with two average-based normalization procedures in situations where there is only a single classification system available (Crespo, Li, et al., 2013; Crespo, Herranz, et al., 2013).

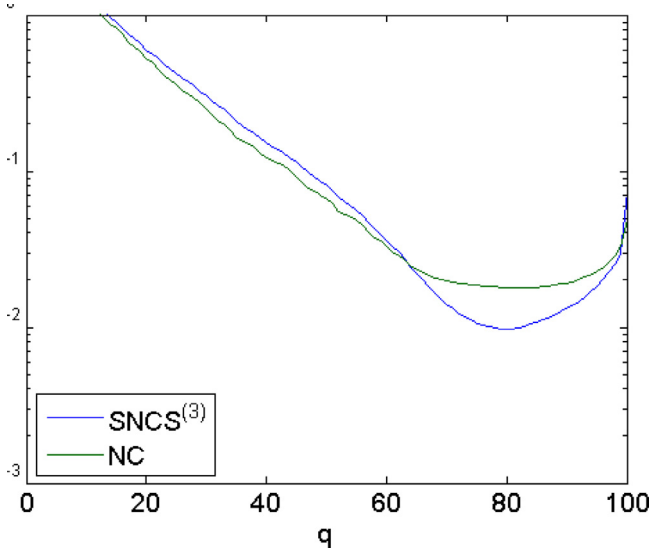

Fig. 1. The comparison of normalization procedures *NC* and $SNCS^{(3)}$ using classification system *WoS* for evaluation purposes.

differences in citation practices across fields generated by readily available normalization procedures may be acceptable in many practical situations.

2. The contrast between citing- and cited-side normalization procedures is very important. However, there is no reason for restricting the analysis to the normalization procedure based on the *WoS* system. One plausible strategy is to choose the best possible procedure among those based on the algorithmic systems *A*, *B*, and *C*. 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, for example, *NC* evaluated in terms of $IDCP^{NC}(C)$ cannot be directly compared with the performance of, for example, *NB* evaluated in terms of $IDCP^{NB}(B)$. The solution to this problem in Li and Ruiz-Castillo (2013) is to use 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 under both classification systems. Taking into account that using a certain classification system for evaluation purposes may favor the normalization procedure based on this system, this is a strong requirement. For example, the results in Table 1 indicate that *NA*, *NB*, and *NC* are non-comparable with *NWoS* according to the double test. However, *NC* dominates both *NB* and *NA* according to this criterion.[5]

Next, note that *NC* performs clearly better than $SNCS^{(3)}$ using system *C* for evaluation purposes. This is also the case using systems *B* or *A*. According to Sirtes (2012) and WVE, the first comparison may be biased in favor of *NC*. Similarly, it might be argued that systems *B* and *A* are not entirely independent of system *C* because, in spite of the fact that they are not hierarchically related, they are all constructed using the same algorithm at different granularity levels. Nevertheless, when we finally use the independent system *WoS* for evaluation purposes we find that *NC* barely dominates $SNCS^{(3)}$.[6] The difference is so small that we may as well conclude that *NC* and $SNCS^{(3)}$ are numerically non-comparable. As illustrated in Fig. 1 (kindly facilitated by Ludo Waltman), the intersections between the two curves indicate that the same conclusion is reached with the graphical approach. However, it must be recognized that this result has not been obtained under ideal conditions: using a classification system at a lower granularity level (such as the *WoS* system) for evaluation purposes might tend to equalize the performance of normalization procedures at a higher granularity level (such as *NC* and $SNCS^{(3)}$). Nevertheless, recall that the *WoS* system consists of 235 categories, not a small number.

In a nutshell, we conclude:

1. When there is only a single classification system available, and no information on the (active) references associated to articles in that dataset, there is ample evidence indicating that, among other alternatives, the procedure that uses field mean citations as normalization factors brings about large reductions in the term capturing the effect on citation inequality of differences in citation practices across fields.

[5] The fact that *NC* dominates *NB* and *NA* does not imply that system *C* is to be preferred to *B* and *A*. As indicated in Li and Ruiz-Castillo (2013), we agree with Zitt, Ramana-Rahari, and Bassecoulard (2005) and WVE that the choice of the best granularity level is an open question left for further research.

[6] Incidentally, note that both *NC* and $SNCS^{(3)}$ perform better than *NA* and *NB* using system *WoS* for evaluation purposes.

2. In agreement with WVE, we have confirmed that $SNCS^{(3)}$ performs numerically better than *NWoS* using independent systems *C* and *B* for evaluation purposes. However, classification-system-based *NC* barely dominates or is non-comparable with $SNCS^{(3)}$ under *WoS*. Therefore, normalization procedures based on algorithmic classification systems at an appropriate granularity level may perform as well as the best source normalization procedures analyzed in WVE, at least when evaluated in terms of a clearly independent system – such as *WoS* – at a lower but not negligible granularity level.

We believe that, taken together, the two conclusions raise some doubts concerning the idea that, as some may sustain based on WVE's results, source normalization procedures are ready to supplant their classification-system-based alternatives.

Acknowledgements

This letter was composed while the author was enjoying the hospitality of the CWTS, in Leiden University, The Netherlands, during the 2013 spring term. Conversations with Ludo Waltman and Nees Jan Van Eck are greatly appreciated. All remaining shortcomings are the sole responsibility of the author.

References

Bornmann, L., & Marx, W. (2013). How good is research really? *EMBO Reports*, *14*, 226–230.
Bornmann, L. (2013). How to analyse percentile citation impact data meaningfully in bibliometrics: The statistical analysis of distributions, percentile rank classes and top-cited papers. *Journal of the American Society for Information Science and Technology* (in press).
Bornmann, L., Bowman, B. F., Bauer, J., Marx, W., Schier, H., & Palzenberger, M. (2013). Standards for using bibliometrics in the evaluation of research institutes. In *Next generation metrics*. Cambridge, MA, USA: MIT Press (in press).
Crespo, J. A., Li, Y., & Ruiz-Castillo, J. (2013). The measurement of the effect on citation inequality of differences in citation practices across scientific fields. *PLoS ONE*, *8*(3), e75872.
Crespo, J. A., Herranz, N., Li, Y., & Ruiz-Castillo, J. (2013). The effect on citation inequality of differences in citation practices at the web of science subject category level. Working Paper 13-03, Universidad Carlos III (http://hdl.handle.net/10016/16327). *Journal of the American Society for Information Science and Technology* (forthcoming).
Li, Y., & Ruiz-Castillo, J. (2013). The comparison of normalization procedures based on different classification systems. *Journal of Informetrics*, *7*, 945–958.
Li, Y., Castellano, C., Radicchi, F., & Ruiz-Castillo, J. (2013). Quantitative evaluation of alternative field normalization procedures. *Journal of Informetrics*, *7*, 746–755.
Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*, 17268–17272.
Radicchi, F., & Castellano, C. (2012). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS ONE*, *7*(e33833), 1–7.
Sirtes, D. (2012). Finding the easter eggs hidden by oneself: Why Radicchi and Castellano's (2012) fairness test for citation indicators is not fair. *Journal of Informetrics*, *6*, 448–450.
Van Eck, N. J., Waltman, L., Van Raan, A. F. J., Klautz, R. J. M., & Peul, W. C. (2013). Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PLoS ONE*, *8*, e62395.
Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*, 2378–2392.
Waltman, L., & Van Eck, N. J. (2013a). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, *7*, 833–849.
Waltman, L., & Van Eck, N. J. (2013b). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison. *Scientometrics* (in press).
Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., & Visser, M. S. (2013). Some modifications to the SNIP journal impact indicator. *Journal of Informetrics*, *7*, 272–285.
Zitt, M., Ramana-Rahari, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalization. *Scientometrics*, *63*, 373–401.

Javier Ruiz-Castillo
Departamento de Economía, Universidad Carlos III, Spain
E-mail address: jrc@eco.uc3m.es

5 September 2013

Available online 12 November 2013