Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

CRIS 2014

A comparison of research output counting methods using a national CRIS – effects at the institutional level

Tomáš Chudlarský*, Jan Dvořák, Martin Souček

Institute of Information Studies and Librarianship, Faculty of Arts, Charles University in Prague, U Kříže 8, 158 00 Praha 5, Czech Republic

Abstract

Recent decades have seen a trend towards scientific publications with many authors. There is not an agreed way of counting co-authored publications. This research in progress contribution compares the behaviour of four representative methods of counting research outputs. We present differences between those methods of counting evaluated at the level of whole universities, their faculties, or non-university research institutes. Our study uses publication metadata from a national CRIS.

Peer-review under responsibility of euroCRIS.

Keywords: bibliometric counting methods, fractional counting, whole counting, scientometric indicator, CRIS

1. Introduction

Recent decades have seen a trend towards scientific publications with many authors. There is not an agreed way of counting co-authored publications. Actually, there are tens of different approaches, see Olesen Larsen 2008[5]. The influence of the counting method was previously studied in Gauffriau et al. 2008[3] in the context of research performance of countries and their groupings. The study Lin et al. 2013[4] concentrates on the ranking of world leading universities in one field of science (physics).

This contribution compares the behaviour of several representative methods of counting research outputs. Our study uses publication metadata from a national Current Research Information System (CRIS), and incorporates a

* Corresponding author.
E-mail address: tomas.chudlarsky@ff.cuni.cz

Peer-review under responsibility of euroCRIS.

two-level hierarchy of organizations. We look into similarities between the studied methods of counting evaluated at the level of whole universities, their faculties, or non-university research institutes. This extends the previous works. We devise an alternative method of measuring differences between counting methods.

Specifically, we consider the following methods of counting research outputs:

- Fractional counting – where credit is distributed uniformly among the contributors. We investigate several variants ways of assigning weights to contributors (to give more credit to internationally co-authored publications, for instance).
- Whole counting – where each collaborator (or collaborating institution) receives a full credit.

These methods are used in well-known university rankings (e.g. the Academic Ranking of World Universities or the Leiden Ranking) as well as in common scientometric practices or recent projects[7]. They were also used, in their specific variants, in large-scale national research evaluation exercises: in the Czech Republic and in Italy (Bonaccorsi 2013[1]), to name a few.

2. Methodology

2.1. The counting methods

Counts of scientific outputs of various types are usually aggregated at the level of departments, faculties, institutions, or national states. It thus provides a basic scientometric indicator. More advanced applications use counts as a factor that multiplies another scientometric indicator, such as citation counts, or points. We only cover the counts in this study.

Suppose an output has n authors. Let m denote the number of authors that are affiliated with any organization unit or institution which belong to a selected set E of organizations that are covered in the CRIS. Credit is defined only for organizations from the set E. We consider the following counting methods:

Whole counting (*W*)

- every organizational unit to which an author is affiliated gets 1 credit
- every institution to which an author is affiliated gets 1 credit
- (if individual authors were considered, they would receive 1 credit each)

Fractional counting (*F*)

- an organisational unit gets k/n of a credit where k is the number of authors affiliated with the organisational unit
- an institution gets k/n of a credit where k is the number of authors affiliated with the institution
- (if individual authors were considered, they would get $1/n$ of a credit each)

Modified fractional counting (*R1*) – based on internal authors only

- an organisational unit gets k/m of a credit where k is the number of authors affiliated with the organisational unit
- an institution gets k/m of a credit where k is the number of authors affiliated with the institution

Second modified fractional counting (*R2*) – counting external authors with a weight of one half

- an organisational unit gets $k/(\frac{m}{2}+\frac{n}{2})$ of a credit where k is the number of authors affiliated with the organisational unit
- an institution gets $k/(\frac{m}{2}+\frac{n}{2})$ of a credit where k is the number of authors affiliated with the institution

The following inequalities hold: $F \leqslant R2 \leqslant R1 \leqslant W$.

The *R1* and *R2* methods are identical with *F* on articles without external authors (authors with affiliations outside of the selected set of organizations). On outputs with external authors, the two methods give higher counts then *F*.

The *R1* method was used in the evaluation of results of research organizations in the Czech Republic in the past, *R2* is used in the recent years. The *W* method on the institutional level is used in the Italian evaluation exercise *Valutazione della Qualità della Ricerca* (VQR).

2.2. The data

The study is based on an extensive set of publication metadata collected in the Czech national CRIS[2,6] (the Czech Research, Development and Innovation Information System). This data source has different characteristics from the conventional Web of Science or Scopus collections:

- The publication metadata is reported by the research organizations themselves, not through publishers.
- The organizations are well identified, the problem of ambiguous institution spelling does not occur. In fact, a two-level hierarchy of universities and their faculties is available.
- The researchers are well identified too.

We used the publicly accessible metadata from the Czech national CRIS of scientific articles published between 2008 and 2012. It covers all research performing organizations in the Czech Republic. The external authors are those affiliated with foreign institutions. We only considered articles published in journals listed in the list of titles of the Scopus database by Elsevier. This represents roughly 60 thousand articles in all scientific disciplines.

The Czech national CRIS maintains a two-level hierarchy of institutions and their organizational units (faculties of universities, institutes of the Academy of Sciences of the Czech Republic). The publication metadata have precise identifications of authors that are affiliated with Czech institutions, and these affiliations are recorded precisely (in our case to roughly five hundreds of institutions and organizational units).

3. Results

Figure 1 illustrates that from our dataset only 12% of articles have a single author. Three percent of articles have more than 20 coauthors; the maximum number of coauthors we encountered was 6,084. With 88% of scientific articles being co-authored, the methods of distributing and counting credit are important.

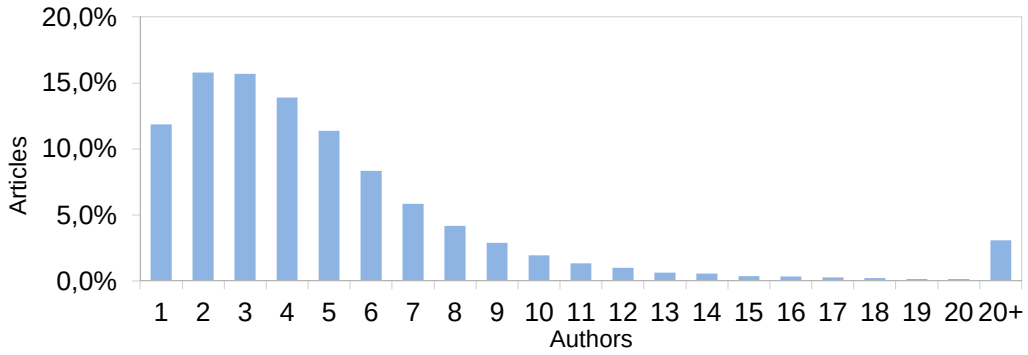

Figure 1. Proportions of articles by number of coauthors.

3.1. Counting method comparison using rankings

Table 1 presents the top institutions with their article counts by different methods. The three methods *F*, *R1*, *R2* seem to produce very similar orders, whereas the order by the *W* method slightly differs. The order of the top five institutions is very stable. Note that these five institutions are responsible for more than 57% of the scientific production of the country, whichever counting method is used.

Table 1. Top 10 institutions by article counts using different methods.

	Order				Share				Count			
	F	*R2*	*R1*	*W*	*F*	*R2*	*R1*	*W*	*F*	*R2*	*R1*	*W*
Academy of Sciences of the Czech Republic	1	1	1	2	21.6%	22.3%	23.9%	21.8%	9151.9	10318.3	14324.7	18059
Charles University in Prague	2	2	2	1	19.2%	19.3%	20.2%	21.9%	8112.3	8927.0	12093.8	18150
Masaryk University, Brno	3	3	3	3	7.2%	7.1%	6.7%	6.2%	3053.0	3292.3	4013.7	5171
Palacky University Olomouc	4	4	4	4	5.3%	5.3%	5.2%	4.7%	2230.4	2442.7	3093.0	3892
Czech Technical University in Prague	5	5	5	5	4.0%	4.0%	4.3%	4.0%	1692.7	1865.8	2561.1	3357
Brno University of Technology	6	6	6	9	3.3%	3.2%	2.8%	2.4%	1415.2	1494.7	1695.1	1956
Mendel University in Brno	7	7	8	11	3.1%	3.0%	2.5%	2.1%	1321.4	1378.8	1523.3	1734
Institute of Chemical Technology, Prague	8	8	7	7	2.9%	2.9%	2.7%	2.6%	1232.2	1348.3	1634.4	2162
Czech University of Life Sciences, Prague	9	9	10	12	2.4%	2.4%	2.1%	1.9%	1030.8	1094.7	1246.1	1560
University of South Bohemia in České Budějovice	10	10	9	8	2.2%	2.2%	2.2%	2.5%	912.1	1015.2	1331.2	2040
Totals of all institutions									42330.6	46351.7	59937.8	83031

Table 2. Top 20 organizational units by article counts using different methods.

	Order				Share				Count			
	F	*R2*	*R1*	*W*	*F*	*R2*	*R1*	*W*	*F*	*R2*	*R1*	*W*
Charles University in Prague / Faculty of Mathematics and Physics	1	1	1	1	4.3%	4.5%	5.5%	4.8%	1831.6	2089.6	3300.5	4204
Charles University in Prague / Faculty of Science	2	2	2	3	3.7%	3.7%	3.7%	3.8%	1547.4	1711.2	2192.6	3313
Charles University in Prague / First Medical Faculty Charles University	3	3	4	2	3.1%	3.1%	3.2%	4.3%	1330.3	1446.9	1904.1	3807
Masaryk University, Brno / Faculty of Science	4	4	5	6	2.9%	3.0%	3.0%	2.5%	1232.3	1367.8	1770.9	2194
Institute of Physics of the AS CR, v.v.i.	5	5	3	4	2.5%	2.7%	3.5%	3.2%	1062.5	1264.5	2110.3	2823
Masaryk University / Medical Faculty of Masaryk´s University	6	6	7	8	2.5%	2.4%	2.2%	2.2%	1049.1	1115.9	1324.4	1901
Palacky University Olomouc / Faculty of Science	7	7	6	7	2.3%	2.3%	2.3%	2.3%	969.5	1072.8	1403.0	1977
Palacký University Olomouc / Medical Faculty UP Olomouc	8	8	8	13	2.3%	2.3%	2.2%	1.9%	959.8	1050.8	1323.6	1663
General University Hospital in Prague	9	9	11	5	1.8%	1.8%	1.7%	2.9%	770.8	822.3	1024.2	2568
Institute of Organic Chemistry and Biochemistry of the AS CR, v.v.i.	10	10	10	14	1.7%	1.8%	1.8%	1.6%	727.5	808.8	1054.8	1423
Charles University in Prague / Third Medical Faculty Charles University	11	11	9	10	1.6%	1.7%	1.9%	2.1%	696.1	773.7	1109.8	1877
Charles University in Prague / Medical Faculty of Charles University	12	12	13	12	1.6%	1.6%	1.5%	2.0%	688.2	732.3	868.2	1716
Czech Technical University in Prague / Faculty of Electrical Engineering	13	14	15	22	1.4%	1.4%	1.3%	1.1%	580.8	633.6	781.0	947
University of Pardubice / Faculty of Chemical Technology	14	15	14	24	1.4%	1.4%	1.3%	1.0%	569.7	630.5	786.9	894
Biology Center of the AS CR, v.v.i.	15	13	12	15	1.3%	1.4%	1.7%	1.6%	569.0	668.0	993.3	1399
University Hospital Hradec Králové	16	17	21	16	1.3%	1.3%	1.1%	1.6%	554.6	578.4	666.4	1382
Institute for Clinical and Experimental Medicine	17	16	17	18	1.3%	1.3%	1.3%	1.2%	542.9	580.4	761.4	1018
Mendel University in Brno / Faculty of Agronomy	18	19	23	29	1.2%	1.2%	1.0%	0.9%	509.3	539.2	607.5	746
University Hospital in Motol	19	18	16	11	1.2%	1.2%	1.3%	2.1%	503.9	548.7	780.1	1807
Charles University in Prague / Second Medical Faculty	20	20	18	9	1.1%	1.1%	1.2%	2.2%	454.9	502.5	736.2	1890

Table 2 presents the top organizational units. Here the *F* and *R2* methods again show a very high degree of similarity in their behaviour. The *R1* method sets itself slightly apart of the previous two, while *W* produces rather from different orderings. Note that neither Table 1 nor Table 2 express a scientometric analysis of the Czech R&D, they only illustrate behaviour of counting methods on the most productive organizations. In this research in progress report we are omitting differences by scientific fields; these are the subject of our continuing research.

Lin et al. 2013[4] express the difference between counting methods using the Spearman's correlation coefficient. That is a valid approach when rankings are in the spotlight of a study. In Table 3 we present the Spearman's correlation coefficients computed on the data. One can see again a difference between *W* and the others methods.

Table 3.Spearman's correlation coefficient

	Institutional level			Organizational Unit level		
Counting	*R2*	*R1*	*W*	*R2*	*R1*	*W*
F	99.8%	98.2%	96.3%	99.9%	99.3%	98.4%
R2	-	99.0%	97.0%	-	99.6%	98.6%
R1	-	-	97.7%	-	-	99.1%

3.2. Counting method comparison using distances

In our opinion the step-wise character of rankings that erases information about quantitative closeness. This calls for a more continuous approach. We propose to measure the differences between counting methods as the distances of relative shares of institutions or organizational units. The motivation is to express the total change of balance between players when choosing one counting method instead of the other.

Let A be the set of all articles. We denote E the set of organizations (whole institutions or individual organizational units), and $A(e)$ the set of articles of organization e (for $e \in E$).

A *counting method* c is a mapping from the Cartesian product of the set of articles and the set of organizations to non-negative rational numbers. *Aggregated count* C is the sum of c over all articles of given organization e: $C(e) = \sum_{a \in A(e)} c(a,e)$.

For each aggregated count C we define the *relative aggregated count* C' by the formula

$$C'(e) = \frac{C(e)}{\sum_{e' \in E} C(e')}.$$

This is a mapping from the set of organizations to rational numbers between zero and one. It has the property that

$$\sum_{e \in E} C'(e) = 1.$$

We define the *distance between counting methods c and d* as the L^1-distance between their relative aggregated counts C' and D':

$$dist(c,d) = \left\| C' - D' \right\| = \sum_{e \in E} \left| C'(e) - D'(e) \right|.$$

We have computed distances in the data on the institutional level as well as on the organization unit level. The results are presented in Table 4. We can conclude that methods *F* and *R2* form the closest pair, they are only 3%

apart in both levels. Counting method *W* is separated from the others by more than 13%. The distances on the institutional level are roughly 2/3 of those on the organizational unit level.

Table 4. Distance matrix

	Institutional level			Organizational Unit level		
Counting	*R2*	*R1*	*W*	*R2*	*R1*	*W*
F	1.9%	8.4%	15.6%	3.0%	12.1%	23.8%
R2	-	6.7%	15.2%	-	9.4%	22.1%
R1	-	-	13.1%	-	-	17.8%

4. Conclusion

We established the counts of articles of a whole country using four different counting methods across a two-level organizational hierarchy. Next to the traditionally used Spearman's correlation coefficient, we propose to measure the differences between counting methods using distances of relative shares of the counts. This method does not erase the quantitative information.

We have observed that the fractional counting method (*F*) and its modification that takes external authors into account with a weight of one half (*R2*) are very similar in both levels. The whole counting method (*W*) differs from the other methods we studied.

Publicly accessible data from the Czech national CRIS was used in the study.

Acknowledgements

We would like to thank Jiří Souček for the fruitful discussions on the problem considered in this contribution. This study was partially supported by the internal development project *Bibliometry and Scientometry* at the Faculty of Art of the Charles University in Prague.

References

1. Bonaccorsi Andrea: Evaluating research on a large scale. The Italian experience. [Presentation] Present and Future National Research Evaluation and the Role of Bibliometrics, Prague, December 11, 2013. Available from URL <http://www.techlib.cz/en/2689-program/>
2. Chudlarský Tomáš, Dvořák Jan: A National CRIS Infrastructure as the Cornerstone of Transparency in the Research Domain. In: *E-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production*: Proceedings of the 11th International Conference on Current Research Information Systems CRIS 2012, Keith G Jeffery, Jan Dvořák (editors), Prague, Czech Republic, June 6-9, 2012. s. 9-17. ISBN 978-80-86742-33-5.
3. Gauffriau Marianne, Olesen Larsen Peder, Maye Isabelle, Roulin-Perriard Anne, Markus von Ins: Comparisons of results of publication counting using different methods. *Scientometrics* 77(1): 147-176 (2008)
4. Lin Chi-Shiou, Huang Mu-Hsuan, Chen Dar-Zen: The influences of counting methods on university rankings based on paper count and citation count. *Journal of Informetrics* 7(3): 611-621 (2013)
5. Olesen Larsen Peder: The state of the art in publication counting. *Scientometrics* 77(2): 235-251 (2008)
6. The Czech Research, Development and Innovation Information System. [database online]. Prague: The Office of The Government of The Czech Republic. Available from URL <http://www.isvav.cz>
7. Snowball Metrics Recipe Book: Their application in The United Kingdom. November 2012. Available from URL <http://www.snowballmetrics.com/wp-content/uploads/snowball-metrics-recipe-book-upd.pdf>