Looking beyond the Italian VQR 2004-2010: Improving the Bibliometric Evaluation of Research

Alberto Anfossi[1,2], Alberto Ciolfi[1], Filippo Costa[1,3]

[1]albertofrancesco.anfossi@anvur.it, [1] alberto.ciolfi@anvur.it, [1,3]filippo.costa@anvur.it
[1]ANVUR, Via Ippolito Nievo 35, 00153 Roma (Italy)
[2] Compagnia di San Paolo Sistema Torino, Piazza Bernini 5, Torino (Italy)
[3]Dipartimento Ingegneria dell'Informazione, Università di Pisa, Pisa (Italy)

Abstract

In the recent Italian Evaluation of Research Quality exercise for the period 2004-2010 (VQR), promoted by the Italian Ministry of Education and carried by the National Agency for Research Evaluation (ANVUR), metrics were massively employed. The use of Impact Factor or article citations (or both) are usually considered a powerful tool for supporting the peer review process but the replacement of the latter with an automatic evaluation tool has been always considered risky. Here we propose a possible prescription to overcome some limitations of the bibliometric evaluation carried out within the context of the VQR, while, at the same time, keeping the main distinctive features of the evaluation approach unchanged, namely, a simple evaluation tool based on the combined use of the CIT and IF variables While maintaining the basic elements of the previous algorithm unchanged and keeping the method simple and feasible on a large scale, we argue that the main flaws and limitations can be overcome.

Conference Topic

University Policy and Institutional Rankings

Introduction

The most popular European national research evaluation is the Research Assessment Exercise (RAE) in Great Britain, which started in 1986 and has been replaced, in 2013, by a new exercise - Research Excellence Framework (REF) - where citation-based metrics were employed to inform and supplement Peer Review (PR) evaluation.

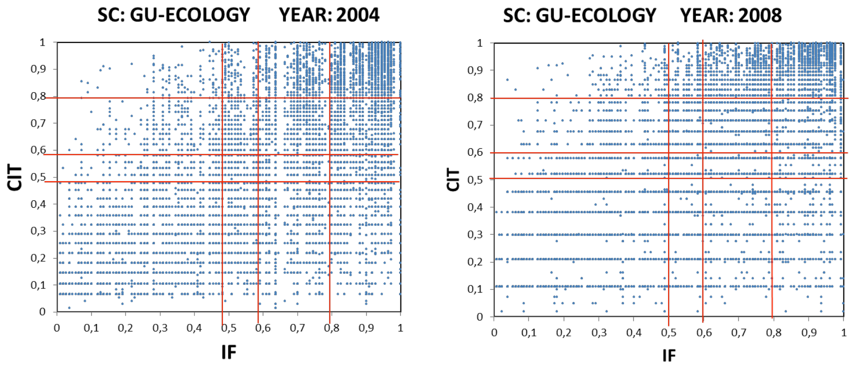
In Italy, the first evaluation exercise was carried out in 2005 by the CIVR with reference to the period 2001-2003 (VTR). The VTR was fully based on the PR evaluation method, each submitted research product being assessed by a pool of experts (Minelli et al., 2008). Some studies (Reale et al. 2007; Abramo et al., 2009; Franceschet et al., 2011) analysed the outputs of the VTR comparing peer quality opinions on papers with metrics based on the Impact Factor of the journals publishing the papers, concluding that the two evaluation methods significantly overlap. However, comparison of PR and bibliometric evaluation methodologies has been largely debated in the literature (Barker, 2007; Moed, 2006; Harnad, 2009; Norris et al. 2003, Butler et al., 2003; Bence et al., 2009, Asknes, et al. 2004) with not always concordant outcomes. The use of Impact Factor or article citations or both are usually considered a powerful tool for supporting the PR process but the replacement of the latter with an automatic evaluation tool has been always considered risky.

In the recent Italian Evaluation of Research Quality exercise for the period 2004-2010 (VQR), promoted by the Italian Ministry of Education and carried by the National Agency for Research Evaluation (ANVUR), metrics were massively employed. Around 200.000 research outputs, mainly journal articles or reviews (both called 'paper' in the following), were evaluated, of which 46,5% by use of a bibliometric algorithm (Ancaiani et al., 2014).

Bibliometric Evaluation in the VQR 2004-2010

According to the Ministerial Decree number 17 of July 15th, 2011 promoting the VQR, each paper submitted for evaluation is classified in one of four possible classes of merit, defined as follows: "Excellent" (E): when the paper falls in the top 20% of the world production in a given Subject Category (SC) and in a given year; "Good" (G): when the paper falls in the following 20%; "Acceptable" (A): when a paper falls in the following 10%; "Limited" (L): when a paper falls in the bottom 50%.

In bibliometric areas, the strategy to assign a paper to a given class was based on the combined use of two variables: (i) CIT: number of citations collected by the paper up to December 31st, 2011 and (ii) IF: Impact Factor (or equivalent indexes) of the Journal in the year of publication of the paper. Each paper was submitted by the Organization (i.e. universities or public research bodies) and then uniquely assigned to a thematic evaluation panel (called Group of Experts for Evaluation, GEV) and to a Subject Category (SC), or All Journal Science Category (ASJC) as defined by ISI Web of Knowledge® or Scopus databases, respectively. In each SC/ASJC and for each year it is possible to construct the cumulative distribution function of the two variables[1], thus assigning to each paper its CIT and IF percentile. In the VQR three thresholds for both IF and CIT were defined to distinguish among the four classes specified in the Ministerial Decree. In the space spanned by IF (x-axis) and CIT (y-axis) it is therefore possible to focus on the region Q = [0,1]x[0,1] and plot the publications distribution defined on the basis of their CIT and IF percentile (Fig. 1, where each dot represents a paper denoted by its CIT and IF percentile). Each GEV had the freedom to assign the "off-diagonal" sub-squares (blocks) of the whole region Q, identified by the intersection of the "threshold segments", to a class of merit, thus completing the automatic phase of the evaluation process. Indeed, the diagonal blocks were quite naturally assigned to the four classes: the intersection of "top 20% for CIT" with "top 20% for IF" was straightforwardly associated to the "Excellent" class of merit, and so on. The choice to assign an off-diagonal block to a class was performed according to basically two drivers: first and foremost, the qualitative insight of the GEV on the scientific field and its publication practices (e. g. lag in citations, etc.) and second, the attempt to keep the final assignment as close as possible to the world distribution D specified in the Ministerial Decree.



(a)

(b)

Figure 1. Papers distribution in a given SC and in two different years.

Such an approach showed some limitations that we summarize schematically:

[1] CIT: by ordering the total number of paper published in that SC and in that year in decreasing order from the highest to the lowest cited; IF: by ordering the Journals belonging to that SC in that year in decreasing order from the highest impact factor to the lowest.

Absence of "micro calibration": all the GEVs except for GEV 02 (Physical sciences) chose a single assignment (typically, one for years 2004-2008 and one for years 2009-2010), i.e., association of blocks to classes of merit, and did not went through a micro calibration at the level of the single SC and single year. Considering that: (i) for each GEV the number of relevant SC[2] was typically of the order of 50 and (ii) the distribution of the papers in Q was totally not uniform and invariant, rather, it varied significantly from one SC to another and form one year to another (see for instance Fig. 4). The absence of a micro calibration affected the possibility to comply with the distribution D punctually (and not only on average).

Structure of the blocks: (i) as showed in Figure 1 the threshold segments are parallel to the x/y axis. This is not convenient given the discrete nature of the two variables under consideration. (i) It can be easily noted in the plot that the points (corresponding to papers) are distributed in rows, according for instance to the limited number of journals present in a SC. As a consequence, the evaluation may not be robust enough, in the sense that a slight perturbation in the thresholds can modify the final class allocation for whole set of papers. (ii) It is quite hard, if not impossible, to comply with the distribution D by leveraging on the sole degrees of freedom given by the possibility to assign the off-diagonal blocks to a final class of merit. In other words, the constraint of assigning to a single class an entire block is too binding and tends to move too many paper from one class of merit to an another. (iii) The degrees of freedom are even reduced by the need to avoid that two non-adjacent classes of merit (say, "Good" and "Limited") can be adjacent in Q, as shown in Fig. 2.

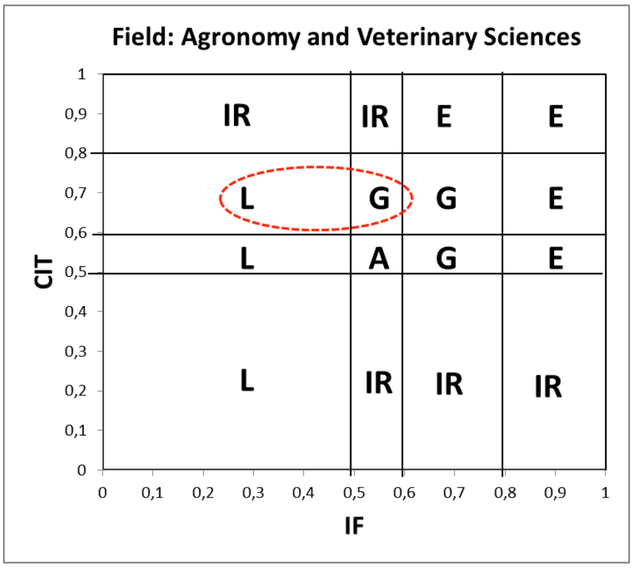


Figure 2. Algorithm used to evaluate research products in the Agronomy and Veterinary Science field: two non-adjacent classes of merit are adjacent in Q (red circle). "IR" indicates products that are lefd undecided by the algorithm and are eventually evaluated by peer review.

The new proposed approach

In the following we discuss a possible prescription to overcome these limitations while, at the same time, keeping the main distinctive features of the evaluation approach unchanged, namely, a simple evaluation tool based on the combined use of the CIT and IF variables. This can be done through the use of three diagonal segments with generic slope (Fig. 3).

[2] By relevant we mean that a great number (more than one hundred) of papers to be evaluated fell under that SC.

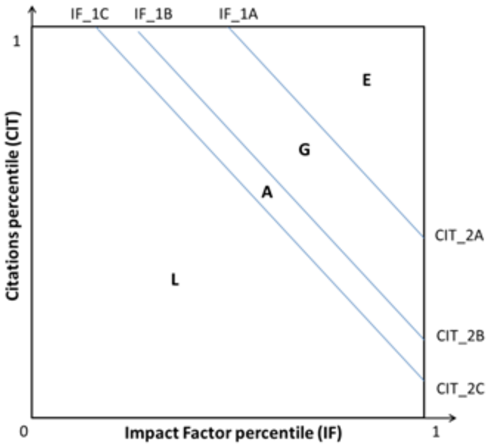
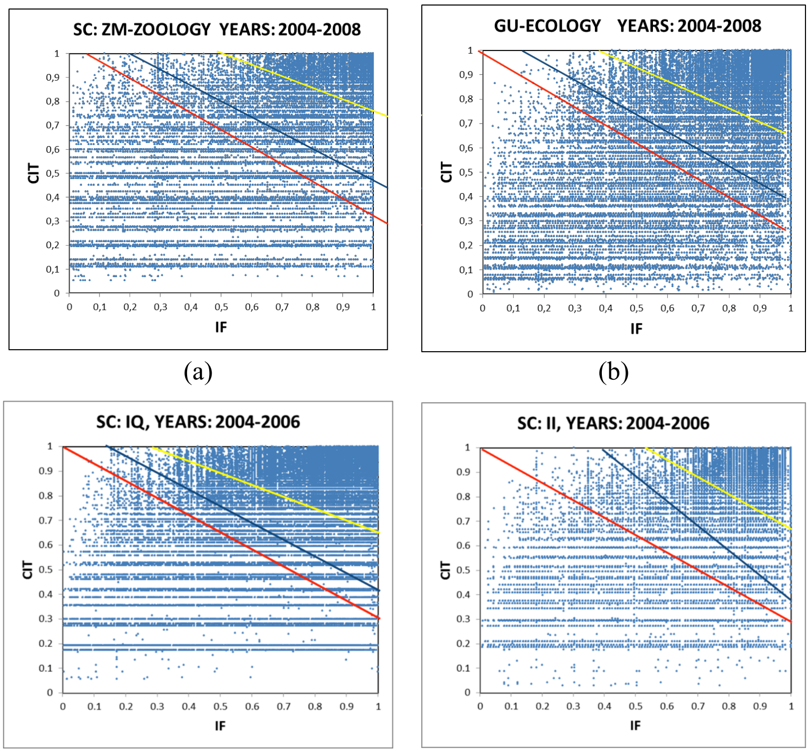


Figure 3. New prescription for combining the CIT and IF variables.

Such a new prescription builds upon three main pillars:

1. The segments identifying the thresholds are now drawn as a linear combination of the CIT/IF thresholds, thus being diagonal and no more parallel to the axes;
2. CIT/IF thresholds do not have to separately satisfy the 20-20-10-50 distribution;
3. The calibration, i.e. where to position the diagonal segments in Q in order to comply with the distribution D, is now performed at the micro level of each SC, for each year and for each GEV (according to general guidelines provided by the GEV itself and based on GEV's proficiency in the specific scientific field);

This would in turn guarantee the effectiveness and the simplicity of the whole process. In Figure 4 we apply this method to some SCs.



(a)

(b)

(c)

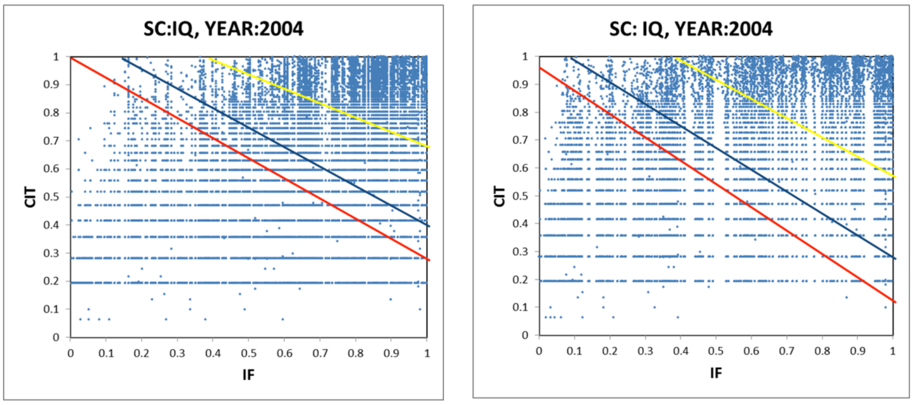
(d)

Figure 4. The application of the new algorithm in various SC and years. IQ stands for Electrical and Electronic Engineering, II stands for Engineering Chemical. The straight lines indicate the thresholds for the four classes of merit.

Comments and future developments

This new approach is characterized by a rather marked level of freedom in the choice of the position of the diagonal segments (or, equivalently, of the CIT/IF thresholds). Indeed, there is typically more than one choice that satisfies the distribution D. On the other hand one could impose additional constraints, such as for instance the parallelism between segments, based on additional empiric work and on scientific validation of the procedure (eg. by a PR comparison of the evaluation outcomes). Furthermore, such a freedom might be exploited to accommodate GEV's requirements. For instance, it would be possible to give more relevance to one of the two dimensions (IF, CIT) depending on, say, the year of publication or the citation praxes of specific disciplines (Mathematics vs Medicine being a paradigmatic example).

A significant possibility to further improve the accuracy of the method we discussed comes from a different definition of the cumulative distribution function for the IF variable. Instead of considering the number of journals belonging to a SC, one could consider the number of items (papers) published in the SC (in a given year). Actually, it is common that some journals host few thousands of items per year while other few tens or units. This induces a possible distortion that is quite evident in the plots shown below. As an example, In Figure 5 we analyze the distribution of the SC Electrical and Electronic Engineering in 2004. The distribution of the papers according to the IF and CIT percentile are depicted both considering only the number of journals in the calculation of the IF percentile and by considering also the number of item for each journal. The distributions are subdivided with different lines in order to obtain the target percentages D. It is evident that the equation of the lines is substantially different to guarantee the same final result. It is worth underlining that the lines used to subdivide the distribution reported in Figure 5(a) would result in very different percentages if applied to the distribution in Figure 5(b).



(a)

(b)

Figure 5. Distribution of the papers according to the number of journals and papers. (a) IF percentile calculated based on the number of journals (b) the IF percentile is calculated considering the number of items. The distributions are subdivided with lines in order to obtain the target percentage D.

Finally, it would be possible to improve also the CIT dimension by overcoming the concept of SC as "reference set" and move on to clustering strategies based on semantic or on citation networks. This would be more rigorous and meaningful considering the existence of a great number of journals that publish very different subjects, but it would come with a significant enhancement of the complexity of the evaluation procedure, probably not feasible for the numbers implied by a national formal evaluation, at the moment.

Results obtained so far are already highly informative about the existing strength and weakness of the Italian University research system, and provide reliable input for policy interventions. Our proposal is intended to further improve the mix of peer review and bibliometric methods through a more precise calibration of the biblio(metrics) used.

The output turns out to be rather general, thus being applicable to other national assessments based on bibliometric analysis.

References

Ancaiani et al. (2014). Evaluating Scientific Research in Italy: the 2004-2010 Research Evaluation Exercise. *Research Evaluation*, *submitted*

Minelli, E., Rebora, G., Turri, M. (2008). The structure and significance of the Italian research assessment exercise (VTR). *European Universities in Transition*. Edward Elgar Publishing.

Reale, E., Barbara, A., & Costantini, A. (2007). Peer review for the evaluation of academic research: lessons from the Italian experience. *Research Evaluation*, *16*(3), 216-228.

Franceschet, M., & Costantini, A. (2011). The first Italian research assessment exercise: A bibliometric perspective. *Journal of Informetrics*, *5*(2), 275-291.

Abramo, G., D'Angelo, C. A., & Pugini, F. (2008). The measurement of Italian universities' research productivity by a non parametric-bibliometric methodology. *Scientometrics*, *76*(2), 225-244.

Barker, K. (2007). The UK Research Assessment Exercise: the evolution of a national research evaluation system. *Research Evaluation*, *16*(1), 3-12.

Moed, H. F. (2006). Citation analysis in research evaluation (Vol. 9). Springer.

Harnad, S. (2009). Open access scientometrics and the UK Research Assessment Exercise, *Scientometrics*, *79*(1), 147-156

Norris, M., & Oppenheim, C. (2003). Citation counts and the Research Assessment Exercise V: Archaeology and the 2001 RAE. *Journal of Documentation*, *59*(6), 709-730.

Butler, L., & McAllister, I. (2009). Metrics or peer review? Evaluating the 2001 UK Research assessment exercise in political science. *Political Studies Review*, *7*(1), 3-17.

Bence, V., & Oppenheim, C. (2004). The influence of peer review on the research assessment exercise. *Journal of Information Science*, *30*(4), 347-368.

Asknes, D. W., Taxt, R. E. (2004). Peer reviews and bibliometric indicators: a comparative study at a Norwegian university. *Research evaluation*, *13*(1), 33–41.