Scientometrics (2016) 107:671–683
DOI 10.1007/s11192-016-1882-9

Large-scale assessment of research outputs through a weighted combination of bibliometric indicators

Alberto Anfossi[1,2] · Alberto Ciolfi[1] · Filippo Costa[1,3] · Giorgio Parisi[4] · Sergio Benedetto[1]

Received: 15 July 2015 / Published online: 15 February 2016

Abstract The paper describes a method to combine the information on the number of citations and the relevance of the publishing journal (as measured by the Impact Factor or similar impact indicators) of a publication to rank it with respect to the world scientific production in the specific subfield. The linear or non-linear combination of the two indicators is represented on the scatter plot of the papers in the specific subfield in order to immediately visualize the effect of a change in weights. The final rank of the papers is therefore obtained by partitioning the two-dimensional space through linear or higher order curves. The procedure is intuitive and versatile since it allows, after adjusting few parameters, an automatic and calibrated assessment at the level of the subfield. The derived evaluation is homogeneous among different scientific domains and can be used to address

This paper is the development of the proceedings paper presented at the 15th Int. Conference on Scientometrics & Informetrics (ISSI) entitled *Looking beyond the Italian VQR 2004–2010: Improving the Bibliometric Evaluation of Research* (Conference Topic: University policy and institutional rankings).

✉ Alberto Ciolfi
alberto.ciolfi@anvur.it

Alberto Anfossi
albertofrancesco.anfossi@anvur.it

Filippo Costa
filippo.costa@anvur.it

Giorgio Parisi
giorgio.parisi@roma1.infn.it

Sergio Benedetto
sergio.benedetto@anvur.it

[1] National Agency for the Evaluation of Universities and Research Institutes (ANVUR), Via Ippolito Nievo 35, 00153 Rome, Italy

[2] Compagnia di San Paolo Sistema Torino, Piazza Bernini 5, Turin, Italy

[3] Dipartimento Ingegneria dell'Informazione, Università di Pisa, Pisa, Italy

[4] Università "La Sapienza" di Roma, Piazzale Aldo Moro 5, 00185 Rome, Italy

the quality of research at the departmental (or higher) levels of aggregation. We apply this method, that is designed to be feasible on a scale typical of a national evaluation exercise and to be effective in terms of cost and time, to some instances of the Thomson Reuters Web of Science database and discuss the results in view of what was done recently in Italy for the Evaluation of Research Quality exercise 2004–2010. We show how the main limitations of the bibliometric methodology used in that context can be easily overcome.

Keywords Bibliometric evaluation · Institutional rankings · Evaluation processes · University policy

Introduction

Ex-post research evaluation, assessing its outputs (typically, publications) and impact, arises once research has been completed, published and a due time has elapsed. The main evaluation methodologies are peer review (which significantly differs from the peer review (PR) process leading to publication of the research results), often in the form of panel rankings, and bibliometric indicators (Moed et al. 2005). The PR process that scholarly publications undergo may be interpreted as an indication of "quality". However, once ex-post impact evaluation comes into play, the reliability of quality assessments based on a pure PR approach is strongly dependent on several factors that influence the accuracy, fairness and timeliness of the judgment. To cite one element for all, when the scientific community underlying a field includes few researchers, avoiding conflicts of interest becomes very difficult, if hardly possible.

To reduce typical problems affecting the PR process the number of peer reviewers *per paper* could be increased and, in parallel, the number of distinct single reviewers be limited (every reviewer has his/her own metric scale and this introduces an additional error). This strategy, however, affects the timeliness of the overall evaluation process, also increasing total costs. The replacement of a pure PR process with an automatic evaluation tool based on bibliometrics has been the object of debate and criticism (HEFCE 2011; Moed 2009). For this reason in most of the recent research evaluation exercises, like for instance in the United Kingdom (UK) Research Excellence Framework (REF) as well as in the Australian Research Quality Framework (RQF), citation-based metrics were employed to inform and supplement PR evaluation (Butler 2008; Oppenheim 2008). However, some studies have shown that the previous UK Research Assessment Exercise (RAE) outputs were characterized by a very high correlation with citation analyses (Smith and Eysenck 2002), even if deviant cases clearly exist (Warner 2000). Other studies (Franceschet and Costantini 2010; Reale et al. 2007), based on the first Italian research assessment exercise, the VTR 2001–2003, demonstrated that metrics based on the Impact Factor (IF) of the journals publishing the papers significantly overlap with the judgements of the peers. However, not all studies (Aksnes and Taxt 2004, 2004; Barker 2007; Bence and Oppenheim 2004) have provided converging outcomes.

On a different aspect, a research quality assessment entirely based on bibliometric data would represent a huge advantage in terms of costs, time, repeatability of the measurements and standardization of the entire process, even in view of the comparison among different scientific fields (Abramo and D'Angelo 2011; Abramo et al. 2011).

A first bibliometric evaluation can be performed with simple methods involving merely the count of publications as done in Australia in the mid-1990s. This method is clearly controversial since it can lead to an increment of papers on low-impact journals (Butler 2003). The use of a single indicator like the journal relevance or the citations count may not be accurate enough to assess the quality of a publication (Eyre-Walker and Stoletzki 2013) for different reasons. The use of journal metrics alone has been discredited for the evaluation of single research papers (Alberts 2013; Bladek 2014; Seglen 1997). At the same time, the use of the sole citation count may not be an appropriate indicator of impact in those cases where the paper is too young, field normalizations are not taken into account or if autocitations strongly affect the final evaluation.

In this paper we discuss a combined use of journal metrics and article citations as a potentially powerful tool for compensating intrinsic flaws of bibliometric analysis and thus for integrating the PR process in the context of a massive and aggregated assessment of research quality, such the ones performed in recent years by an increasing number of public agencies at national level. The criticism usually raised on the use of journal-based metrics to measure the quality of a single researcher or a single paper should be put into the following perspective: first, in what we are discussing here the information provided by the journal metric is combined with citation count; second, the outputs generated by these evaluation processes are typically the result of field-dependent averages over several hundreds of single scores and/or rankings of departments/institutes (or similar aggregations).

In general, there can be different methods to combine the information on the quality of the journal with the impact of the publication to work out an indicator (better, a proxy) of quality. The rank of the journal in which the paper is published (as measured by the IF or similar variables (Bergstrom and West 2008; Falagas et al. 2008)) and its citation count (or its refinements (Glänzel and Thijs 2004)) can be weighted giving more importance either to the former or to the latter. One can develop an algorithm which combines those parameters automatically, as it happens for instance with the principal component analysis (PCA) (Bollen et al. 2009), or can find a way to directly choose the weight. Our view, also based on our experience of the recent Italian Evaluation of Research Quality exercise for the period 2004–2010 (VQR) (Ancaiani et al. 2015), is that this degree of freedom should be left to the panel of experts because of different habits of scientific communities and because of different significance of citation count when applied to recent papers. A possibility for doing that is to quantize the IF percentile and the citation percentile into quartiles, as done in the Italian VQR, and thus partition the bi-dimensional space spanned by IF and CIT containing the papers into blocks. In the VQR, at each of these blocks a specific evaluation by the panel was then assigned. However, such an approach showed some limitations in terms of flexibility and easiness of calibration. Motivated by these arguments, a new approach to the design of a simple algorithm capable of calibration at the level of scientific areas and subfields is described in the paper.

The proposed approach

We propose simple tool suitable to large-scale evaluation exercises where the quality of the research output has to be assessed not at the individual level but at some level of aggregation (departments, institutes, universities, etc.; in the following, we will refer in general

to "universities"). In addition to a few scientific and methodological requirements, such a tool has to meet a minimal set of practical constraints:

- it has to be simple enough and intuitive to be communicated to—and managed by—a broad array of disciplines and panel members with different backgrounds;
- it has to be cost effective when a large number of publications (even of the order of hundreds of thousands) is considered;
- it must be time-effective and fully predictable.

We restrict ourselves to "bibliometric" fields of sciences and consider the following situation: each university would submit a given number of papers to be evaluated, specifying for each of them the most appropriate Subject Category, or All Journal Science Category[1] (denoted generically SC hereafter) among those associated to the publishing journal, and the most qualified thematic evaluation panel.

Two variables are associated to each paper: the citation count (CIT), i.e., the number of citations collected by the paper up to a given point in time) and a journal-based metric (IF for simplicity hereafter), i.e., the Impact Factor—or similar indices (Bergstrom and West 2008; Fersht 2009; Setti 2013)—of the Journal in the year of publication of the paper, which we will consider both in terms of their percentile representation. Indeed, for each SC—and for each year—it is possible to construct the cumulative distribution function (CDF) of these two variables,[2] thus assigning to each paper its CIT and IF percentile. In the space spanned by IF (X-axis) and CIT (Y-axis) it is then possible to focus on the subspace $Q = [0,1] \times [0,1]$ and plot the publications distribution (see for instance Fig. 2, where each dot represents a paper identified by its CIT and IF percentile).

Building on the experience of the Italian Evaluation of Research Quality exercise for the period 2004–2010 (VQR), carried by the National Agency for Research Evaluation (ANVUR), we propose an improved algorithm for accomplishing a massive evaluation of research products. The VQR assessed around 200.000 research outputs, mainly journal articles or reviews, of which 46.5 % by use of a bibliometric algorithm (Ancaiani et al. 2015). In the VQR, the region Q was partitioned into blocks by using discrete thresholds (top 20 %, top 40 %, top 50 %) for both citations and journal metrics *separately*. The diagonal blocks were quite naturally assigned to the four classes of merit (see Fig. 1a): the intersection of "top 20 % for CIT" with "top 20 % for IF" was associated to the "Excellent" class of merit, and so on for the subsequent classes ("Good", "Acceptable", "Limited"). Each panel had the freedom to assign *autonomously* the "off-diagonal" blocks of the whole region Q to a class of merit, thus completing the automatic phase of the evaluation process. The choice to assign an off-diagonal block to a class was performed according to two drivers: first and foremost, the qualitative insight of the thematic panel on the scientific field and its publication practices (e.g. lag in citations, etc.) and second, the attempt to keep the ex-ante probability of assigning a paper to a class of merit as close as possible to the world distribution specified in the Ministerial Decree.

Here a still simple but more convenient partitioning of the above-mentioned space Q is proposed. It is straightforward to note that a paper located at the top-right corner of Q is of high quality in terms of both variables under consideration, while a paper lying at the

[1] As defined by Web of Science by Thomson Reuters® or Scopus by Elsevier® databases, respectively.

[2] 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 IF to the lowest. This is not the only strategy to build the cumulative distribution function for the IF variable, as we will discuss later in the paper.

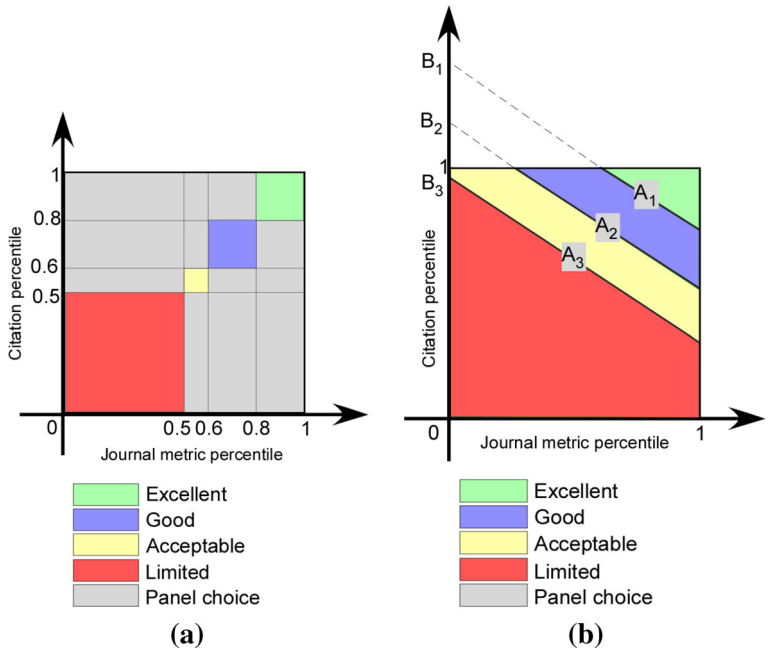


Fig. 1 Combination of the CIT and IF variables to obtain a weighted evaluation of papers. **a** The method employed in the past Italian Research Assessment Exercise (VQR 2004–2010) and **b** the method proposed here

bottom-left of Q is performing poorly. Following this intuition we can partition Q into $n + 1$ sub regions by drawing n curves or "thresholds". In principle, the *n-th* threshold can be identified by setting equal to zero a generic function $f_n(CIT, IF)$ of the two variables CIT and IF:

$$f_n(\mathrm{CIT}, \mathrm{IF}) = \mathrm{Const}_n + a_{1n} \cdot \mathrm{CIT} + a_{2n} \cdot \mathrm{IF} + a_{3n} \cdot \mathrm{CIT} \cdot \mathrm{IF} + a_{4n} \cdot \mathrm{CIT}^2 + a_{5n}\mathrm{IF}^2 + \cdots; \tag{1}$$

satisfying basic requirements of Pareto dominance. The high degree of flexibility of the approach (1) to thresholds would allow shaping them according to a defined calibration method (say a high-quality peer review of a different bunch of papers). However, to keep the discussion as simple as possible we restrict ourselves to the linear truncation of (1) and, moreover, we consider parallel thresholds with slope A, i.e.,

$$\begin{array}{ll} a_{in} = 0 & \forall i > 2; \\ a_{2n}/a_{1n} = A_n = A & \forall n; \end{array} \tag{2}$$

We thus end up with the following set of equations (where $B_n = -\mathrm{Cost}_n/a_{1n}$):

$$\mathrm{CIT} = A \cdot \mathrm{IF} + B_n \tag{3}$$

In Fig. 1b we draw the case for $n = 3$ where we can choose the free parameters, namely the slope A and the three coefficients B_1, B_2, B_3, to satisfy a given density distribution (fraction of the total number of papers in Q) in the four sub regions that are thus identified.

Such a density distribution D represents straightforwardly the ex-ante probability for a paper submitted for the evaluation to fall in one of the four sub regions. The latter and the distribution D could be for instance defined as follows:

1. "Excellent" (E): the paper falls in the top 20 % of the world production in a given Subject Category (SC) and in a given year;
2. "Good" (G): the paper falls in the following 20 %;
3. "Acceptable" (A): the paper falls in the following 10 %;
4. "Limited" (L): the paper falls in the bottom 50 %.

The free parameters are then fixed by imposing that 20 % of the total number of papers in Q fall into the region E of Fig. 1b, and similarly for the regions G, A and L.

It is important to note that this approach is characterized by a rather marked degree of freedom in the choice of the free parameters. Indeed, there is typically more than one choice that satisfies the distribution D and this can be exploited to impose additional constraints based on empirical considerations. In our view the choice of the slope of the lines should be left to the panels since it imposes the relative weight of citations and journal metrics. Clearly, a slope greater than one of the threshold lines means that the evaluation of the publications is carried out by giving more relevance to the journal metric than to the citation count, whereas a slope less than one means that citations play a greater role in defining the final evaluation of the paper in the specific subfield. It is therefore possible to assign more relevance to one of the two dimensions (IF, CIT) depending on, say, the year of publication or the citation habits of specific disciplines (Mathematics vs Medicine being a paradigmatic example). We finally remark that the algorithm, independently from the autonomous choices of each panel of experts, ensures that for each specific subfield and year the fraction of papers falling in each class of merit (or region) is always the same when applied to the whole set of publications in a bibliometric data base. As a consequence, the algorithm is consistently calibrated and comparisons among different scientific subfields and/or fields are feasible.

To summarize, the tool we are proposing builds upon few main pillars:

1. A normalized distribution of the two variables CIT and IF, and the intuitive representation of the papers in the (CIT, IF) space as a scatter plot;
2. The partition of the whole space Q into regions by drawing thresholds as weighted linear combinations of the CIT and IF variables;
3. The calibration, i.e., where to position the thresholds in Q to comply with a given a priori constraint in terms of a probability distribution D, performed at the micro level of each SC, for each year and for each thematic panel of experts (according to general guidelines provided by the panel itself and based on its proficiency in the specific scientific field).

The costs involved in implementing this tool reduce to the acquisition of few data for each paper present in the chosen database(s), being therefore far less expensive (at least one order of magnitude less) than a PR approach where the work of the reviewers is remunerated. As far as timing is concerned, the most demanding part of the evaluation is the pinning down of the thresholds (in our example, three for each SC and for each year). Once the thresholds are defined, the evaluation is almost immediate and basically independent from the number of papers to assess.

Compared to the former algorithm adopted in VQR 2004–2010, in which the stringent deadlines did not allow for a sufficient time to devise calibration procedures applicable at

all levels of aggregation, the present approach allows to cope with success with several issues.

Absence of "micro calibration": all panels[3] chose a single global 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 pursue a micro calibration at the level of the single SC and single year. Considering that: (1) for each thematic panel of experts the number of relevant SC[4] was typically of the order of 50 and (2) the distribution of the papers in Q could change from one SC to another and from one year to another (see for instance Fig. 5). The absence of a micro calibration affected the possibility to comply with the distribution D punctually (and not only on average) and made it improper to compare results among different scientific fields.

Structure of the blocks: (1) as showed in Fig. 1a the threshold segments were parallel to the XY axis. This is not convenient given the discrete nature of the two variables under consideration. (2) It can be easily noted in the scatter plots that the points (corresponding to papers) are distributed in rows/columns, 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 sets of papers. (3) 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 papers from one class of merit to another. (4) "corner effects": two papers characterized by a slight difference in one variable and a marked difference in the other may end into two different classes of merit.

All the aforementioned limitations are overcome by the method proposed in this paper since, as already said, the calibration can be performed at the level of each SC and for each year (thus complying with a given probability distribution and allowing for a consistent comparison of results among disciplines) and Q is partitioned by means of thresholds that are smooth combinations of the CIT and IF variables (thus avoiding corner effects, adjacent sub regions assigned to non-adjacent classes of merit, and granting a sufficient number of degrees of freedom to accommodate both general external conditions, such as given probability distributions, and field-specific requirements).

The evaluation tool we are discussing grants also the possibility to identify those papers characterized by high IF and very low citation percentile, and the opposite case of papers with a high citation count published on low impact journals. Such a selection can be done by tracing two oblique lines in correspondence of the top-left and bottom-right corners of the space Q. For these papers a completely automatic evaluation could be considered not sufficiently reliable, so that the panel may decide to apply an informed PR procedure. An example of this approach is shown in Fig. 2, where the area outlined by the red dashed lines singles out less than one percent of the papers in the specific SC.

Finally, it is worth mentioning that the algorithm could be further improved and customized if the concept of SC as "reference set" would be overcome, moving to clustering strategies based on semantic or on citation networks. This approach would be more rigorous and meaningful considering the existence of a great number of journals that publish

[3] Except for the Physical Sciences one ("GEV 02").

[4] By relevant we mean that a great number (more than 100) of papers to be evaluated fell under that SC.

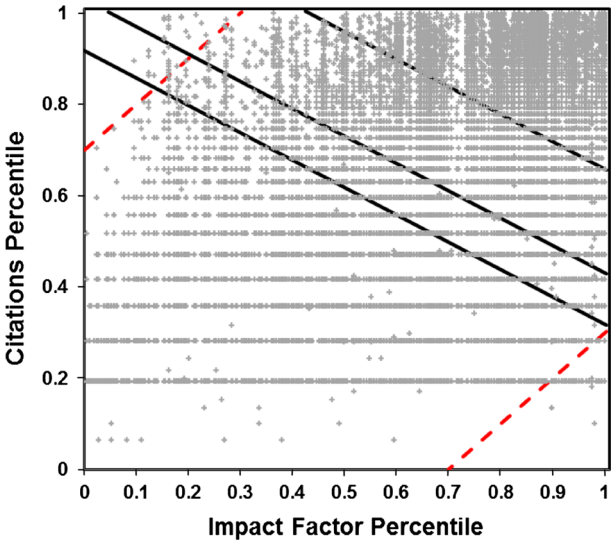


Fig. 2 Assignment of papers with "uncertain" metric to peer review

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.

Examples on real data

We now focus on some instances of SCs coming from the Thomson Reuters Web of Science database and use the discussed bibliometric evaluation method to divide the papers in each SC into four classes of merit. In particular, Fig. 3 reports the scatter plot of the papers in the SC Physics, Atomic, Molecular and Chemical in the year 2004. Starting from Fig. 3a–c it is shown how the target of reproducing the ex-ante probability distribution D can be actually accomplished by choosing different slopes of the set of threshold lines. This means that the panel has the degree of freedom of selecting the weight of citation and journal metric in the linear combination reported in (3) without compromising the calibration of the algorithm. This is an important point since each scientific community needs a certain degree of freedom in choosing the evaluation criteria without compromising the possibility of comparing and aggregating the results at an institutional level.

A second example, reported in Fig. 4, is finalized to highlight how the choice of the same threshold lines (i.e., same slope and same coefficients) can bring to very different evaluation results even if the selected SC belong to the same area, Engineering in this case. This example reinforces the idea that the calibration of the algorithm has to be performed at the level of the single SC and for every year (micro-calibration).

The effect of considering different years is reported in Fig. 5 where the scatter plot of the papers in the same SC (Electrical and Electronic Engineering in this specific example) can be very different both quantitatively and qualitatively. In particular, the distribution is very sparse and collapsed to high percentiles in the axis of citations if the

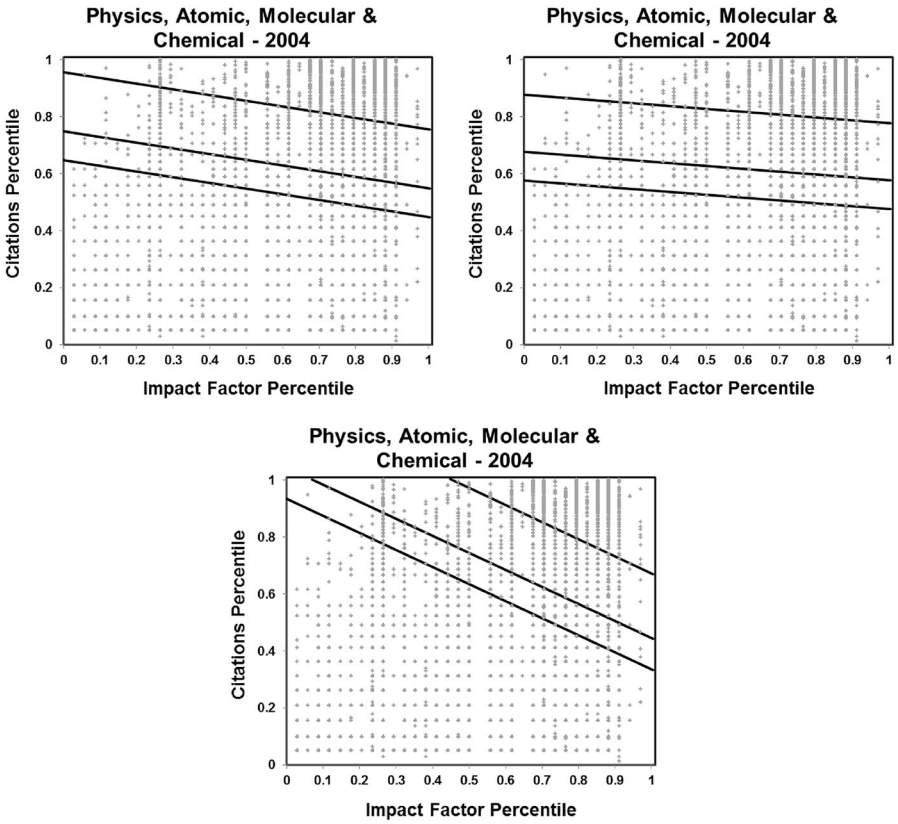
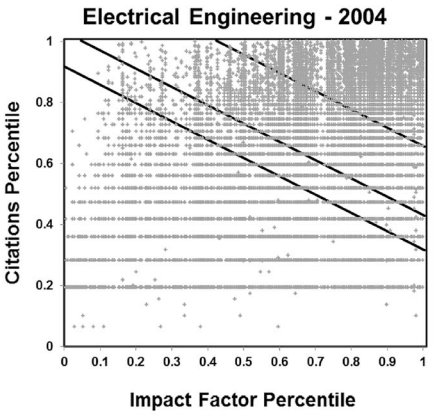


Fig. 3 Application of the algorithm for the SC Physics, Atomic, Molecular and Chemical in the year 2004. The *straight lines* indicate the thresholds for the four classes of merit. The figure shows how the target percentage at a world level (20, 20, 10, 50 in this example) can be obtained by choosing different slopes ($A = 0.2$; $A = 0.1$; $A = 0.6$ respectively)

papers are 'too young'. Percentiles reported in Fig. 5b where computed in 2012, which means only 2 years after the publication of the papers. The result is that papers published in recent years are favored with respect to older papers if the same evaluation tool is adopted.

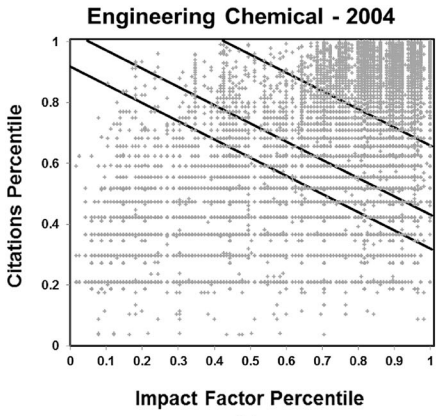
An additional interesting aspect concerns the calculation of the CDF for the IF variable. It is common that some journals host few thousands of items per year while other few tens or units. To build the CDF, instead of considering the number of journals belonging to a SC, one could consider the number of items (papers) published in the SC (in a given year), always ordered according to the IF of the journal. In Fig. 6 we show the SC Electrical and Electronic Engineering in 2004; the distribution of the papers according to the IF and CIT percentile are depicted considering the number of journals in the calculation of the IF percentile (Fig. 6a) and by considering the number of item for each journal (Fig. 6b). It is evident that by using the same set of thresholds in the two cases one would obtain very different ex-ante probabilities (calibration).



(a)

Category	Percentage
Excellent	20.01%
Good	19.96%
Acceptable	10.01%
Limited	50.02%

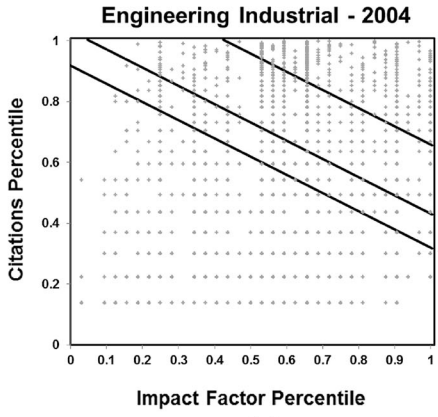
Number of items: 27,333



(b)

Category	Percentage
Excellent	25.77%
Good	20.79%
Acceptable	8.91%
Limited	44.63%

Number of items: 18,446

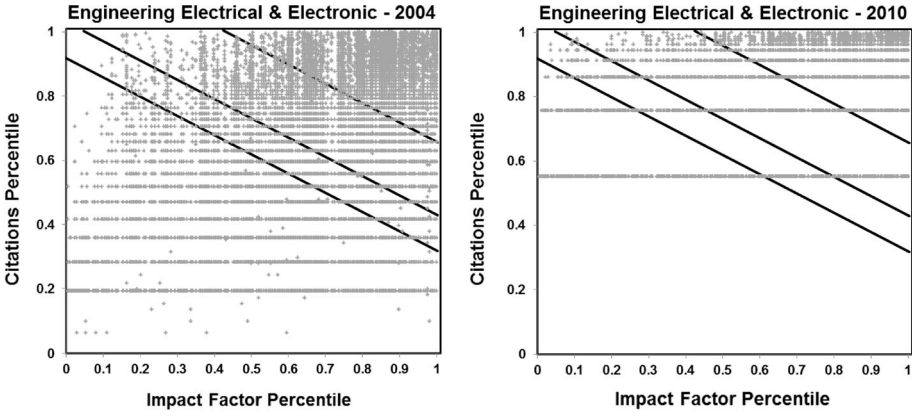


(c)

Category	Percentage
Excellent	13.96%
Good	21.29%
Acceptable	13.34%
Limited	51.37%

Number of items: 2,893

Fig. 4 Application of the algorithm for three different SC belonging to the field of Engineering in the year 2004. The *straight lines* indicate the thresholds for the four classes of merit. The same set of thresholds is employed for all the three SC. The figure shows that very different rakings would be obtained in absence of a specific calibration for each category



Category	Percentage
Excellent	20.01%
Good	19.96%
Acceptable	10.01%
Limited	50.02%

Number of items: 27,333

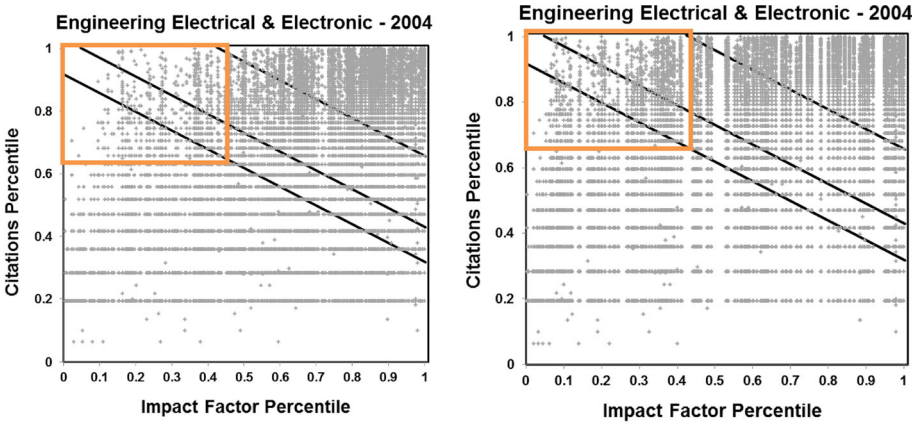
Category	Percentage
Excellent	22.87%
Good	27.84%
Acceptable	12.52%
Limited	36.76%

Number of items: 40,131

Fig. 5 Application of the algorithm for the SC Electrical & Electronic Engineering in two different years: 2004 in (**a**) and 2010 in (**b**). The *straight lines* indicate the thresholds for the four classes of merit. The figure shows that very different ex-ante probabilities are obtained without a specific calibration for each year even in a specific subject category. The paper distribution in the year 2010 is characterized by a few quantized states along citation percentile, as expected. The lowest row of points in (**b**) corresponds to papers with zero citations; indeed, we build the CDF of the CIT variable by assigning percentile 1 to the paper(s) with the highest number of citations (and obviously assigning the same percentile to papers with the same number of citations)

Conclusions

A method to use bibliometrics indicators to assess aggregated research outputs in the context of a large-scale evaluation exercise has been presented in the paper. The algorithm allows an easy calibration of the evaluation procedure at the level of scientific areas or narrower subfields. The calibration procedure, when applied to assign all papers contained in a large data base to a given number of "quality" classes of merit, permits to cope with a priori probability distributions of the classes of merit. Examples of application of the algorithm have been shown based on the recently performed Italian Evaluation of Quality of Research 2004–2010. A by-product of the algorithm is the so-derived fairness in comparing different scientific areas (where bibliometrics indicators are available).



Category	Percentage
Excellent	20.01%
Good	19.96%
Acceptable	10.01%
Limited	50.02%

Category	Percentage
Excellent	15.68%
Good	18.64%
Acceptable	9.5%
Limited	56.17%

Fig. 6 Distribution of the papers according to the number of journals and papers. **a** IF CDF calculated on the basis of number of journals **b** the IF CDF is calculated considering the number of items

Acknowledgments The authors would like to thank Dr. Marco Malgarini for useful discussions.

References

Abramo, G., & D'Angelo, C. A. (2011). Evaluating research: From informed peer review to bibliometrics. *Scientometrics, 87*, 499–514.

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2011). National research assessment exercises: A comparison of peer review and bibliometrics rankings. *Scientometrics, 89*, 929–941.

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

Alberts, B. (2013). Impact factor distortions. *Science, 340*, 787–787.

Ancaiani, A., Anfossi, A. F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., Cicero, T., Ciolfi, A., Costa, F., & Colizza, G., et al. (2015). Evaluating scientific research in Italy: The 2004–2010 research evaluation exercise. *Research Evaluation*. doi:10.1093/reseval/rvv008

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

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

Bergstrom, C. T., & West, J. D. (2008). Assessing citations with the Eigenfactor™ metrics. *Neurology, 71*, 1850–1851.

Bladek, M. (2014). DORA San Francisco declaration on research assessment (May 2013). *College and Research Libraries News, 75*, 191–196.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE, 4*, e6022.

Butler, L. (2003). Modifying publication practices in response to funding formulas. *Research Evaluation, 12*, 39–46.

Butler, L. (2008). Using a balanced approach to bibliometrics: Quantitative performance measures in the Australian Research Quality Framework. *Ethics in Science and Environmental Politics, 8*, 83–92.

Eyre-Walker, A., & Stoletzki, N. (2013). The assessment of science: the relative merits of post-publication review, the impact factor, and the number of citations. *PLoS Biology, 11*, e1001675.

Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., & Karageorgopoulos, D. E. (2008). Comparison of SCImago journal rank indicator with journal impact factor. *FASEB J, 22*, 2623–2628.
Fersht, A. (2009). The most influential journals: Impact Factor and Eigenfactor. *PNAS, 106*, 6883–6884.
Franceschet, M., & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics, 4*, 540–553.
Glänzel, W., & Thijs, B. (2004). The influence of author self-citations on bibliometric macro indicators. *Scientometrics, 59*, 281–310.
HEFCE. (2011). REF2014 impact pilot exercise. Www.hefce.ac.uk/research/ref/impact/. Accessed October 2011.
Moed, H. F. (2009). New developments in the use of citation analysis in research evaluation. *Archivum Immunologiae et Therapiae Experimentalis, 57*, 13–18.
Moed, H. F., Glänzel, W., & Schmoch, U. (2005). Editors' introduction. Berlin: Springer.
Oppenheim, C. (2008). Out with the old and in with the new: The RAE, bibliometrics and the new REF. *Journal of Librarianship and Information Science, 40*, 147–149.
Reale, E., Barbara, A., & Costantini, A. (2007). Peer review for the evaluation of academic research: Lessons from the Italian experience. *Research Evaluation, 16*, 216–228.
Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ, 314*, 497.
Setti, G. (2013). Bibliometric indicators: Why do we need more than one? *IEEE Access, 1*, 232–246.
Smith, D. A. T., & Eysenck, P. M. (2002). *The correlation between RAE ratings and citation counts in psychology*. Royal Holloway: University of London.
Warner, J. (2000). A critical review of the application of citation studies to the Research Assessment Exercises. *Journal of Information Science, 26*, 453–459.