STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Relationship between field-normalized indicators calculated with different approaches of field-categorization[1]

Robin Haunschild*$, Werner Marx*, Bernie French**, and Lutz Bornmann***

$ Corresponding author
* *R.Haunschild@fkf.mpg.de; W.Marx@fkf.mpg.de*
Max Planck Institute for Solid State Research, Heisenbergstr. 1, Stuttgart, 70569 (Germany)

** *bfrench4255@gmail.com*
CAS (Chemical Abstracts Service), a division of the American Chemical Society, CAS Innovation LAB, 2540 Olentangy River Road, Columbus, Ohio 43202-1505 (USA)

*** *bornmann@gv.mpg.de*
Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, Munich, 80539 (Germany)

Introduction

It is one principle in the Leiden manifesto for the professional application of bibliometrics to use field-normalized scores instead of simple citation counts (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015). These scores reflect the impact of papers against the backdrop of their reference sets – papers published at the same time and in the same field. An important topic in the calculation of these scores is the definition of fields, which are used as reference sets (Wilsdon et al., 2015; Wouters et al., 2015). Three different approaches of field-categorization are currently (mainly) used for normalizing impact without a clear preference for one alternative: (1) journal sets, (2) intellectual assignments, and (3) citation relations.

In this study, we compare normalized citation scores, which have been calculated based on the three approaches to build reference sets. We are interested whether they lead to the same, similar, or different scores for the same papers – if the formula of calculating the scores is held constant. Since all approaches are in use for field-normalization in similar research evaluation contexts, we expect similar scores. Great differences would question the use of field-normalized scores in research evaluation, as long as no standard approach has been established.

This study focusses on chemistry and related sciences, because we have access to a comprehensive dataset from Chemical Abstracts Service (CAS), a division of the American

[1] The bibliometric data used in this paper are from an in-house database developed and maintained by the Max Planck Digital Library (MPDL, Munich) and derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI) prepared by Clarivate Analytics, formerly the IP & Science business of Thomson Reuters (Philadelphia, Pennsylvania, USA). We would like to thank the Centre for Science and Technology Studies (CWTS) for making their ACCS assignments to Web of Science (WoS) UTs available. Parts of this work were performed during a research visit of one of the co-authors (RH) with the CAS Innovation Lab, Columbus, Ohio). RH thanks CAS for support during his stay.

Chemical Society (ACS). CAS offers the largest database of the literature in these fields including intellectual assignments of fields to papers.

Methods

Approaches of field-classification

This study compares the agreement of normalized citation scores for the same papers, which have been calculated based on the following three field-categorization approaches:

(1) The most frequent approach in bibliometrics is to use subject categories that are defined by Clarivate Analytics for Web of Science (WoS) or by Elsevier for Scopus to assign papers to fields. The subject categories pool journals to sets, which publish papers in similar research areas (e.g. biochemistry or economics). It is an advantage of journal sets that they define a multidisciplinary classification system covering all research areas (Wang & Waltman, 2016). It is a disadvantage of the sets that they stretch to their limits with multi-disciplinary journals (e.g. *Nature* or *Science*) and journals covering many subfields (e.g. *Physical Review Letters*, or *The Lancet*). These journals cannot be reliably and validly assigned to one field (Haddow & Noyons, 2013) on the journal basis. However, the papers of multidisciplinary journals can be reassigned on a paper-level (Evidence, 2009),

(2) To overcome the limitations of journal sets, Bornmann, Mutz, Neuhaus, and Daniel (2008) propose to use mono-disciplinary classification systems (Waltman, 2016), e.g., Chemical Abstracts^{TM} (CA) sections in chemistry and related areas (Bornmann & Daniel, 2008; Bornmann, Schier, Marx, & Daniel, 2011), MeSH (Medical Subject Headings) terms in biomedicine (Bornmann, et al., 2008; Leydesdorff & Opthof, 2013; Strotmann & Zhao, 2010), or PACS (Physics and Astronomy Classification Scheme) codes in physics and related areas (Radicchi & Castellano, 2011). In these systems, experts in the field or the authors themselves assign each specific paper to the corresponding subfield, highlighting the most important aspects of the papers. It is an advantage of these systems that they have been introduced to reflect the subfield patterns in specific fields. Their disadvantage is that they can only be used for the normalization of papers from one discipline (and related areas).

(3) Waltman and van Eck (2012) introduced a multi-disciplinary classification system, which is based on direct citation relations between papers. The algorithm for computing the classification system needs three basic parameters as input in addition to the direct citation network: (i) the number of levels of the system, (ii) the resolution parameter, and (iii) the minimum number of papers per class (field). The approach is already in use in the Leiden ranking (see http://www.leidenranking.com/) for the calculation of normalized impact scores. The empirical results of Klavans and Boyack (2017) indicate that algorithmically constructed classifications are more accurate than classifications based on journal sets. Similar positive results have been published by Perianes-Rodriguez and Ruiz-Castillo (2016). Leydesdorff and Milojević (2015) criticize the classification system as follows: “Because these ‘fields’ are algorithmic artifacts, they cannot easily be named (as against numbered), and therefore cannot be validated. Furthermore, a paper has to be cited or contain references in order to be classified, since the approach is based on direct citation relations” (p. 201).

Statistics

The overview of Waltman (2016) demonstrates that several different approaches of calculating field-normalized scores have been developed. In this study, we use the normalized citation score (NCS) to compare normalized scores, since it is still the most frequently used approach. For the calculation of the NCS, each paper's citation count is divided by the average citation count in a corresponding reference set. The reference sets are defined by the papers, which belong to the same field (as defined by the field categorization approach) and publication year as the focal paper. If, for example, the paper has 3 citations and the average in the field is 10.67, the NCS of the paper is 3/10.67=0.28. The NCS is formally defined as

$$NCS = \frac{c_i}{e_i}$$

where c_i is the citation count of a focal paper and e_i is the corresponding citation rate in the field (Lundberg, 2007; Rehn, Kronman, & Wadskog, 2007; Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). Since the number of citations received by a paper depends on the time since publication, the NCS is calculated for publications from the same year. Using the different approaches of field-categorization, we calculated three NCS for every paper: NCS_{JS} (based on journal sets), NCS_{CA} (based on CA sections), and NCS_{CR} (based on citation relations).

In this study, we are interested in the relationship between NCS_{JS}, NCS_{CA}, and NCS_{CR}. To investigate the extent of agreement and disagreement between the different NCS, we group the papers in our dataset according to the Characteristics Scores and Scales (CSS) method. This method was proposed by Glänzel, Debackere, and Thijs (2016). For each NCS separately (NCS_{JS}, NCS_{CA}, and NCS_{CR}), the normalized scores are obtained by (1) truncating the publications at their mean and (2) recalculating the mean of the truncated part. Performing this procedure three times leads to four impact classes. Following Glänzel, et al. (2016), we labeled the four classes with "poorly cited", "fairly cited", "remarkably cited", and "outstandingly cited". The poorly cited papers are below the average impact of all papers; the other three classes are above this average and further differentiate the papers in the high impact sectors.

We undertook three pairwise comparisons to investigate the differences between the three NCS variants. Each pair is compared in a 4 x 4 contingency table. The cells in the diagonal of the table reveal the papers, which have been assigned to a CSS class in agreement of both NCS, and the share of papers assigned in agreement can be calculated. We further calculated the Kappa coefficient in this study, which is a robust alternative to the share of agreement, since the possibility of agreement occurring by chance is taken into consideration (Gwet, 2014). A further advantage of using the Kappa coefficient is that guidelines by Landis and Koch (1977) are available for the proper interpretation of the level of agreement: <0.00 "poor", 0.00-0.20 "slight", 0.21-0.40 "fair", 0.41-0.60 "moderate", 0.61-0.80 "substantial", and 0.81-1.00 "almost perfect".

We additionally calculated concordance coefficients for continuous variables following Lin (1989, 2000) to measure the agreement between NCS_{JS}, NCS_{CA}, and NCS_{CR}. We abstained from calculating correlation coefficients in this study, because we are interested in the agreement between two NCS (Lowenstein, Koziol-McLain, & Badgett, 1993). Correlation is a poor substitute for agreement. For example, systematic bias might be ignored. Suppose

NCS_{CA}, and NCS_{JS} have a perfect correlation, but the NCS_{CA} consistently measures citation impact 0.5 levels lower than the NCS_{JS}.

Data sets used

Database for calculating NCS_{JS}: The WoS journal sets are available in our in-house database developed and maintained by the Max Planck Digital Library (MPDL, Munich) and derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI) provided by Clarivate Analytics (Philadelphia, Pennsylvania, USA). We calculated the NCS_{JS} values by using the journal sets and the citation counts from the WoS in-house database. The journal set classification of the WoS, however, assigns multiple fields to many publications without any priority. Therefore, we calculated for every paper an average of the NCS_{JS} values in each field to receive an overall score.

Database for calculating NCS_{CA}: The CAplus^{SM} database accessible to us contains 8,219,858 journal articles published between 2000 and 2014. CAS uses a three-level field classification scheme to assign the publications into five broad headings of chemical research (section headings), which are further separated into 80 subject areas named as Chemical Abstracts sections. Most publications are assigned to only one section based on the main subject field; some publications are also assigned to a secondary section. To avoid multiple classifications of publications in this study, only the primary section assignment is used following previous studies (Bornmann & Daniel, 2008; Bornmann, et al., 2011). Although the section assignments are intellectually made by CAS, the classification does not seem to be affected by the "indexer effect": according to Braam and Bruil (1992), the indexer classification accords with author preferences for 80% of the publications. We calculated the $NCS_{\mathbf{CA}}$ values using the CA sections and the citation counts from the CAplus database.

Database for calculating NCS_{CR}: The algorithmically constructed classifications by Waltman and van Eck (2012) have been made freely available. The field classifications are uniquely assigned to papers: each paper is assigned to only one field. We downloaded the classifications of the papers and the corresponding WoS UTs on November 7th, 2014 from http://www.ludowaltman.nl/classification_system. The CR classification is available on three different levels. We used the third level. The classifications were matched via the WoS UT to the data in our in-house database. The NCS_{CR} values have been calculated by using these classifications and the citation counts from the WoS in-house database.

Only journal articles were included in the calculations of NCS_{JS}, NCS_{CA}, and NCS_{CR}. The NCS_{CA} values for each paper where matched with the NCS_{JS} and NCS_{CR} values via the DOI. Only matched publications with DOI (n=2,690,143) have been used in the statistical analysis.

Results

The 4x4 contingency table for the comparison of NCS_{CA} and NCS_{JS} is shown in Table 1. The level of agreement between NCS_{CA} and NCS_{JS} is 82.2%. Lin's concordance coefficient amounts to 0.67 [0.671, 0.672]; the Kappa coefficient is 0.61±0.001. According to the guidelines by Landis and Koch (1977) the Kappa coefficient indicates a substantial agreement between the two sets of normalized citation scores.

Table 1. 4x4 contingency table for NCS_{CA} and NCS_{JS}

		NCS_{JS}			
		poorly cited	fairly cited	remarkably cited	outstandingly cited
NCS_{CA}	poorly cited	1,722,880	167,075	21,135	5,350
	fairly cited	150,571	362,507	45,110	3,894
	remarkably cited	3,018	51,423	85,690	14,679
	outstandingly cited	245	2,752	14,755	39,059

Table 2 and Table 3 present the 4x4 contingency tables for NCS_{CA} and NCS_{CR} (see Table 2) and NCS_{JS} and NCS_{CR} (see Table 3).

Table 2. 4x4 contingency table for NCS_{CA} and NCS_{CR}

		NCS_{CR}			
		poorly cited	fairly cited	remarkably cited	outstandingly cited
NCS_{CA}	poorly cited	1,576,980	282,501	45,383	11,576
	fairly cited	179,628	277,771	86,317	18,366
	remarkably cited	12,542	55,827	56,987	29,454
	outstandingly cited	1,573	7,585	15,101	32,552

Table 3. 4x4 contingency table for NCS_{JS} and NCS_{CR}

		NCS_{CR}			
		poorly cited	fairly cited	remarkably cited	outstandingly cited
NCS_{JS}	poorly cited	1,600,214	251,196	22,957	2,347
	fairly cited	158,190	307,911	98,972	18,684
	remarkably cited	10,942	56,936	65,190	33,622
	outstandingly cited	1,377	7,641	16,669	37,295

The level of agreement between NCS_{CA} and NCS_{CR} is 72.3% and between NCS_{JS} and NCS_{CR} 74.7%. Lin's concordance coefficients are 0.50 [0.502, 0.503] and 0.43 [0.428, 0.430]. The corresponding Kappa coefficients amount to 0.42±0.001 and 0.48±0.001. According to the guidelines by Landis and Koch (1977) the Kappa coefficients indicate a moderate agreement between the two sets of normalized citation scores.

All three statistics – the level of agreement, the Kappa coefficients, and Lin's concordance coefficients – order the normalized citation score pairs the same way: NCS_{CA} and NCS_{JS} > NCS_{JS} and NCS_{CR} > NCS_{CA} and NCS_{CR} (in the order of decreasing similarity). The results further reveal that NCS_{CA} and NCS_{JS} are more in agreement than NCS_{JS} and NCS_{CR} and NCS_{CA} and NCS_{CR}.

Discussion

According to Ioannidis, Boyack, and Wouters (2016) "the basic premise of normalization is that not all citations are equal. Therefore, normalization can be seen as a process of benchmarking". Although it is standard in bibliometrics to use field-normalized citation scores for cross-field comparisons (of universities, for example), different approaches exist of calculating these scores. The differences refer either to the method of calculating the scores (percentiles have been proposed as an alternative to scores based on average citations, Bornmann & Marx, 2015) or to the approach of field categorization which are used to build the reference set for each paper. In this study, we addressed the second aspect by comparing the normalized scores, which have been calculated based on three different approaches.

The analysis of the scores basically reveals an agreement which is at least at the moderate level. Since we used the same method for calculating the scores based on the different approaches, the moderate level is lower than that level which we expected. The parallel use of the different approaches in the current research evaluation practice should have led to a generally higher level of agreement. However, our results also show that normalized scores based on intellectual field assignments are more in agreement with scores based on journal sets than with scores based on citation relations. Thus, one can expect more similar scores based on intellectual assignments and journal sets than on citation relations. The reason for the similarity might be that intellectual assignments and journals are better rooted in the disciplines than virtual constructs based on citation relations. CA sections, which are used by CAS indexers, have been developed by specialists in the discipline. According to Sugimoto and Weingart (2015), the establishment of new journals is a sign of emerging new disciplines.

The results of this study should be interpreted against the backdrop that the study focusses on one discipline only: chemistry and related areas. Furthermore, other statistical analyses could be performed. It is not clear whether our results can be generalized. Thus, we encourage similar studies with data from other disciplines using different statistical methods and as many classification schemes as possible.

References

Bornmann, L., & Daniel, H.-D. (2008). Selecting manuscripts for a high impact journal through peer review: a citation analysis of Communications that were accepted by *Angewandte Chemie International Edition*, or rejected but published elsewhere. *Journal of the American Society for Information Science and Technology, 59*(11), 1841-1852. doi: 10.1002/asi.20901.

Bornmann, L., & Marx, W. (2015). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts? *Journal of Informetrics, 9*(2), 408-418.

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H.-D. (2008). Use of citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics, 8*, 93-102. doi: 10.3354/esep00084.

Bornmann, L., Schier, H., Marx, W., & Daniel, H.-D. (2011). Is interactive open access publishing able to identify high-impact submissions? A study on the predictive validity of *Atmospheric Chemistry and Physics* by using percentile rank classes. *Journal of the American Society for Information Science and Technology, 62*(1), 61-71.

Braam, R. R., & Bruil, J. (1992). Quality of indexing information: authors views on indexing of their articles in Chemical Abstracts Online CA-File. *Journal of Information Science, 18*(5), 399-408. doi: Doi 10.1177/016555159201800508.

Evidence. (2009). Report on the citation database for the Human Frontier Science Program. Retrieved from http://www.hfsp.org/sites/www.hfsp.org/files/webfm/Executive/HFSP%20Bibliometrics%202010.pdf

Glänzel, W., Debackere, K., & Thijs, B. (2016). Citation classes: a novel indicator base to classify scientific output. Retrieved October, 21, 2016, from https://www.oecd.org/sti/051%20-%20Blue%20Sky%20Biblio%20Submitted.pdf

Gwet, K. L. (2014). *Handbook of Inter-Rater Reliability, 4th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters*: Advanced Analytics, LLC.

Haddow, G., & Noyons, E. (2013). *Misfits? research classification in research evaluation: Visualizing journal content within fields of research codes.* Paper presented at the Proceedings of ISSI 2013 - 14th International Society of Scientometrics and Informetrics Conference.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429-431.

Ioannidis, J. P. A., Boyack, K., & Wouters, P. F. (2016). Citation Metrics: A Primer on How (Not) to Normalize. *PLoS Biol, 14*(9), e1002542. doi: 10.1371/journal.pbio.1002542.

Klavans, R., & Boyack, K. W. (2017). Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge? *Journal of the Association for Information Science and Technology, 68*(4), 984-998. doi: 10.1002/asi.23734.

Landis, J. R., & Koch, G. G. (1977). Measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Leydesdorff, L., & Milojević, S. (2015). The Citation Impact of German Sociology Journals: Some Problems with the Use of Scientometric Indicators in Journal and Research Evaluations. *Soziale Welt, 66*(2), 193-204.

Leydesdorff, L., & Opthof, T. (2013). Citation analysis with medical subject Headings (MeSH) using the Web of Knowledge: A new routine. *Journal of the American Society for Information Science and Technology, 64*(5), 1076-1080. doi: 10.1002/asi.22770.

Lin, L. I. (1989). A CONCORDANCE CORRELATION-COEFFICIENT TO EVALUATE REPRODUCIBILITY. *Biometrics, 45*(1), 255-268. doi: 10.2307/2532051.

Lin, L. I. (2000). A Note on the Concordance Correlation Coefficient. *Biometrics, 56*(1), 324-325. doi: 10.1111/j.0006-341X.2000.00324.x.

Lowenstein, S. R., Koziol-McLain, J., & Badgett, R. G. (1993). Concordance versus correlation. *Annals of Emergency Medicine, 22*(2), 269. doi: 10.1016/S0196-0644(05)80225-2.

Lundberg, J. (2007). Lifting the crown - citation *z*-score. *Journal of Informetrics, 1*(2), 145-154.

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2016). A comparison of the Web of Science with publication-level classification systems of science. In I. Ràfols, J. Molas-Gallart, E. Castro-Martínez & R. Woolley (Eds.), *Proceedings of the 21 ST International Conference on Science and Technology Indicator*. València, Spain: Universitat Politècnica de València.

Radicchi, F., & Castellano, C. (2011). Rescaling citations of publications in physics. *Physical Review E, 83*(4). doi: 10.1103/PhysRevE.83.046116.

Rehn, C., Kronman, U., & Wadskog, D. (2007). *Bibliometric indicators – definitions and usage at Karolinska Institutet*. Stickholm, Sweden: Karolinska Institutet University Library.

Strotmann, A., & Zhao, D. (2010). Combining commercial citation indexes and open-access bibliographic databases to delimit highly interdisciplinary research fields for citation analysis. *Journal of Informetrics, 4*(2), 194-200. doi: 10.1016/j.joi.2009.12.001.

Sugimoto, C. R., & Weingart, S. (2015). The kaleidoscope of disciplinarity. *Journal of Documentation, 71*(4), 775-794. doi: 10.1108/jd-06-2014-0082.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365-391.

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*(12), 2378-2392. doi: 10.1002/asi.22748.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: some theoretical considerations. *Journal of Informetrics, 5*(1), 37-47. doi: 10.1016/j.joi.2010.08.001.

Wang, Q., & Waltman, L. (2016). Large-Scale Analysis of the Accuracy of the Journal Classification Systems of Web of Science and Scopus. *Journal of Informetrics, 10*(2), 347-364.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., . . . Johnson, B. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. Bristol, UK: Higher Education Funding Council for England (HEFCE).

Wouters, P., Thelwall, M., Kousha, K., Waltman, L., de Rijcke, S., Rushforth, A., & Franssen, T. (2015). *The Metric Tide: Literature Review (Supplementary Report I to the Independent Review of the Role of Metrics in Research Assessment and Management)*. London, UK: Higher Education Funding Council for England (HEFCE).