



SAPIENZA
UNIVERSITÀ DI ROMA

2019

**17th INTERNATIONAL CONFERENCE ON
SCIENTOMETRICS & INFORMETRICS**

ISSI2019

with a Special STI Indicators Conference Track

2-5 September 2019

Sapienza University of Rome, Italy

PROCEEDINGS

VOLUME II

PROCEEDINGS OF THE 17TH CONFERENCE OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS

- © Authors. No part of this book may be reproduced in any form without the written permission of the authors.
- © International Society for Scientometrics and Informetrics
- © Edizioni Efesto - ISBN 978-88-3381-118-5 - August 2019
Printed in Italy

Editors: *Giuseppe Catalano, Cinzia Daraio, Martina Gregori,
Henk F. Moed and Giancarlo Ruocco*

Graphic cover design: *Francesco Manzo* | graframan.com

Cover photo: ©*Fayee* - stock.adobe.com

Measurement variation in bibliometric impact indicators

Stephan Stahlschmidt¹ and Marion Schmidt²

¹*stahlschmidt@dzhw.eu*

Department 2 – Research System and Science Dynamics, German Centre for Higher Education Research and Science Studies (DZHW), Schützenstr. 6a, Berlin, 10117 (Germany)

²*schmidt@dzhw.eu*

Department 2 – Research System and Science Dynamics, German Centre for Higher Education Research and Science Studies (DZHW), Schützenstr. 6a, Berlin, 10117 (Germany)

Abstract

The bibliometric measurement process transfers scientific publications and citations into indicators on scientific impact. In defining specific measurement paths researchers hold several degrees of freedom as various methodical decisions are scarcely founded on stringent criteria and their respective implications are not fully understood. These diverse measurement paths result in varying measurements. We propose to compute many possible measurement paths and to analyse the resulting measurement variation. On the one hand, effects of decisions can thus be better understood and on the other hand the resulting measurement variation should be taken into account when using impact values for e.g. funding decisions.

Introduction

Bibliometric indicators result from a measurement process which converts scientific publications and the included references into aggregate values. While the particular selection of citations arises from complex and partially unobserved mechanisms and consequently could be modelled via a stochastic approach, the data generating process of the deterministic bibliometric measurement is perfectly known. The measuring researcher decides upon the process and defines a specific measuring approach, e.g. the choice between Web of Science and Scopus. Figure 1 presents three exemplary measurement decisions, which lead via different measurement paths to eight varying values on the same bibliometric impact indicator.

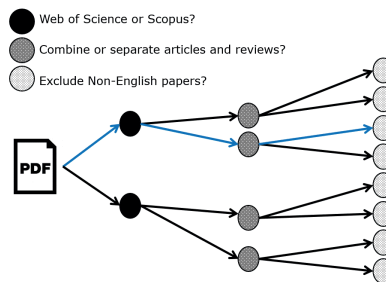


Figure 1: Measurement paths in an exemplified garden of forking paths for a bibliometric analysis.

In defining a particular measurement approach the researcher holds several degrees of freedom, as the implications of each single measurement decision are not fully understood and furthermore the content to be measured, e.g. scientific productivity or impact, consists of latent constructs. The exact extent to which scientific impact is comprehensively covered via citations is disputed (MacRoberts and MacRoberts, 2018), while at the same time citations to an unknown degree occur due to aspects unrelated to scientific impact (Latour & Woolgar, 1986). Citations and scientific impact overlap, but they are not congruent - as any missing Mertonian citation and any existing non-Mertonian citation decrease the signal-to-noise ratio of a citation

based measure of scientific impact. Due to the lacking justification of measurement decisions and the lacking clearness about the content to be measured, no – in a statistical sense – true or optimal measurement path might be identified. Instead, several diverging measurement paths through the so-called garden of forking paths (Gelman and Loken, 2014) co-exist. However, most measuring authorities opt for a single measurement path (e.g. CWTS, 2018). In contrast, we propose to embrace this variety and compute several measurement paths. The resulting fuzziness might be understood as a measurement variation. Due to its known and deterministic origin it might not be qualified as a stochastically induced variance, but highlights uncertainty of the measured values caused by the measurement process.

A conceptual model of variation in bibliometrics

The analysis of variation in bibliometrics is predominately discussed in terms of a stochastic variance. In a recent contribution Williams and Bornmann (2016) assume randomness in bibliometric citation counts and propose frequentist statistical inference techniques to quantify its magnitude. For example Bornmann (2017) proposes the estimation of parametric confidence intervals for Journal Impact Factors. Thelwall and Fairclough (2017) extend this approach by proposing a partially randomly determined capacity to produce impactful research observed via publications. Apart from these parametric approaches also non-parametric techniques like bootstrapping (Waltman et al., 2012) and the jackknife (Sağlam and Friggens, 2018) have been applied. However, these modelling approaches mostly do not account for the data generating process, but rely solely on the observed cross-sectional variation arising in a single measurement path. Furthermore Schneider (2016) argues that frequentist statistical inference on populations is inappropriate. Recently the debate on p-hacking and the reproducible crisis has inspired research to focus on modelling decisions (e.g. Gelman and Loken, 2014; Rohrer et al., 2018). In bibliometrics, the synchronous presentation of different measurement approaches is usually restricted to a limited subset of single parameters, like database coverage (Archambault et al., 2009, Struck et al, 2018), self-citations (Mittermaier et al., 2016) or fractional and whole counting (Waltman et al., 2012).

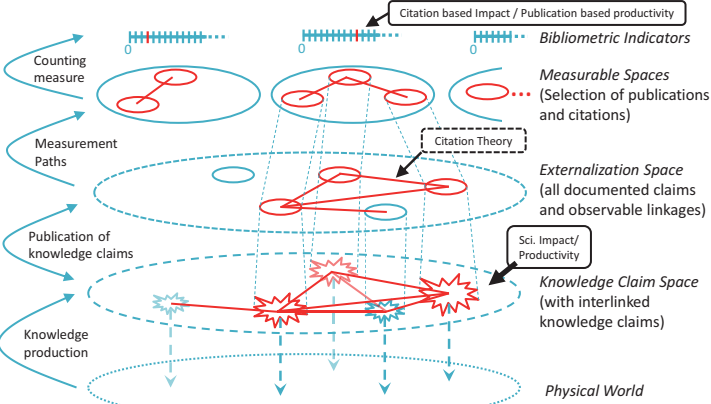


Figure 2: Conceptual model of measurement variation

Figure 2 presents a model detailing how such different measurement paths are embedded in the wider mechanism of the science system. We assume several distinct layers which are inter- and intra-related. The bottom layer of the physical world is related to the space of knowledge claims understood as the (debated and extending) knowledge on the physical world. The next layer of

the externalization space depicts observable artefacts i.e. documented claims as publications and explicit (e.g. citations) and implicit (e.g. author keywords) links between them. The topmost layer of the measurable spaces is central to our analysis. The transfer of artefacts from the externalization space to diverging measurable spaces is thought to be governed by diverse measurement paths. As the externalization space is too large to be observed in its entirety, the measurement path extracts a subset of artefacts and consequently renders them accessible to the bibliometric measurement. However, different measurement paths equally co-exist and each path defines a separate although potentially overlapping measurable space. Any such measurable space might be equipped with a counting measure and consequently bibliometric statistics might be calculated. Hence any variation in the measurable spaces consequently causes variation in the values of bibliometric indicators. Given the unknown ground truth of the scientific impact each of these single incarnations of the same indicator maps the respective scientific impact of an entity to a different value constructing en passant separate realities (Desrosières, 1998), upon which funding and policy decision are made.

A quantitative description of measurement variation

For this research-in-progress paper we focus our analysis on eight binary measurement decisions, which result in 256 different measurement paths and values of the same indicator for the same analysed entity. As publications and citations have to be counted separately for every path and are to be supplemented with field-specific expectancy values, we limit our analysis to a computationally feasible random sample of 25% of all potential measurement paths, i.e. we compute 64 parallel bibliometric worlds defined by the respective randomly drawn measurement paths. The following measurement decisions were taken into account: (1) Web of Science or Scopus, (2) include or exclude Non-English publications, (3) combine or separate reviews and articles in normalization, (4) include or exclude self-citations, (5) include or exclude Social Sciences and Humanities (via OECD Fields of Sciences), (6) apply fractional or whole counting to multi-author papers, (7) three-year or five-year citation window and (8) discipline classification by database provider or OECD Disciplines of Sciences.

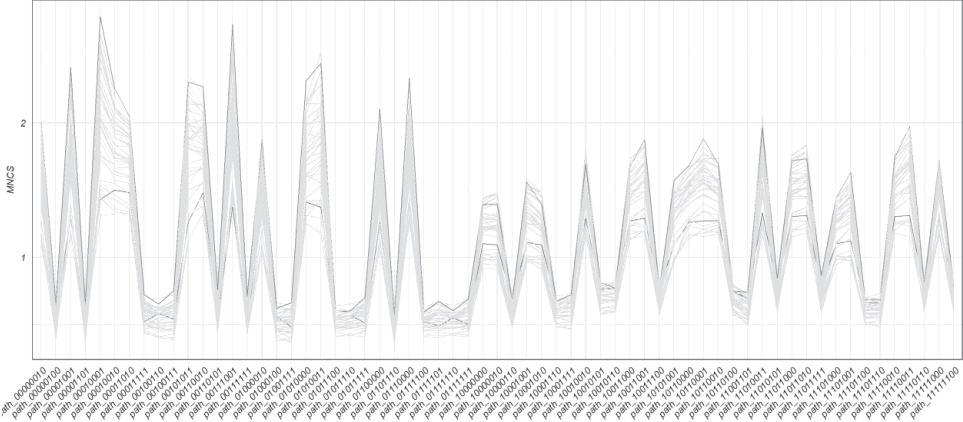


Figure 3: MNCS for 37 German universities across 64 measurement paths in 2012.

For each of these constructed bibliometric realities we compute the *Mean Normalised Citation Score (MNCS)* for every German university. Figure 3 illustrates a preliminary description of the resulting measurement variation. Every line indicates the respective MNCS values (y-axis) across the measurement paths (x-axis) for one of the 37 universities. For example, the

Universität Heidelberg observes a more pronounced variation with values between 0.6 and 2.75, while the equivalent interval for the Freie Universität Berlin starts at 0.5 and ends at 1.5. Given that the value of 1 is commonly understood as the world average, the Universität Heidelberg (Freie Universität Berlin) might find its citation impact in the interval of being 175% (50%) better or 40% (50%) worse than the world average. Consequently the information value on the actual bibliometric impact state seems dubious. Obviously this substantial variation might not be observed if the analysis is limited to a single measurement path. Apart from this institutional perspective on the variation the relative position of these two universities among all other 35 German universities does not vary to a large degree.

In order to gauge the stability across the measurement paths Figure 4 presents the rank-based Spearman correlation matrix between the 37 German universities' MNCS in each of the 64 parallel bibliometric worlds, i.e. every small square colour codes the Spearman correlation in the ranking of the same universities across two measurement paths. Dark blue symbolizes a strong accordance in the ranking of the respective measurement paths, while white denotes no accordance. Negative correlation does not arise.

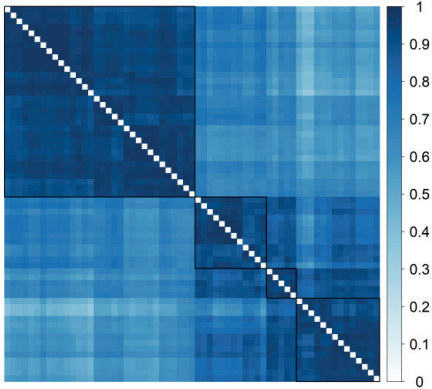


Figure 4: Spearman correlation matrix across the MNCS of 37 German universities resulting from 64 measurement paths.

The correlation matrix has been ordered according to a hierarchical Ward clustering to allow an inspection of the structure causing the aforementioned stability. Four dominant cluster can be identified, the first and biggest one in the upper left corner, the next one in the lower right corner, a third one in the middle and a comparable small one just below on the diagonal, all marked by black lines. The biggest cluster at the upper left corner is constituted by all Spearman correlations between measurement paths based on whole counting, which seem to render greater stability than fractional counting in this preliminary state. Consequently the square to the right (and bottom) of this cluster denotes the rather inconsistent Spearman correlation between whole and fractional counting. On the lower right corner, we find all measurement paths applying fractional counting for the Web of Science, while the equivalent paths for Scopus are subdivided into two clusters. Thus Web of Science seems to be a more stable and probably coherent base less affected by single measurement decisions. The cluster in the middle incorporates measurement paths based on Scopus and fractional counting and the OECD classification, while the smallest cluster holds measurement paths based on Scopus, fractional counting, the Scopus ASJC discipline classification, only English publications and including the social sciences and humanities (SSH). Hence German universities with their rather large corpus of non-English publications in the SSH seem to be uniformly affected by the enlarged

database of Scopus in the SSH and Scopus’s ASJC discipline classification. These structural observations will be modelled in the next section to infer the direction and size of these effects.

Modelling measurement variation

We model intrinsic values independent of the measurement process by employing a linear mixed model

$$Y_{ij} = \alpha_i + x_{ij}^t \beta + u_{ij}^i \gamma_i + \epsilon_{ij}$$

where

- Y_{ij} indicates the MNCS of university i corresponding to measurement path j
- $i \in [1, \dots, m]$ denotes the $m = 37$ German universities
- $j \in [1, \dots, n_i]$ states the balanced size of $n_i = 64$ observations per university arising from the diverse measurement paths
- α_i denotes the university specific (random) intercept
- β describes the vector of fixed effects of the eight binary measurement decisions and
- γ_i details the random effects of these measurement decisions on the university i .

Hence the effect of the measurement decision on the universities’ MNCS is composed of an overall effect β and an individual effect γ_i allowing for a large degree of flexibility. As we analyse the same 37 German universities in the 64 different measurement paths, we assume the respective MNCS values of the same university to be related throughout all measurement paths and hence obtain a cluster of related MNCS values for every university. The university specific intercept denotes the constant, unaffected part across the observed MNCS values of all measurement paths and accordingly might be understood as the citation-based latent scientific impact irrespective of the variation caused by the measurable spaces of Figure 2.

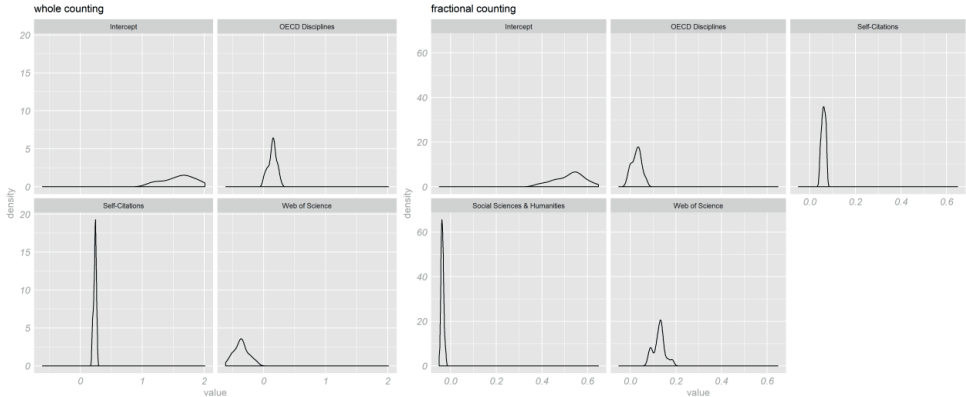


Figure 5: Distribution over German universities of the composed effect of the measurement decisions on the respective MNCS by fractional and whole counting.

The direction and size of the diverse measurement decisions are depicted in Figure 5. Due to fitting issues in this research-in-progress paper we still not present an overall model, but separate models for whole and fractional counting. However, the effect of whole and fractional counting might be inferred from a comparison of the intercepts. Fractional counting reduces the MNCS to a large degree, as observable by the different scales in the x-axis. The application of the broader OECD discipline classification improves the MNCS values slightly, but to a varying degree. Some universities gain more than others from applying the OECD disciplines, possibly due to different disciplinary profiles. The positive effect of including self-citations does not vary to a large degree across universities. However its positive effect allows to infer that

German authors cite themselves more often than the average global author. The inclusion of the SSH reduces the MNCS slightly for fractional counting, while we observe no such effect for whole counting. We also observe no overall effect of non-English papers, separating reviews from articles and the size of the citation window. Although contradictory results for these negligible effects might have been observed in particular measurement paths, they do not hold any overall effect in the multivariate regression framework. The application of Web of Science instead of Scopus has contradictory results, whose reason is still to be investigated.

Outlook

We are currently extending the current analysis by (1) refining the measurement decisions, (2) including further measurement decisions, (3) computing the PP(top10) indicator and (4) drafting a Bayesian model to show how measurement paths in bibliometrics carry considerable consequences for the analysed entities and how “model-based” descriptive statistics might help to alleviate these.

References

- Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the Association for Information Science and Technology*, 60(7), 1320-1326
- Bornmann, L. (2017). Confidence intervals for Journal Impact Factors. *Scientometrics*, 111, 1869-1871
- CWTS (2018). CWTS Leiden Ranking 2018: Methodology, Leiden: Universiteit Leiden
- Desrosières, A. (1998). *The politics of large numbers: A history of statistical reasoning*. Cambridge: Harvard University Press.
- Gelman, A. and Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102, 460-465.
- Latour, B., & Woolgar, S. (1986). *Laboratory Life: The Construction of Scientific Facts* (2nd Edition with a New Postscript). London: Sage Publications.
- MacRoberts, M.H., MacRoberts, B.R. (2018). The mismeasure of science: Citation analysis. *Journal of the Association for Information Science and Technology*, 69, 474-482
- Mittermaier, B., Tunger, D., Meier, A., Glänzel, W., Thijs, B. & Chi, P.-S. (2016). Erfassung und Analyse bibliometrischer Indikatoren für den PFI-Monitoringbericht 2017; <http://hdl.handle.net/2128/15276>
- Rohrer, J.M., Egloff, B. & Schumke, S.C. (2018). Run all the analyses. 51. Kongress der Deutschen Gesellschaft für Psychologie, 15.-20. Sept 2018, Frankfurt.
- Sağlam S.Y and Friggens, D. (2018). Cut Your Bootstraps: Use a Jackknife. In: STI 2018 Conference Proceedings (eds. Wouters, P., Costas, R., Franssen, T. and Yegros-Yegros, A.). Leiden: Centre for Science and Technology Studies (CWTS).
- Schneider, J. (2016). The imaginary of statistical inference when data are the population: Comments to Williams and Bornmann, *Journal of Informetrics*, 10, 1243-1248
- Struck B., Durning M., Roberge G., & Campbell D. (2018). Modelling the effects of open access, gender and collaboration on citation outcomes: Replicating, expanding and drilling. In: STI 2018 Conference Proceedings (eds. Wouters, P., Costas, R., Franssen, T. and Yegros-Yegros, A.). Leiden: Centre for Science and Technology Studies (CWTS).
- Thelwall, M. and Fairclough, R. (2017). The accuracy of confidence intervals for field normalised indicators. *Journal of Informetrics*, 11, 530-540
- Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C., Tijssen, R. J., Eck, N. J., Leeuwen, T. N., Raan, A. F., Visser, M. S. and Wouters, P. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the Association for Information Science and Technology*, 63: 2419-2432.

Williams, R. and Bornmann, L. (2016). Sampling issues in bibliometric analysis. *Journal of Infometrics*, 10(4), 1225-1232.

Printed in Italy, August 2019