Article sous presse dans *Journal of Informetrics*, doi:10.1016/j.joi.2011.02.001

Averages of ratios vs. ratios of averages: an empirical analysis of four levels of aggregation

Vincent Larivière[1, 2] and Yves Gingras[1]

[1] Observatoire des sciences et des technologies (OST), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, CP 8888, Succ. Centre-Ville, Montréal, QC. H3C 3P8, Canada

[2] Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, 10th Street & Jordan Avenue, Wells Library, Bloomington, IN. 47405, USA

[lariviere.vincent; gingras.yves]@uqam.ca

Abstract

In the recent debate on the use of Averages of Ratios (AoR) and Ratios of Averages (RoA) for the compilation of field-normalized citation rates, little evidence has been provided on the different results obtained by the two methods at various levels of aggregation. This paper provides such an empirical analysis at the level of individual researchers, departments, institutions and countries. Two datasets are used: 147,547 papers published between 2000 and 2008 and assigned to 14,379 Canadian university professors affiliated to 508 departments, and all papers indexed in the Web of Science for the same period (N=8,221,926) assigned to all countries and institutions. Although there is a strong relationship between the two measures at each of these levels, a pairwise comparison of AoR and RoA shows that the differences between all the distributions are statistically significant and, thus, that the two methods are not equivalent and do not give the same results. Moreover, the difference between both measures is strongly influenced by the number of papers published as well as by their impact scores: the difference between AoR and RoA is greater for departments, institutions and countries with low RoA scores. Finally, our results show that RoA relative impact indicators do not add up to unity (as they should by definition) at the level of the reference dataset, whereas the AoR does have that property.

Introduction

Although field-normalized citations rates have been used for almost 25 years (Braun and Schubert, 1986), several recent papers have suggested new manners for normalizing citations and impact factors (Lundberg, 2007; Moed, 2010b; Zitt & Small, 2008). Another group of papers have discussed the proper manner of compiling field-normalized citation indicators (Leydesdorff and Opthof, 2010; 2011; Leydesdorff and Opthof, 2010b; Opthof and Leydesdorff, 2010; Moed, 2010a; van Raan *et al.*, 2010a; van Raan *et al.*, 2010b; Waltman *et al.*, 2010a; Waltman *et al.*, 2010b)[1]. The central focus of the latter set of papers is the order of operations that lead to the field-normalization of a given group of papers. Typically, field-normalized citation rates imply a ratio between the number of citations received by a given paper or set of papers and the average (or median) number of citations received by all papers of the same field and publication year. When this ratio is above one, it means that the papers considered have received, on average, more citations than the average of the papers of reference (usually at the world level); when it is below one, it is the opposite. Field-normalized impact indicators should thus shave the property of adding to unity at the world level and not all countries can be above or below one.

The issue at stake here is whether this field normalization for the chosen group of papers should be performed *before* averaging the citations received by each paper—and hence be calculated at the paper level—or *after* these citations have been averaged, that is at the group level. While the first calculation is an average of ratios (AoR), the second type is a ratio of averages (RoA). It is clear for us that the former method is the correct one as 1) papers are discrete units of knowledge receiving citations which cannot be blended with those of other papers as if they were mixing fluids coming to an equilibrium (Gingras and Larivière, 2011), 2)

[1] Prior to the current debate, these calculation methods were also analyzed by Egghe and Rousseau (1996 and 2002) and Vinkler (1996).

it follows the usual order of operations (Opthof and Leydesdorff, 2010), that is we make a ratio for each unit and then average the results over all the units, 3) it allows for statistical analysis of differences (Opthof and Leydesdorff, 2010), 4) it does not intervenes in an *a priori* and unpredictable manner to reduce or increase the weight of some papers depending on their citation rates.

The goal of this paper is to provide an empirical analysis of the differences observed between the results obtained by using these two types of calculations for the case of a large dataset of papers assigned to individual researchers and research groups. We do not discuss here other aspects of the field normalization such as the field definitions, fractioning of citations, citation, window, etc.—which are all kept constant in this paper—because these are questions different and independent from the one concerning the order of operations which has recently been much debated. Only scarce empirical evidence has been provided so far on the differences between the results obtained by these two averaging methods and protagonists agree that more empirical analysis would be welcome to clarify and finally settle the situation (Moed, 2010a; Opthof and Leydesdorff, 2010)[2]. Both Opthof and Leydesdorff (2010) and van Raan *et al.* (2010a) have used the same dataset of papers authored by researchers from the Amsterdam Medical Center. While the former analyzes only 232 of these researchers, the latter limit their analysis to the 190 of the 232 researchers with at least 20 publications over the 1997-2006 period. In another paper, van Raan *et al.* (2010b) also provided a scatter plot for 158 Dutch research groups in chemistry and chemical engineering.

This paper provides an empirical analysis of the differences between AoR and RoA at four different levels of aggregation: individuals, departments, institutions and countries. Two datasets are used: 1) 147,547 papers published between 2000-2008 and assigned to 14,379 Canadian university professors (grouped into 508 departments) and 2) all 2000-2008 papers (N=8,221,926) indexed in the Web of Science, which were assigned to countries based on their institutional addresses.

Methods

This paper uses the Web of Science to assess the differences between AoR and RoA for individuals, departments, institutions and countries. Data for individuals and departments are a subset of Canadian papers, and consist of 147,547 papers published between 2000 and 2008 assigned to 14,379 individual researchers in all disciplines. Parts of this dataset have previously been analyzed in Larivière *et al.* (2010) and Gingras *et al.* (2008). The overall dataset comprises 213,514 author-article combinations. The manual assignation of papers and removal of false-positives was performed according to the method described in Larivière *et al.* (2010). The list of researchers provided information on their departmental affiliation, which was used to compile the research impact of 508 departments located in 22 Canadian universities. Unsurprisingly, distributions of research output at these lower levels of aggregation are highly skewed (Larivière et al., 2010). At the individual level, the mean number of papers was 14.8—with a standard deviation of 20.9—and the median number was 8. The mean number of papers of departments was 364.3—with a standard deviation of 693.8—and the median 131.5.

At the level of institutions and countries, all 2000-2008 papers (N=8,221,926) were attributed to 739,753 institutions and 219 countries based on their institutional addresses. These distributions are also much skewed: the mean number of papers by institution is 20.19, the median 1 and the standard deviation 483.7. Given the very large number of institutions with very low numbers of papers—which are, in many cases, institutions with spelling mistakes—, we limited our analysis to the top 3,236 institutions with at least 500 papers over the period. For this subset, the mean number of papers was 3,451.6—with a standard deviation of 6,446.7—and the median 1,314. The mean number of papers at the level of countries was 46,552.8—with a standard deviation of 203,039.8—and a median of 985.

[2] This is one aspect on which both groups of authors agree: "It would be interesting to see how this might work for larger (e.g., institutionally defined) datasets." (Opthof and Leydesdorff, 2010); "…I would strongly encourage conducting more research on the differences between globalized and averaged impact ratios at the level of research groups and other aggregations." (Moed, 2010a)

The journal classification created by the firm The Patent Board (formerly CHI Research) and used by the US National Science Foundation[3] was used for the field normalization of citations. This classification has an important advantage over that of Thomson Reuters, as it categorizes journals in only one category and, thus, removes any overlap between categories. It does not, however, solve the problem of multidisciplinary journals such as *Science* or *Nature*, which are categorized in the *General Biomedical Research* category. Similar categories also exist for other disciplines, such mathematics, chemistry, physics, social science, etc. For each researcher and research group, the two methods (AoR and RoA) were used to compile field-normalized citation indicators. In both cases, citations are counted from 2000 to 2009, which means that papers from 2008 have, at least, a citation window of one complete year.

Results

Figure 1 present scatterplots of the relationship between AoR and RoA at the level of individual researchers (A), departments (B), institutions (C) and countries (D). In all of these figures, a threshold in terms of number of papers was set, although measures of correlation are presented for both the complete distributions and its upper end in figures' inset. Figures 1.A presents the relationship between the scores obtained for individual researchers with at least 20 publications (N=3,449). It shows that the two measures are highly correlated, with Pearson's R and spearman's Rho (ϱ) of about 0.95. Despite these correlations—which are not surprising given that both calculation methods are performed on the same set of papers—Wilcoxon signed-ranks test[4] performed using the PASW statistics software (v. 18.0) showed that the two distributions were statistically different at $p< 0.001$. Our data also show that AoR scores are generally higher than RoA, which suggests that RoA generally underestimates the impact of individuals or, as others have claimed, that it reduces the weight of highly cited papers. More specifically, in 43.8% of the cases, AoR > RoA, while the opposite is true in for 37.7% of the researchers. Finally, in 18.5% of the cases (2,655), AoR and RoA obtained are identical—at the fourth digit. In this latter case, 81.2% of the 2,655 researchers have only one paper (N=2,157).

Figures 1.B present the correlations between AoR and RoA at the level of departments with at least 50 papers. Because the spectrum of disciplines in which departments can publish is broader than that of individual researchers—and hence, the probability of publishing a high impact paper in a journal categorized in specialty with low impact is greater—, the correlation between AoR and RoA is much smaller when all departments are considered, both in terms of values (R=0.755) and ranks (ϱ=0.845). When departments with at least 50 papers are considered (N=337), the Pearson's R drop to 0.680, while Spearman's ϱ remains quite high (0.877). Despite this agreement, Wilcoxon signed-rank test shows that the difference between the two distributions is significantly different at $p<0.001$. Also, we see, again, that AoR are more often greater than RoA, and in a proportion that is much more important than that observed in the case of individual researchers. More specifically, in 68.1% of the cases (346), AoR is greater than RoA; in 28.2% of the cases (143), AoR is smaller than RoA and in 3.7% of the cases (19), AoR and RoA are equal at the fourth digit.

Figure 1.C shows the correlation at the level of institutions with 500 papers or more (N=3,236). Again, the relation between the two indexes is quite high (≈0.98), but the two distributions remain significantly different at $p<0.001$ using the Wilcoxon signed-rank test. We also observe at this level that, in most of the cases, AoR are greater than RoA (76%, N=2,470), while the opposite is true in 24% of the cases (N=766). At the level of countries, (1.D) with at least 1,000 papers (N=109), both measures are also highly correlated (≈0.98) although this relation is lower when all countries are considered (≈0.94). Still, both distributions are statistically different at $p<0.001$ as measured by Wilcoxon's signed-rank test. Despite the very high number of papers involved at this level of aggregation, we still observe a clear tendency of RoA scores to be quite lower than AoRs. When all countries are considered, 76% (N=166) of countries have greater AoR than RoA. When only

[3] More details on the classification scheme can be found at: http://www.nsf.gov/statistics/seind06/c5/c5s3.htm#sb1

[4] Wilcoxon signed-ranks test compare the difference between the two measurements for each unit analyzed. In other words, "… the Wilcoxon signed ranks test is a non parametric statistical procedure for comparing two samples that are parired or related." (Corder and Foreman, 2009, p. 38.)

countries with at least 1,000 papers over the period are considered (N=109), 84% of countries (N=92) obtain AoR greater than RoA.

As mentioned earlier in the paper, relative impact indicators aggregated at the world level are , by construction, equal to 1. Hence calculating, in a reverse engineering manner, the weighted world average using the impact scores of all countries should give us 1. Using only the first address of 2000-2008 papers in order to remove the assignation of papers to more than one country—and count the impact of papers only once—we calculated the weighted world average using countries AoR and RoA. While the weighted world average of AoR equals 1.00000, the same calculation performed using countries' RoA results in a value of 0.97600—which becomes the baseline for determining, in this dataset, if countries' impact is below or above the world average. This shows that inverting the order of operations and performing the average before doing the ratio creates an inconsistency since, by definition, the normalized weighted world average should be equal to 1.0. This peculiar situation is caused by the—generally—lower weight attributed, by RoA, to highly cited papers, which 'lowers' the scores of the numerators but does not touch the denominators.

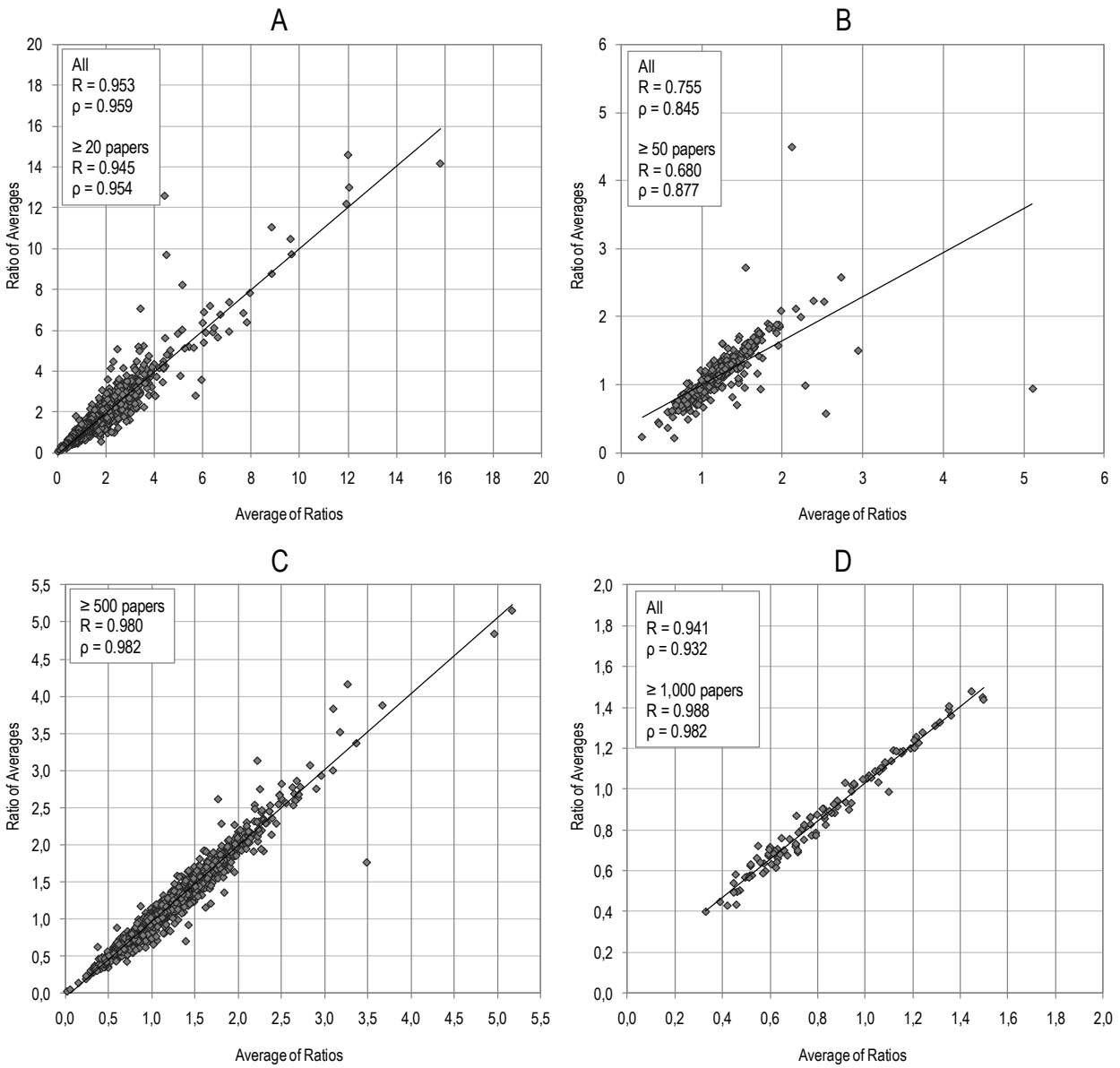

Figure 1. Relation between RoA and AoR field normalized citation indicators at the level of A) individual researchers (≥20 papers), B) departments (≥50 papers), C) institutions (≥500 papers) and D) countries (≥1000 papers).

Figure 2 presents, for the various levels of aggregation, the difference between AoR and RoA as a function of the number of papers. When the number of the Y-axis is negative, AoR is smaller than RoA, which implies that the position of the researchers is underestimated by using the deficient (RoA) method; when it is positive, the opposite is true. It shows that the number of papers by researcher has a strong influence on the relationship between AoR and RoA. Despite the fact that the large differences as seen at the left of Figure 2 become less on a percentage scale when the number of papers per author increases, deviations of more than 20% remain prominent even with authors with more than 100 papers. However, for 'ordinary' researchers with less than 10 papers (58% of researchers), the difference is greater, and even exceeds 50% in 9% of the cases (excluding researchers who have only one paper). On the other hand, researchers who publish more than one paper, but in only one specialty for one given year also obtain identical AoR and RoA, as the denominator of their papers is always the same. There is thus a 'tension' between on the one hand having few papers—and, hence an increased probability of publishing in only one specialty—and, on the other, having

many papers, for which the law of large numbers tends to diminish the difference between both measures though the effect is not systematic and large random fluctuations exist for many researchers when using RoA.

The researcher labeled 'A' in Figure 2.A is worth a little more investigation. Even if this researcher has authored a fairly high number of papers (121), the agreement between AoR and RoA obtained is very small (2.09 vs. 0.95). This difference is mainly due to one paper published in a high impact medical journal in 2008, which attracted 429 citations in 2008 and 2009. Since that journal is categorized in the general & internal medicine field category—for which the average number of citations received is about 3 for the papers published the same year—the field normalized citation rate of this paper is 143. This, of course, increases the average impact of the researcher when compiled as an AoR, but has little effect on the RoA since the 'raw' number of citations of this paper is 'blended' with the citations received by all other papers he has published. Because the resulting index is an average—and not a median—this highly cited publication has a strong effect on the AoR of this researcher. In order to illustrate how using RoA lead to arbitrary results, let us suppose that our researcher would have received 4290 citations instead of 429, and that the field average was 30. Then the AoR of the researcher would be identical as it should be (2.09) while her RoA, which averages citations to all papers and then divides by the average of the number of citations received by papers of the same specialty, would jump to 2.44. It should be obvious that the correct measure here is AoR and that RoA produces an arbitrary value which depends on the absolute values of the citations used in the numerator and dominator.

The agreement between AoR and RoA measures is better for departments with a larger number of papers (Figure 2.B), there are, however, some departments—labeled as 'B' and 'C' on Figure 2.B—with a fairly large number of papers (260 and 480 respectively) that obtain very different values of AoR and RoA: 2.12 vs. 4.50 for department B and 1.55 vs. 2.72 for department C. The distribution of citations received by papers from these departments is very much skewed: 5 papers account for 72% of all citations received by department B and 5 papers account for 51% of citations of department C. More specifically, one paper they have co-authored has received 2,421 citations. Because of this important skewness, the very large number of citations received by this paper alone—which accounts for an important part of all citations summed in the numerator—as well as the fact that these departments publish papers in specialties where the average denominator is relatively low (4.31 and 5.11), makes RoA much larger than AoR. In these two cases, thus, highly cited publications clearly have more weight in the global RoA index than in the AoR. This clearly shows that the objective of the RoA to control the effect of highly cited publications is not obtained and the results are unpredictable and depend in a complex manner on the distributions of the papers among different specialties.

At the level of world institutions (Figure 2.C), we observe that the two indicators tend to converge as the number of papers increase. However, we also observe that the two institutions with the largest number of papers—the Chinese Academy of Sciences (labeled 'D', N=105,319 papers) and the Russian Academy of Sciences (labeled 'E', N=101,447 papers)—are quite underestimated by the use of RoA. More specifically, while the AoR of the Chinese Academy of Sciences is close to the world average (0.96), its RoA is much lower 0.81. The Russian Academy of Sciences obtains a 0.44 AoR, but a 0.42 RoA. In fact, many of these institutions which obtain much lower RoA are from emerging nations, which generally obtain low citation scores. A similar phenomenon is also observed at the level of countries (Figure 2.D): despite the very large number of papers, we observe important differences in countries' AoR and RoA scores, especially for emerging nations. Two of the most productive countries, India (labeled 'F') and China (labeled 'G') obtain drastically lower RoA impact scores, and so do several other countries with more than 10,000 papers who generally obtain low impact scores.

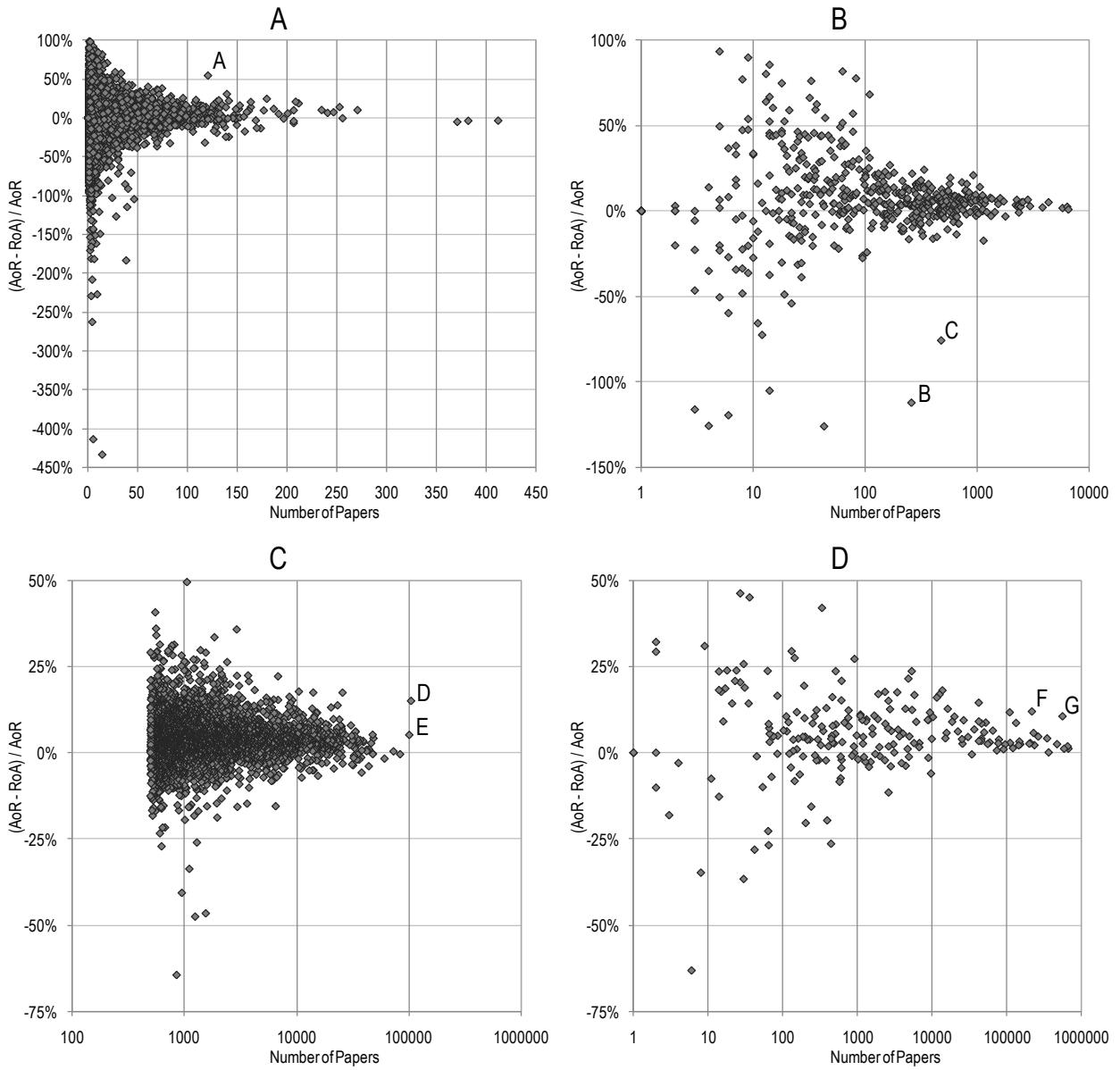

Figure 2. Relationship between (AoR – RoA) / AoR and the number of papers at the level of A) individual researchers, B) departments, C) at the level of institutions (≥500 papers), D) countries.

Figure 3 provides more insight on the relationship between countries' impact and the difference between the two calculation methods. At each of the levels except that of individuals, we observe a moderate negative correlation between the AoR/RoA difference and their impact as measured by RoA. Hence, departments, institutions and countries with low RoA scores are more likely to be affected negatively by the use of this method of calculation and, thus, have their impact underestimated. In other words, the lower the RoA score is, the higher the difference between AoR and RoA will be. On the other hand, aggregates with high impact will obtain high scores on both methods, while aggregates with lower impact will obtain even lower impact using RoA.

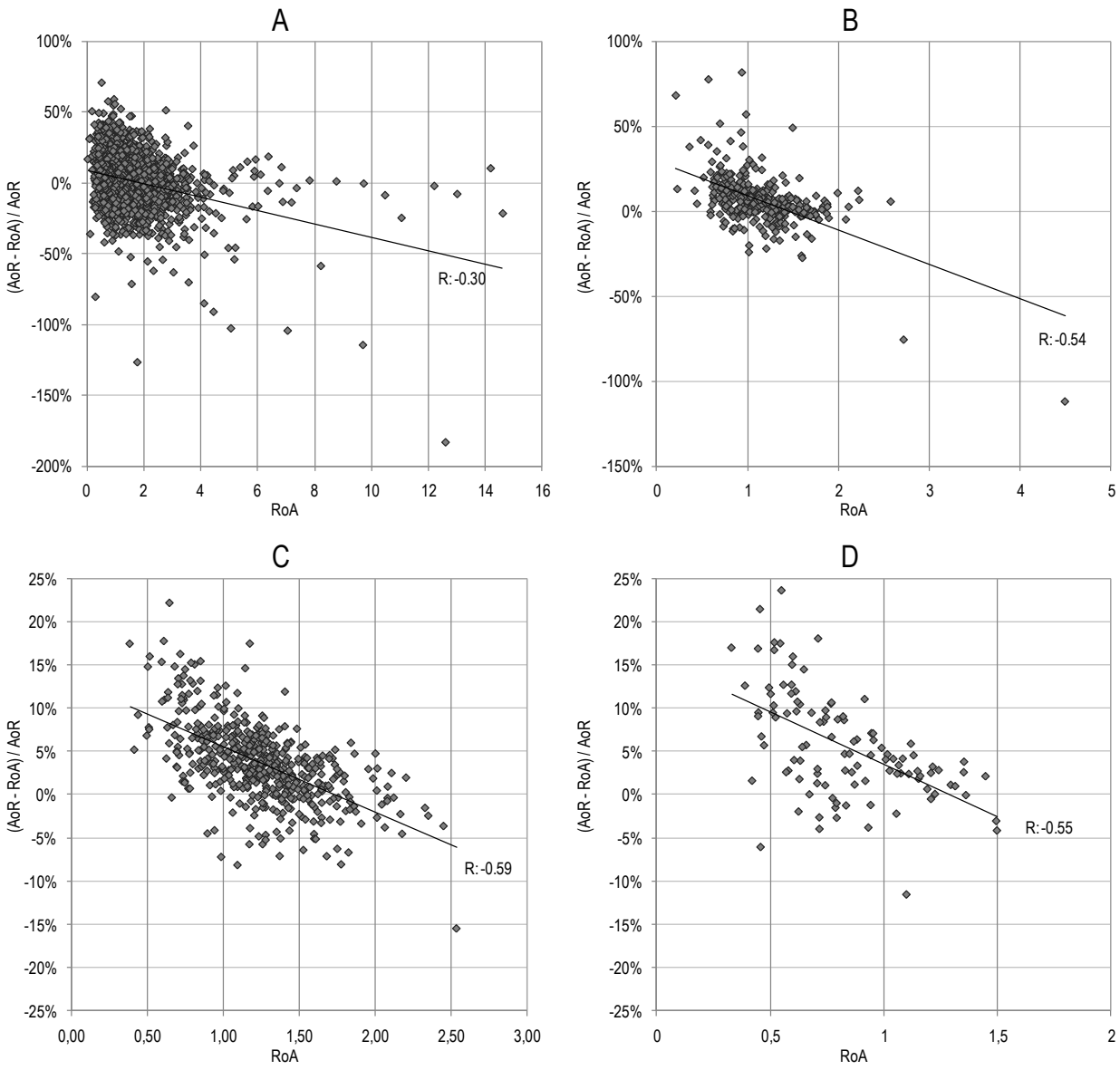

Figure 3. Relation between ((AoR – RoA) / AoR) and RoA at the level of A) individual researchers (≥20 papers), B) departments (≥50 papers), C) at the level of institutions (≥5000 papers), D) countries (≥1000 papers).

Discussion and Conclusion

The data presented in this paper shows that RoA and AoR impact score obtained by researchers produce statistically different results ($p<0.001$) at all levels of aggregation using Wilcoxon signed-ranks test, even though the measures are highly correlated. Our results also show that the use of RoA makes it impossible to calculate a 'global' weighted average consistent with the use of relative indicators which should add up to 1 when compiled at the level of the entire reference dataset. While this is true of countries' AoR, the weighted average of their RoA results in a value of 0.97600, which does not really make sense.

In the dataset of Canadian researchers and departments, the difference between both calculations generally depend on the number of papers published: at one end of the spectrum, individuals with only one paper or who publish in only one discipline for a specific year have by definition identical AoR and RoA, as the denominator is always identical. At the other end, having a large number of papers generally makes the two measures comparable since large numbers of papers and citations are involved and variations tend to average out. Researchers and departments that are most likely to obtain different impact scores depending on the method used are those with papers in between these two limits, those with low and average numbers of papers. Interestingly, the differences between AoR and RoA are greater in the case of departments than of individuals, as the 'breadth' of specialties in which departments publish is generally greater than that of individuals. This suggests that the scores of 'multidisciplinary' researchers or research units working in several research areas will get different evaluations whether it is the AoR or the RoA that is used.

In some of cases, using the AoR instead of the RoA can change drastically the impact score of a researcher or a department. In most of these cases, AoR scores are greater than RoAs because in the former, citations to highly cited publications are weighted only against the average number of citations received by papers of the same specialty, while in the latter case, citations received by these highly cited publications are blended – without justification – with citations received by all other research papers, and so is the average number of citations of the papers of their specialty. The idea of constructing an indicator that would 'correct' for these cases is not consistent with the fact that one wants to evaluate the real dynamic of the different units and it is thus normal to take into account these highly cited papers in the overall impact score of a department or an individual, since they actually represent what one wants to achieve through research planning and funding. The same reasoning applies to uncited papers: an uncited paper in a discipline where the average value is 50 citations will curiously 'weigh' more in the overall score of the RoA that an uncited paper published in a field with an average number of citations of 5. However, in the case of AoR, both papers will have the same weight, as one should expect. This likely explains why countries, institutions and departments with lower impact scores obtain even lower scores with RoA.

Finally, at the level of institutions and countries, the discrepancies between the two measures tend to diminish for entities with large number of papers, but there are also cases of countries with very high number of papers for which the impact is greatly underestimated by the use of RoA (China, India, Russia and Turkey, among others). More specifically, there is a tendency for departments, institutions and countries with low RoA scores to be underestimated by the use of this deficient calculation method: the lower RoA scores are, the higher the difference between AoR and RoA will be.

Acknowledgements

We thank Loet Leydesdorff, Tobias Opthof, Jean-Pierre Robitaille and Matthew Wallace as well as the three anonymous referees for their useful comments and suggestions.

References

Corder, G.W., and Foreman D.I. (2009). *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. Hoboken: Wiley.

Egghe, L. and Rousseau, R (1996) Average and global impact of a set of journals, *Scientometrics*, 36(1), 97-107.

Egghe, L. and Rousseau, R (2002) A general framework for relative impact indicators. *Canadian Journal of Information and Library Science*, 27(1), 29-48.

Gingras, Y. and Larivière, V. (2011). There are neither "king" nor "crown" in scientometrics: Comments on a supposed "alternative" method of normalization. *Journal of Informetrics*, 5(1), 226-227.

Gingras, Y., Lariviere, V., Macaluso, B., and Robitaille, J.P. (2008) The effects of aging on researchers' publication and citation patterns. PLoS ONE, 3(12): e4048. arXiv:0810.4292

Larivière, V., Macaluso, B., Archambault, É. and Gingras, Y. (2010). Which scientific elites? On the concentration of research funds, publications and citations. *Research Evaluation*, 19(1), 45-53.

Leydesdorff, L. and Opthof, T. (2010). Normalization at the field level: Fractional counting of citations, *Journal of Informetrics*, 4 (4), 644-646.

Leydesdorff, L. and Opthof, T. (2011). Remaining problems with the "New Crown Indicator" (MNCS) of the CWTS. *Journal of Informetrics*, 5(1), 224-225.

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

Moed, H.F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.

Moed, H.F. (2010a). CWTS crown indicator measures citation impact of a research group's publication oeuvre. *Journal of Informetrics*, 4(3), 436-438.

Moed, H.F. (2010b) Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, 4 (3), pp. 265-277

Opthof, T. and Leydesdorff, L. (2010). Caveats for the journal and field normalizations in the CWTS ("Leiden") evaluations of research performance. *Journal of Informetrics*, 4(3), 423–430.

Schubert A. and Braun T (1986) Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics,* 9, 281–291.

Spaan, J.A.E. (2010). The danger of pseudoscience in Informetrics, *Journal of Informetrics*, 4(3), 439–440

Van Raan, A. F. J., Van Leeuwen, T. N., Visser, M. S., Van Eck, N. J. and Waltman, L. (2010a). Rivals for the crown: Reply to Opthof and Leydesdorff. *Journal of Informetrics*, 4(3), 431–435.

Van Raan, A. F. J., Van Eck, N. J., Van Leeuwen, T. N, Visser, M. S. and Waltman, L. (2010b). The new set of bibliometric indicators of CWTS, *Book of Abstracts STI Conference*, 9-11 September 2010, Leiden, Universiteit Leiden, p. 291-293.

Vinkler, P. (1996). Model for quantitative selection of relative scientometric impact indicators. *Scientometrics* 36(2), 223-236.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S. and van Raan, A.F.J. (2010a) Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics* (in press), arXiv:1003.2167.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S. and van Raan, A.F.J. (2010b) Towards a new crown indicator: An empirical analysis. arXiv:1004.1632.

Article sous presse dans *Journal of Informetrics*, doi:10.1016/j.joi.2011.02.001

Zitt, M. and Small, H. 2008 Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology* 59 (11), 1856-1860.