STI 2018 Leiden

23rd International Conference on Science and Technology Indicators
"Science, Technology and Innovation Indicators in Transition"

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

Cut Your Bootstraps: Use a Jackknife

Ṣenay Yaṣar Saḡlam[*] and David Friggens[**]

[*]*senay.yasarsaglam @mbie.govt.nz;* [**]*david.friggens@mbie.govt.nz*
Ministry of Business, Innovation and Employment, Wellington, 6011 (New Zealand)

Introduction

In the last decade there has been increasing interest in using statistical inference or descriptive-based techniques to quantify error and uncertainty in bibliometrics (see e.g. Bornmann and Williams (2016) and the eight replies in the same volume). Colliander and Ahlgren (2011) proposed the concept of "stability intervals" (SI) for assessing the vulnerability of citation-based indicators to the underlying data, using repeated subsampling without replacement. A stability interval evaluates the degree of sensitivity of a parameter estimate to the underlying data. This approach has been adopted by other authors (e.g., Waltman et al, 2012; Schneider & van Leeuwen, 2014; Andersen et al., 2018) with the notable difference of sampling with replacement, known as "bootstrapping". The change in the sampling technique results in practical differences that have not previously been explored.

In this study we investigate to what extent the stability intervals differ across the random subsampling (without replacement) and bootstrap methods, alongside a third approach, which we propose as the delete 10% jackknife. This third approach is similar to the delete-n Jackknife or K-fold cross-validation method. After highlighting the differences among these three approaches, we will first analyse the issue of pre-estimation error in the bootstrap approach. Then, we will run simulations to evaluate the stability intervals by the three methods.

Finally, we will apply our approach to the empirical data. Our dataset is Scopus custom data, which was extracted in June 2017. For the empirical analysis, we restrict our attention to New Zealand and Australia, so the dataset contains articles, conference papers, and reviews for the data period from 2002 to 2015.

Our main finding is that the bootstrap method provides wider stability intervals throughout the simulations as well as the Scopus data, compared to the random subsampling and the delete-10% jackknife methods. Meanwhile, random subsampling performs slightly worse than delete-10% jackknife in the Scopus data, even though this difference is not significant in the simulations.

Methodology

Citation Indicators

In bibliometrics, stability and confidence intervals are increasingly used to obtain robust sample estimates. In this context, Thelwall (2017a, 2017b) investigated the accuracy of confidence intervals for field normalised indicators (e.g. MNCS) and different sample sizes. To construct accurate stability intervals and provide stronger evidences at different units of analysis, we utilise two citation indicators to assess the scientific impact of the analysed unit (e.g., country). The first indicator is the *mean normalised citation score (MNCS)*. Normalised citation score (NCS_{pub}) divides a publication's citations by the average of the citations for all publications of the same type from the same year in the same subfield. The MNCS of the unit is the average across NCS_{pub} of the publications of the unit.

The second measure is the *proportion of publications among the 10% most cited* (PPtop10%). To determine which publications are among the 10% most cited, publications of the same type from the same year in the same subfield are ranked according to the raw citation counts and the percentiles are then calculated. Where an exact 10% division is not possible, the multiple publications at the threshold will be fractionally counted towards the top 10%. Furthermore, if a publication is classified by multiple subfields, it is fractionally counted toward the top 10% according to the number of the subfields it ranks in the top 10%. Finally, PPtop10% of a unit (e.g., country) is the fraction of publications that are among the top 10% most cited publications of similar characteristics.

Sampling Strategies

In this section, we discuss the sampling strategies for assessing the dependency of the indicators to the underlying data of the unit of analysis: bootstrap (Waltman 2012, Schneider & van Leeuwen 2014, Andersen et al. 2018) and random sampling without replacement (Colliander and Ahlgren 2011) along with our proposed approach.

Bootstrap is the most commonly used sampling strategy proposed by Brandon Efron (1981). Each bootstrap sample is drawn from the original sample with replacement. The original bootstrap method requires estimation of parameters (e.g., mean) for each bootstrap sample to create a distribution for that parameter. With bibliometric data, there is a complex relationship among the observations, and bibliometrics databases (e.g., Scopus) are selective samples of the super population (e.g., decision for indexing a journal is not random). We believe this method is not suitable for generating stability intervals. The citation-based indicator (e.g., MNCS_{pub}) values are computed based on the data available in the bibliometric database (Scopus or Web of Science) which creates dependencies among publications. This is because multiple occurrences of the same observation may change the citation counts, as well as the author information, and in general the entire population. This violates one of the major assumptions of the bootstrap: the independence of observations. In the bibliometrics literature, the bootstrap method is modified from its original version and these values are not computed for the bootstrap samples. Instead, the initial values from the database are used. It is assumed that the distortion introduced by bootstrapping the unit level data (e.g., New Zealand publications) is small for the aggregated indicators and the pre-estimation error can be ignored. In Section *Experiments*, we perform Monte Carlo simulation to assess whether this error can be ignored.

Subsampling has two key differences to the bootstrap: (i) the resample size is smaller than the original sample size and (ii) resampling is done without replacement; see Horvitz and Thompson (1952), McMurdie and Holmes (2014), Sampford (1962), and Strobl et al. (2007) for details and variations on this method. Sam One of the advantages of this method is that the data structure and all the dependencies are still maintained at the publication level since it assumes that the publications that are not in the subsample are still part of the bibliometric database. Unlike bootstrapping, there is no artificial introducing/removing of publications from the database. Hence the calculated indicator values (e.g., MNCS_{pub}) are still valid since there are no changes to the database. However, this is still not a robust method to fully assess the effect of each publication on the average-based indicator as the method does not guarantee the removal of influential observations.

Delete10%-Jackknife is similar to the delete-n Jackknife or K-fold cross-validation method, splitting the data into 10 subsets, each with 10% of the original sample size. In each case, one of the 10 parts is held out and we calculate the indicator values for the unit (e.g., country, institution, researcher) using the remaining 9 parts. This method makes the same assumptions as subsampling (i.e. the observations that are left out are still part of the population but just not part of the unit (e.g., New Zealand publication set)). It also addresses the issue around not drawing some of the influential publications during sampling, since the procedure guarantees that every observation is removed once. In this way, we can assess how each publication affects the results. It is a what-if scenario: what-if this publication was not part of the researchers' publication set. It is worth noting that there are several combinations of observations and it will be computationally expensive to assess them all. We believe that if this procedure is repeated many times, it will provide better sensitivity analysis than *Subsampling*. Additionally, in this paper, we demonstrate our results assuming 10% of the sample is held out systematically. However, we can alternatively set this proportion to 5% or 1, and these results are not shown here for brevity of the paper (available from authors on request).

Experiments

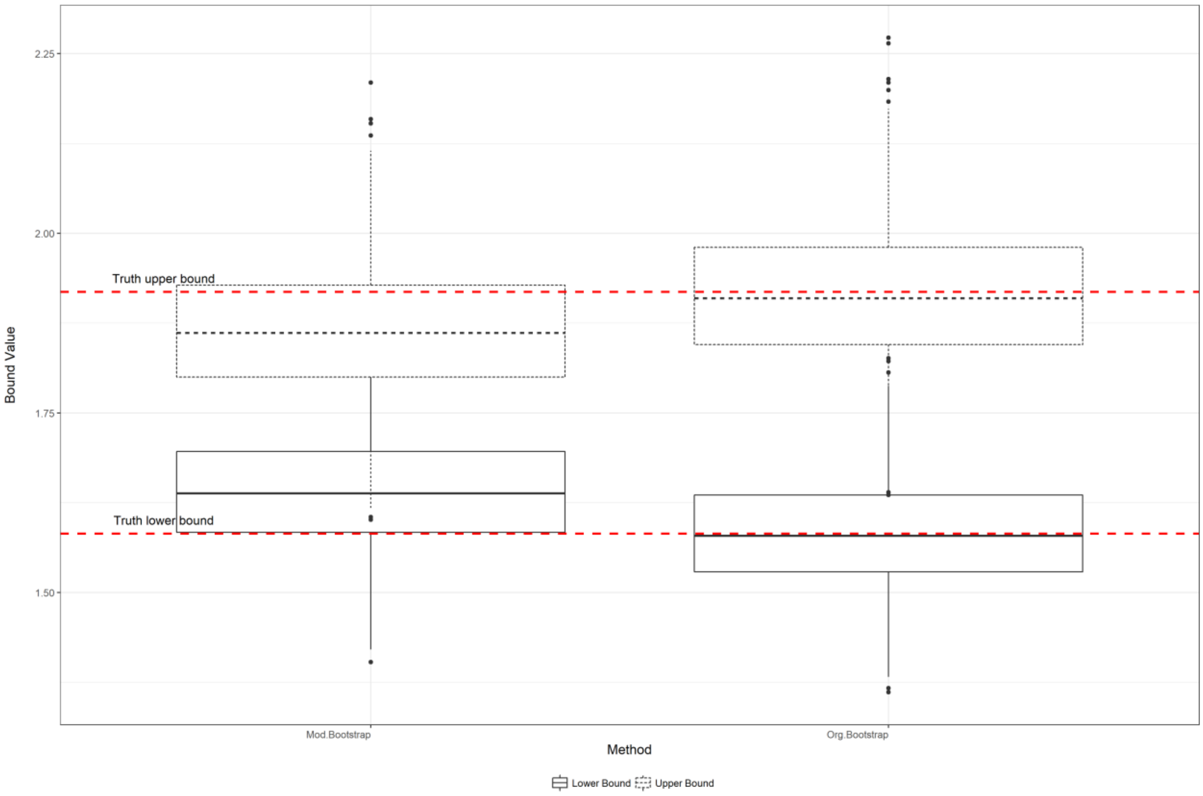
In this section, we aim to compare and contrast the three approaches discussed in Section *Sampling Strategies*. We use Monte Carlo simulations and then our Scopus Custom Data for the comparison.

First simulation is performed to address the question of whether re-calculation of parameters (e.g., MNCS_{pub} or the 90th percentile) is required for each bootstrap sample. In other words, can we ignore the pre-estimation error? The second simulation compares the modified bootstrap, random sampling and delete-10% jackknife to show to what extent their stability intervals differ.

Simulations: Pre-Estimation Error in Bootstrap Method

The citation-based indicators (e.g., MNCS, Top 10%) create dependencies among publications. Several studies (Waltman et al. 2012; Chen, Jen, & Wu 2014; Schneider & van Leevun 2014, Andersen et al. 2018) adopted the modified bootstrapping approach to calculate stability intervals for the unit of analysis they are interested in. In bibliometrics literature, the modified bootstrap approach does not calculate the indicator values for the publications in the bootstrap sample. Instead, they use the indicator values from the original sample. However, these approaches ignored the dependencies created by these indicators. The main question

Figure 1: Two bootstrap approaches: Distribution of lower and upper bounds for N = 500.



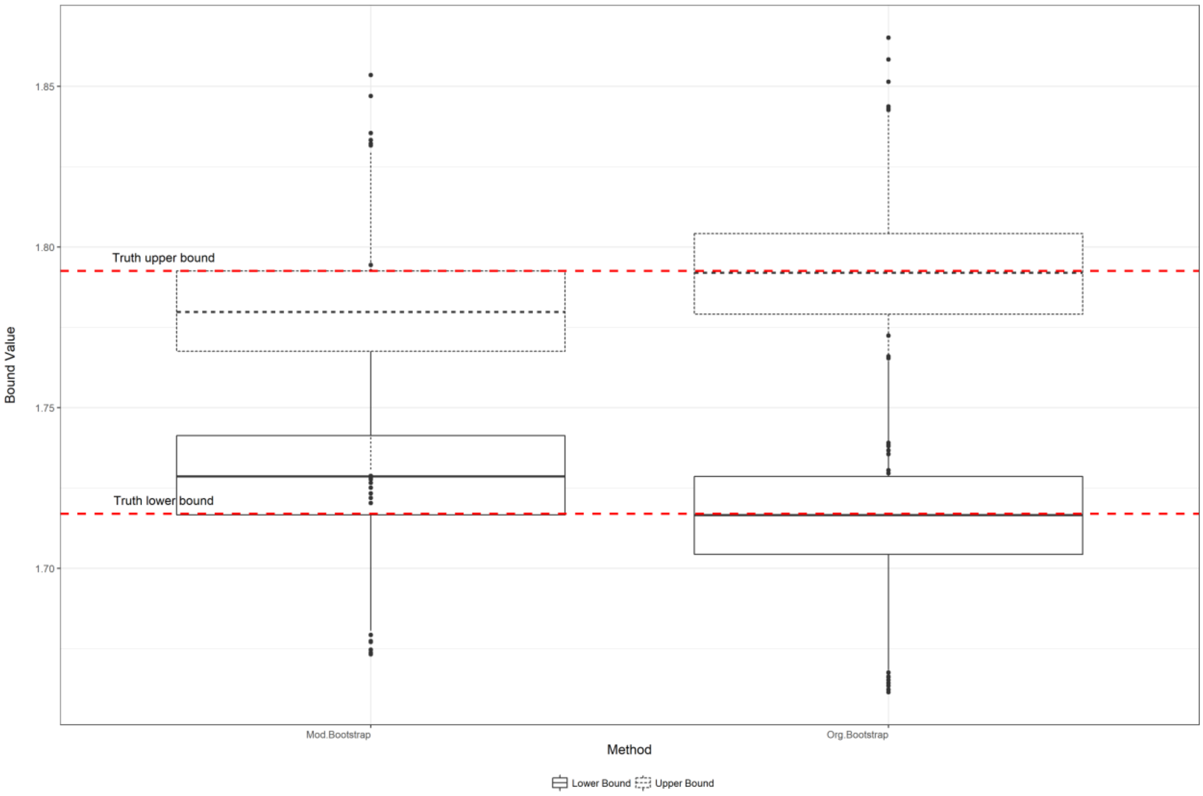
here is whether we can ignore the error introduced by not calculating the indicators for each bootstrap sample (pre-estimation error).

Let's consider the mean of the values above 90th percentile as our random variable. Our Monte Carlo simulation is described as follows:

1. **True Values:** Draw a sample of size N from the standard normal distribution and calculate the mean of the values above 90th percentile ($\mu_{0.9}$).
2. Repeat step (1) M times to construct the sampling distribution for our random variable ($\mu_{0.9}$); calculate the true mean and standard deviation for this distribution.
3. **Original Sample:** Draw a sample of size N from the standard normal distribution, set this sample as the "*original sample*", and calculate the 90th percentile, $\widehat{Q}(0.9)$ and then $\hat{\mu}_{0.9}$.
4. **Bootstrapping:** Form a bootstrap sample of size N by drawing with replacement from the original sample to get $\{Y_n^b\}_{n=1}^N$.
5. Calculate $\widehat{Q^b}(0.9)$ using the bootstrap sample $\{Y_n^b\}_{n=1}^N$.
6. Calculate $\hat{\mu}_{0.9}^b$ based on $\widehat{Q^b}(0.9)$ - ${}_{org}\mu_{0.9}^b$; then $\hat{\mu}_{0.9}^b$ based on $\widehat{Q}(0.9)$ - ${}_{mod}\mu_{0.9}^b$.
7. Go to step (4) to repeat this procedure b = 1, 2, . . . , B times.
8. Calculate the lower and upper bounds for the 95% stability interval.
9. Go to step (3) and repeat this procedure D times.

In our simulation, we set M = 100,000, B = 1,000, and D = 1,000 times. The simulation is conducted for N = 500, 1,000, 10,000, and 100,000.

Figure 2: Two bootstrap approaches: Distribution of lower and upper bounds for N = 10,000.



Figures 1 and 2 illustrate the distribution of the lower and upper bounds of the stability intervals after the Monte Carlo iterations for both bootstrap approaches. According to these two figures, under the modified bootstrap method, a 95-percent stability interval only covers the true value around 80 percent of the time, while under the second method, the 95-percent SI covers the true value around 94.3 percent of the time for sample size N = 500. Similar results are obtained for higher values of N.

The result of this simulation suggests that keeping the publications' indicator values constant for the bootstrap samples may provide biased stability intervals.

Simulations: Modified Bootstrap, Random Subsampling, and Delete-10%-Jackknife

We perform another simulation to assess whether the stability intervals generated using different sampling approaches would produce similar results. In this simulation, instead of $\mu_{0.9}$ we look at PPtop10% and the true value in this case is 10% (or 0.1).

1. **True Value:** Draw an original sample of size N from the standard normal distribution and calculate the 90th percentile, $\widehat{Q}(0.9)$.
2. **Bootstrapping:** Form a bootstrap sample of size N by drawing with replacement from the original sample to get $\{Y_n^b\}_{n=1}^N$.
3. Calculate ${}_{\text{mod}}\widehat{PP}_{\text{top10}}^{b}$ based on $\widehat{Q}(0.9)$ for the bootstrap sample $\{Y_n^b\}_{n=1}^N$.
4. Go to step (2) to repeat this procedure: b = 1, 2, . . . , B times.
5. Calculate the lower and upper bounds for the 95% stability interval.
6. **Random Subsampling:** Form a random subsample of size (0.9 N) by drawing without replacement from the original sample.
7. Calculate $\widehat{PP}_{\text{top10}}^{rs}$ based on $\widehat{Q}(0.9)$ for the random subsample $\{Y_n^{rs}\}_{n=1}^{N*0.9}$.

Figure 3: Distribution of lower and upper bounds for B = 100.

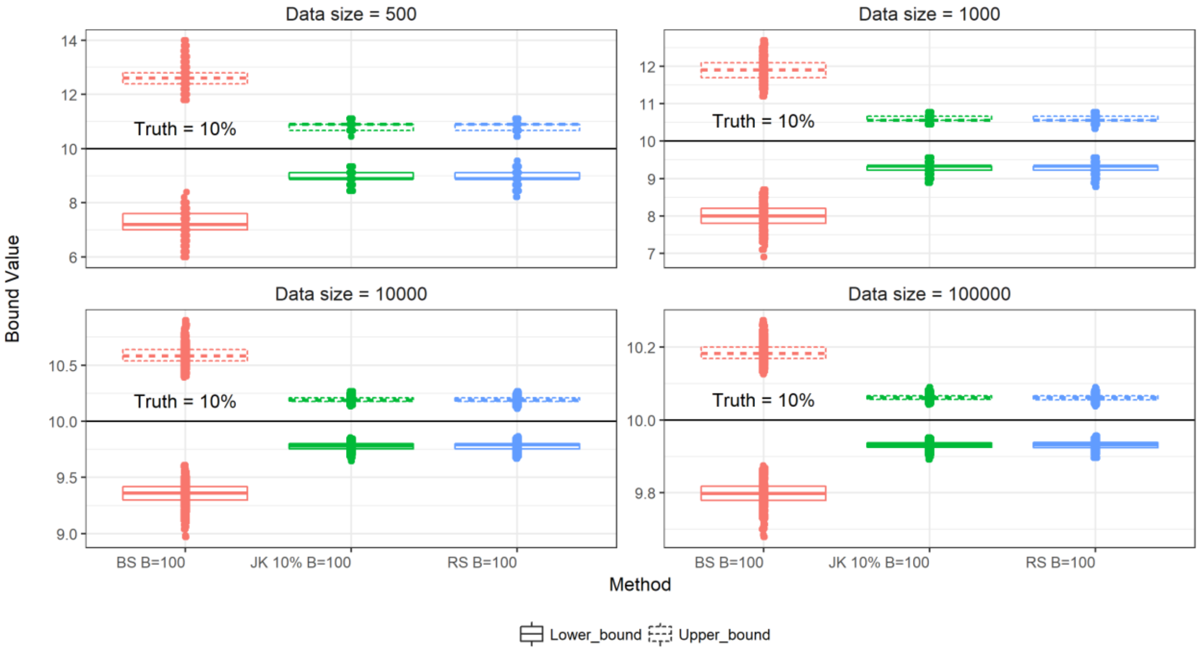
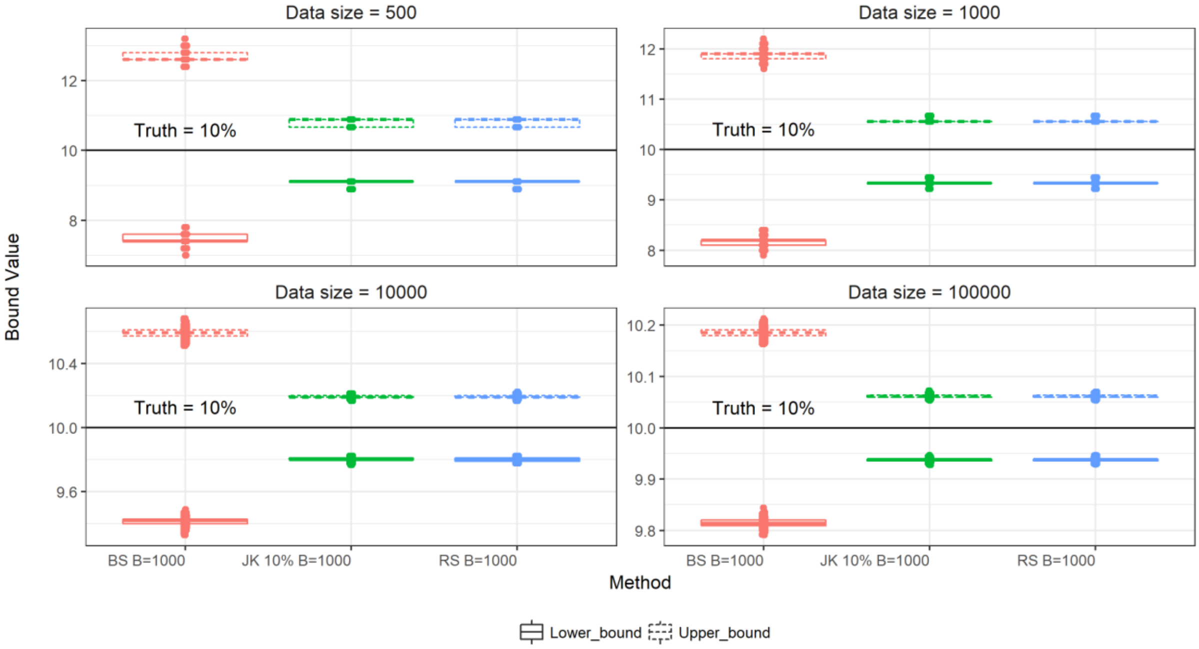


Figure 4: Distribution of lower and upper bounds for B = 1,000.



8. Go to step (6) to repeat this procedure: rs = 1, 2, . . . , B times.
9. Calculate the lower and upper bounds for the 95% stability interval for random subsampling.
10. **Delete-10% Jackknife:** Sort the data in the original sample in random order for the delete-10% jackknife method and split it into 10 parts.
11. Remove part *i* from the sorted sample and calculate $\widehat{\mathrm{PP}}^{\mathbf{jk}}_{\mathrm{top10}}$, repeat this step for *i* = 1,…, 10.
12. Go to step (10) and repeat this procedure (B/10) times.

13. Calculate the lower and upper bounds for the 95% stability interval for delete 10% jackknife.
14. Go to step (1) and repeat this procedure D times.

Colliander and Ahlgren (2011) suggested sampling 90 percent of the underlying data randomly without replacement when reporting the stability intervals. Therefore, we decided to sample 90 percent of the data for random subsampling (and so the delete-10% jackknife method to be consistent with this approach). In our simulations, we set D=1,000, B = {100, 500, 1000}, N = {500, 1000, 10,000, 100,000}.

Figure 3 and 4 illustrate the distribution of the lower and upper bounds of the stability intervals for these 3 methods. Our results indicate that the bootstrap method provides a wider range. The closest range to the true value is achieved by delete-10% jackknife and random subsampling for the number of variations we considered.

Scopus Custom Data Results

In addition to the simulations, we also compare these methods on Scopus data from 2002-2015 for New Zealand and Australia. The Scopus custom data was extracted in June 2017. The data used in this analysis contain articles, conference papers, and reviews. Australia and New Zealand publication set sizes are 757,223 and 137,474, respectively.

Trend analysis based on some of the citation-based indicators is very common to show how science and innovation system of a country performs over the years. Figure 5 and 6 illustrate that the 95 percent stability intervals for New Zealand and Australia using modified bootstrapping, delete-10% jackknife and random subsampling of 90% of the observations. In these figures, the points indicate the actual values, whereas the error bars show the stability intervals (for each method). The results in these figures are based on 10,000 samples (with or without replacement) drawn from the publication set in consideration. It is worth noting that for delete-10% jackknife, 1 iteration consists of dividing data into 10 parts, removing each part and calculating the indicator values on the remaining 90%. For consistency, this is repeated 1000 times.

Similar to the simulation results discussed above, the stability interval based on bootstrap method is wider than the other two methods. Subsampling method generates slightly wider intervals than delete-10% jackknife does. This is true for both indicators we considered in this study.

In this what-if scenario of New Zealand and Australia comparison (expecting 10% error in the country assignment), the bootstrap method and delete-10% jackknife would result in different conclusions.

Conclusion

Stability intervals analysis have been used to analyse the vulnerability of the indicators to the underlying publications. It is a descriptive statistic rather than an inferential one. However, studies in this field have been using different sampling strategies to create the stability intervals. In this study, we compare bootstrap, random subsampling, and delete-10% jackknife methods for the generation of stability intervals.

Figure 5: Average MNCS trend for New Zealand and Australia

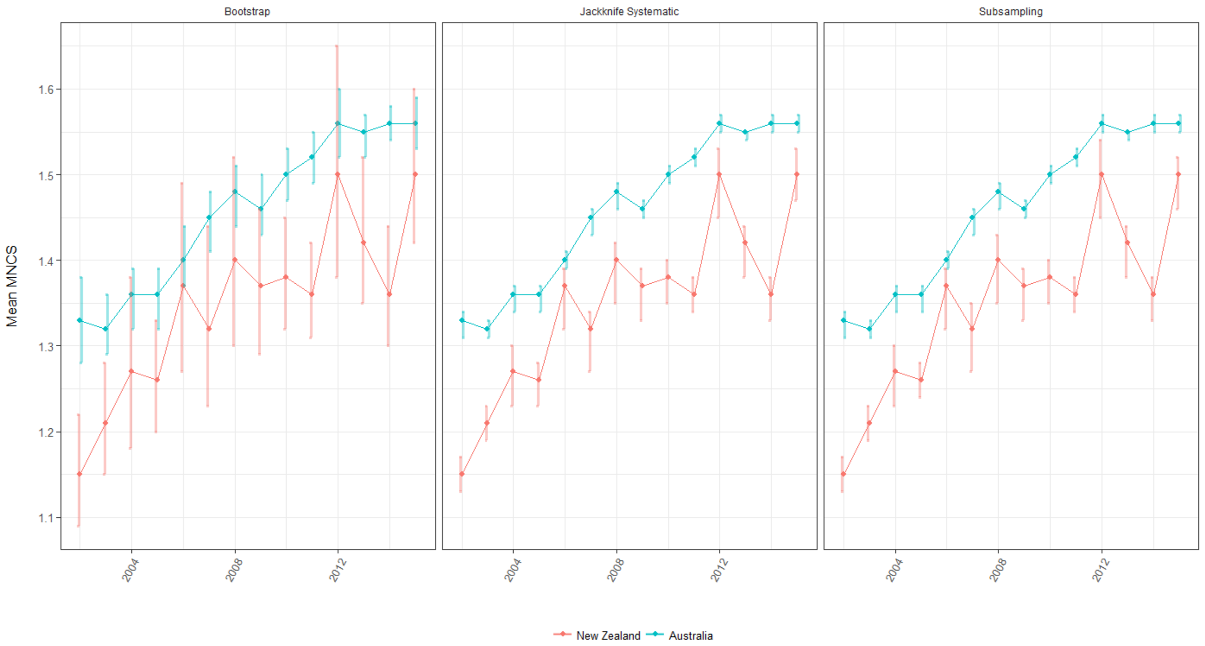
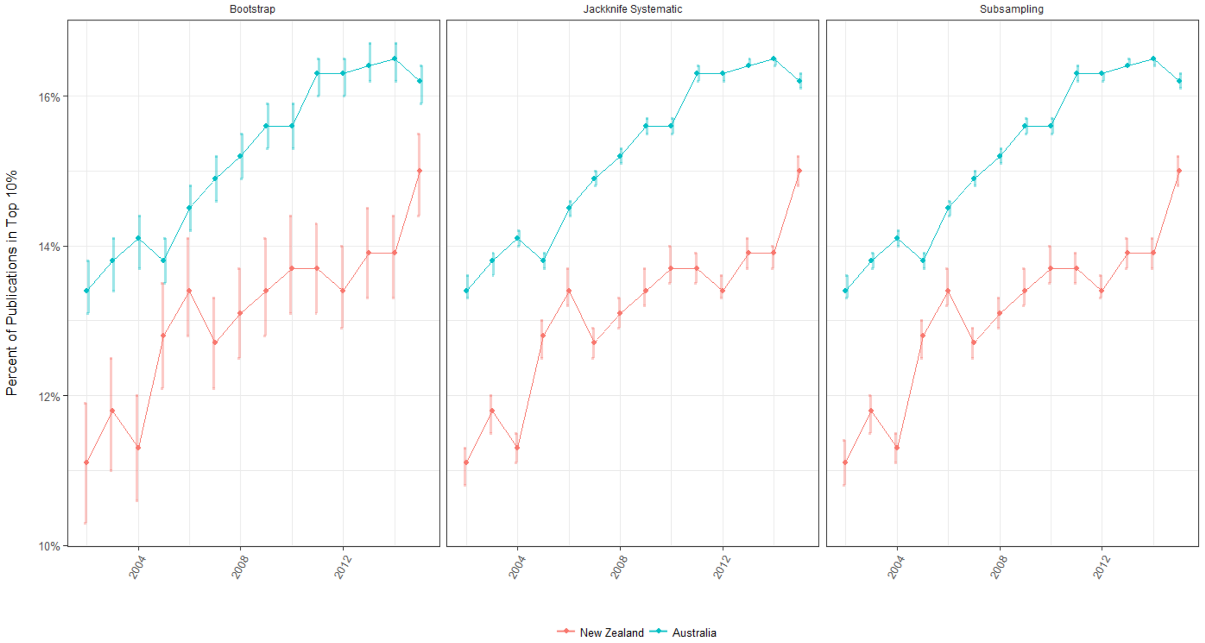


Figure 6: PPtop10% trend for New Zealand and Australia



Simulation results have shown that the pre-estimate errors cannot be ignored for the modified bootstrap method. Furthermore, it generates wider ranges than sampling without replacement. We confirm this by applying these three methods to New Zealand and Australian publication sets.

Stability intervals reflect the vulnerability of the indicator to the size of the data. In other words, the range gets wider as the data size gets smaller which is a behaviour that is expected and desired.

We conclude that the bootstrap method is not ideal for constructing stability intervals. Instead, delete-X% jackknife or random subsampling without replacement should be preferred. While the delete-10% jackknife provides tighter stability intervals than random subsampling suing the Scopus data, the two methods perform about the same throughout the simulations.

Further research is required to investigate whether there is an optimal value for the delete-X% jackknife method. To do so, we will carry on our analysis for various values and compare our results against other methods.

References

Andersen, J. P., Krogsgaard, K., Engel, A. M., & Schneider, J. W. (2018). Mapping international impact of Danish neuroscience from 2004 to 2015 using tailored scientometric methodology. *European Journal of Neuroscience*, *47*(3), 193-200.

Colliander, C., & Ahlgren, P. (2011). The effects and their stability of field normalization baseline on relative performance with respect to citation impact: A case study of 20 natural science departments. *Journal of Informetrics*, *5*(1), 101-113.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, *68*(3), 589-599.

Horvitz, D. G., Thompson, D. J. (1952). A Generalization of Sampling without Replacement from a Finite Universe. Journal of American Statistical Association, 47, 663 – 685.

McMurdie, P. J., Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Computational Biology*, 10(4): e1003531.

Sampford, M. (1962). Methods of Cluster Sampling with and without Replacement for Clusters of Unequal Sizes. *Biometrika*, 49(1/2), 27 – 40.

Schneider, J. W., & van Leeuwen, T. N. (2014). Analysing robustness and uncertainty levels of bibliometric performance statistics supporting science policy. A case study evaluating Danish postdoctoral funding. *Research Evaluation*, *23*(4), 285-297.

Strobl, C., Boulesteix, A. L., Zeileis, A. & Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, 8 (25).

Thelwall, M. (2017a). Confidence Intervals for Normalised Citation Counts: Can They Delimit Underlying Research Capability? *Journal of Infometrics*, 11 (4), 1069 – 1079.

Thelwall, M., Fairclough, R. (2017b). The Accuracy of Confidence Intervals for Field Normalised Indicators. *Journal of Infometrics*, 11 (2), 530 – 540.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E., Tijssen, R. J., Eck, N. J., ... & Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the Association for Information Science and Technology*, *63*(12), 2419-2432.