Journal of Informetrics 11 (2017) 622–624

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Correspondence

Comments on the paper "Critical remarks on the Italian assessment exercise", Journal of Informetrics, 11 (2017) and pp. 337–357

1. Introduction

Constructive criticism is a key factor to advance scientific knowledge. The accent here is on the adjective "constructive". It is rather surprising then that almost all criticisms to the evaluation procedures adopted in the two Italian research assessments VQR 2004–2010 and 2011–2014 limit themselves to criticize the procedures without proposing anything new and more apt to the scope. The paper by Franceschni and Miasano (2017) represents no exception to this rule. In the following, we will briefly address the criticisms expressed in the paper.

2. Errors

Before responding to the criticism of Franceschni and Miasano (2017), let us first point out some trivial errors incurred by the authors in describing the main aspects of the VQR 2011–2014, showing a poor reading of the main documents.

1. The authors claim (page 338) that "25% derived from a composition of other indicators (capacity to attract resources, mobility of research staff, internationalization, Ph.D. programs, etc.)." This statement is not true, since the criterion on internationalization is not used in the VQR 2011–2014.
2. The authors affirm (page 339) that "The papers were then submitted to the appropriate GEVs based on the researcher's identification of the more pertinent research areas for them (ANVUR, 2015a, 2015b)." This statement is not true. The papers were submitted to the GEVs based on the SSD (scientific-disciplinary sector) of the researcher to whom the papers are associated, and not based on the researcher's identification of the research areas.
3. The authors claim (page 339) that "The institutions are also subject to potential penalties: (i) in proven cases of plagiarism or fraud". This statement is not true. There are no penalties in this case contrary to what happened in the VQR 2004–2010.
4. The authors affirm (page 339) that "For each SC and issue year, the GEV has to identify the most pertinent journal metric, among the possible ones (see Table 2)." This statement is not true. In fact, the choice of the journal metric to be adopted, together with the choice of the data base between Scopus and ISI Web of Science, has been left to the researchers.
5. The authors claim (Page 340) that: "The choice of the w value is left to the GEV. In general, ANVUR (2015a, 2015b, 2015c) recommends to use relatively higher values for older articles (e.g., those issued in 2011–2012), as they are likely to be mature enough in terms of citation impact." This statement is not true. In fact, in Ref. [2015a] of the paper there is no mention on how to attribute more or less importance to either indicator. References [2015a] and [2015b] of the paper are the evaluation criteria approved by two GEVs: strangely enough, the authors have chosen the only two in which, due to the characteristics of the respective scientific communities, there is a mention toward attributing more importance to the journal metrics. The two documents are also available in English.
6. When the authors discuss the compatibility between peer review and bibliometric analysis, the authors write that "ANVUR (...) claims that the previous VQR 2006–2011 [sic] met this requirement (Bertocchi, Gambardella, Jappelli, Nappi, & Peracchi, 2015)." (page 352). However, the article they cite exclusively focuses on results concerning the Economics area, without any generalization to the other areas.
7. In both the summary and introduction the authors claim that "The penultimate one – i.e., the VQR 2004–2010, which adopted a hybrid evaluation approach based on bibliometric analysis and peer review – suffered heavy criticism at a national and international level. The architecture of the subsequent exercise – i.e., the VQR 2011–2014, still in progress – is partly similar to that of the previous one, except for a few presumed improvements. Nevertheless, this other exercise

DOI of the original article:http://dx.doi.org/10.1016/j.joi.2017.02.005.

http://dx.doi.org/10.1016/j.joi.2017.03.005

is suffering heavy criticism too." The citations that should support this strong statement are for the VQR 2004–2010 four: two of them stem from the Italian website ROARS, and the others are by Italian authors, and thus do not support the claim of "international" criticism.

3. Response to the main criticisms

Let us now move to the three main criticisms raised by Franceschni and Miasano (2017), which will be dealt with according to their order in the paper.

3.1. The low coverage of scientific publications

The authors claim that requesting researchers to present only 2 (or 3 for those affiliated with research centers) publications in four years does not allow to capture any of these aspects: productivity, average quality/impact and excellence. They say that the only captured feature would be "researchers decency".

To support their statement they propose a trivial, ad hoc example showing that under certain invented conditions more than 90% of the researchers would be able to propose for evaluation at least two papers reaching the merit classes of excellent or good (classes A and B).

To show the inconsistency of this exercise suffices to say that the VQR 2011–2014 results published on February 22nd, 2017, show that only 32.6% of the proposed products were assessed as excellent, and a total of 63.4 are in the two classes A or B. The authors also propose to evaluate all the products published in the VQR period, at least for the bibliometric scientific areas. This is contrary to the principle of treating all areas equally and, by the way, has never been adopted in RAEs or REF assessments in the United Kingdom, precisely those that the critics of ANVUR cite as the example to follow.

3.2. Incorrect and anachronistic use of journal metrics

The authors criticize the use of the journal impact as one of the two bibliometric indicators. They cite several criticisms directed toward the Impact Factor as a proxy of the article impact. We do not disagree on this in general, and would never recommend using the journal impact when evaluating the career of individual researchers. However, the story is significantly different here: first, the VQR is evaluating significantly large communities of researchers, second, to do so, it has to evaluate relatively recent publications (as an example, articles published in 2014 based on citations received up to February 2016), and third, it uses two bibliometric indicators rather than one.

The authors say that the experts in bibliometry suggest using only the number of citations as a proxy of the article impact and, also, point to the fact that the Impact Factor can be easily manipulated. Referring to what has been suggested and published in the past as the "Truth" has always prevented science to progress, and it represents a conservative attitude that should not belong to scientists. Moreover, in less than two years the number of citations in any discipline is far from reaching its steady state, and cannot be used alone. By the way, the number is generally so low that its manipulation through citation stacking or inappropriate self-citation is much easier than manipulating the journal impact. Using two indicators is generally more robust than using one, and given the time constraints of the VQR assessment (very "young" publications) appears indeed the best way to go.

Referring to the correlation between journal impact and number of citations with the impact or "quality" of an article, we invite the authors to read carefully the results of a deep study performed on the correlation between bibliometric indicators and Panels peer review in the recent REF (Wilsdon et al., 2015). This study indicates that in almost all disciplines and for both recent and old articles the SJR journal impact indicator (in REF only the Scopus data base was used) is the best predictor of the panels' classification, better than the citations count.

The authors say that "A less debatable solution could be complementing Ci with the so-called altmetrics–i.e., alternative metrics related to individual papers, such as the count of the number of views, downloads, blogs, media coverage, etc. (Bornmann, 2014; Costas, Zahedi, & Wouters, 2015; Thelwall, Haustein, Larivière, & Sugimoto, 2013); (. . .)" (page 343). This is a rather surprising proposal, since it is well known that such metrics are easily prone to manipulation. Moreover, in the already mentioned REF Report (Wilsdon et al., 2015), almost no correlation is found between the Panels' classification and that induced by Altmetrics indicators. Indeed, the authors continue by saying "however, it is still necessary to investigate the potential of altmetrics and their benefits and disadvantages for measuring impact."

It is also worth noticing that the two criticisms discussed above are intrinsically in contradiction with each other: in fact in the international literature (see for instance Siversten, 2016) there are examples of evaluation systems based on the analysis of the whole scientific production, but in those cases the evaluators make extensive use of journal metrics.

Last but not least, since research assessments influence researchers' behavior, the use of the journal impact as one of the two bibliometric indicators constitutes a clear invitation to (especially young) authors to privilege quality over quantity, and to submit their papers to the best journals, which are undoubtedly identified by high journal impact metrics.

3.3. Conceptually misleading criteria for normalizing bibliometric indicators

To support their criticisms, the authors treasure and expand a criticism of the ROARS web site on the "mistake" of adding percentiles. An answer to this criticism has been published by the web site "La Voce".[1]

In essence, combining the cumulative distributions functions of the two indicators as is done in the VQR 2011–2014 boils down to using one of the many possible nonlinear transformations that modify the values of the two indicators allowing the transformed variables to lie both in a normalized interval (0,1) within which it is easier to perform a partition in merit classes. Adding the values of the two transformed variables is nothing different than adding nondimensional numbers.

It has been shown[2] that the number of possible classification inconsistencies of such nonlinear transformation is statistically irrelevant, and much lower than those produced by an alternative proposed by the ROARS web site.

Finally, the authors claim (page 351) that "Also, even if (erroneously) deciding to combine Ci and Ji, we believe that this could be done avoiding dubious transformations/normalizations that alter the scales of the initial data." We would be eager to see the authors' elaboration on better, alternative ways of doing the combination. Unfortunately, as already pointed out, constructive criticism does not pertain to this paper.

References

ANVUR (2015a). Valutazione della qualità della ricerca 2011-2014 (VQR 2011–2014) – Bando di partecipazione, Roma, http://www.anvur.org/attachments/article/825/Bando%20VQR%202011-2014_secon~.pdf [retrieved on July 2016].

ANVUR (2015b). Valutazione della qualità della ricerca 2011-2014 (VQR 2011–2014) – Criteri per la valutazione dei prodotti di ricerca Gruppo di Esperti della Valutazione dell'Area 01 Scienze Matematiche e Informatiche (GEV01), Roma, https://www.anvur.it/attachments/article/842/Criteri%20GEV%2001.pdf [retrieved on July 2016].

ANVUR (2015c). Valutazione della qualità della ricerca 2011-2014 (VQR 2011–2014) – Criteri per la valutazione dei prodotti di ricerca Gruppo di Esperti della Valutazione dell'Area Ingegneria Civile (GEV08b), Roma, http://www.anvur.it/attachments/article/850/Criteri%20VQR%202011-2014%20GEV~.pdf [retrieved on July 2016].

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., & Peracchi, F. (2015). Bibliometric Evaluation vs. Informed Peer Review: Evidence from Italy. *Research Policy*, *44*, 451–466.

Franceschni, F., & Miasano, D. (2017). Critical remarks on the Italian assessment exercise. *Journal of Informetrics*, *11*, 337–357.

Siversten, G. (2016). Publication based funding: the Norwegian model. In M. Ochsner, S. E. Hug, & H. D. Daniel (Eds.), *Research assessment in the humanities – Towards criteria and procedures* (p. 201). Springer. http://dx.doi.org/10.1007/978-3-319-29016-4p [ISBN 978-3-319-29014-0 ISBN 978-3-319-29016-4 (eBook)]

Wilsdon J. et al. (2015). The metric tide: report of the independent review of the role of metrics in research assessment and management. doi:10.13140/RG.2.1.4929.1363.

Sergio Benedetto
Politecnico di Torino, Torino, Italy

Daniele Checchi
Andrea Graziosi*
Marco Malgarini
ANVUR, Rome, Italy

* Corresponding author.
E-mail address: andrea.graziosi@anvur.it (A. Graziosi)
Available online 8 May 2017

[1] See http://www.lavoce.info/archives/41481/valutazione-della-ricerca-quellalgoritmo-e-affidabile/ and for details http://www.lavoce.info/wp-content/uploads/2016/06/algoritmo-analisi-empirica.pdf.

[2] See http://www.lavoce.info/wp-content/uploads/2016/06/algoritmo-analisi-empirica.pdf.