Critical remarks on the Italian research assessment exercise VQR 2011–2014

OPINION PAPER

Critical remarks on the Italian research assessment exercise VQR 2011-2014

Fiorenzo Franceschini[1] and Domenico Maisano[2]

[1] *fiorenzo.franceschini@polito.it* [2] *domenico.maisano@polito.it*
Politecnico di Torino, DIGEP (Department of Management and Production Engineering),
Corso Duca degli Abruzzi 24, 10129, Torino (Italy)

Abstract

For nearly a decade, several national exercises have been implemented for assessing the Italian research performance, from the viewpoint of universities and other research institutions. 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 is suffering heavy criticism too.

This paper presents a structured discussion of the VQR 2011-2014, collecting and organizing some critical arguments so far emerged, and developing them in detail.

Some of the major vulnerabilities of the VQR 2011-2014 are: (1) the fact that evaluations cover a relatively small fraction of the scientific publications produced by the researchers involved in the evaluation, (2) incorrect and anachronistic use of the journal metrics (i.e., ISI Impact Factor and similar ones) for assessing individual papers, and (3) conceptually misleading criteria for normalizing and aggregating the bibliometric indicators in use.

Keywords: Research assessment exercise, Italian VQR, Bibliometric evaluation, Peer review, Journal metric, Percentile rank.

1. Introduction and literature review

In the latter 10-20 years, a growing number of countries have been implementing national exercises for assessing the performance of research institutions, with five key objectives (Schotten and El Aisati, 2014; Abramo and D'Angelo, 2015):

1. Guiding merit-based allocation of public funding;
2. Stimulating continuous improvement in research productivity, through comparative analysis of performance;
3. Identifying the strengths and weaknesses in disciplines and geographic areas, so as to support formulation of research policy and management strategies at a governmental and institutional level;
4. Providing convincing information to tax payers on the effectiveness of research management and

delivery of public benefits;

5. Reducing the information asymmetry between knowledge users (i.e., students, enterprises, and funding agencies) and suppliers (i.e., individual scientists).

Although the shares of overall public funding and the criteria for assigning them tend to vary from nation to nation, the number of countries that conduct regular comparative performance evaluations of universities and link the results to public financing seems to increase gradually (Hicks, 2012). Focusing on Italy, the first research evaluation exercise – denominated VTR (Triennial Evaluation Exercise) 2001-2003 – was launched in 2004, and used a pure peer-review approach of a limited portion of the publications produced by researchers affiliated to universities and other research institutions (Abramo et al., 2011).

After about seven years, a new assessment exercise was launched: the VQR (Research Quality Evaluation) 2004-2010, which marked an important turning point due to (i) the introduction of bibliometric criteria and (ii) the fact that – unlike the previous exercise – the results determine allocation of an important share of financing for individual institutions. The implementation of the VQR 2004-2010 has been entrusted by the Italian Ministry of Education, University and Research (Ministero dell'Istruzione, dell'Università e della Ricerca, hereafter abbreviated as MIUR) to the newly formed Agency for the Evaluation of University and Research Systems (Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca, hereafter abbreviated as ANVUR).

In a nutshell, the VQR 2004-2010 was a hybrid type of evaluation exercise, based primarily on bibliometric analysis for the so called *bibliometric* areas (i.e., hard sciences) and on peer review for the so called *non-bibliometric* ones (i.e., social sciences and humanities). For details, see (ANVUR, 2011; Ancaiani et al., 2015).

Since the time of its introduction, the VQR 2004-2010 had been receiving heavy criticism by part of the Italian scientific community. One of the targets of this criticism was the mechanism for determining the merit class of scientific papers, i.e., a bibliometric assessment combining (i) the number of citations obtained and (ii) a metric of the journal impact (ISI Impact Factor or similar ones) publishing the papers examined. Some Italian scientists considered the criteria by ANVUR as the product of *Do-It-Yourself Bibliometrics*[1] (ROARS, 2016), for their anachronistic disregard of the basic rules of this discipline. Several Italian bibliometricians expressed similar criticism to the attention of the international scientific community (Abramo and D'Angelo, 2015; Baccini and De Nicolao, 2016; Geuna and Piolatto, 2016).

After the "stormy" VQR 2004-2010, a new assessment exercise, denominated VQR 2011-2014, has been recently implemented and is still in progress. Despite the criticism to the evaluation criteria of the VQR 2004-2010, the architecture of the new exercise is rather similar to that of the previous one. The most noticeable difference is the new criterion for determining the merit class of the

[1] In Italian, "bibliometria fai-da-te".

papers examined. Details on this and other differences are contained in the conference paper by Anfossi et al. (2015), later published in extended form on a Scientometrics special issue (Anfossi et al., 2016). As with the VQR 2004-2010, the VQR 2011-2014 has also been receiving heavy criticism (ROARS, 2016).

The aim of this paper is to discuss the current research assessment exercise, collecting and organizing some critical arguments directed to the previous exercise and developing other arguments in detail. The discussion will address the conceptual/methodological aspects of the VQR 2011-2014, without investigating the practical implications of this exercise for the future of research in Italy.

The remainder of the paper is organized into three sections. Sect. 2 recalls the main features of the VQR 2011-2014 and provides a simplified description of the relevant bibliometric criteria, so as to prepare the ground for understanding the subsequent analysis. Sect. 3, which represents the core of this paper, discusses in detail five vulnerabilities of the VQR 2011-2014; description is supported by several pedagogical examples. Sect. 4 summarizes and comments the main findings of our critical analysis.

2. Description of the VQR 2011-2014

This section presents a "pedagogical" description of the current Italian assessment exercise (VQR 2011-2014), which is propaedeutic for understanding the contents of Sect. 3.

As mentioned in Sect. 1, the VQR 2011-2014 represents the "third act" of research assessment exercises in Italy. The purpose of this exercise is to evaluate the research activity carried out over the 2011-2014 period in public universities, legally recognized private universities and other research institutions under the responsibility of the MIUR. Apart from research institutions, objects of the evaluation are their macro-disciplinary areas and departments but not individual researchers. The results may influence two areas of future action: (1) overall institutional evaluations will guide allocation of the merit-based share of the so-called Ordinary Finance Funds (FFO), i.e., the core government funding for Italian universities; (2) evaluation of the macro areas and departments can be used by research institutions to guide internal allocation of the acquired resources.

The evaluation of the whole institutions is determined by the weighted sum of a number of indicators: 75% based on a score for the quality of the research output and 25% derived from a composition of other indicators (capacity to attract resources, mobility of research staff, internationalization, Ph.D. programs, etc.).

Let us now focus the attention on the evaluation of the so-called research *products*, namely articles, books, book chapters, conference proceedings, critical reviews, commentaries, book translations, patents, prototypes, project plans, software, databases, exhibitions, works of art, compositions and thematic papers. The term "product" is used in the official ANVUR documents, indicating entities

of different nature. Since our study will consider almost exclusively articles in scientific journals, conference proceedings and book chapters, this term will be hereafter replaced with the terms "paper", "article" or "publication".

ANVUR nominated 16 evaluation panels, i.e., the so-called Groups of Evaluation Experts (GEVs), including national and foreign experts, one for each research area composing the national academic system (details on the research areas and relevant GEVs are reported in Tab. A1, in the appendix). The institutions subject to evaluation should submit a specific number of papers for each researcher with a permanent position, based on his/her academic rank and period of activity over the four years considered. Simplifying, the requirement for university staff is two papers per researcher, whereas that for other research institutions is three papers per researcher. 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).

The 16 research areas, are divided into *bibliometric* and *non-bibliometric* ones, depending on their peculiarities (see Tab. A1, in the appendix). In the latter ones (i.e., typically social sciences and humanities) papers are evaluated exclusively through *peer review*, while in the former ones (i.e., typically hard sciences, such as engineering and life sciences) papers are evaluated using a mixed approach consisting of bibliometric analysis, for those indexed by Scopus and WoS, and peer review for the other papers or even for the indexed papers, when expressly requested by the institution.

Consistently with the Ministerial Decree of 27 June 2015 by MIUR (2015), the (bibliometric or peer-review) evaluation of the quality of each paper should result into five merit classes (A, B, C, D and E), as described in Tab. 1.

Tab. 1. Classes of merit and relevant score, in which papers evaluated are classified.

Class	Score (S_i)	Description
A. Excellent	1	The paper places in the top 10% of the so-called "distribution of the international scientific production"[2], for the specific area of interest and issue year.
B. Good	0.7	The paper places in the top 10–30% range of the same distribution.
C. Fair	0.4	The paper places in the top 30–50% range of the same distribution.
D. Acceptable	0.1	The paper places in the top 50–80% range of the same distribution.
E. Limited	0	The paper places in the bottom 20% of the same distribution or cannot be evaluated because it does not conform to the types of acceptable papers.

The institutions are also subject to potential penalties: (i) in proven cases of plagiarism or fraud, (ii) for paper types not admitted by the GEV, or lack of relevant documentation, or produced outside the 2011–2014 period, and (iii) for failure to submit the requested number of papers.

We now focus the attention on the paper evaluation in the bibliometric areas. Simplifying, each research institution submits the papers to be evaluated, specifying (1) the most appropriate *subject categories* – as defined by the Thomson Reuters WoS database – or the most appropriate *all journal*

[2] The Ministerial Decree of 27 June 2015 by MIUR (2015) is quite nebulous on this point; for instance, it is not clear which aspects should this (presumed) distribution consider (e.g., impact/diffusion, quality, originality, etc.).

science categories – as defined by Scopus Elsevier (for simplicity, both these groups of categories will be hereafter referred to as *SC*), among those associated to the publishing journals, and (2) the most pertinent GEV panels.

Two indicators are associated with each *i*-th paper: the citation count (C_i), i.e., the number of citations accumulated by the paper up to a given point in time (e.g., 29 February 2016 for some research areas (ANVUR, 2015b)), according to the WoS or the Scopus database, and a journal metric (J_i) related to the publishing journal. For each *SC* and issue year, the GEV has to identify the most pertinent journal metric, among the possible ones (see Tab. 2). GEVs can sometimes admit journal metrics related to other specialized databases, different from WoS and Scopus, such as the Mathematics Citation Quotient (MCQ) for journals indexed by the MathSciNet database.

Tab. 2. Major journal metrics used for the VQR 2011-2014 evaluation procedure, in the bibliometric areas.

Journal metric	Description
ISI Impact Factor (*IF*)	Average number of times articles from the journal, published in the past two years, have been cited in the year of interest, according to the WoS database.
5-year Impact Factor	Average number of times articles from the journal, published in the past five years, have been cited in the year of interest, according to the WoS database.
Impact per Publication (*IPP*)	Average number of times articles from the journal, published in the past three years, have been cited in the year of interest, according to the Scopus database.
Article Influence (*AI*)	Indicator obtained weighing the citations received by the articles (in a specific time period), depending on the rank of the relevant journals, i.e., citations from highly ranked journals are weighted to make a larger contribution than those from poorly ranked journals. This journal metric can therefore be considered as normalized on the basis of the *prestige* of citing journals. *AI* is pre-calculated based on the WoS citation statistics.
SCImago Journal Ranking (*SJR*)	Indicator similar to *AI* but pre-calculated according to the citation statistics by Scopus.
Source Normalized Impact per Paper (*SNIP*)	Indicator similar to *IPP* but normalized based on the different citation propensity of the citing articles. *SNIP* is therefore a *field-normalized* indicator, pre-calculated using the Scopus citation statistics.

Having determined the reference database and the journal metric to be used, the evaluation procedure concerning each *i*-th paper is based on the following steps:

- Normalization of C_i, considering the cumulative probability or *percentile rank* $F_C(C_i) \in [0, 100\%]$ related to the distribution of the C_i values of the totality[3] of the papers issued by journals in the same *SC* and issue year of the *i*-th article of interest.
- Normalization of J_i, considering the cumulative probability or percentile rank $F_J(J_i) \in [0, 100\%]$ related to the distribution of the J_i values of the totality of the papers issued by journals in the same *SC* and issue year of the *i*-th article of interest.
- Construction of a $F_J(J_i)$-$F_C(C_i)$ map related to the papers issued by journals in a certain *SC* and issue year.
- Definition of an aggregate indicator, given by the linear combination of $F_J(J_i)$ and $F_C(C_i)$:

$$Y_i = w \cdot F_C(C_i) + (1-w) \cdot F_J(J_i), \quad (1)$$

[3] The term "totality" indicates that the distribution is built considering all the literature production within that *SC* and issue year, i.e., also including papers different from those submitted by Italian researchers to the VQR 2011-2014 (ANVUR, 2015a).

where

$w \in [0, 1]$ is a weight used for giving more/less importance to the $F_C(C_i)$ and $F_J(J_i)$ contributions, in their aggregation by a weighted sum[4].

The choice of the w value is left to the GEV. In general, ANVUR (2015a, 2015b, 2015c) recommends to use relatively higher w values for older articles (e.g., those issued in 2011-2012), as they are likely to be mature enough in terms of citation impact. On the other hand, it recommends to use relatively lower w values for more recent articles (such as those issued in 2014), in order to give more weight (i.e., $1 - w$) to the journal metric, which is used as a proxy of the future impact of these articles.

- Normalization of Y_i, considering the cumulative probability or percentile rank $F_Y(Y_i) \in [0, 100\%]$ related to the distribution of the Y_i values of the papers issued by journals in the same *SC* and issue year of the *i*-th article of interest.
- For each combination of *SC* and issue year, the distribution of the Y_i values is supposed to represent the (so-called) "distribution of the international scientific production" (MIUR, 2015). Consistently with what reported in Tab. 1, papers can be classified into the five merit classes, depending on their $F_Y(Y_i)$ values: A-Excellent ($0.9 \leq F_Y \leq 1$, score 1), B-Good ($0.7 < F_Y \leq 0.9$, score 0.7), C-Fair ($0.5 < F_Y \leq 0.7$, score 0.4), D-Acceptable ($0.2 < F_Y \leq 0.5$, score 0.1), E-Limited ($0 < F_Y \leq 0.2$, score 0).

The bibliometric evaluation procedure of VQR 2011-2014 largely follows that one of VQR 2004-2010. Apart from some differences – such as (1) the number of papers submitted by researchers, (2) the possible journal metrics (i.e., only *IF* and *IPP* for the VQR 2004-2010), and (3) the number of merit classes representing the quality of each paper and the relevant scores – the greatest difference between the two exercises concerns the aggregation of $F_C(C_i)$ and $F_J(J_i)$ and the subsequent determination of the merit classes. The penultimate exercise adopted a technique based on partitioning the $F_J(J_i)$–$F_C(C_i)$ plane into rectangular areas, as shown in Fig. 1(a). Papers included in the four squared zones positioned around the diagonal are uniquely assigned to four merit classes; the papers positioned in the remaining zones (highlighted in grey) can be assigned by GEVs to the classes that they considered as appropriate or can be subject to an additional informed peer-review procedure. For details, see (ANVUR, 2011; Abramo et al., 2015; Ancaiani, 2015). On the other

[4] We remark that Y_i and w are not explicitly defined in the official documents by ANVUR (2015a; 2015b; 2015c), which hint at *partitioning of the F_C-F_J space into sub-regions delimited by parallel lines (i.e., with same slope), defined by equations:*

$$F_J(J_i) = A \cdot F_C(C_i) + B_n, \quad \text{(n1)}$$

where A is the (fixed) slope of the lines and B_n is the relevant angular coefficient.

Comparing Eq. 1 with Eq. n1, we obtain:

$$F_J(J_i) = -\frac{w}{1-w} \cdot F_C(C_i) + \frac{Y_i}{1-w} \quad \Rightarrow A = -\frac{w}{1-w}, B_n = \frac{Y_i}{1-w} \cdot \quad \text{(n2)}$$

Thus, setting A corresponds to setting w uniquely, while setting a B_n value corresponds to setting a Y_i value uniquely.

hand, the VQR 2011-2014 adopts a technique based on partitioning the $F_J(J_i)$-$F_C(C_i)$ plane into oblique stripes, as shown in Fig. 1(b).

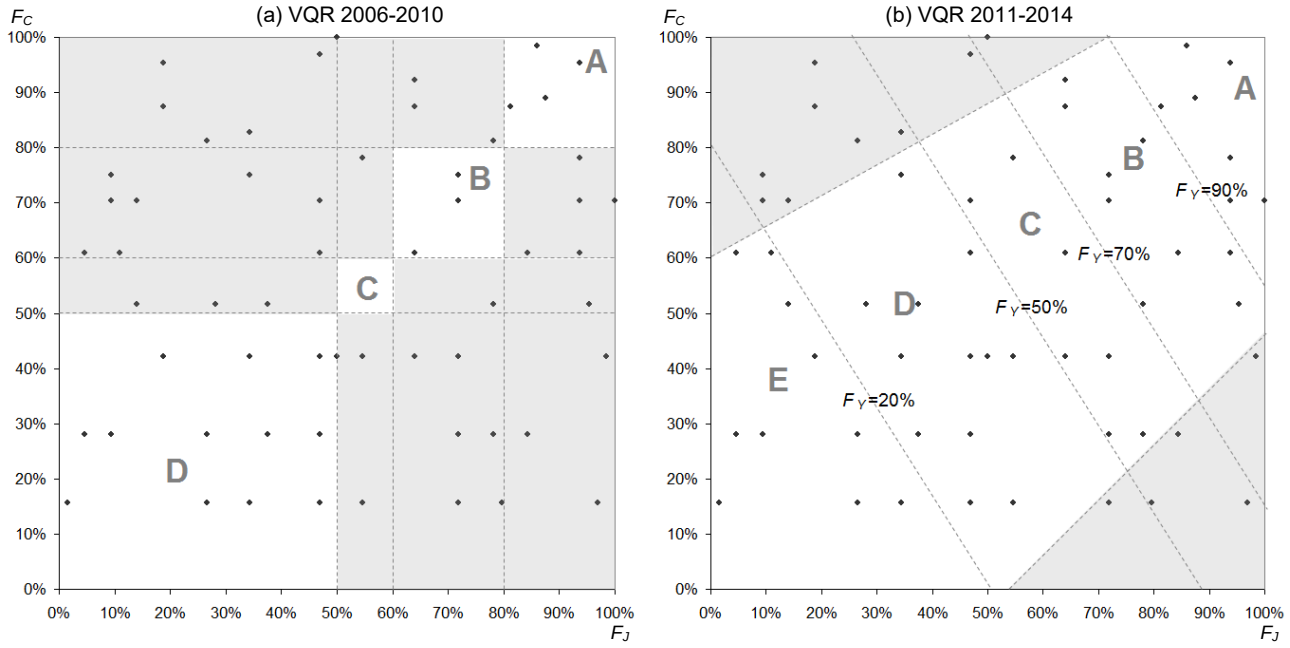

Fig. 1. Partitioning of the F_J–F_C space for determining the merit classes of the papers examined, in the bibliometric evaluation of VQR 2004-2010 and VQR 2011-2014. (a) For VQR 2004-2010, the papers included in the four squared zones positioned around the diagonal are uniquely assigned to four merit classes; the papers positioned in the remaining zones (highlighted in grey) can be assigned by GEVs to the classes that they considered as appropriate or can be subject to an additional informed peer-review procedure. (b) For the VQR 2011-2014, the merit classes correspond to oblique stripes, whose slope depends on the *w* value in use (0.4 in this case); papers positioned in the top-left and bottom-right zones (highlighted in grey) are assigned to an additional informed peer-review procedure. Charts have been built considering the J_i and C_i values of the 64 fictitious papers reported in Tab. A2 (in the appendix).

According to ANVUR (2015a), the bibliometric evaluation is not sufficiently reliable for papers with relatively high $F_J(J_i)$ values and relatively low $F_C(C_i)$ values, and vice versa. In other words, for papers positioned in the top-left and bottom-right corner of the $F_J(J_i)$–$F_C(C_i)$ plane, the GEV may decide to complement the result of the automatic evaluation with an additional informed peer-review procedure; we will return to this point later in Sects. 3.4 and 3.5.

Having determined the merit classes of the individual papers – regardless whether through the bibliometric or peer-review procedure – predetermined scores (S_i) are assigned to them. Next, the S_i values related to the totality of the papers submitted by each research institution are added up and combined with other indicators, determining a single overall performance indicator; for details, see the official documents by ANVUR (2015a; 2015b; 2015c).

3. Critical analysis

This section is divided into five subsections, dealing with the major vulnerabilities of the bibliometric evaluation procedure of the VQR 2011-2014; a synthetic description of these vulnerabilities is reported in Tab. 3).

Tab. 3. Brief description of the major vulnerabilities of the bibliometric evaluation procedure in the VQR 2011-2014.

Vulnerabilities	Brief description
1. Evaluation of a small number of papers.	Even assuming that the (bibliometric and non-bibliometric) evaluation procedure is methodologically impeccable, the evaluation of just two/three papers per researcher represents a serious limitation for assessing the performance of research institutions.
2. (Mis)use of journal metrics.	Using journal metrics (even when combined with other indicators) to evaluate the quality of individual papers is potentially misleading.
3. Normalization/combination of indicators.	The normalization of C_i and J_i through the F_C and F_J percentile ranks, their subsequent aggregation into Y_i, and the normalization of Y_i through the F_Y percentile rank are conceptually questionable operations.
4. Decisional autonomy to GEVs.	Several operations of "calibration" of the metrics (e.g., setting w, choosing the more appropriate journal metric, etc.) are entrusted to GEVS; in the absence of solid guidelines, this freedom can be counterproductive.
5. Compatibility between peer review and bibliometric analysis.	According to the VQR 2011-2014, the output of the bibliometric and peer-review evaluation should be mutually compatible. This assumption does not seem to be supported by adequate empirical evidence.

3.1 Evaluation of a small number of papers

As anticipated, the VQR 2011-2014 evaluates a relatively small number of papers per researcher, i.e., two or three. This limitation – which generally characterizes peer-review based exercises, due to the considerable effort required to read and (manually) evaluate the examined papers – may represent a critical concern for the reliability of results; Abramo et al. (2014) justly state that it could be reasonable to extend the evaluation to the totality of the papers produced, at least for bibliometric areas.

In light of the previous considerations, a question arises: *Which research-performance features can the VQR 2011-2014 depict*? Proceeding by elimination, we believe that this exercise does not allow to depict *productivity*, due to the relatively low number of papers evaluated. Also, it does not seem appropriate to assess the *average quality/impact* of the research, since it ignores a significant portion of the papers produced during the evaluation period. It does not even seem appropriate to assess the research *excellence*, defined as the ability to produce high-level research with a certain regularity (Franceschini and Maisano, 2011); in fact, the production of two high quality/impact papers in four years does not seem a sufficient condition to prove the excellence of a generic researcher (*one swallow does not make a summer*).

Let us present a simple numerical example to clarify the last point: consider a generic *mid-level* researcher (X) with a scientific production in line with the so-called "distribution of the international scientific production". We hypothesize that this researcher is able to produce about three papers per year, therefore, about 12 papers in the 2011-2014 time window. Consistently with the information contained in Tab. 1, only 10% of the papers will (on average) achieve the highest class (A), while only 30% will (on average) achieve class A or B. The probability that this (mid-level) researcher has at least two papers of class A or B will be:

$$Pr = 1 - \sum_{k=0}^{1} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \approx 91.5\% \qquad \text{being } n = 12 \text{ and } p = 30\%, \tag{2}$$

Let us consider a second *excellent* researcher (Y), who is able to produce about 12 papers (in the same time window), all of which of class A or B. Researcher Y will obviously have at least two papers of class A or B (i.e., $Pr = 100\%$).

The previous example shows that, in spite of the obvious superiority of the excellent researcher (Y), even the mid-level one (X) has a very high probability (91.5%) to have at least two papers of class A or B (see also the chart in Fig. 2). It therefore is very difficult to discriminate between these two researchers when considering two papers only. As an alternative example, it is trivial to demonstrate that it would be impossible to discriminate between a researcher with two-and-only-two papers of class A and a researcher with a plethora of papers of class A.

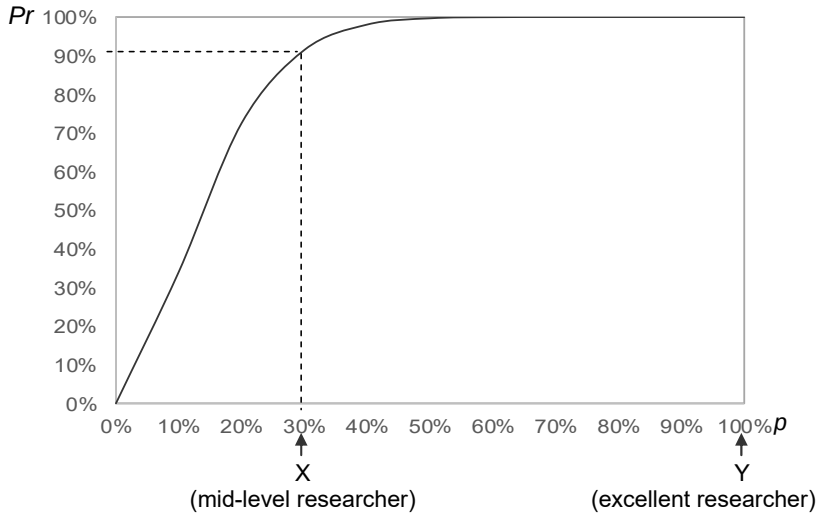

Fig. 2. Graph showing the probability (*Pr*) of a researcher to have at least two articles of class A or B, assuming that he/she has produced 12 papers, each with a probability (*p*) to be in these classes; *Pr* was calculated using the model in Eq. 2. For example, the (mid-level) researcher X has a probability $p = 30\%$ to produce articles of class A or B, while the (excellent) researcher Y exclusively produces papers of class A or B ($p = 100\%$). Despite this large gap, the *Pr* values related to the two researchers are not much different (i.e., 91% against 100%).

In view of the fact that the assessment of entire research institutions is performed by aggregating the contributions from individual researchers, our considerations on the poor discrimination power in the identification of excellent researchers can be extended to the identification of excellent research institutions. Returning to the initial example, let us assume that there are two research institutions: the first mostly consists of mid-level researchers (as researcher X), while the second mostly consists of excellent researchers (as researcher Y). The big gap between these two populations could not necessarily be caught when applying an evaluation system based on the submission of two papers per head only. Thus, we believe that it would be unwise to use the results of the VQR 2011-2014 exercise to estimate the level of excellence of research institutions.

Having said that, a new question arises: *Is there any reasonable use of the results of the proposed exercise*? With some effort of imagination, it seems that the results of this exercise can only depict the level of research *decency*, meaning the ability to produce – in the relatively long time period of four years – a low number of papers with relatively high impact/quality. It is not unrealistic to assume that, for a research institution in which researchers are (on the average) active, it would not

be so difficult to "saturate" the expected scores for the papers submitted, i.e., most researchers would be able to submit papers classified in relatively high merit classes (e.g., A or B, as also illustrated in the previous examples). Inverting the reasoning, this exercise could allow to find out institutions with relatively high incidence of "lazy" researchers, i.e., unable to produce at least two/three papers with relatively high impact, in four years. Let us clarify this through a metaphor: if the students of a middle-school class were evaluated through a very permissive test, most of them would be likely to pass it with a high score, except for the least prepared.

In conclusion, authors believe that this type of evaluation could be effective for identifying the less virtuous research institutions but could be ineffective for identifying the excellent ones.

3.2 (Mis)use of journal metrics

As previously described, the bibliometric classification of a generic *i*-th paper is based on the combination of the C_i and J_i indicators; this subsection focuses the attention on the latter one. According to ANVUR, journal metrics can be especially useful to support the evaluation of relatively recent papers, which are not so mature in terms of citation impact (Anfossi et al., 2015); following this reasoning, when evaluating these papers, ANVUR suggest to decrease *w*, in order to give more weight to J_i (which is implicitly used as a proxy of the future citation impact of the papers) with respect to C_i (ANVUR 2015a; 2015b; 2015c).

For many years now, a large number of contributions in the scientific literature prove the diffused misuse of journal metrics for assessing individual articles (Seglen, 1997; Lozano et al., 2012; IEEE, 2013; Marx and Bornmann, 2013; Ware and Mabe, 2015); according to Van Raan, this would be a "mortal sin" (Levine, 2011). The reason, almost universally acknowledged among bibliometricians, is that the variability in the number of citations received by articles published by the same journal is generally high; as a consequence, the use of central tendency indicators – as journals metrics – is inappropriate for estimating the citation impact of individual papers. To use a metaphor, it would be like predicting the future height of a specific individual, using the average height of the population (being the human height relatively dispersed).

It matters little that the results of the national exercise will not be used to evaluate individual researchers but entire research institutions or perhaps portions of them (Ancaiani et al., 2016): the use of journal metrics remains incorrect, as it is directed to the evaluation of individual articles. It can be also said that the combination of a correct metric (C_i) with a distorted one (J_i) can only produce a new distorted metric (Y_i in the case of the VQR 2011-2014).

Also, the fact of giving more merit to papers published in journals with relatively high J_i values is questionable for two reasons:

- Journals with higher J_i values are not necessarily more stringent and rigorous in the selection of the papers to be published, also due to the diffusion of techniques for manipulating journal metrics (Martin, 2016);
- Papers published on journals with higher J_i values tend to have a higher propensity (on average) to be cited than papers (of similar quality) published on journals with lower J_i values, due to a sort of "showcase effect" (Didegah and Thelwall, 2013; Franceschini and Maisano, 2014). It is therefore debatable that such papers should receive a further advantage.

Although we are aware of the difficulties in estimating the future citation impact of recent papers, we believe that the use of journal metrics as predictors represents an illusory and distorting solution. This is confirmed by several authoritative scientific contributions (Lett, 2013; Bohannon, 2016).

A less debatable solution could be complementing C_i 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. (Thelwall et al., 2013; Bornmann, 2014; Costas et al., 2015); however, it is still necessary to investigate the potential of altmetrics and their benefits and disadvantages for measuring impact.

3.3 Normalization/combination of indicators

The bibliometric evaluation of individual papers is based on the normalization of the two indicators C_i and J_i, through the percentile ranks F_C and F_J, and their subsequent aggregation into Y_i, through a weighted sum. Even assuming that combining C_i and J_i is meaningful (see the criticism in Sect. 3.2), this section shows that the proposed normalization and consequent combination is conceptually misleading. The remainder of this section is divided into five sub-sections: Sect. 3.3.1 recalls some basic properties of the scales of measurement, which are functional to the understanding of the subsequent criticism, Sect. 3.3.2 criticizes the normalization of J_i and C_i, Sect. 3.3.3 criticizes the combination of F_C and F_J, Sect. 3.3.4 criticizes the score assignment to merit classes, and Sect. 3.3.5 summarizes the criticism in Sects. 3.3.2 to 3.3.4.

3.3.1 Basic properties of the scales of measurement

A largely accepted classification of the scales of measurement was proposed by Stevens (1946). In this proposal, measurements/indicators can be classified into four different types of scales: nominal, ordinal, interval and ratio (see Tab. 4).

It will be convenient to illustrate this scheme with a variable X and two objects, say A and B, whose scores on X are x_A and x_B, respectively.

1. A *nominal* scale merely distinguishes between classes (*equivalence* relationship). That is, with respect to A and B one can only say $x_A = x_B$ or $x_A \neq x_B$.

Tab. 4. Classification scheme of measurements/indicators depending on their scale types (Stevens, 1946; Roberts, 1979).

Scale Type	Empirical Properties	Permissible Statistics	Permissible scale-transformation	Examples
Nominal	Equivalence	Mode, chi square	Permutation (one-to-one substitution)	Eye colour, place of birth, etc…
Ordinal	Equivalence, order (greater or less)	Median, percentile	Monotonic increasing function	Surface hardness, military rank, etc…
Interval	Equality, order, distance (addition or subtraction)	Mean, standard deviation, correlation, regression, analysis of variance	Linear function: $\Phi(x) = a \cdot x + b$, being $a > 0$	Temperature in °C, serial numbers, etc…
Ratio	Equality, order, distance, ratio (multiplication or division)	All statistics permitted for interval scales plus the following: geometric mean, harmonic mean, coefficient of variation, logarithms	Similarity: $\Phi(x) = a \cdot x$, being $a > 0$	Temperature in K, weight, age, number of children, etc…

2. An *ordinal* scale induces an ordering of the objects (*order* relationship). In addition to distinguishing between $x_A = x_B$ and $x_A \neq x_B$, the case of inequality is further refined to distinguish between $x_A > x_B$ and $x_A < x_B$.
3. an *interval* scale assigns a meaningful measure of the difference between two objects (*distance* relationship). One may say not only that $x_A > x_B$, but also that A is $x_A - x_B$ units different than B.
4. a *ratio* scale is an interval scale with a meaningful zero point (which allows *ratio* relationship). If $x_A > x_B$ then one may say that A is x_A / x_B times superior to B.

From the viewpoint of the scale properties, the above types of measurement scales are ordered from "less powerful" to "more powerful". In particular, the more powerful scales (interval and ratio) provide more information and are generally preferred for measurement purposes. It is often a goal of measurement to obtain scales that are as much powerful as possible, but – unfortunately – this is not always so straightforward (Franceschini et al., 2007).

As a general rule, numbers should be analysed on the basis of the properties of the scale with which they are gathered (Roberts, 1979). Consequently, one may obtain results that do not make sense by applying arithmetic operations to measurements/indicators with scales in which these operations are inadmissible (see the second column of Tab. 4).

3.3.2 Normalization using percentile ranks

In light of the classification in Sect. 3.3.1, C_i and J_i are defined on ratio scales since they both have a meaningful zero, corresponding to the absence of the measured manifestation (i.e., the citations obtained by an article or an entire journal), and an objective and precise unit. Thus, they allow relationships of *equivalence*, *order*, *distance* and *ratio* among the objects represented; the only permissible scale transformation, which preserves the above relationships among the objects, is that of *similarity*.

The normalizations through the percentile ranks F_C, F_J and F_Y can be interpreted as non-necessarily-linear monotonically increasing transformations, which turn the realizations of the

variables of interest (C_i, J_i and Y_i) into the cumulative probabilities (or percentile ranks) of the corresponding distributions (see the example in Fig. 3).

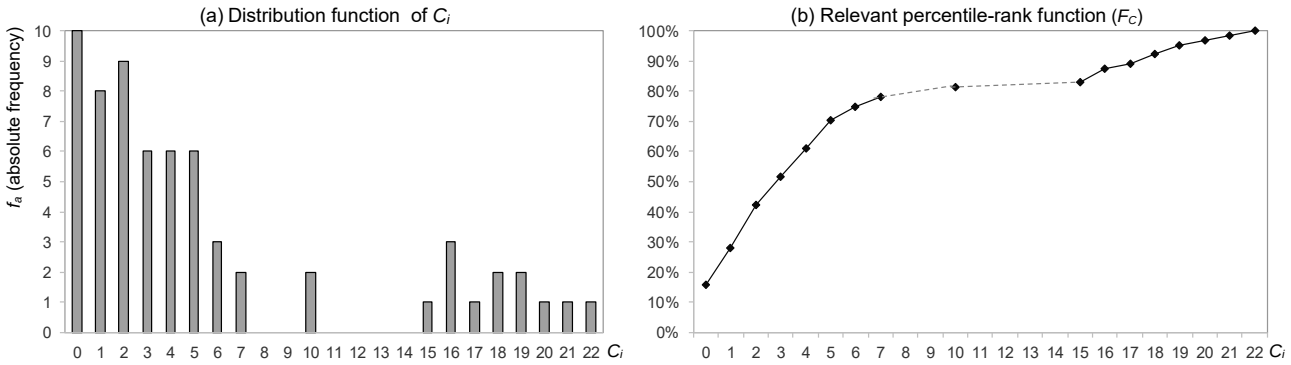

Fig. 3. (a) Example of distribution function of a fictitious variable $C_i \in N$ included between 0 and 22 and (b) relevant percentile-rank function F_C. The distribution exemplified is that of the C_i values of 64 fictitious papers, as reported in Tab. A2 (in the appendix).

Of course, depending on the distributions of interest, the percentile rank functions (F_C, F_J and F_Y) will be different. Only in the special, and very unlikely, case in which C_i, J_i and Y_i were uniformly distributed, these monotonically increasing functions would degenerate into similarity functions ($\Phi(x) = a \cdot x$, being $a > 0$). In general, the F_C, F_J and F_Y transformations would distort both the interval and ratio relationships among the initial objects (C_i, J_i and Y_i values), preserving only the equivalence and order relationships (Roberts, 1979; Kreifeldt and Nah, 1995; Thompson, 1993 Bornmann et al., 2013).

Let us provide a practical example, considering three fictitious papers (P_α, P_β and P_δ) published by two journals in the same *SC* and issue year. The three papers respectively received $C_\alpha = 5$, $C_\beta = 10$ and $C_\delta = 15$ citations. Since C_i is defined on a ratio scale and is typically used to evaluate the citation impact of a paper[5], the following statements are meaningful (see the first two columns of Tab. 6):

1. *Equivalence* relationship: all the three papers have different citation impact;
2. *Order* relationship: the citation impact of P_δ is higher than that of P_β, which is in turn higher than that of P_α;
3. *Distance* relationship: the difference (in terms of citation impact) between P_δ and P_β is equal to that between P_β and P_α;
4. *Ratio* relationship: the citation impact of P_β is twice that of P_α, while that of P_δ is three times that of P_α.

[5] Since the citations associated to a certain paper can be interpreted as countable characteristics reflecting its impact in the scientific community (De Bellis, 2009), it could be argued that this impact is (at least roughly) proportional to the citations obtained; many classical bibliometric indicators (such as the ISI Impact Factor or other journal metrics) rely on this assumption. Although the authors are aware that this (presumed) proportionality may be sometimes questionable (Wang, 2014), they believe that it is not unreasonable.

Let us now consider the empirical distribution of the C_i values reported in Tab. 5, which is also represented graphically in Fig. 3(a); the percentile ranks related to the C_α, C_β, and C_δ values are $F_C(C_\alpha = 5) = 70.3\%$, $F_C(C_\beta = 10) = 81.3\%$ and $F_C(C_\delta = 15) = 82.8\%$ (see Tab. 5).

Tab. 5 – Absolute/relative frequencies and percentile ranks related to the C_i values of 64 fictitious papers, published by journals in the same *SC* and issue year.

C_i	0	1	2	3	4	5	6	7	8	9	10	11
f_a	10	8	9	6	6	6	3	2	0	0	2	0
f_r	15.6%	12.5%	14.1%	9.4%	9.4%	9.4%	4.7%	3.1%	0.0%	0.0%	3.1%	0.0%
F_C	15.6%	28.1%	42.2%	51.6%	60.9%	70.3%	75.0%	78.1%	78.1%	78.1%	81.3%	81.3%

C_i	12	13	14	15	16	17	18	19	20	21	22	Total
f_a	0	0	0	1	3	1	2	2	1	1	1	64
f_r	0.0%	0.0%	0.0%	1.6%	4.7%	1.6%	3.1%	3.1%	1.6%	1.6%	1.6%	100%
F_C	81.3%	81.3%	81.3%	82.8%	87.5%	89.1%	92.2%	95.3%	96.9%	98.4%	100.0%	N/A

f_a is the absolute frequency related to a certain C_i value;
f_r is the relative frequency related to a certain C_i value;
F_C is the cumulative probability (or percentile rank) related to a certain C_i value.

Returning to the four previous statements about the relationship among objects, it can be seen that the application of the F_C transformation does not alter the relationships for the less "powerful" scales, i.e. the *categorical* scales (nominal and ordinal), but it may alter the relationships among objects for the *cardinal* scales (i.e., interval and ratio) (see Tab. 6). In other words, the application of the F_C transformation downgrades the initial (ratio) scale of C_i to an ordinal scale, preserving the relationships of equivalence and order but distorting those of distance and ratio.

The above considerations can be extended to J_i and Y_i and the respective transformation/normalization through the F_J and F_Y functions.

Tab. 6. Example of statements preserved and distorted, after having applied the F_C transformation function in Tab. 5 to $C_\alpha = 5$, $C_\beta = 10$ and $C_\delta = 15$.

Relationship	Initial statement	After the F_C transformation	Statement preserved?
1. Equivalence	$C_\alpha \neq C_\beta \neq C_\delta$	$F_C(C_\alpha) \neq F_C(C_\beta) \neq F_C(C_\delta)$	Yes
2. Order	$C_\delta > C_\beta > C_\alpha$	$F_C(C_\delta) > F_C(C_\beta) > F_C(C_\alpha)$	Yes
3. Distance	$C_\delta - C_\beta = C_\beta - C_\alpha$	$F_C(C_\delta) - F_C(C_\beta) \neq F_C(C_\beta) - F_C(C_\alpha)$	No
4. Ratio	$C_\beta / C_\alpha = 2$ $C_\delta / C_\alpha = 3$	$F_C(C_\beta) / F_C(C_\alpha) = 81.3\% / 70.3\% = 1.16$ $F_C(C_\delta) / F_C(C_\alpha) = 82.8\% / 70.3\% = 1.18$	No

3.3.3 Combination of F_C and F_J

Being defined in the same [0, 100%] range, the normalized indicators $F_C(C_i)$ and $F_J(J_i)$ may seem comparable. $F_C(C_i)$ and $F_J(J_i)$ are then combined into the synthetic indicator Y_i, through a polynomial function. Anfossi et al., 2016 state that the aggregation function could be a generic

polynomial function – even of order higher than one – satisfying the basic requirement of *Pareto dominance*; then, for the purpose of simplicity, they suggest to use linear functions as modelled in Eq. 1. It seems that this hint has been followed by most of the GEVs (ANVUR, 2015b; 2015c).

Having said that, the proposed combination of F_C and F_J is questionable for (at least) four reasons:

1. Although the authors share the opinion of Anfossi et al. (2016), regarding the fact that Pareto dominance would be a desirable property, they point out that any convex combination of F_C and F_J, like the one in Eq. 1, cannot satisfy Pareto dominance. In fact a pair ($F_C(C_1)$ and $F_J(J_1)$) is said to be *Pareto dominating* another pair ($F_C(C_2)$ and $F_J(J_2)$) whenever both the conditions $F_C(C_1) \geq F_C(C_2)$ and $F_J(J_1) \geq F_C(J_2)$ hold. Obviously, since Y_i is a linear combination of F_C and F_J, there are situations in which $Y_1 \geq Y_2$ but Pareto dominance does not hold. By the way, it can be noticed that the classification adopted in the VQR 2004-2011 satisfies the requirement of Pareto dominance (see Fig. 1(a)). In other words, all publications in the merit class A are Pareto dominant to all lower classes (and similarly the merit class B is Pareto dominant to C and D, etc.).
2. The aggregation model in Eq. 1 is based on the weighted sum of objects (i.e., the F_C and F_J percentile ranks), which are defined on ordinal scales (see Sect. 3.4.2). This aggregation is therefore prohibited (cf. Tab. 4) and conceptually misleading (Roberts, 1979); to confirm this, the scientific literature includes several contributions indicating that percentile ranks cannot be added, such as (Thompson, 1993; Kreifeldt and Nah, 1995).
3. The proposed aggregation presupposes the existence of questionable equivalence classes for the papers examined, depending on the J_i and C_i values.

Let us develop the fourth point with an example. Considering the C_i and J_i values related to 64 fictitious papers in a certain *SC* and issue year, we can represent them in the J_i-C_i plane with the relevant distributions (see Fig. 4).

By applying the empirical transformations $F_C(C_i)$ and $F_J(J_i)$ to the initial data (see the relevant columns in Tab. A2, in the appendix), the initial J_i-C_i plane is "deformed" into the new F_J-F_C plane in Fig. 5(b) (ROARS, 2016). Comparing the graphs in Fig. 5(a) and Fig. 5(b), we note that these transformations may cause an uncontrollable variation in the point positioning (numeric labels refer to the paper ID numbers reported in Tab. A2, in the appendix).

The *loci* of the points with the same Y_i value, i.e., the so-called *equivalence classes* or *iso-Y_i contour lines*, can be represented on the F_J-F_C plane. When adopting a linear aggregation model (like the one in Eq. 1), iso-Y_i are straight lines (see also Fig. 1(b)). For the purpose of example, Fig. 6 shows four lines for the Y_i values corresponding to $F_Y \approx 20\%$, 50%, 70% and 90% respectively; in this case, w was set to 0.4.

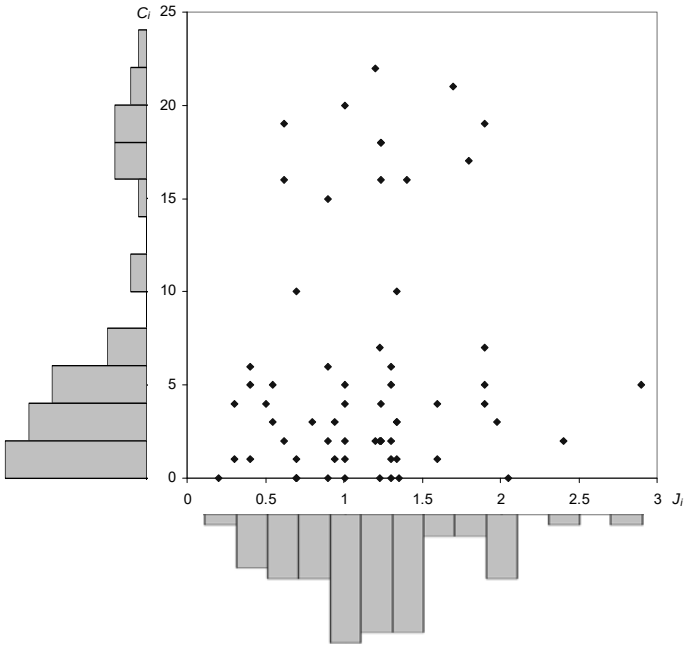

Fig. 4. Representation of the J_i and C_i values and relevant distributions, related to the 64 fictitious scientific papers in Tab. A2 (in the appendix).

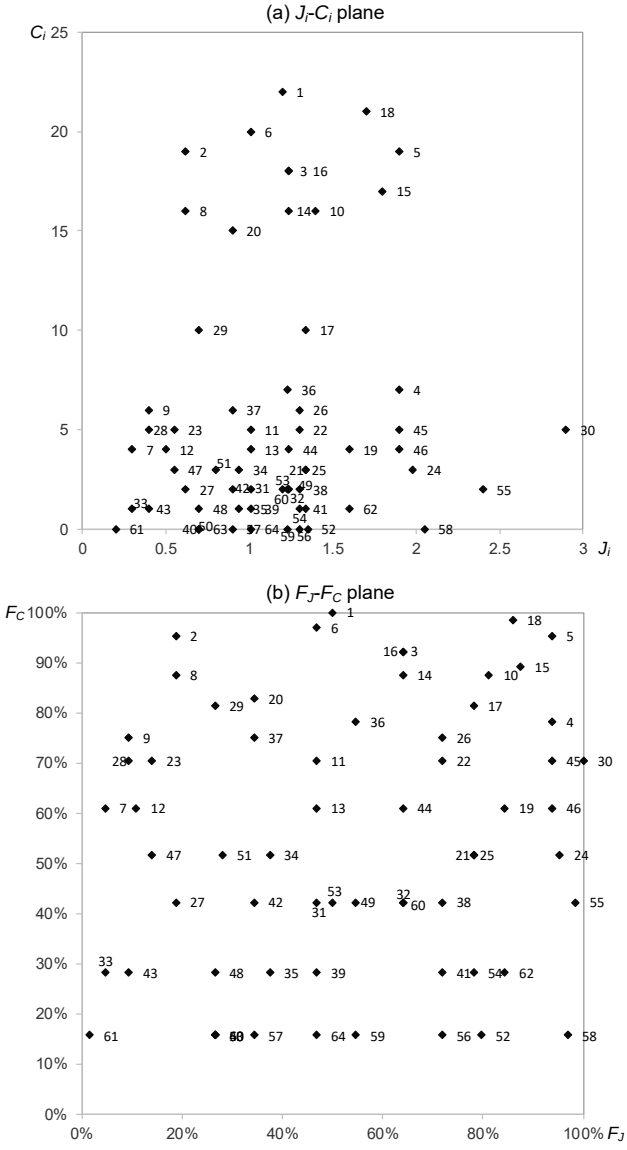

Fig. 5. Comparison of the J_i-C_i map and F_J-F_C map for 64 fictitious scientific papers, in the same *SC* and issue year. Numeric labels refer to the paper ID numbers reported in the first column of Tab. A2, in the appendix.

From the perspective of the Y_i indicator, two (or more) points/papers on the same oblique line (Fig. 6) are considered equivalent. Although this may sound reasonable, it is a source of possible distortions. In fact, referring to the initial scales of C_i and J_i, the iso-Y_i contour lines have unpredictable form, as they are influenced by the empirical distributions of the C_i and J_i values (see the representation in Fig. 7) (ROARS, 2016).

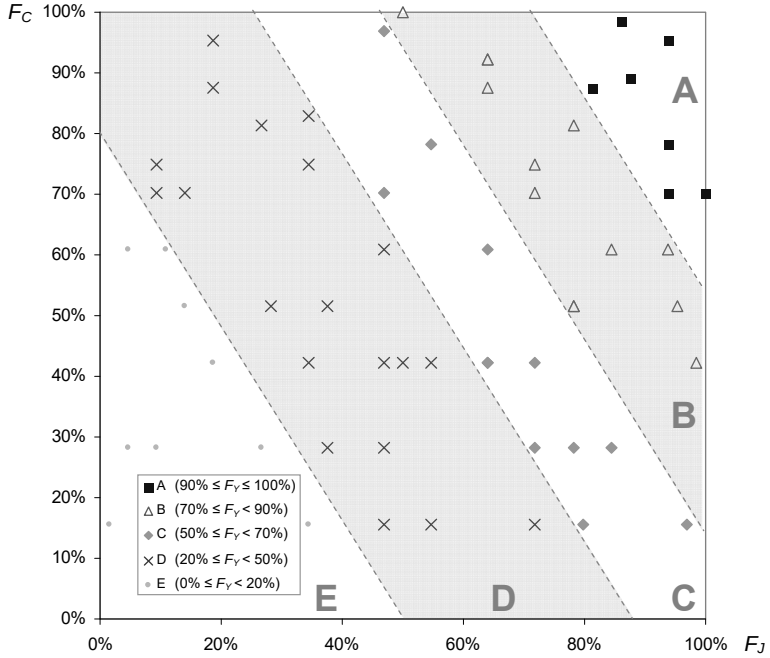

Fig. 6. Iso-Y_i contour lines for the F_J-F_C plane, relating to the data shown in Fig. 5(b). The original data are reported in Tab. A2 (in the appendix). Y_i has been calculated setting w = 0.4.

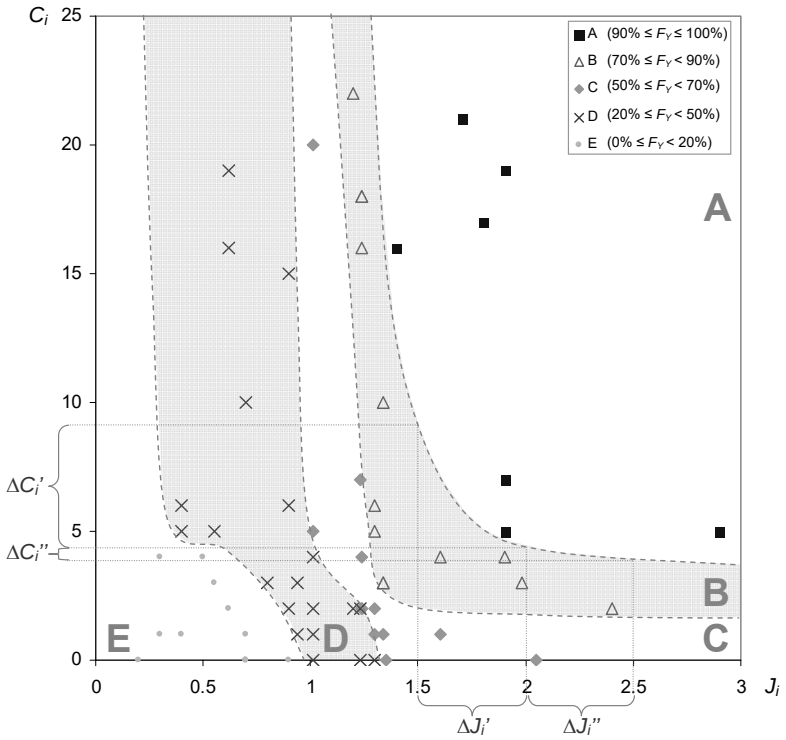

Fig. 7. Representation of the iso-Y_i contour lines in the J_i–C_i plane, for the 64 fictitious papers reported in Tab. A2 (in the appendix). Y_i has been calculated setting w = 0.4. For a generic iso-Y_i line, e.g., that in the borderline between the A and B merit classes – identical variations in J_i (e.g., $\Delta J_i' = \Delta J_i'' = 0.5$) may correspond to very different variations in C_i (i.e., $\Delta C_i' \approx 4.5 \neq \Delta C_i'' \approx 0.5$) and vice versa.

For example, assuming that the distribution of the C_i values changes into that of the C_i' values reported in Tab. A3 (in the appendix), while that of the J_i values remains unchanged, the new contour lines would be deformed significantly with respect to the initial ones (see Fig. 8(a) and (b)). Similar uncontrolled variations can result when introducing small changes in w; for example, Fig. 8(c) represents new iso-Y_i contour lines, when using w' = 0.6 instead of w = 0.4.

In light of the above observations, a new question arises: *what is the rationale for considering two points laying on the same line as equivalent*? We believe that there is no convincing conceptual or empirical reason that can justify this kind of equivalence. The "instability" related to the equivalence classes is simply a negative consequence of the above-described improper aggregation of C_i and J_i.

It can also be noticed that the *substitution rate* between C_i and J_i – defined as *the rate at which the C_i value can be increased/decreased in exchange for a decrease/increase in the J_i value, maintaining the same Y_i value* – is not constant. The example in Fig. 7 shows that – for a generic iso-Y_i line, e.g., that in the borderline between the A and B merit classes – identical variations in J_i (e.g., $\Delta J_i' = \Delta J_i'' = 0.5$) may correspond to very different variations in C_i (i.e., $\Delta C_i' \approx 4.5 \neq \Delta C_i'' \approx 0.5$) and vice versa. In other words, the substitution rate is not constant over the J_i-C_i plane, as it depends on the C_i and J_i values related to the i-th paper of interest. In addition, the fact that the distributions of the C_i and J_i values tend to vary depending on the *SC* and issue year of the papers of interest, makes substitution rates even more variable in an uncontrolled way. *What is the rationale behind*?

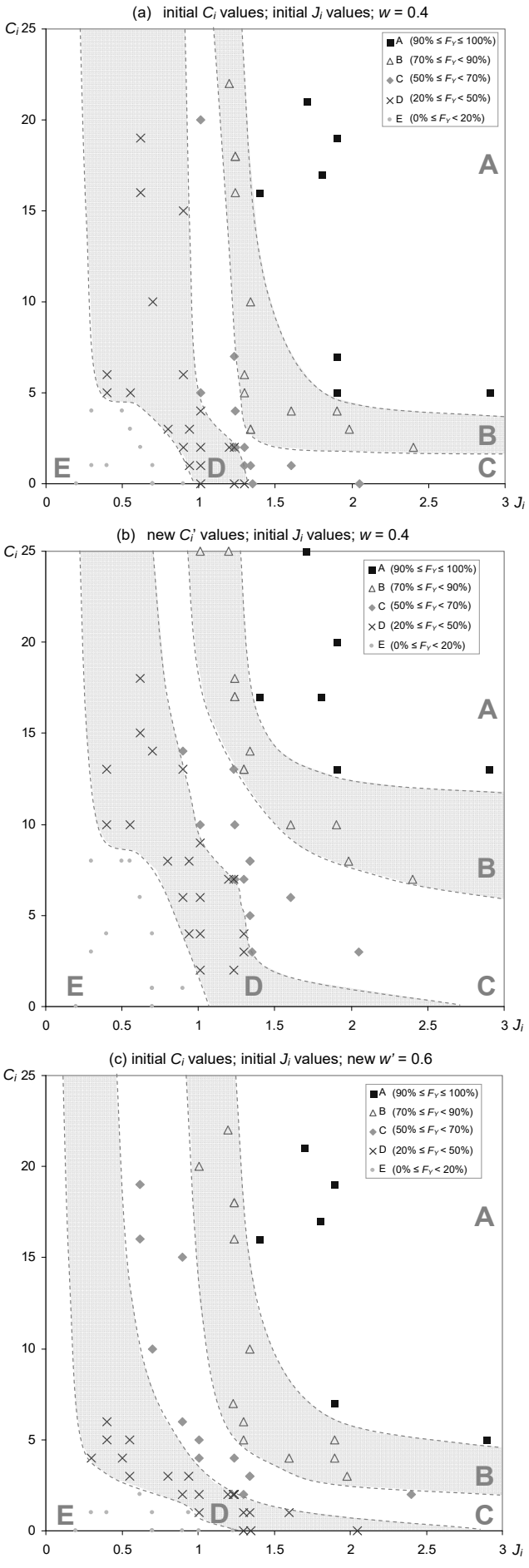

Fig. 8. Variation of the iso-Y_i contour lines when varying the C_i values and the w value in use.

3.3.4 Score assignment to merit classes

Having associated each paper with a Y_i value, the corresponding percentile rank $F_Y(Y_i) \in [0, 100\%]$ can be determined. Consistently with the description in Sect. 3.3.2, this operation may distort the distance relationships (if any) among the initial Y_i values, generating a new indicator $F_Y(Y_i)$ that only preserves the equivalence and order relationships. Next, each paper receives a score (S_i) depending on the merit class related to the relevant $F_Y(Y_i)$ values; this operation can be represented graphically through the function in Fig. 9, which is weakly monotonically increasing. This transformation further degrades the scale of $F_Y(Y_i)$ to another ordinal scale (of S_i) with much lower resolution, due to the limited number of levels (i.e., five only); e.g., papers with different Y_i and therefore different $F_Y(Y_i)$ values can be mapped into the same merit class, obtaining the same S_i score.

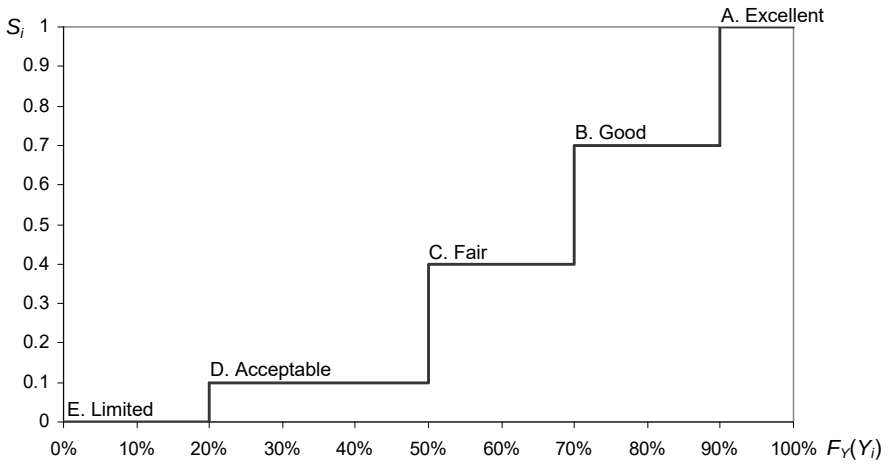

Fig. 9. Graphical representation of the transformation adopted to determine the score of each *i*-th paper (S_i), depending on the relevant $F_Y(Y_i)$ value.

This other mapping function is questionable for three reasons:

1. The order relationships among papers with different Y_i values but in the same merit class are partly lost;
2. The S_i score assigned to each class is purely conventional and therefore arbitrary;
3. The scores related to papers from the same institution are then summed up; this operation is not permissible for indicators defined on ordinal scales (cf. Sect. 2). In other words, the initial ordinal scale of S_i is unduly promoted to a cardinal one (interval or ratio scale).

Despite the criticism reported in this sub-section, we understand that the questionable discretization of the Y_i percentile ranks into corresponding classes is an operation that ANVUR was forced to introduce, in order to fulfil the Ministerial Decree of 27 June 2015 by MIUR (2015) (see Tab. 1 in Sect. 2). For this reason, we think the vulnerability described in this section is considerably less serious than those reported in Sects. 3.3.2 and 3.3.3.

Finally, we note that the fact that the criterion for assigning merit classes adopted in the VQR 2011-2014 is significantly different from that one adopted in the VQR 2004-2010 (ANVUR, 2011) makes

any direct comparison between the results of the two different exercises difficult and, in some ways, incorrect.

3.3.5 Summary of the dubious operations

Summarizing the content of the previous sub-sections, we present a brief outline of the afore-described dubious operations in the bibliometric evaluation procedure of the VQR 2011-2014 (see Tab. 7).

Tab. 7. Summary of the questionable normalization/aggregation operations, in the bibliometric evaluation procedure of individual papers, according to the VQR 2011-2014.

Critical issues	Short description
1. Normalization of C_i and J_i using F_C and F_J.	These operations downgrade the initial ratio scales of C_i and J_i to ordinal scales (i.e., F_C and F_J).
2. Aggregation of $F_C(C_i)$ and $F_J(J_i)$ through a weighted sum.	This operation, which is prohibited for indicators defined on nominal or ordinal scales, has some distorting effects: - unpredictable equivalence classes iso-Y_i; - unpredictable and variable substitution rate between C_i and J_i.
3. Normalization of Y_i through F_Y.	This operation may distort the distance relationships (if any) among the initial Y_i values.
4 Score assignment to the (initial) merit classes.	This transformation deteriorates the resolution of the F_Y indicator. The aggregation of the S_i scores by a sum is incorrect, as these scores are defined on an ordinal scale.

The authors are aware that defining adequate indicators is a difficult task (Franceschini data et al., 2007); nevertheless, they believe that the bibliometric evaluation process of the VQR 2011-2014 contains too many questionable operations. Also, even if (erroneously) deciding to combine C_i and J_i, we believe that this could be done avoiding dubious transformations/normalizations that alter the scales of the initial data.

3.4 Decisional autonomy to GEVs

A presumed improvement of the VQR 2011-2014 with respect to the previous exercise is the increased decisional autonomy to the panel of experts (GEVs), in defining some parameters/indicators related to the bibliometric evaluation procedure (Benedetto, 2016; Benedetto and Setti, 2016; Anfossi et al., 2016). In our opinion, in the absence of solid and reasonable guidelines, this autonomy may sound like "abandoning GEVs to their fate". Our concerns stem from two different reasons: first, since it is (implicitly) assumed that GEV members necessarily have specialized competences in bibliometric evaluation in their research areas (Abramo and D'Angelo, 2015), and secondly, since several operations of selection and "calibration" of the metrics may be tricky, even assuming that GEV members really have those competences.

Although much will depend on how GEVs will work and the assistance that they will receive by ANVUR, we believe that three potentially tricky operations are:

1. Selection of appropriate journal metrics (see Tab. 2), to be combined with C_i for the bibliometric evaluation of the papers in a certain *SC* and issue year. The GEVs' freedom to choose between different types of journal metrics seems pointless: given that the C_i values are neither field-

normalized nor normalized according to the scientific reputation of the citing papers (Franceschini and Maisano, 2014), it is "asymmetric" to combine them with journal metrics implementing a field normalization (such as SNIP) or a normalization based on the reputation of the authors (such as SJR or AF). For this reason, we are quite surprised to read some statements by presumed experts stating that a certain journal metric is "totally inadequate", while another one is appropriate for a bibliometric evaluation of the papers presented in a certain area[6] (ANVUR, 2015b, page 13).

2. Choice of the weight (w) to be used when aggregating the F_C and F_J values, through the model in Eq. 1. Given the conceptual problems highlighted in Sect. 3.3, choosing the "right" value of w seems rather adventurous. One of the obstacles is the uncontrollability of the substitution rate between the C_i and J_i indicators, as discussed in Sect. 3.3.3. ANVUR does not provide precise guidelines for choosing the values of w, probably because it would be very difficult to formulate them. The only indication[7] is that, for the more recent papers, it would be appropriate to give greater weight to J_i than C_i.
3. In cases of wide discrepancy between the C_i and J_i values (see the grey areas in Fig. 1(b)), GEVs may decide to complement the automatic classification of papers with an additional *informed-peer-review* assessment[8] (ANVUR, 2015b; Anfossi et al., 2016). This probably makes sense for papers with low C_i and high J_i values, since they can be seen as papers of little impact with the sole merit of being published in journals generally containing papers of high impact. Conversely, it does not seem reasonable that papers with high C_i and low J_i values are re-assessed, as they have the merit of having achieved a relatively high impact, although being part of off-peak journals. Moreover, the right to "amend" the result of the bibliometric classification through the *informed-peer-review* assessment seems a further way to reduce the repeatability and increase the subjectivity of the whole evaluation process.

[6] A document describing the evaluation criteria that GEVs are going to use for the "Mathematics and Computer Science" area (ANVUR, 2015b, page 13) reports (translated from Italian): *We excluded the IF and IPP because it was verified that the indications provided by pure impact indicators, i.e., non field-normalized (SNIP) or calculated without a selection of the journals in the area of interest, are totally inadequate to measure the impact of the journals in that area.*

[7] *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. […] It is therefore possible to assign more relevance to one of the two dimensions depending on, say, the year of publication or the citation habits of specific disciplines* (Anfossi et al., 2016, page 676).

[8] The basic concept of *informed peer review* is that a judicious application of specific bibliometric indicators and other data concerning the papers examined (e.g., abstract, brief description, any awards/reviews received by these papers, ORCID of the co-authors, etc.) may inform the process of peer review, depending on the exact goal and context of the assessment. According to Moed (2007), both metrics and peer review have their strengths and limits. The challenge is to combine the two methodologies in such a way that the strengths of the first compensates for the limitations of the second and vice versa. However, it matters a lot exactly which forms of peer review and which specific dimensions of peer review are being related to exactly which bibliometric indicators. It is also important to define exactly how these bibliometric indicators are being measured and on the basis of which data sets. Bibliometric measures ought not by definition to be seen as the objective benchmark against which peer review is to be measured (Wouters et al., 2015, page 65).

3.6 Compatibility between peer review and bibliometric analysis

A very delicate point of the VQR 2011-2014, which has been inherited from the VQR 2004-2010, is the presumed "interchangeability" between the assessment through bibliometric indicators and that through peer review, for bibliometric areas. As described in Sect. 2, researchers in these areas may choose the type of evaluation for each of the papers submitted. Moreover, some papers subject to bibliometric assessment may be evaluated through an additional informed-peer-review assessment (ANVUR, 2015a; 2015b; 2015c).

According to some bibliometricians, the problem of the correlation between the results of the bibliometric evaluation and those of the peer review process is controversial and, to date, the alignment between the results of peer review and bibliometric analysis is still an open question (Wouters et al., 2015). ANVUR declares the importance of this presumed correlation for the effectiveness of hybrid research evaluation exercises like the VQR, and claims that the previous VQR 2006-2011 met this requirement (Bertocchi et al., 2016). On the other hand, Baccini and De Nicolao (2016a; 2016b) argue that the results related to the VQR 2004-2010 show a rather poor correlation, except in a specific area (i.e., Economics); they also argue that, in this specific case, results of the peer review were influenced by those of bibliometric evaluation, leading to abnormally high correlation.

4. Conclusions

The major original contribution of this paper is to collect, organize and develop the criticism directed to the bibliometric assessment procedure of the VQR 2011-2014, with the aim of encouraging the debate on how to improve future research assessment exercises in Italy and, maybe, in other countries. Several pedagogical examples were introduced to support the description. Three of the more critical methodological vulnerabilities of the VQR 2011-2014, partly inherited from the VQR 2004-2010, are:

1. The small number of papers evaluated for each researcher makes the results of the whole exercise inappropriate to assess the average quality nor the level of *excellence* of research institutions.
2. Incorrect and anachronistic use of journal metrics for assessing individual papers (Seglen, 1997; Levine, 2011; DORA 2013; IEEE, 2013; Marx and Bornmann, 2013; Ware and Mabe, 2015; Bornmann and Marx, 2016; Mingers, 2016).
3. Misleading normalization and composition of C_i and J_i. These operations may cause additional distortion and lead to the classification into doubtful and not very controllable merit classes.

In light of the arguments gathered and developed in this paper, we are doubtful whether the whole procedure – once completed thanks to the participation of tens of thousands of individuals, including evaluation experts, researchers, administrative staff, government agencies, etc. – will lead

to the desired results, i.e., providing reliable information to rank universities and other research institutions, depending on the quality of their research. We understand the importance of national research assessment exercises for guiding strategic decisions, however, we believe that the VQR 2011-2014 has too many vulnerabilities that make it unsound and often controversial.

We believe that the major vulnerabilities of the VQR 2011-2014 can be (at least partly) solved by (1) extending the bibliometric evaluation procedure to the totality of the papers, (2) avoiding the use of journal metrics in general, and (3) avoiding questionable normalizations/combinations of the indicators in use. It might also be appropriate to introduce consolidated indicators that allow practical comparisons of papers from different areas, such as the so-called "success indicators" (Franceschini et al., 2013; Bornmann and Haunschild, 2016; Rousseau and Rousseau, 2016).

Finally, the introduction of the so-called altmetrics could be a way to solve (at least partly) the old problem of estimating the impact of relatively recent articles, without (mis)using journal metrics.

References

Abramo, G., D'Angelo, C.A., Di Costa, F. (2011). National research assessment exercises: A comparison of peer review and bibliometrics rankings. Scientometrics, 89(3), 929–941.

Abramo, G., D'Angelo, C.A., Di Costa, F. (2014). Inefficiency in selecting products for submission to national research assessment exercises. Scientometric, 98(3), 2069–2086.

Abramo, G., D'Angelo, C.A. (2015) The VQR, Italy's second national research assessment: Methodological failures and ranking distortions. Journal of the Association for Information Science and Technology, 66(11): 2202-2214.

Ancaiani, A., Anfossi, A. F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., et al. (2015). Evaluating scientific research in Italy: The 2004–10 research evaluation exercise. Research Evaluation, 24(3): 242-255.

Anfossi, A., Ciolfi, A., Costa, F. (2015). Looking beyond the Italian VQR 2004-2010: Improving the Bibliometric Evaluation of Research, Proceedings of the 15th International Society of Scientometrics and Informetrics (ISSI) Conference, 1200-1207, 29 June - 3 September 2015, Istanbul, Turkey, ISBN: 978-975-518-381-7.

Anfossi, A., Ciolfi, A., Costa, F., Parisi, G., Benedetto, S. (2016). Large-scale assessment of research outputs through a weighted combination of bibliometric indicators. Scientometrics, 107(2): 671-683.

ANVUR (2011). Valutazione della Qualità della Ricerca 2004-2010 (VQR 2004-2010) – Bando di partecipazione, Roma, http://www.anvur.org/attachments/article/122/bando_vqr_def_07_11.pdf [retrieved on July 2016].

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].

Baccini, A., De Nicolao, G. (2016a). Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. Scientometrics, 108(3): 1651-1671.

Baccini, A., De Nicolao, G. (2016b). Reply to the comment of Bertocchi et al. Scientometrics, 108(3), 1675-1684.

Benedetto, S. (2016) Valutazione della ricerca, quell'algoritmo è affidabile. Lavoce.info, http://www.lavoce.info/archives/41481/valutazione-della-ricerca-quellalgoritmo-e-affidabile/ [retrieved on July 2016].

Benedetto, S., Setti, G. (2016) Un'analisi empirica dell'algoritmo di classificazione bibliometrica della VQR 2011-2014, available at: http://www.lavoce.info/wp-content/uploads/2016/06/algoritmo-analisi-empirica.pdf.
Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C.A., Peracchi, F. (2016) Comment to: Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. Scientometrics, 108(1): 349-353.
Bohannon, J. (2016). Hate journal impact factors? New study gives you one more reason. Science, posted in Scientific CommunityTechnology, DOI: 10.1126/science.aag0643.
Bornmann, L., Leydesdorff, L., Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. Journal of Informetrics, 7(1): 158-165.
Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. Journal of Informetrics, 8(4): 895-903.
Bornmann, L., Marx, W. (2016). The journal Impact Factor and alternative metrics. EMBO reports, 17(8), 1094-1097.
Bornmann, L., Haunschild, R. (2016). Citation score normalized by cited references (CSNCR): The introduction of a new citation impact indicator. Journal of Informetrics, 10(3): 875-887.
Costas, R., Zahedi, Z., Wouters, P. (2015). Do “altmetrics” correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. Journal of the Association for Information Science and Technology, 66(10): 2003-2019.
De Bellis, N. (2009). Bibliometrics and citation analysis: From the science citation index to cybermetrics, Scarecrow Press, Lanham, MD (2009).
Didegah, F., Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. Journal of Informetrics, 7(4): 861-873.
DORA (2013). San Francisco Declaration on Research Assessment, http://www.ascb.org/dora/ [retrieved on July 2016].
Franceschini, F., Galetto, M., Maisano, D. (2007) Management by Measurement: Designing Key Indicators and Performance Measurement Systems. Springer, Berlin.
Franceschini, F., Maisano, D. (2011). Proposals for evaluating the regularity of a scientist's research output. Scientometrics, 88(1), 279-295.
Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2013). Evaluating research institutions: The potential of the success-index. Scientometrics, 96(1): 85-101.
Franceschini, F., Maisano, D. (2014). Sub-field normalization of the IEEE scientific journals based on their connection with Technical Societies. Journal of Informetrics, 8(3): 508-533.
Geuna, A., Piolatto, M. (2016). Research assessment in the UK and Italy: Costly and difficult, but probably worth it (at least for a while). Research Policy, 45(1): 260-271.
Hicks, D. (2012). Performance-based university research funding systems. Research Policy, 41(2): 251-261.
Kreifeldt, J.G., Nah, K. (1995). Adding and Subtracting Percentiles—How bad can it be?. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 39, No. 5, pp. 301-305). SAGE Publications.
IEEE (2013) Appropriate use of bibliometric indicators for the assessment of journals, research proposals, and individuals,” September 2013, http://www.ieee.org/publications_standards/publications/rights/ieee_bibliometric_statement_sept_2013.pdf [retrieved on July 2016].
Lett, N. (2013) Beware the impact factor, Nature Materials, 12(89), doi:10.1038/nmat3566.
Levine, J. (2011) IEEE, Personal Communication, 23 September 2011.
Lozano, G. A., Larivière, V., Gingras, Y. (2012). The weakening relationship between the impact factor and papers' citations in the digital age. Journal of the American Society for Information Science and Technology, 63(11): 2140-2145.
Martin, B.R. (2016). Editors’ JIF-boosting stratagems–Which are appropriate and which not?. Research Policy, 45(1): 1-7.
Marx, W., Bornmann, L. (2013). Journal Impact Factor:“the poor man’s citation analysis” and alternative approaches. European Science Editing, 39(3): 62-63.
Mingers, J., & Yang, L. (2016). Evaluating Journal Quality: A Review of Journal Citation Indicators and Ranking in Business and Management. arXiv preprint arXiv:1604.06685.
MIUR (2015). Decreto Ministeriale 27 giugno 2015 n. 458, Linee guida valutazione qualita' della ricerca (VQR) 2011-2014. Roma, http://attiministeriali.miur.it/anno-2015/giugno/dm-27062015.aspx [retrieved on July 2016].
Moed, H.F. (2007). The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. Science and Public Policy, 34 (8): 575-583.

ROARS (2016). Return on Academic ReSearch, http://www.roars.it [retrieved on July 2016].
Roberts, F.S. (1979) Measurement theory: with applications to decisionmaking, utility, and the social sciences, Encyclopedia of Mathematics and its Applications, vol. 7, Addison-Wesley, Reading, MA.
Rousseau, R., Rousseau, S. (2016). From a success index to a success multiplier. Theories of Informetrics and Scholarly Communication, Ed. by Sugimoto, Cassidy R., De Gruyter Press, pages:148-164
Schotten, M., El Aisati, M. (2014). The Rise of National Research Assessments, and the Tools and Data That Make Them Work, Elsevier Connect, posted on 17 December 2014, https://www.elsevier.com/connect/the-rise-of-national-research-assessments-and-the-tools-and-data-that-make-them-work [retrieved on July 2016].
Seglen, P.O. (1997). Why the impact factor of journals should not be used for evaluating research. BMJ: British Medical Journal, 314(7079): 498-502.
Stevens, S.S. (1946) On the Theory of Scales of Measurement, Science, 103(2684): 677-680, DOI: 10.1126/science.103.2684.677
Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. PloS one, 8(5): e64841.
Thompson, B. (1993). GRE Percentile Ranks Cannot Be Added or Averaged: A Position Paper Exploring the Scaling Characteristics of Percentile Ranks, and the Ethical and Legal Culpabilities Created by Adding Percentile Ranks in Making, Annual Meeting of the Mid-South Educational Research Association (New Orleans, LA, November 12, 1993).
Voorneveld, M. (2003). Characterization of Pareto dominance. Operations Research Letters, 31(1): 7-11.
Ware, M., Mabe, M. (2015). The STM report: An overview of scientific and scholarly journal publishing (4th edition), http://digitalcommons.unl.edu/scholcom/9/?utm_source=digitalcommons.unl.edu%2Fscholcom%2F9&utm_medium=PDF&utm_campaign=PDFCoverPages [retrieved on July 2016].
Wang, J. (2014). Unpacking the Matthew effect in citations. Journal of Informetrics, 8(2): 329-339.
Wouters, P., Thelwall, M., Kousha, K., Waltman, L., De Rijcke, S., Rushforth, A., Franssen, T. (2015). The Metric Tide: Literature Review (Supplementary Report I to the Independent Review of the Role of Metrics in Research Assessment and Management). HEFCE. DOI: 10.13140/RG.2.1.5066.3520

Appendix

A.1 Additional material

See the following tables.

Tab. A1. Research areas identified in the VQR 2011-2014 and number of members in the relevant evaluation panels (GEVs).

Area	B/NB[(a)]	Description	No. of GEV members
Area 1	B	Mathematics and Computer Science	22
Area 2	B	Physics	33
Area 3	B	Chemistry	22
Area 4	B	Earth Science	15
Area 5	B	Biology	33
Area 6	B	Medicine	58
Area 7	B	Agricultural and Veterinary Sciences	20
Area 8a	B	Architecture	14
Area 8b	B	Civil Engineering	9
Area 9	B	Industrial and Information Engineering	33
Area 10	NB	Antiquities, Philological-Literary and Historical-Artistic Sciences	36
Area 11a	NB	Historical, Philosophical and Pedagogical Sciences	25
Area 11b	B	Psychology	6
Area 12	NB	Law	32
Area 13	B	Economics and Statistics	31
Area 14	NB	Political and Social Sciences	11

[(a)] "B" stands for *bibliometric* while "NB" for *non-bibliometric* area.

Tab. A2. Data concerning C_i, J_i and other indicators related to 64 fictitious papers of a specific *SC* and issue year. Y_i is calculated using the relationship in Eq. 1, having set $w = 40\%$.

Paper ID.	J_i	J_i-rank	$F_J(J_i)$	C_i	C_i-rank	$F_C(C_i)$	Y_i	Y_i-rank	$F_Y(Y_i)$	Merit class
P_1	0.2	1	1.6%	0	10	15.6%	0.10	1	1.6%	E
P_2	0.7	17	26.6%	0	10	15.6%	0.20	3	4.7%	E
P_3	0.9	22	34.4%	0	10	15.6%	0.23	7	10.9%	E
P_4	1.3	46	71.9%	0	10	15.6%	0.38	15	23.4%	D
P_5	0.7	17	26.6%	0	10	15.6%	0.20	3	4.7%	E
P_6	0.7	17	26.6%	0	10	15.6%	0.20	3	4.7%	E
P_7	2.05	62	96.9%	0	10	15.6%	0.48	29	45.3%	D
P_8	1.23	35	54.7%	0	10	15.6%	0.31	10	15.6%	E
P_9	1.35	51	79.7%	0	10	15.6%	0.41	19	29.7%	D
P_{10}	1.01	30	46.9%	0	10	15.6%	0.28	9	14.1%	E
P_{11}	1.3	46	71.9%	1	18	28.1%	0.46	23	35.9%	D
P_{12}	1.34	50	78.1%	1	18	28.1%	0.48	28	43.8%	D
P_{13}	1.6	54	84.4%	1	18	28.1%	0.51	31	48.4%	D
P_{14}	0.94	24	37.5%	1	18	28.1%	0.32	11	17.2%	E
P_{15}	0.7	17	26.6%	1	18	28.1%	0.28	8	12.5%	E
P_{16}	0.4	6	9.4%	1	18	28.1%	0.21	6	9.4%	E
P_{17}	1.01	30	46.9%	1	18	28.1%	0.36	13	20.3%	D
P_{18}	0.3	3	4.7%	1	18	28.1%	0.19	2	3.1%	E
P_{19}	1.23	35	54.7%	2	27	42.2%	0.47	26	40.6%	D
P_{20}	0.62	12	18.8%	2	27	42.2%	0.33	12	18.8%	E
P_{21}	1.24	41	64.1%	2	27	42.2%	0.51	32	50.0%	D
P_{22}	1.2	32	50.0%	2	27	42.2%	0.45	22	34.4%	D
P_{23}	1.01	30	46.9%	2	27	42.2%	0.44	21	32.8%	D
P_{24}	1.24	41	64.1%	2	27	42.2%	0.51	32	50.0%	D
P_{25}	1.3	46	71.9%	2	27	42.2%	0.54	34	53.1%	C
P_{26}	0.9	22	34.4%	2	27	42.2%	0.39	17	26.6%	D
P_{27}	2.4	63	98.4%	2	27	42.2%	0.65	44	68.8%	C
P_{28}	0.8	18	28.1%	3	33	51.6%	0.42	20	31.3%	D
P_{29}	0.55	9	14.1%	3	33	51.6%	0.37	14	21.9%	D
P_{30}	0.94	24	37.5%	3	33	51.6%	0.46	25	39.1%	D
P_{31}	1.34	50	78.1%	3	33	51.6%	0.62	40	62.5%	C
P_{32}	1.34	50	78.1%	3	33	51.6%	0.62	40	62.5%	C
P_{33}	1.98	61	95.3%	3	33	51.6%	0.69	47	73.4%	B
P_{34}	1.01	30	46.9%	4	39	60.9%	0.55	35	54.7%	C
P_{35}	1.9	60	93.8%	4	39	60.9%	0.74	51	79.7%	B
P_{36}	0.3	3	4.7%	4	39	60.9%	0.38	16	25.0%	D
P_{37}	0.5	7	10.9%	4	39	60.9%	0.41	18	28.1%	D
P_{38}	1.24	41	64.1%	4	39	60.9%	0.62	40	62.5%	C
P_{39}	1.6	54	84.4%	4	39	60.9%	0.70	48	75.0%	B
P_{40}	2.9	64	100.0%	5	45	70.3%	0.82	59	92.2%	A
P_{41}	1.9	60	93.8%	5	45	70.3%	0.80	54	84.4%	B
P_{42}	1.01	30	46.9%	5	45	70.3%	0.61	39	60.9%	C
P_{43}	1.3	46	71.9%	5	45	70.3%	0.71	49	76.6%	B
P_{44}	0.55	9	14.1%	5	45	70.3%	0.48	27	42.2%	D
P_{45}	0.4	6	9.4%	5	45	70.3%	0.46	24	37.5%	D
P_{46}	0.4	6	9.4%	6	48	75.0%	0.49	30	46.9%	D
P_{47}	0.9	22	34.4%	6	48	75.0%	0.59	36	56.3%	C
P_{48}	1.3	46	71.9%	6	48	75.0%	0.74	50	78.1%	B
P_{49}	1.23	35	54.7%	7	50	78.1%	0.69	46	71.9%	B
P_{50}	1.9	60	93.8%	7	50	78.1%	0.84	60	93.8%	A
P_{51}	0.7	17	26.6%	10	52	81.3%	0.59	37	57.8%	C
P_{52}	1.34	50	78.1%	10	52	81.3%	0.80	55	85.9%	B
P_{53}	0.9	22	34.4%	15	53	82.8%	0.63	43	67.2%	C
P_{54}	1.24	41	64.1%	16	56	87.5%	0.78	53	82.8%	B
P_{55}	0.62	12	18.8%	16	56	87.5%	0.60	38	59.4%	C
P_{56}	1.4	52	81.3%	16	56	87.5%	0.85	61	95.3%	A
P_{57}	1.8	56	87.5%	17	57	89.1%	0.88	62	96.9%	A
P_{58}	1.24	41	64.1%	18	59	92.2%	0.81	57	89.1%	B
P_{59}	1.24	41	64.1%	18	59	92.2%	0.81	57	89.1%	B
P_{60}	0.62	12	18.8%	19	61	95.3%	0.65	44	68.8%	C
P_{61}	1.9	60	93.8%	19	61	95.3%	0.95	64	100.0%	A
P_{62}	1.01	30	46.9%	20	62	96.9%	0.77	52	81.3%	B
P_{63}	1.7	55	85.9%	21	63	98.4%	0.93	63	98.4%	A
P_{64}	1.2	32	50.0%	22	64	100.0%	0.80	55	85.9%	B

J_i is the value of journal metric related to the publishing journal of the *i*-th paper;
J_i-rank is the corresponding rank position, having sorted the (64) papers of interest increasingly with respect to their J_i values.
$F_J(J_i)$ is the corresponding cumulative probability, considering the distribution of the (64) J_i values available;
C_i is the number of citations accumulated by the *i*-th paper;

C_i-rank is the corresponding rank position, having sorted the (64) papers of interest increasingly with respect to their C_i values.
$F_C(C_i)$ is the corresponding cumulative probability, considering the distribution of the (64) C_i values available;
Y_i is a composite indicator combining C_i and J_i, according to Eq. 1;
Y_i-rank is the corresponding rank position, having sorted the (64) papers of interest increasingly with respect to their Y_i values.
$F_Y(Y_i)$ is the corresponding cumulative probability, considering the distribution of the (64) Y_i values available;
The merit class of each i-th paper depends on the relevant $F_Y(Y_i)$ value, according to the conventions in Tab. 1.

Tab. A3. Data concerning C_i', J_i, and other indicators related to 64 fictitious papers of a specific *SC* and issue year. The J_i values are the same ones reported in Tab. A2, while the C_i' values replace the corresponding C_i ones. Y_i is calculated using the relationship in Eq. 1, having set w = 40%.

Paper ID.	J_i	J_i-rank	$F_J(J_i)$	C_i	C_i-rank	$F_C(C_i)$	Y_i	Y_i-rank	$F_Y(Y_i)$	Merit class
P_1	0.2	1	1.6%	0	1	1.6%	0.02	1	1.6%	E
P_2	0.7	17	26.6%	0	1	1.6%	0.17	4	6.3%	E
P_3	0.9	22	34.4%	1	3	4.7%	0.23	7	12.5%	E
P_4	1.3	46	71.9%	3	8	12.5%	0.48	27	45.3%	D
P_5	0.7	17	26.6%	1	3	4.7%	0.18	4	7.8%	E
P_6	0.7	17	26.6%	1	3	4.7%	0.18	4	7.8%	E
P_7	2.05	62	96.9%	3	8	12.5%	0.63	43	67.2%	C
P_8	1.23	35	54.7%	2	6	9.4%	0.37	20	32.8%	D
P_9	1.35	51	79.7%	3	8	12.5%	0.53	33	57.8%	C
P_{10}	1.01	30	46.9%	2	6	9.4%	0.32	15	25.0%	D
P_{11}	1.3	46	71.9%	4	12	18.8%	0.51	34	48.4%	D
P_{12}	1.34	50	78.1%	5	17	26.6%	0.58	38	60.9%	C
P_{13}	1.6	54	84.4%	6	18	28.1%	0.62	40	65.6%	C
P_{14}	0.94	24	37.5%	4	12	18.8%	0.30	13	21.9%	D
P_{15}	0.7	17	26.6%	4	12	18.8%	0.23	8	15.6%	E
P_{16}	0.4	6	9.4%	4	12	18.8%	0.13	3	4.7%	E
P_{17}	1.01	30	46.9%	4	12	18.8%	0.36	21	31.3%	D
P_{18}	0.3	3	4.7%	3	8	12.5%	0.08	2	3.1%	E
P_{19}	1.23	35	54.7%	7	22	34.4%	0.47	29	40.6%	D
P_{20}	0.62	12	18.8%	6	18	28.1%	0.23	10	12.5%	E
P_{21}	1.24	41	64.1%	7	22	34.4%	0.52	35	53.1%	C
P_{22}	1.2	32	50.0%	7	22	34.4%	0.44	25	37.5%	D
P_{23}	1.01	30	46.9%	6	18	28.1%	0.39	23	34.4%	D
P_{24}	1.24	41	64.1%	7	22	34.4%	0.52	35	53.1%	C
P_{25}	1.3	46	71.9%	7	22	34.4%	0.57	39	59.4%	C
P_{26}	0.9	22	34.4%	6	18	28.1%	0.32	18	25.0%	D
P_{27}	2.4	63	98.4%	7	22	34.4%	0.73	54	78.1%	B
P_{28}	0.8	18	28.1%	8	28	43.8%	0.34	18	29.7%	D
P_{29}	0.55	9	14.1%	8	28	43.8%	0.26	11	18.8%	E
P_{30}	0.94	24	37.5%	8	28	43.8%	0.40	22	35.9%	D
P_{31}	1.34	50	78.1%	8	28	43.8%	0.64	45	68.8%	C
P_{32}	1.34	50	78.1%	8	28	43.8%	0.64	45	68.8%	C
P_{33}	1.98	61	95.3%	8	28	43.8%	0.75	55	84.4%	B
P_{34}	1.01	30	46.9%	9	36	56.3%	0.51	31	48.4%	D
P_{35}	1.9	60	93.8%	10	37	57.8%	0.79	57	89.1%	B
P_{36}	0.3	3	4.7%	8	28	43.8%	0.20	9	10.9%	E
P_{37}	0.5	7	10.9%	8	28	43.8%	0.24	12	17.2%	E
P_{38}	1.24	41	64.1%	10	37	57.8%	0.62	41	64.1%	C
P_{39}	1.6	54	84.4%	10	37	57.8%	0.74	51	82.8%	B
P_{40}	2.9	64	100.0%	13	43	67.2%	0.87	61	95.3%	A
P_{41}	1.9	60	93.8%	13	43	67.2%	0.83	59	90.6%	A
P_{42}	1.01	30	46.9%	10	37	57.8%	0.51	37	51.6%	C
P_{43}	1.3	46	71.9%	13	43	67.2%	0.70	48	75.0%	B
P_{44}	0.55	9	14.1%	10	37	57.8%	0.32	17	23.4%	D
P_{45}	0.4	6	9.4%	10	37	57.8%	0.29	14	20.3%	D
P_{46}	0.4	6	9.4%	13	43	67.2%	0.33	16	28.1%	D
P_{47}	0.9	22	34.4%	13	43	67.2%	0.48	30	42.2%	D
P_{48}	1.3	46	71.9%	13	43	67.2%	0.70	49	75.0%	B
P_{49}	1.23	35	54.7%	13	43	67.2%	0.60	42	62.5%	C
P_{50}	1.9	60	93.8%	13	43	67.2%	0.83	60	90.6%	A
P_{51}	0.7	17	26.6%	14	51	79.7%	0.48	26	43.8%	D
P_{52}	1.34	50	78.1%	14	51	79.7%	0.79	56	87.5%	B
P_{53}	0.9	22	34.4%	14	51	79.7%	0.53	32	56.3%	C
P_{54}	1.24	41	64.1%	17	55	85.9%	0.73	50	78.1%	B
P_{55}	0.62	12	18.8%	15	54	84.4%	0.45	24	39.1%	D
P_{56}	1.4	52	81.3%	17	55	85.9%	0.83	58	90.6%	A
P_{57}	1.8	56	87.5%	17	55	85.9%	0.87	62	96.9%	A
P_{58}	1.24	41	64.1%	17	55	85.9%	0.73	52	78.1%	B
P_{59}	1.24	41	64.1%	18	59	92.2%	0.75	52	85.9%	B

P_{60}	0.62	12	18.8%	18	59	92.2%	0.48	28	46.9%	D
P_{61}	1.9	60	93.8%	20	61	95.3%	0.94	64	100.0%	A
P_{62}	1.01	30	46.9%	25	62	96.9%	0.67	44	71.9%	B
P_{63}	1.7	55	85.9%	25	62	96.9%	0.90	63	98.4%	A
P_{64}	1.2	32	50.0%	25	62	96.9%	0.69	47	73.4%	B

J_i is the value of journal metric related to the publishing journal of the i-th paper;
J_i-rank is the corresponding rank position, having sorted the (64) papers of interest increasingly with respect to their J_i values.
$F_J(J_i)$ is the corresponding cumulative probability, considering the distribution of the (64) J_i values available;
C_i is the number of citations accumulated by the i-th paper;
C_i-rank is the corresponding rank position, having sorted the (64) papers of interest increasingly with respect to their C_i values.
$F_C(C_i)$ is the corresponding cumulative probability, considering the distribution of the (64) C_i values available;
Y_i is a composite indicator combining C_i and J_i, according to Eq. 1;
Y_i-rank is the corresponding rank position, having sorted the (64) papers of interest increasingly with respect to their Y_i values.
$F_Y(Y_i)$ is the corresponding cumulative probability, considering the distribution of the (64) Y_i values available;
The merit class of each i-th paper depends on the relevant $F_Y(Y_i)$ value, according to the conventions in Tab. 1.

Tab. A4. Data concerning C_i, J_i, and other indicators related to 64 fictitious papers of a specific *SC* and issue year. C_i and J_i values are the same ones reported in Tab. A2, while Y_i is calculated (using the relationship in Eq. 1), having set w = 60%.

Paper ID.	J_i	J_i-rank	$F_J(J_i)$	C_i	C_i-rank	$F_C(C_i)$	Y_i	Y_i-rank	$F_Y(Y_i)$	Merit class
P_1	0.2	1	1.6%	0	10	15.6%	0.10	1	1.6%	E
P_2	0.7	17	26.6%	0	10	15.6%	0.20	3	4.7%	E
P_3	0.9	22	34.4%	0	10	15.6%	0.23	7	10.9%	E
P_4	1.3	46	71.9%	0	10	15.6%	0.38	15	23.4%	D
P_5	0.7	17	26.6%	0	10	15.6%	0.20	3	4.7%	E
P_6	0.7	17	26.6%	0	10	15.6%	0.20	3	4.7%	E
P_7	2.05	62	96.9%	0	10	15.6%	0.48	29	45.3%	D
P_8	1.23	35	54.7%	0	10	15.6%	0.31	10	15.6%	E
P_9	1.35	51	79.7%	0	10	15.6%	0.41	19	29.7%	D
P_{10}	1.01	30	46.9%	0	10	15.6%	0.28	9	14.1%	E
P_{11}	1.3	46	71.9%	1	18	28.1%	0.46	23	35.9%	D
P_{12}	1.34	50	78.1%	1	18	28.1%	0.48	28	43.8%	D
P_{13}	1.6	54	84.4%	1	18	28.1%	0.51	31	48.4%	D
P_{14}	0.94	24	37.5%	1	18	28.1%	0.32	11	17.2%	E
P_{15}	0.7	17	26.6%	1	18	28.1%	0.28	8	12.5%	E
P_{16}	0.4	6	9.4%	1	18	28.1%	0.21	6	9.4%	E
P_{17}	1.01	30	46.9%	1	18	28.1%	0.36	13	20.3%	D
P_{18}	0.3	3	4.7%	1	18	28.1%	0.19	2	3.1%	E
P_{19}	1.23	35	54.7%	2	27	42.2%	0.47	26	40.6%	D
P_{20}	0.62	12	18.8%	2	27	42.2%	0.33	12	18.8%	E
P_{21}	1.24	41	64.1%	2	27	42.2%	0.51	32	50.0%	D
P_{22}	1.2	32	50.0%	2	27	42.2%	0.45	22	34.4%	D
P_{23}	1.01	30	46.9%	2	27	42.2%	0.44	21	32.8%	D
P_{24}	1.24	41	64.1%	2	27	42.2%	0.51	32	50.0%	D
P_{25}	1.3	46	71.9%	2	27	42.2%	0.54	34	53.1%	C
P_{26}	0.9	22	34.4%	2	27	42.2%	0.39	17	26.6%	D
P_{27}	2.4	63	98.4%	2	27	42.2%	0.65	44	68.8%	C
P_{28}	0.8	18	28.1%	3	33	51.6%	0.42	20	31.3%	D
P_{29}	0.55	9	14.1%	3	33	51.6%	0.37	14	21.9%	D
P_{30}	0.94	24	37.5%	3	33	51.6%	0.46	25	39.1%	D
P_{31}	1.34	50	78.1%	3	33	51.6%	0.62	40	62.5%	C
P_{32}	1.34	50	78.1%	3	33	51.6%	0.62	40	62.5%	C
P_{33}	1.98	61	95.3%	3	33	51.6%	0.69	47	73.4%	B
P_{34}	1.01	30	46.9%	4	39	60.9%	0.55	35	54.7%	C
P_{35}	1.9	60	93.8%	4	39	60.9%	0.74	51	79.7%	B
P_{36}	0.3	3	4.7%	4	39	60.9%	0.38	16	25.0%	D
P_{37}	0.5	7	10.9%	4	39	60.9%	0.41	18	28.1%	D
P_{38}	1.24	41	64.1%	4	39	60.9%	0.62	40	62.5%	C
P_{39}	1.6	54	84.4%	4	39	60.9%	0.70	48	75.0%	B
P_{40}	2.9	64	100.0%	5	45	70.3%	0.82	59	92.2%	A
P_{41}	1.9	60	93.8%	5	45	70.3%	0.80	54	84.4%	B
P_{42}	1.01	30	46.9%	5	45	70.3%	0.61	39	60.9%	C
P_{43}	1.3	46	71.9%	5	45	70.3%	0.71	49	76.6%	B
P_{44}	0.55	9	14.1%	5	45	70.3%	0.48	27	42.2%	D
P_{45}	0.4	6	9.4%	5	45	70.3%	0.46	24	37.5%	D
P_{46}	0.4	6	9.4%	6	48	75.0%	0.49	30	46.9%	D
P_{47}	0.9	22	34.4%	6	48	75.0%	0.59	36	56.3%	C
P_{48}	1.3	46	71.9%	6	48	75.0%	0.74	50	78.1%	B
P_{49}	1.23	35	54.7%	7	50	78.1%	0.69	46	71.9%	B
P_{50}	1.9	60	93.8%	7	50	78.1%	0.84	60	93.8%	A
P_{51}	0.7	17	26.6%	10	52	81.3%	0.59	37	57.8%	C

P_{52}	1.34	50	78.1%	10	52	81.3%	0.80	55	85.9%	B
P_{53}	0.9	22	34.4%	15	53	82.8%	0.63	43	67.2%	C
P_{54}	1.24	41	64.1%	16	56	87.5%	0.78	53	82.8%	B
P_{55}	0.62	12	18.8%	16	56	87.5%	0.60	38	59.4%	C
P_{56}	1.4	52	81.3%	16	56	87.5%	0.85	61	95.3%	A
P_{57}	1.8	56	87.5%	17	57	89.1%	0.88	62	96.9%	A
P_{58}	1.24	41	64.1%	18	59	92.2%	0.81	57	89.1%	B
P_{59}	1.24	41	64.1%	18	59	92.2%	0.81	57	89.1%	B
P_{60}	0.62	12	18.8%	19	61	95.3%	0.65	44	68.8%	C
P_{61}	1.9	60	93.8%	19	61	95.3%	0.95	64	100.0%	A
P_{62}	1.01	30	46.9%	20	62	96.9%	0.77	52	81.3%	B
P_{63}	1.7	55	85.9%	21	63	98.4%	0.93	63	98.4%	A
P_{64}	1.2	32	50.0%	22	64	100.0%	0.80	55	85.9%	B

J_i is the value of journal metric related to the publishing journal of the i-th paper;
J_i-rank is the corresponding rank position, having sorted the (64) papers of interest increasingly with respect to their J_i values.
$F_J(J_i)$ is the corresponding cumulative probability, considering the distribution of the (64) J_i values available;
C_i is the number of citations accumulated by the i-th paper;
C_i-rank is the corresponding rank position, having sorted the (64) papers of interest increasingly with respect to their C_i values.
$F_C(C_i)$ is the corresponding cumulative probability, considering the distribution of the (64) C_i values available;
Y_i is a composite indicator combining C_i and J_i, according to Eq. 1;
Y_i-rank is the corresponding rank position, having sorted the (64) papers of interest increasingly with respect to their Y_i values.
$F_Y(Y_i)$ is the corresponding cumulative probability, considering the distribution of the (64) Y_i values available;
The merit class of each i-th paper depends on the relevant $F_Y(Y_i)$ value, according to the conventions in Tab. 1.