Bootstrapping to evaluate accuracy of citation-based journal indicators

Jens Peter Andersen[1] and Stefanie Haustein[2]

[1] *jepea@rn.dk*
Medical Library, Aalborg University Hospital, Sdr. Skovvej 15, 9000 Aalborg (Denmark)

[2] *stefanie.haustein@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal (Canada)

Introduction

Bibliometric indicators ranking aggregate units have a long tradition, including criticisms of methodology, interpretation and application. Despite the criticism, there is a demand for these indicators, and recent developments have led to improvements of methodology and interpretation. An essential element of these interpretations is to provide estimates of the accuracy, robustness, stability and confidence of bibliometric indicators, thereby providing the reader with data required to interpret results. This has, for example, been demonstrated for the set of indicators in the Leiden ranking (Waltman et al., 2012), the Journal Impact Factor (Chen, Jen, & Wu, 2014) and other journal indicators (Andersen, Christensen, & Schneider, 2012) as well as author metrics (Lehmann, Jackson, & Lautrup, 2008). The present study applies the same type of bootstrapping technique to estimate stability, as is used in the Leiden ranking (Waltman et al., 2012), on an array of citation-based journal indicators. The purpose of this analysis is to compare recent methodological advances, as well as traditional approaches. The study is based on clinical medicine journals in the Web of Science (WoS).

Methods

Data acquisition

The dataset contains all articles and reviews in the WoS, published in 2012 in journals classified as clinical medicine according to the National Science Foundation (NSF) classification system. This amounts to 362,556 papers and 2,699 journals from 34 different specialties within the discipline of clinical medicine. Each journal and paper is assigned to exactly one specialty. Citations are observed for a two-year window. In order to account for field differences in citation patterns, relative citations, $\hat{c}$, are computed by normalising observed against expected citations per specialty and year.

Journal indicators

The journal citation indicators selected for this study represent both traditional (means and medians of observed and relative) and novel (percentile) approaches. For a given journal *j*, we calculate the mean citations, μ_c, median citations, M_c, mean relative citations, $\mu_{\hat{c}}$, median relative citations, $M_{\hat{c}}$, top decile ratio of citations, N_{D10}, and relative citations. The top decile ratio for a journal is the percentage of papers present in the overall set of papers with citations in the highest decile range.

Indicator evaluation

Each indicator is evaluated for every journal by performing bootstrapping (Efron & Tibshirani, 1993). The technique involves resampling with replacement, i.e. for a given sample, all observed values are resampled so that a new sample of the same size is drawn randomly, but with the possibility that the same observation can be drawn multiple times. When repeating this resampling numerous times, we can calculate stability intervals to estimate how accurately the observed indicator value describes the underlying observations or whether it is influenced by outliers and thus less robust. To make our results comparable to those reported in the Leiden ranking, we have chosen to iterate each bootstrap 1,000 times and calculate 95% confidence intervals. In addition to this confidence interval we also calculate the standard deviation for each distribution. As the values of the different indicators are observed in very different ranges, we provide an additional mean-standardised version of every indicator. All calculations are performed using the *boot* package (Canty & Ripley, 2015) for *R* version 3.0.3 x64 (R Development Core Team, 2010).

Results and Discussion

We find that bootstrapping can identify outlying indicator scores within a specialty, by showing stability intervals (95% confidence intervals) for every indicator. As exemplified in Figure 1 for the subset of dentistry journals, the stability intervals demonstrate the robustness of rankings based on particular indicators. While, for example, the stability intervals indicate that the citation impact of the 1st journal in Figure 1 is higher than that of the 5th, the first four journals cannot be clearly distinguished in terms of mean citation impact.

Their mean citation rates are heavily influenced by a few highly cited papers.

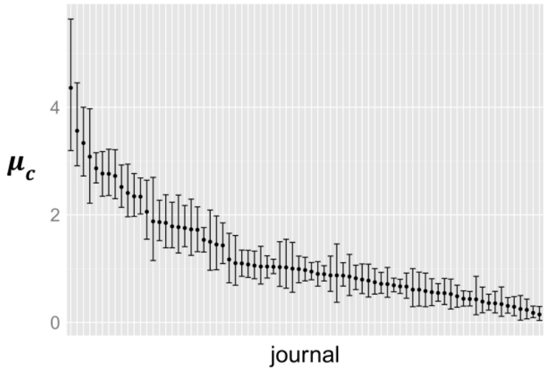

Figure 1. μ_c with stability intervals for all journals in the dentistry specialty.

The study also shows that the percentile-based indicators perform considerably better regarding stability than both mean- and median-based indicators (Figure 2 and Table 1). It is particularly interesting that the medians indicators do not seem to be more stable than the means.

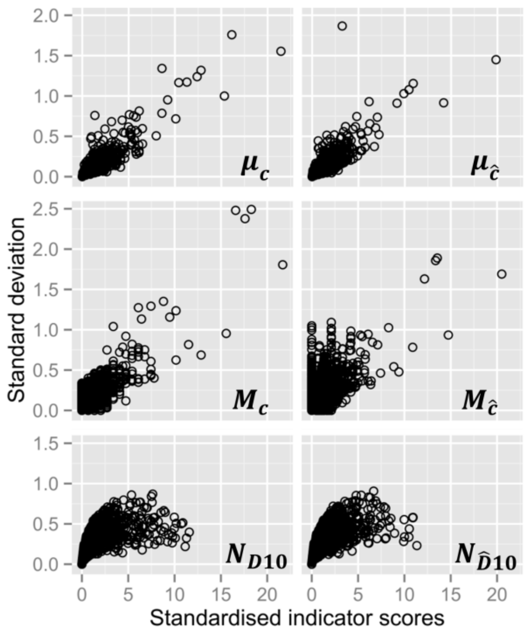

Figure 2. Standard deviation of bootstrapped scores as a function of standardised indicator scores, limited to journals with at least 50 papers.

Finally, we show that indicators are extremely sensitive to sample sizes. Journals with less than 50 papers published in the observation period show significantly larger variance than those publishing at least 50 papers (Table 1). Our results reiterate the importance of testing indicators and providing stability intervals to improve their interpretability. This would identify the limitations of rankings and avoid cases like the 24-fold increase of *Acta Crystallographica A*'s impact factor in 2009 (Haustein, 2012).

Table 1. Mean indicator values and standard deviations for all journals ("All") and journals publishing 50 or more papers ("≥50").

	All				**≥50**	
	Raw		**Standardised**			
Indi-cator	*mean*	*SD*	*mean*	*SD*	*mean*	*SD*
μ_c	2.321	3.897	1.000	1.679	1.052	1.261
M_c	1.477	2.278	1.000	1.543	1.079	1.471
$\mu_{\hat{c}}$	0.835	1.107	1.000	1.326	1.053	1.076
$M_{\hat{c}}$	0.520	0.717	1.000	1.381	1.075	1.297
N_{D10}	0.081	0.131	1.000	1.625	1.107	1.640
$N_{\hat{D}10}$	0.078	0.119	1.000	1.536	1.090	1.513

Further research will include in-depth analyses of multiple indicators and differences of stability intervals across specialties.

References

Andersen, J. P., Christensen, A. L., & Schneider, J. W. (2012). An approach for empirical validation of citation-based journal indicators. In E. Archambault, Y. Gingras, & V. Lariviére (Eds.), *Proceedings of STI 2012* (pp. 71–81). Montréal, Canada: 17th International Conference on Science and Technology Indicators.

Canty, A., & Ripley, B. (2015). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-15.

Chen, K. M., Jen, T. H., & Wu, M. (2014). Estimating the accuracies of journal impact factor through bootstrap. *Journal of Informetrics*, *8*(1), 181–196.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the Bootstrap* (p. 456). New York: Chapman & Hall.

Haustein, S. (2012). *Multidimensional Journal Evaluation. Analyzing Scientific Periodicals beyond the Impact Factor*. Berlin / Boston: De Gruyter Saur.

Lehmann, S., Jackson, A. D., & Lautrup, B. E. (2008). A quantitative analysis of indicators of scientific performance. *Scientometrics*, *76*(2), 369–390.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., … Wouters, P. F. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*(12), 2419–2432.