VQR 2011-2014 (Bibliometric Algorithm): a comparison with the past

Gianluca Setti
Dipartimento di Ingegneria, Università di Ferrara
GEV09 Coordinator
(2013-2014 IEEE Vice President for Publications)
gianluca.setti@unife.it

Rome, 14 December 2015
(some slides adapted from a presentation by Alberto Anfossi - ANVUR)

Outline

- Research products evaluation in VQR2011-2014
- Bibliometric classification of journal papers
  - VQR2004-2010 algorithm and its drawbacks
  - The solution for VQR2011-2014
- Use of more than one Journal Metric: why do we need to do this?
- A few final remarks

Research Product Evaluation

- GEV09 is **globally responsible** for the evaluation of the products in its own area. In other words, the final responsibility is **not** on the shoulder of each individual GEV member to which a particular product is assigned but of the **entire GEV.**
- For the evaluation process GEV09 uses an *informed peer review,* based on:
  - Bibliometric indicators;
  - Report of *peer reviewers* who are not member of GEV09;
  - A direct evaluation made by members of GEV09;
- At least 51% of the products submitted to VQR2011-2014 must be evaluated using *peer review* (both external or internal). This figure **must not be respected by each GEV individually**
- 10% of the products evaluated trough the bibliometric algorithms is also evaluated trough external *peer review* to study the level of correlation between the two different methodologies. **Such a peer review evaluation does not influence the final classification of the product**

Research Product Evaluation - cont

- The institution submitting the product **specify the scientific area (SSD) which is most appropriate for the evaluation of the product**. This is use to assign the product to the most competent members. **NOTE**: GEV can change the SSD if not considered appropriate
- For GEV09 the choice of the products which need to be evaluated trough *external peer review* is based on:
  - reliability of the bibliometric algorithm (f.i. IR zones)
  - motivated request for an external peer review made by the institution (case of a product in a particularly innovative area)
  - product typology (conference proceedings, books, book chapters, …), i.e. those for which the bibliometric information is incomplete
  - …
- GEV members to which the products is assigned will choose the *peer reviewers,* at least (and typically) two for each product.

Research Product Evaluation - cont

(*from VQR official call*)

The final evaluation is based on

a. ***Originality***
b. ***Methodological rigor*** (clear descript of objectives wrt to the state of the art and clear demonstration trough an appropriate methodology that the results have been obtained)
c. ***Demonstrated or Potential Impact*** in the respective scientific community (capability to have a present or future theoretical or applicative impact in the scientific community)

Research Product Evaluation - cont

After evaluation each product is classified in one of the following 5 (+1) categories:

a. ***Excellent* (weight 1):** the publication is in the **top 10%** of the distribution of all products in the same scientific area
b. ***High quality* (weight 0.7):** the publication is in the **top 10%-top 30%** segment of the distribution of all products in the same scientific area
c. ***Fair* (weight 0.4):** the publication is in the **top 30%-top 50%** segment of the distribution of all products in the same scientific area
d. ***Acceptable* (weight 0.1):** the publication is in the **top 50%-top 80%** segment of the distribution of all products in the same scientific area
e. ***Limited* (weight 0):** the publication is in the **bottom 20%** of the distribution of all products in the same scientific area
f. ***Non-classifiable* (weight 0):** the publication is missing, or it is not published in 2011-2014, or it of a non-admissible kind (i.e. internal report)…

Research Product Evaluation - cont

After evaluation each product is classified in one of the following 5 (+1) categories:

a. ***Excellent* (weight 1):** the publication is in the **top 10%** of the distribution

IMPORTANT – IMPORTANT – IMPORTANT

This division in classes **does not imply anything in terms of classification of the products submitted for evaluation**. It is simply used for the calibration of the bibliometric algorithm or for the calibration of the peer review report

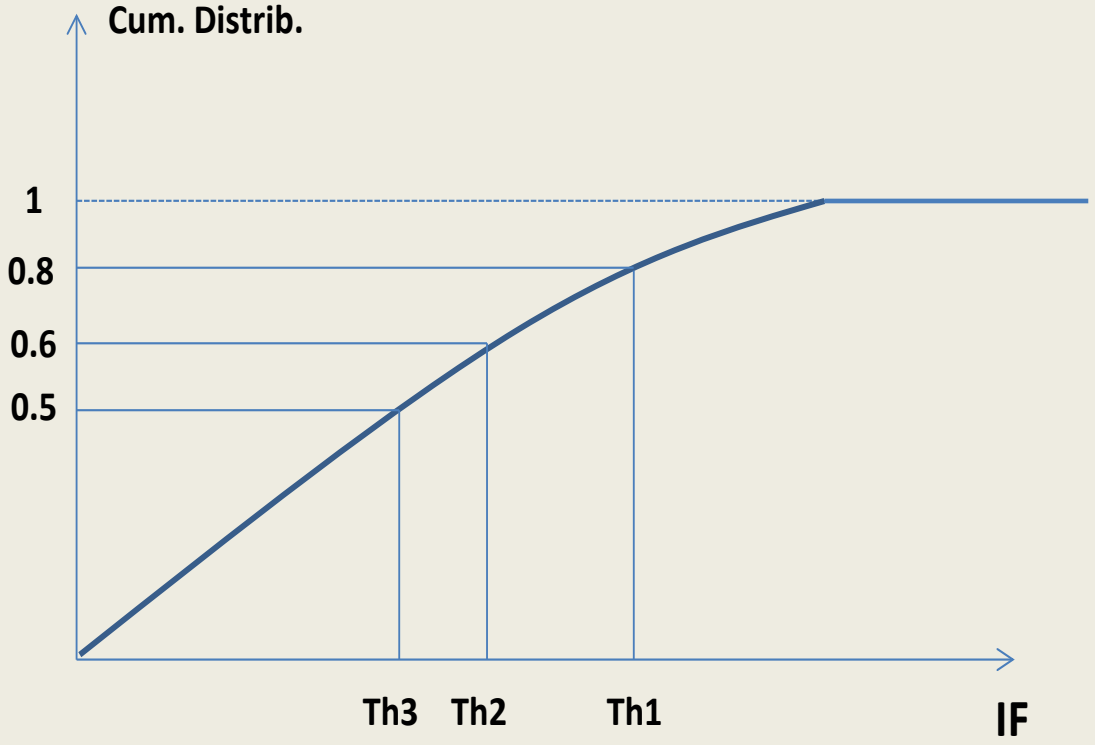
(in other words a result were all product submitted to VQR2011-2014 are excellent **is possible**, even if unlikely)

e. ***Limited* (weight 0):** the publication is in the **bottom 20%** of the distribution of all products in the same scientific area

f. ***Non-classifiable* (weight 0):** the publication is missing, or it is not published in 2011-2014, or it of a non-admissible kind (i.e. internal report)…

VQR2004-10: How did bibliometrics work? - 1

- For each article at submission time the structure indicated: **coordinates of the paper** (journal, ISSN, pages, number, year, DOI, …); **typology of paper** (article/letter, review,…); **scientific sector** (both according to the Italian university classification "settore scientifico disciplinare" and **WoS subject category, Scopus ASJC**)
- FIRST STEP: For **each year and SC/ASJC** one computed the **empirical cumulative distribution function** (CDF) of a bibliometric indicator of journal impact (see later)



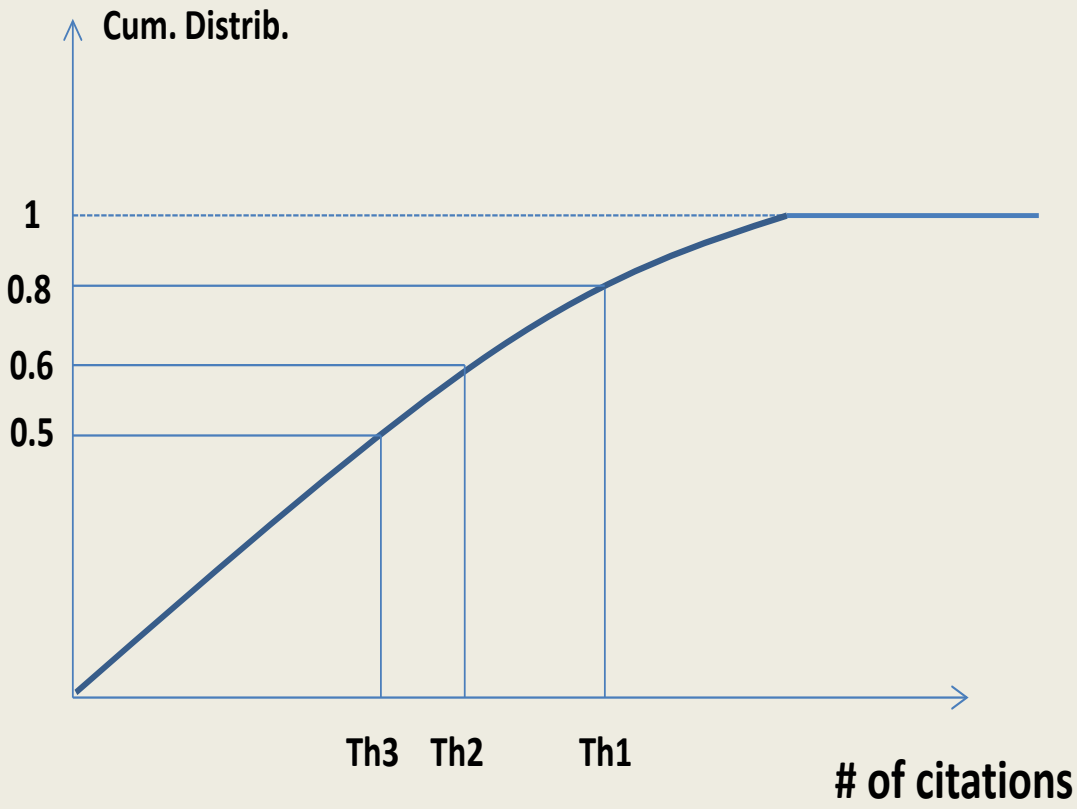
VQR2004-10: How did bibliometrics work? - 2

- The choice of the bibliometric indicator for Scopus was (when the database was used) SJR;
- The choice of the bibliometric indicator for WoS has always been the Impact Factor (IF) apart from:
  - The area of Mathematics used also MCQ (bibliometric indicator from MathSciNet). Two sub-GEVs ("Probability and Mathematical Statistics" and "Applied Mathematics") used a combination of MCQ and 5YIF with **"arbitrary" weights** (4/5 and 1/5 or 2/3 and 1/3). Also value of the indicators was **not normalized** (range of variation could be different) and the 2 indicators were taken **from different databases.**
  - The area of GEV09 (with the exception of SSD ING-INF/05 – Computer Engineering) used a combination of 4 indicators (IF, 5YIF, AI and EF) from the **same database** with **agnostic weights** determined using a Principal Component Analysis of the indicators (once transformed into standardized variables to eliminate the "variability of the dynamic ranges" of the indicators due to their different definition).
  - The SSD ING-INF/05 collaborated with sud-GEV1.1 (SSD INF/01 – Computer Science) to define a unique journal ranking

VQR2004-10: How did bibliometrics work? - 3

- SECOND STEP: One computed the CDF of the number of citations considering all articles presents **in each SC/ASJC and for each year**

- The distribution was divided in **4 regions (from top to bottom called 1, 2,3,4)**, with inclusion probability that was determined to be equal to 0,2; 0,2; 0,1 e 0,5 (thresholds Th1, Th2, Th3).

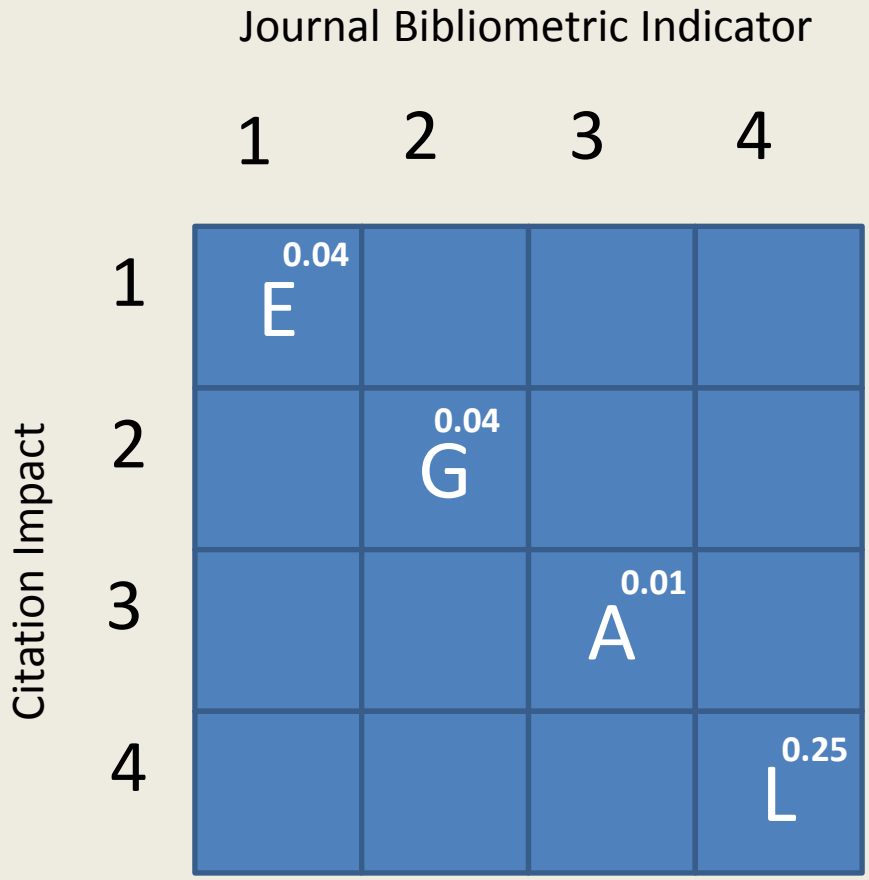


VQR2004-10: How did bibliometrics work? - 4

- THIRD STEP: Each GEV had 2 different classifications available for each journal paper: **one based on the impact of the journal** and one based on the **citation impact of the individual paper**.

- To combine those, the different GEVs constructed a **4x4 classification matrix**

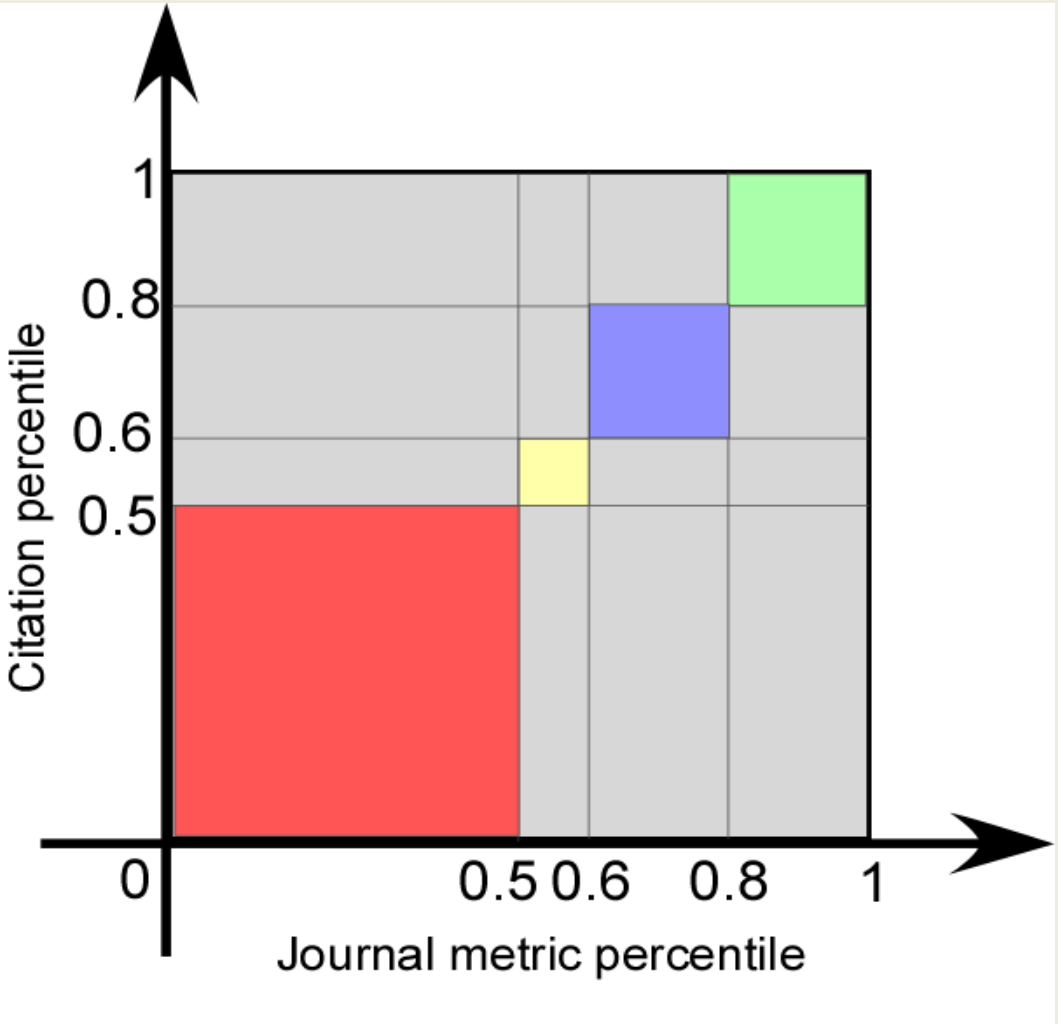
- For the classes where the 2 indicators gave the **same classification**, the final result was obvious (classes Excellent, Good, Acceptable, Limited)

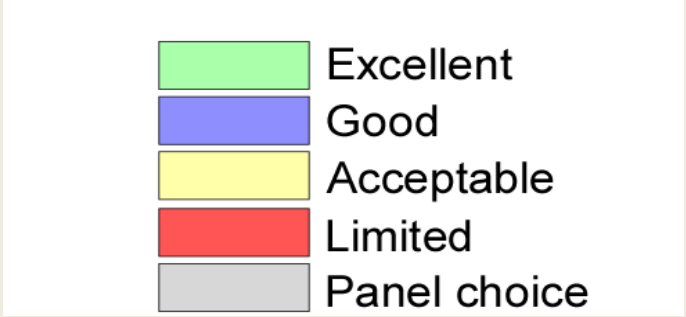


The area of the elements of the matrix in not constant:
{1,1}, {2,2} area = 0,2*0.2 = 0,04 (GEV09 = 0.25*0.2=0.05)
{3,3} area = 0.1*0.1 = 0.01 (GEV09=0.25*0.1=0.025)
{4,4} area = 0,5*0,5 = 0,25 (GEV09=0.25*0.5=0.125)

VQR2004-10: How did bibliometrics work? - 5

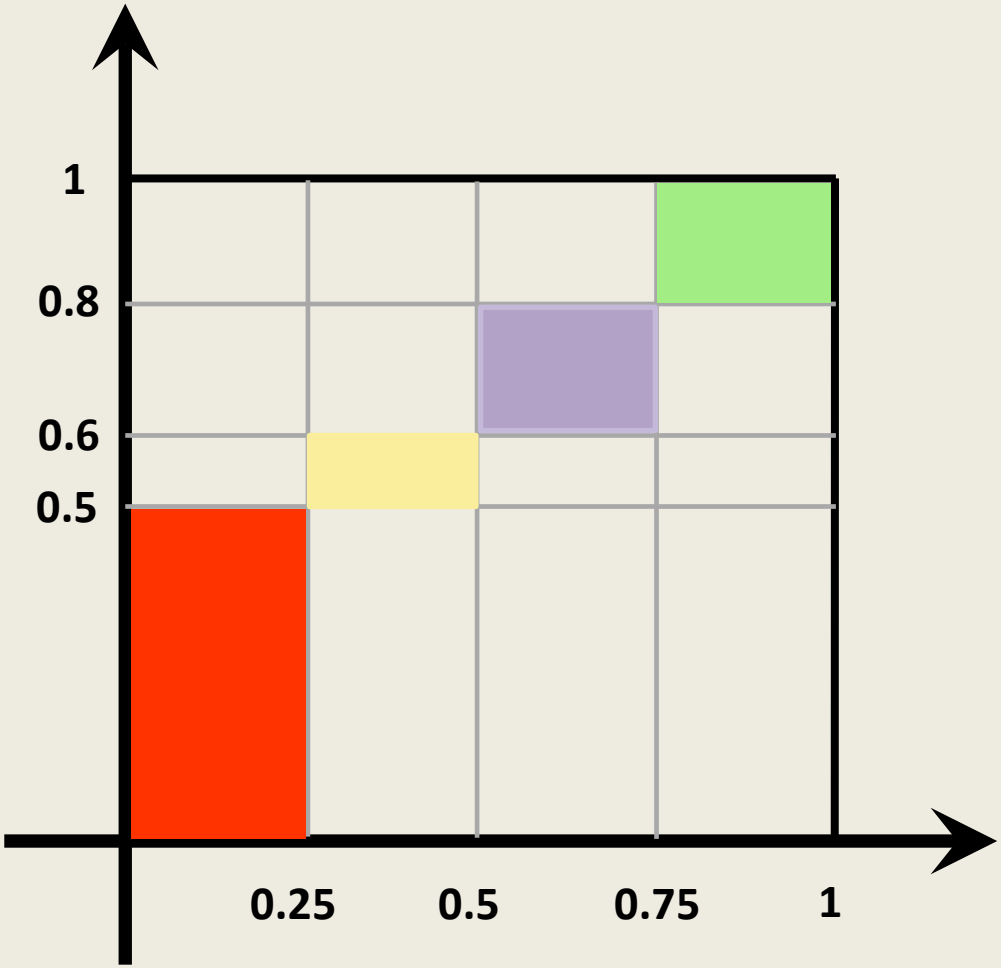
Percentiles representation of the vQr 2004-10 classification matrix for all GEVs but GEV09

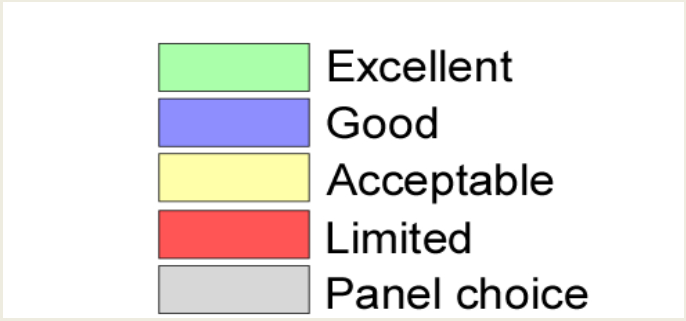




VQR2004-10: How did bibliometrics work? - 5

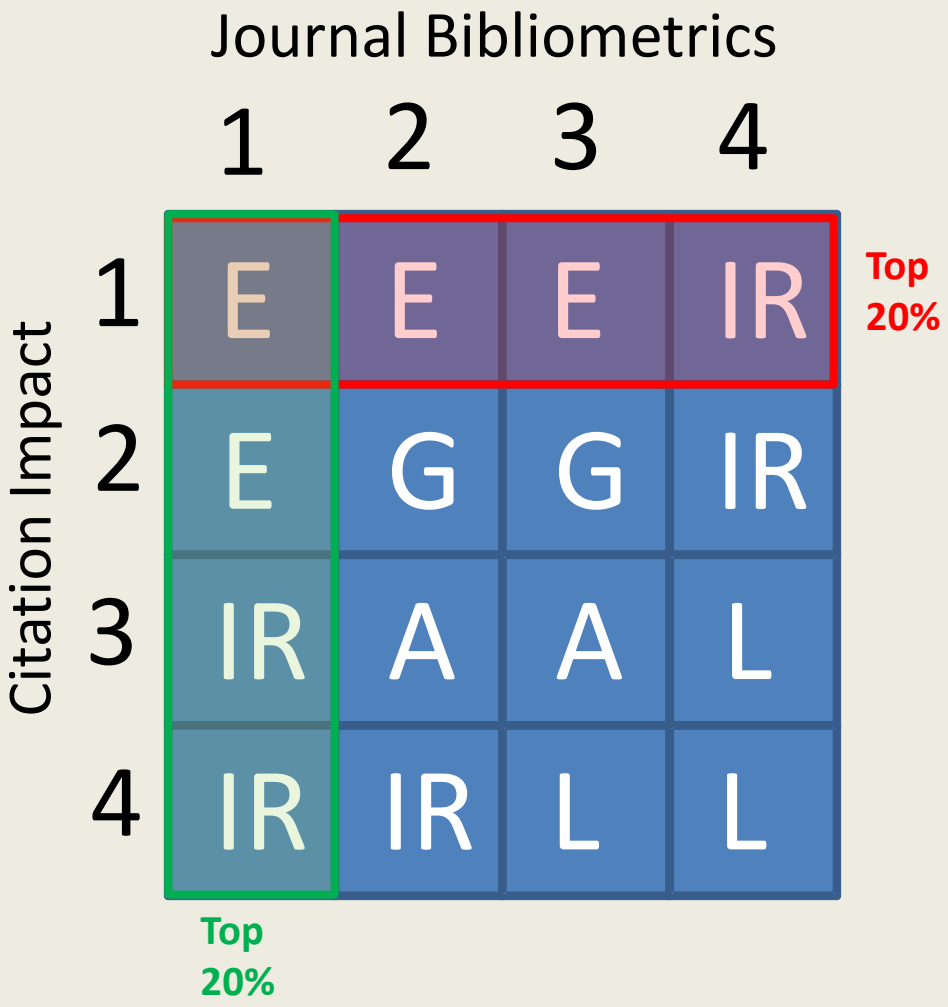
Percentiles representation of the vQr 2004-10 classification matrix for GEV09





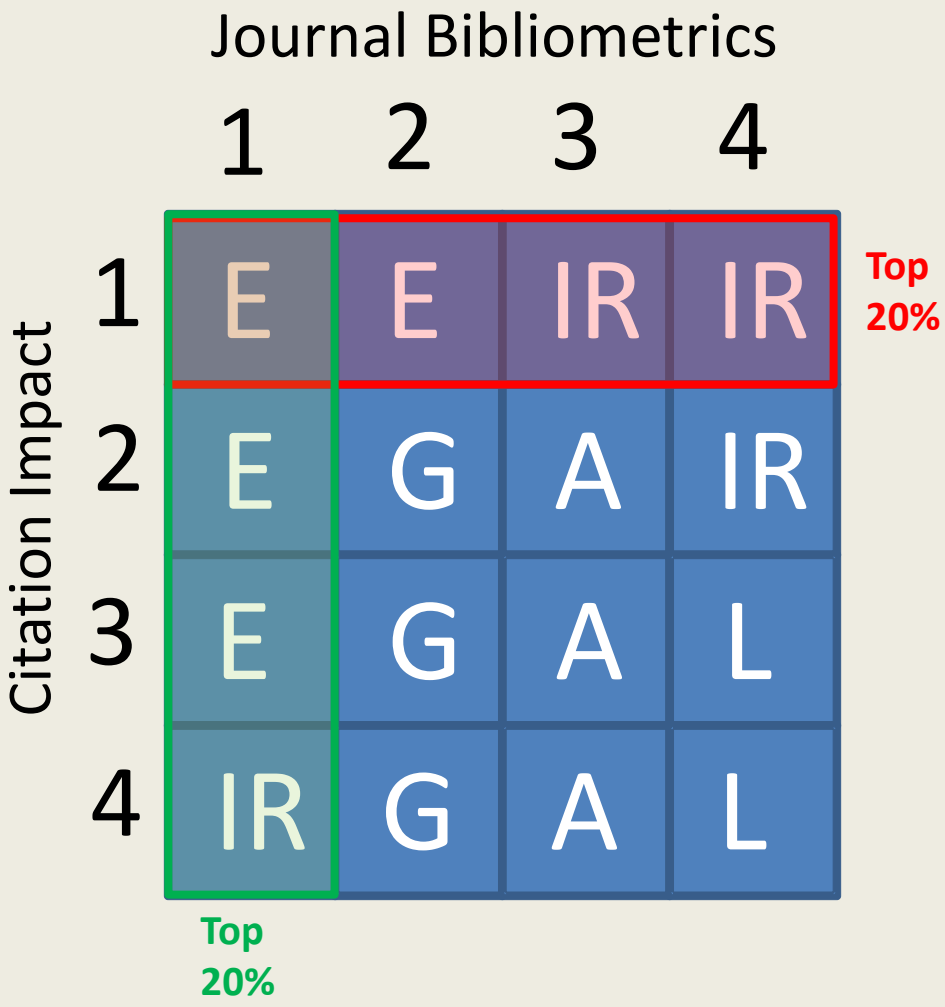
VQR2004-10: How did bibliometrics work? - 6

- If the 2 classifications are not identical the final class will depend:
  - On the choice to give more emphasis to one of the 2 indicators (citations impact vs journal indicator)
  - On the difference between the classification offered by the single indicators (distance from the diagonal elements)
- The choice how to fill the matrix was given to the individual GEVs
- **GEV09**: one matrix for 2004-2007 with more emphasis on citation impact;



VQR 2004-10: How did we operate? - 7

- If the 2 classifications are not identical the final class will depend:
  - On the choice to give more emphasis to one of the 2 indicators (citations impact or journal indicator)
  - On the difference between the classification offered by the single indicators (distance from the diagonal elements)
- The choice how to fill the matrix was given to the individual GEVs
- GEV09: one matrix for 2004-2007 with more emphasis on citation impact; one matrix for 2008-2010 with more emphasis on the journal bibliometric indicator



VQR 2004-10: How did we operate? - 7

- If the 2 classifications are not identical
  - O

- The c
given
- GEV09:
more em
one mat
emphasi
indicato

Journal Bibliometrics

1 2 3 4

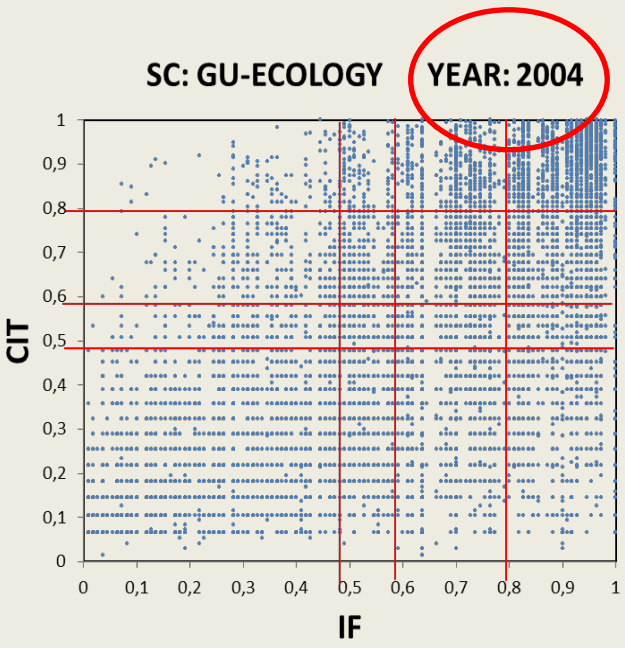
Top 20%

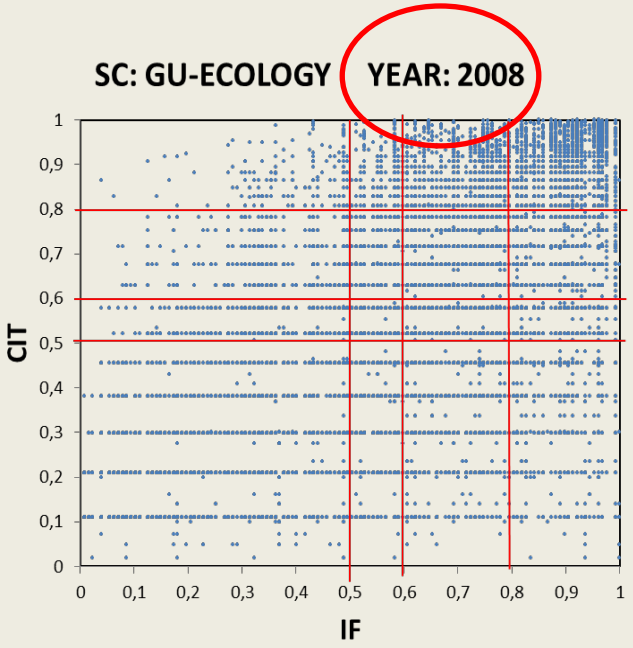
Warnings:

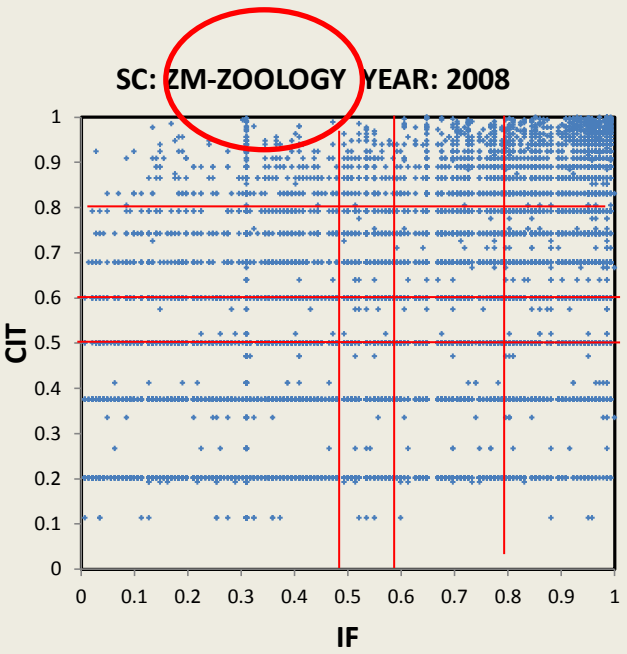
1. No guarantee that the a-priori choice of the element of the matrix **will respect any desired final distribution** in terms of classification (E=top20%;G=second 20% from the top;A=further 10%; L=bottom 50%)
2. The matrix in unique for any given year. Variation among differen SC/ASJC **may be quite large**

The problem is the a-priori choice of the classification matrix without any knowledge of the citation impact and journal bibliometric indicators CDFs

VQR 2004-10: Examples of Classification







- Same SC (Ecology) in two different years (2004 and 2008)
- Two different SC (Ecology and Zoology) in the same year (2008)

As it can be noted the resulting distributions are visibly rather different, despite the classification matrix was the same

Difference Among Areas (again)

- In VQR2004-2010 GEVs were individually responsible for the final compositions of the classification matrices (ANVUR simply guaranteed a general level of uniformity)
- Therefore the classification matrices varied significantly from area to area
- This **did not bias the evaluation within the same area**, but introduced a area-dependent bias, which make impossible to compare different areas **wrt the global research output in the same area** (f.i. to find out if Italian Zoologists compare better that Italian Mechanical Engineers wrt to raking to the rest of the world)

	Area 1	Area 2	Area 3	Area 4	Area 5	Area 6	Area 7	Area 8	Area 9	Area 11
E	37,27%	24,59%	23,43%	23,54%	22,10%	21,95%	35,15%	37,92%	39,67%	36,34%
G	22,32%	27,50%	28,26%	24,59%	21,24%	21,04%	23,52%	23,29%	24,98%	12,46%
A	10,62%	19,82%	10,66%	11,64%	10,68%	12,57%	4,81%	5,27%	13,80%	12,32%
L	29,78%	28,09%	37,64%	40,23%	45,97%	44,44%	36,52%	33,53%	21,55%	38,88

VQR 2011-14: Improved Bibliometric Algorithm

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

Alberto Anfossi[1,2]; Alberto Ciolfi[3]; Filippo Costa[4,5]; Giorgio Parisi[6]; Sergio Benedetto[7]

Fundamental constraints:

- Guarantee **some similarities in terms of methodology** with respect to VQR2004-2010.
- Guarantee the **same classification output** in terms of distribution of the new five categories (Excellent, High Quality, Fair, Acceptable, Limited, that is 10%-20%-20%-30%-20% of the distribution of the research products of the same area)

Main cornerstones:

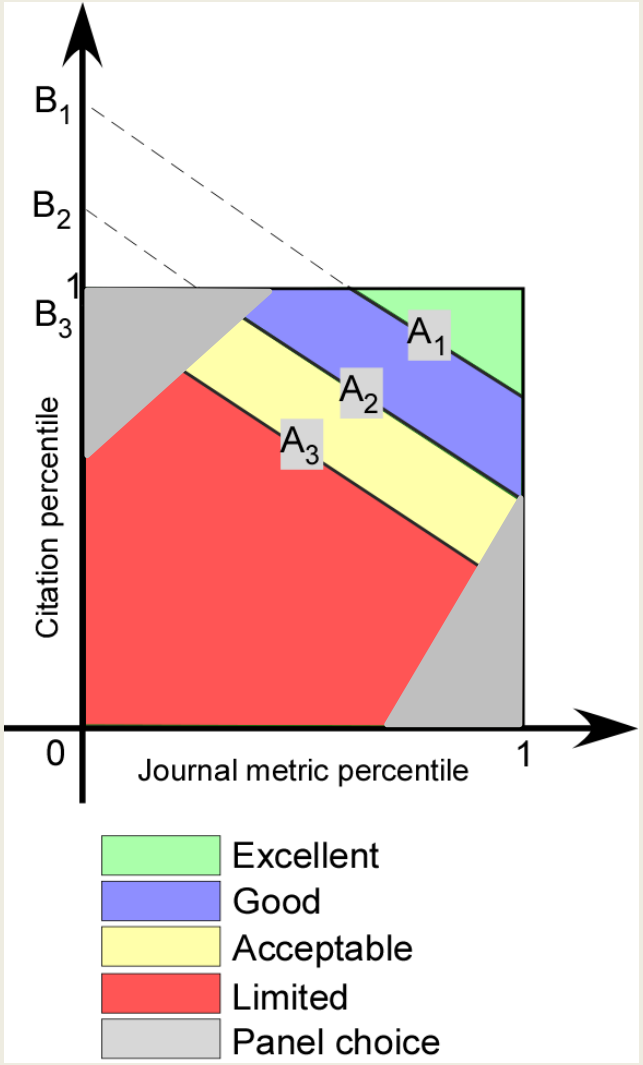
- Maintain the use of the CDF of the 2 variables CIT=citation impact and JB=Journal bibliometric indicator and the intuitive representation of the of publications as points in a scatter plot {CIT,JB}
- Partition the unitary square [0,1]x[0,1] in regions using thresholds lines obtained as a linear combination of CIT and JB;
- Calibrate the position of these thresholds to obtain the desired distribution for each *subject category*, for each year, and according to a degree of freedom specified by the individual GEV (see next few slides)

Improved Bibliometric Algorithm - 1

- The thresholds **are now expressed as linear combination** of CIT and JB

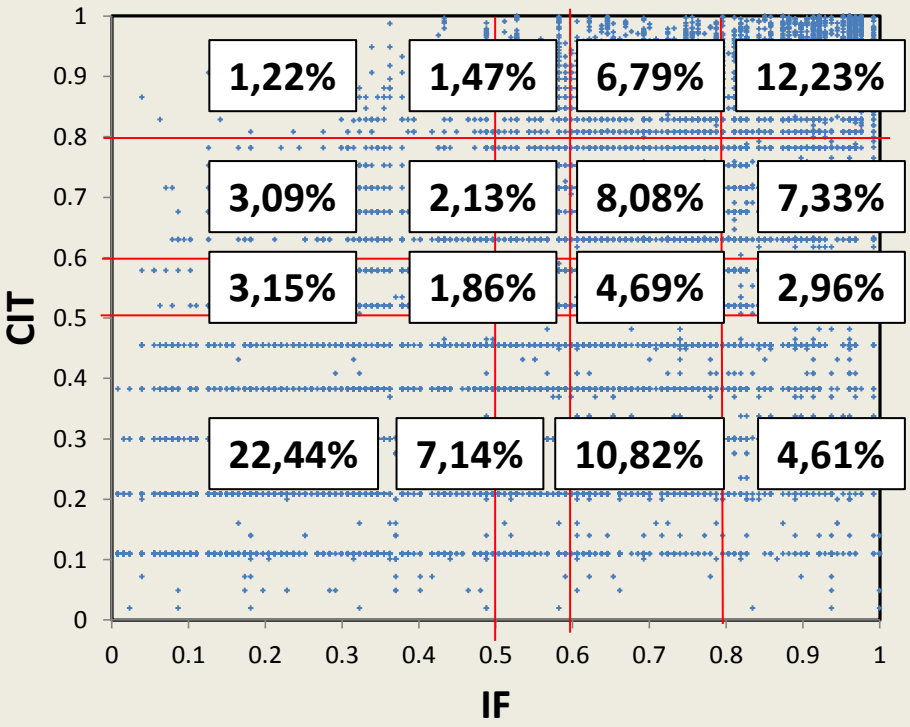
$$CIT = A \times JB + B_n$$

- **Slopes** A_1, A_2, A_3 could be different (but it would be simpler if they are the same, i.e. if thresholds are parallel lines)
- B_1, B_2, B_3 are the corresponding **constant terms**
- A_j and B_j are the degree of freedom that we have at our disposal to calibrate the classification of each SC/ASJC
- These degree of freedom have been experimentally shown to be sufficient to satisfy the desired distribution
- The gray areas represents those of large conflicts between JB and CIT which need to be evaluated **with IR** (which case by case **may involve or not the opinion of external experts**)



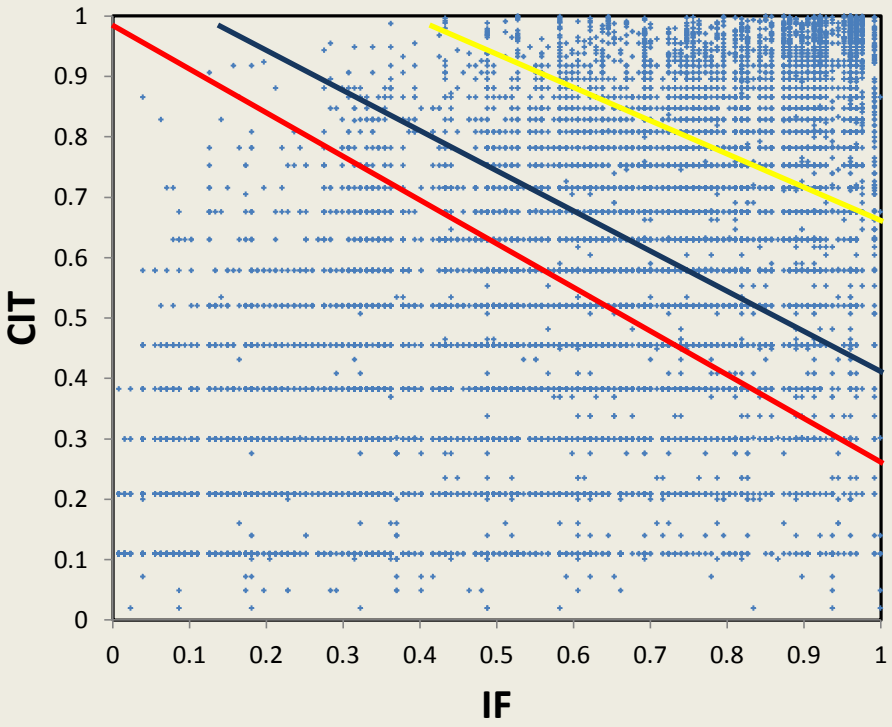
Improved Bibliometric Algorithm: An example

VQR 2004-2010 algorithm



Class of merit	%
E	29,31
G	14,90
A	1,86
L	28,68

Proposed algorithm



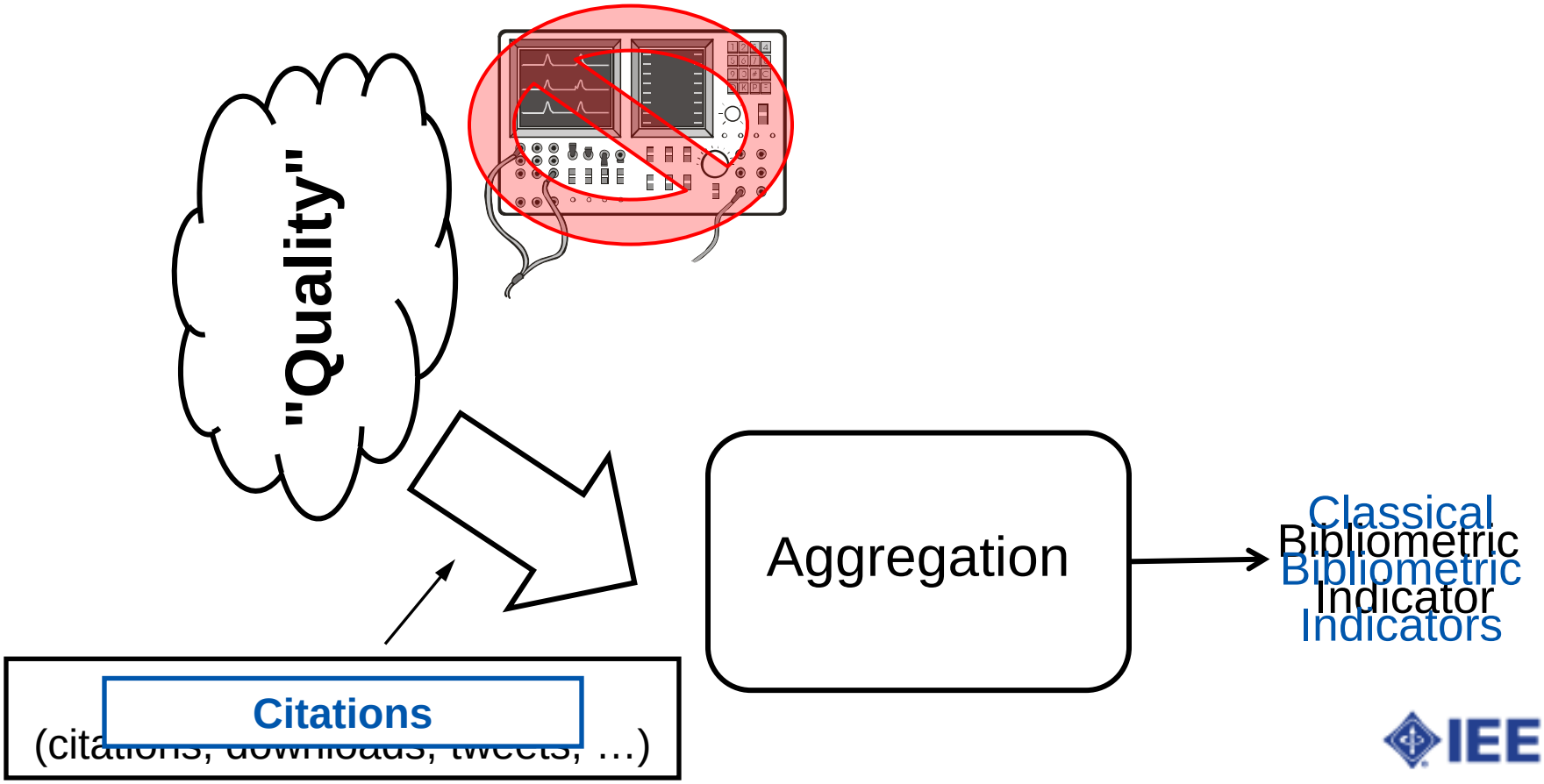
Class of merit	%
E	20,19
G	20,08
A	9,16
L	49,33

Which JB shall we use?

1. Overview on journal bibliometric indicators
2. Show that the "quality" of a **journal** as measured by journal bibliometric indicators is a multidimensional concept which **cannot be captured by any single indicator**
3. Show that the bibliometric indicators should not be misused by giving them **"more significance than they have":**
   a) the impact of an **individual paper cannot be measured** by the impact of the journal in which it has appeared
   b) there is **no strong correlation** between the Impact Factor of a journal and its **selectivity** (rejection rate)
   c) the Impact Factor of a journal **is not a good proxy** for the probability that **an individual paper will be highly cited**
4. Highlight that the misuse of journal bibliometric indicators has **undesired consequences** in the behavior of editors and individuals

Bibliometrics

- Definition: **Bibliometrics** is a set of methods to **quantitatively analyze** scientific and technological literature (it is part of Informetrics, which does the same for all information)



Journal Bibliometric Indicators, i.e. …numbers, numbers, numbers…

Many bibliometric indicators exist, each aiming to measure "journal quality"; they should:

1. ***Give a result which corresponds to the technical quality of the papers published in that journal***: *Nature*, *Science* or *Proceedings of the IEEE* and the "*Journal of Obscurity*" should have a very different value of the indicator
2. ***Be "fair" if applied to different areas***: different areas/communities may have different citation practices (e.g., long/short citation list)
3. ***Be immune to external manipulation***: it should be very difficult to artificially manipulate its value

Impact Factor and its criticisms - I

- Introduce by Eugene Garfield in 1972 to help librarians understand how much a **journal** was being used (useful in renewal process)
- It is an **average** measure of usage across **an entire journal**
- It contains **no information** on the impact of an **individual paper**

- For a journal J_i in a year n

$$\mathrm{IF}(J_i, n) = \frac{\#\{\text{ citations to all items published in } J_i \text{ in } n-1 \text{ and } n-2 \}}{\#\{\text{articles and letter published in } J_i \text{ in } n-1 \text{ and } n-2 \}}$$

- **Pros**: simple, easy to compute, known and disseminated

Impact Factor and its criticisms - II

- **Cons/criticisms**:
  1. Only 2 years of data to account for citations may not be enough in some areas to reach the citation peak ⇒ **IF varies very significantly among (sub)areas**

Ex: In SC Eng. E&E, $\boldsymbol{E}[IF_{2011}]$= 1.32; $\max[IF_{2011}]$=7

In SC Biology, $\boldsymbol{E}[IF_{2011}]$= 2.10; $\max[IF_{2011}]$=11.45

In SC Bioch and Molec. Bio $\boldsymbol{E}[IF_{2011}]$= 3.78; $\max[IF_{2011}]$=34.31

  2. Citations are counted in the same way **independently of the source** (i.e. a citation obtained from *Science* is the same as the "*Journal of Obscurity*")
  3. IF has an "non-consistent" definition: elements considered at the numerator are different than the denominator
  4. IF is **liable to active manipulation**

Impact Factor: manipulation (1/3)

- How has IF been manipulated?

1. **Inconsistent definition**: citations to notes/"letters to the editor"/editorials count in the numerator but the same items are not counted in the denominator. They can be cited and, even more importantly, their citations count normally.


Neth Heart J (2012) 20:481–482
DOI 10.1007/s12471-012-0336-0
Netherlands Heart Journal

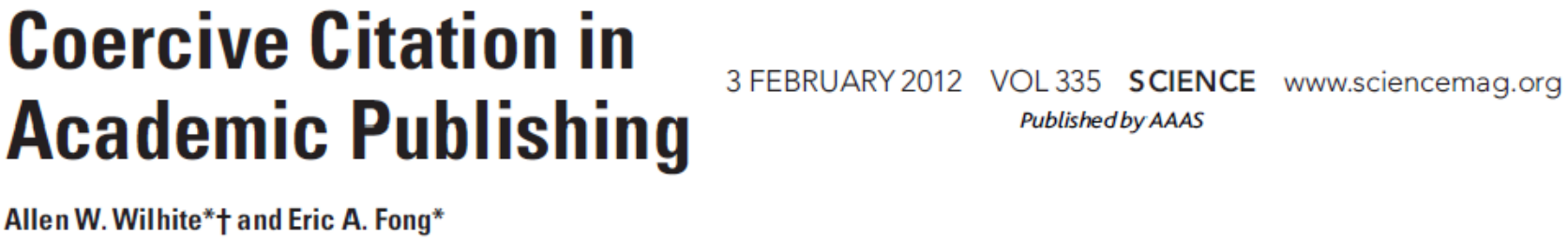
EDITOR'S COMMENT

The NHJ 2012 in retrospect: which articles are cited most?

Its bibliography contains 25 citations to the same journal, 24 of which count toward the 2012 IF

Impact Factor: manipulation (2/3)

2. **Coerce self-citations**: EiCs "force" authors to add citations to their journal (not necessarily to the authors) to increase IF



- EICs of 175/832 journals in the area of economics, sociology, psychology, and multiple business disciplines were found to "coerce" self-cites
- Coercing was more frequent with young authors than experienced ones
- Relation to area: if one journal coerces its authors other journals will most likely follow

Impact Factor: manipulation (3/3)

3. **Citation Cartel/Stacking**: EiCs or other members of editorial board of J_A and J_B:
   - publish in J_A a paper with (several) tens of citation to J_B
   - publish in another journal as authors to do the same

Brazilian citation scheme outed

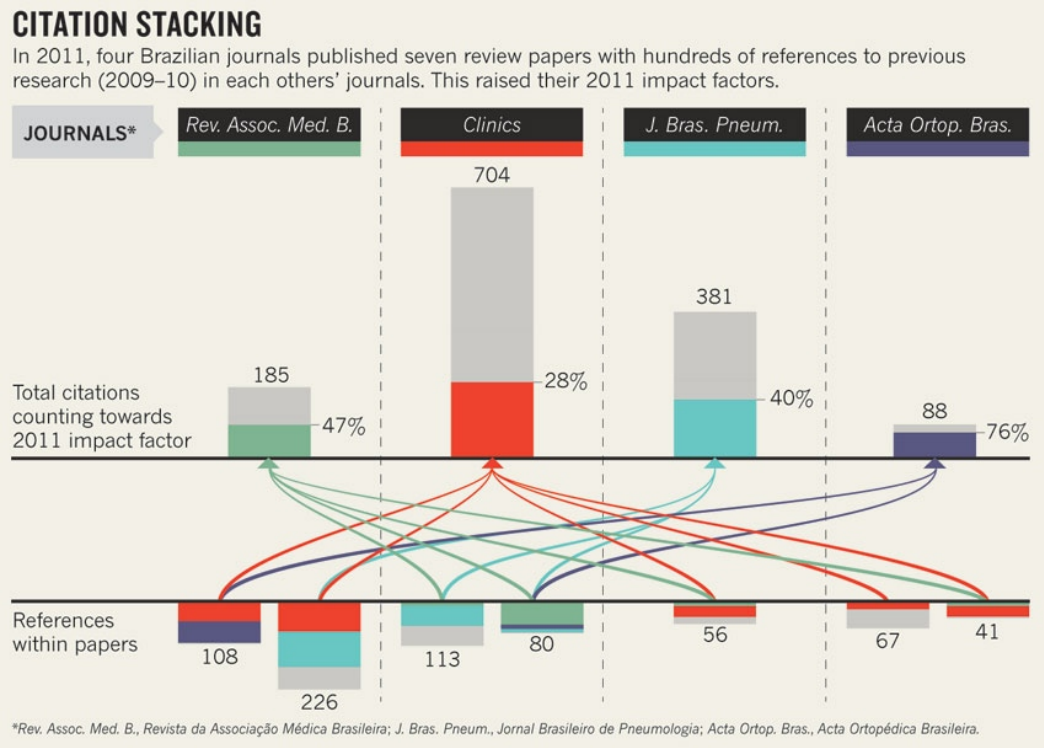
Thomson Reuters suspends journals from its rankings for 'citation stacking'.

Richard Van Noorden

nature

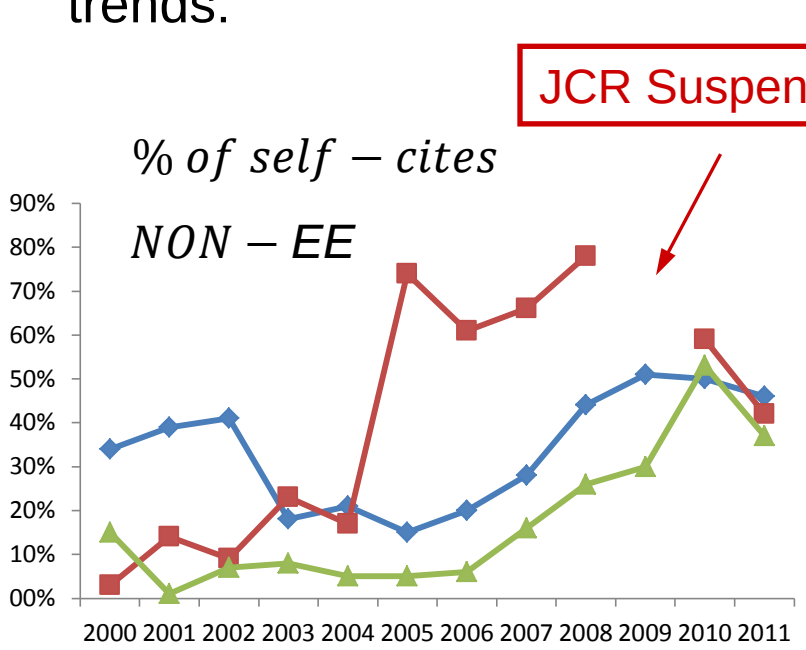
27 August 2013

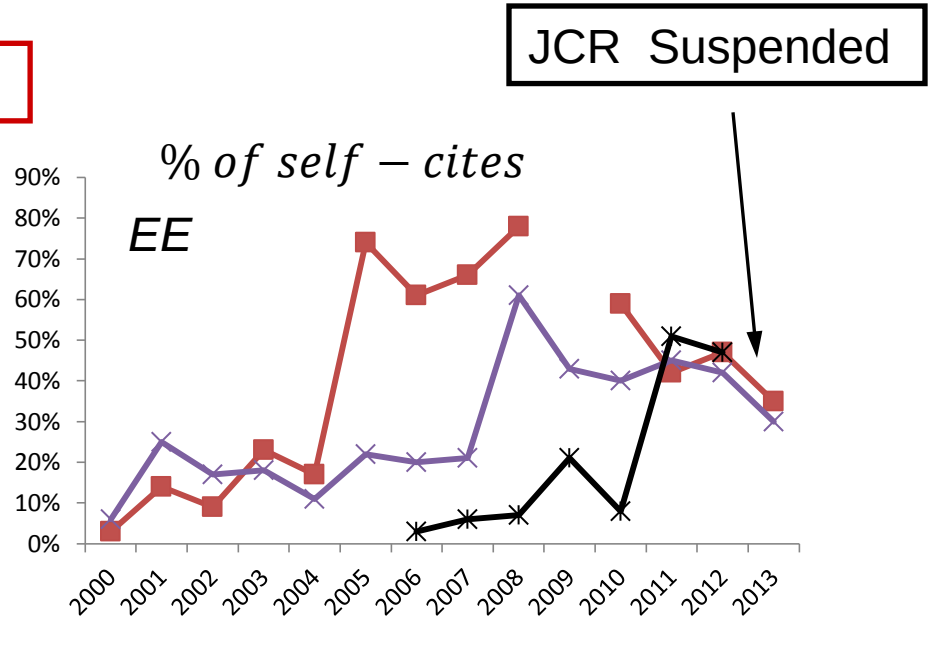
- Four Brazilian journals (Rev Assoc. Medic B, Clinics, J. Bras. Pneum, Acta Ortop Bras.) were found to establish a citation cartel
- Three Italian journals in the area of medicine (with the same EiC!)



Is the phenomenon widespread?

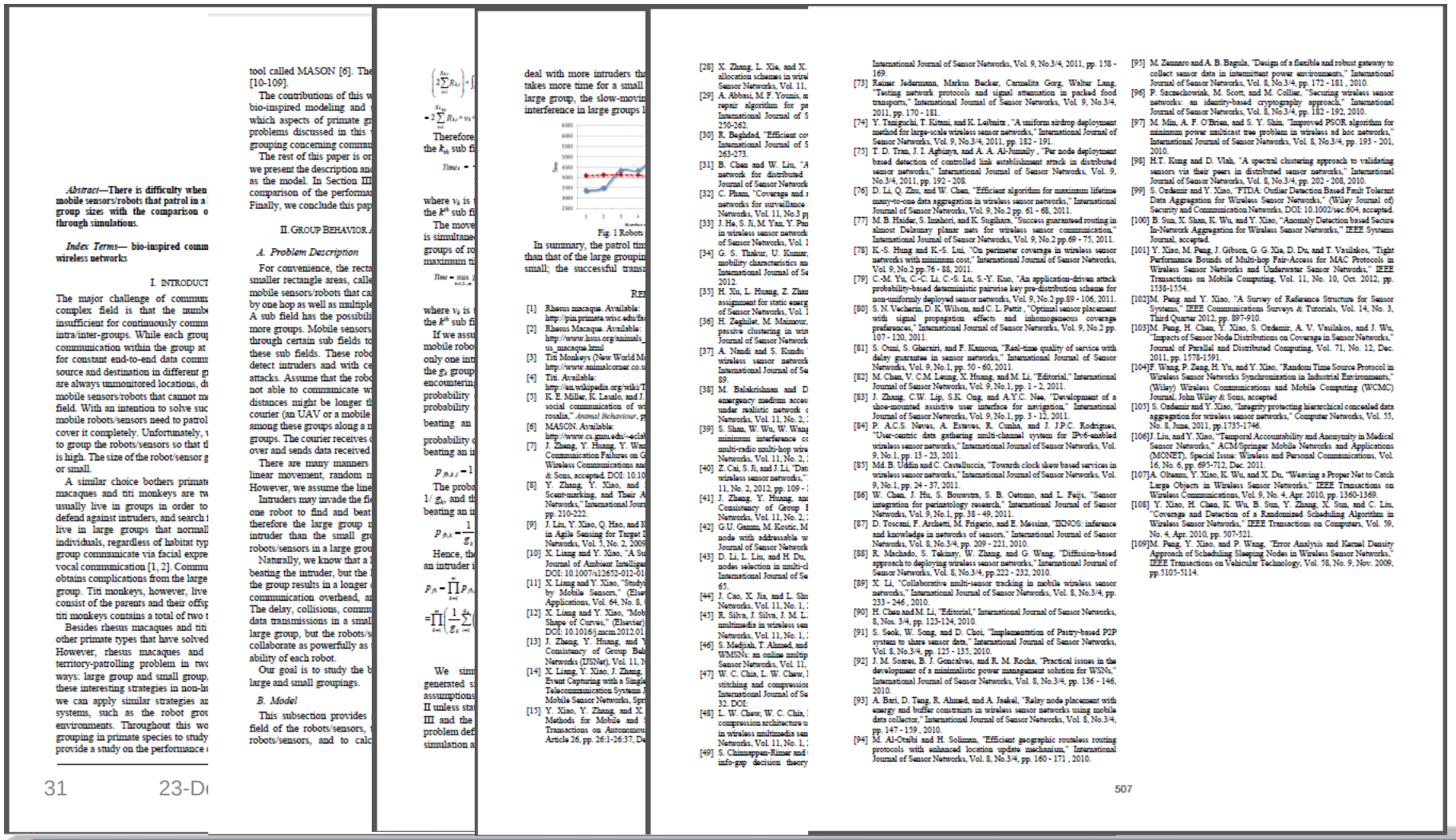
- No systematic study yet: one must use JCR data: For **citation cartels** the systematic analysis is very difficult, but one can rely on **self-citation** trends:





- Laser and Particles Beams (Phy Applied), Cortex (Neuroscience), Int. Journal of Hydrogen Energy (Energy and Fuels) show an increasing self-citation trend (and similar examples exist in many more areas)
- **E&E Engineering**: Int J. Circuit Theory and Applications and Asian Journal of Control shows that we are not immune.

What is wrong with this conference paper?



What is wrong with this conference paper? - II

- The authors published 2 conference papers with 100+109 items in the reference list.
- There are 74+82 citations to the International Journal of Sensor Networks (IJSN)
- One of the 2 authors is the EiC of the IJSN
- IJSN was not included by Thomson in the 2013 Journal Citation Report since the above citations account for 82% of the total citations to IJSN.
- The addition of the citation was done **after** the review process was completed

What is wrong with this conference paper? - II

- The authors published 2 conference papers with 100+109 items in the reference list.
- There are 74+82 citations to the International Journal of Sensor Networks (IJSN)
- One of the 2 a
- IJSN was not nal Citation Report since t he total citations to IJ
- The addition c process was completed

Why this is happening?

- The IF was historically created **to give librarians tools for deciding renewals**, yet…
- It is currently **more and more used as the gold standard to evaluate the impact of an individual's research activity** (for hiring, tenure, promotion, salary increase…).
  - As an example, the Chinese government pays scientists for publication in high IF journals (see http://scholarlykitchen.sspnet.org/2011/04/07/paying-for-impact-does-the-chinese-model-make-sense/)

IF range	(0,1)	[1,3)	[3,5)	[5,10)	>10	Nature/Science
Increase in salary	$306	$458	$611	$764	$2139	$30562

Why this is happening?

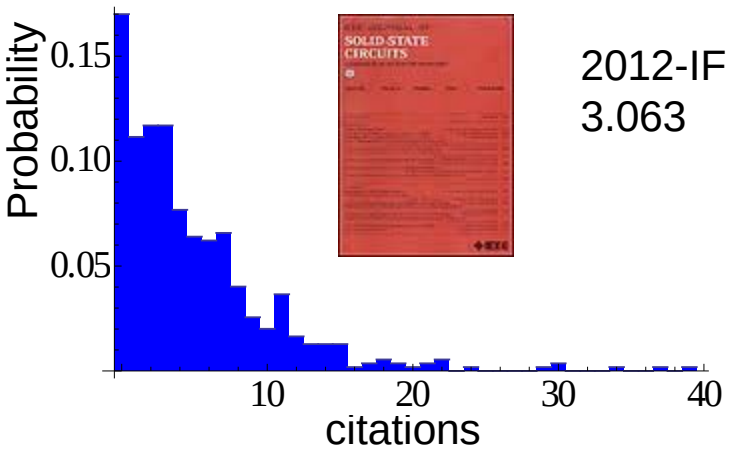
- The IF was historically created **to give librarians tools for deciding renewals**, yet…
- It is currently **more and more used as the gold standard to evaluate the impact of an individual's research activity** (for hiring, tenure, promotion, salary increase…)
- This use is commonly based on 2 main "assumptions". Assume that J_A has $IF_A \gg IF_B$ of J_B, then
  1. Any paper published in J_A has more impact (has received more citations) than any paper published in J_B
  2. The review process of J_A is more stringent than the one of J_B

Are these assumptions supported by data?

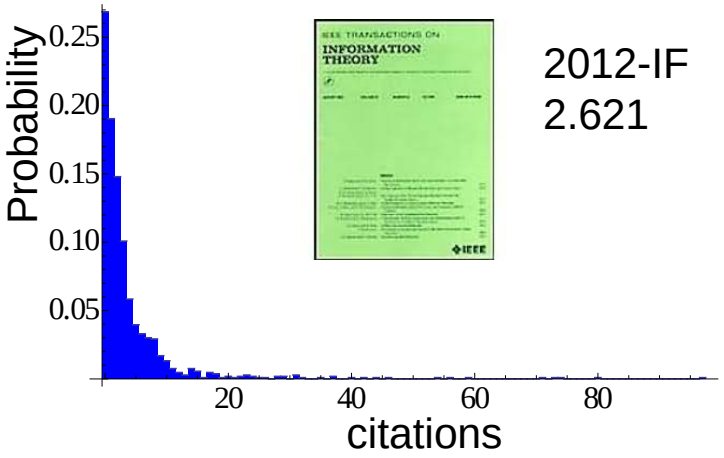
NO

Some data - I

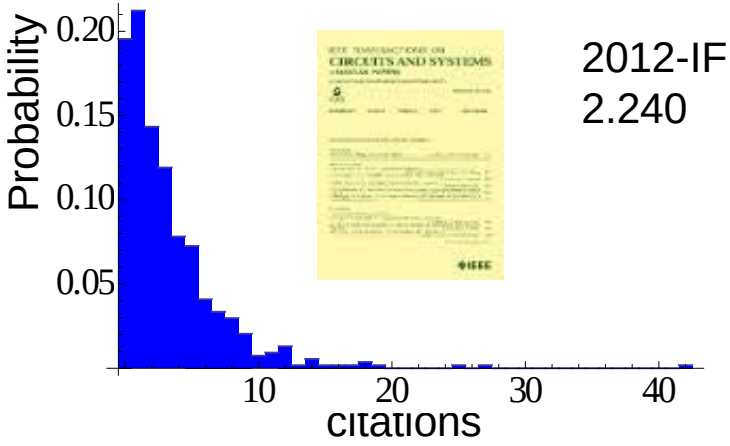
1. Evaluation of the impact of a **single** paper in a journal



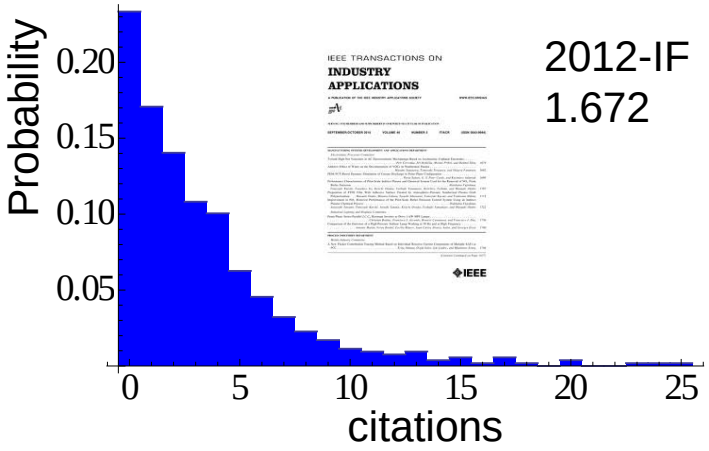
2012-IF
3.063



2012-IF
2.621



2012-IF
2.240



2012-IF
1.672

- JSSC, TIT, TCAS-I, and TIA distributions of citations for 2012 to papers of 2011 and 2010 show the same shape: **most papers are cited only a few times or never cited and only very few have high impact**

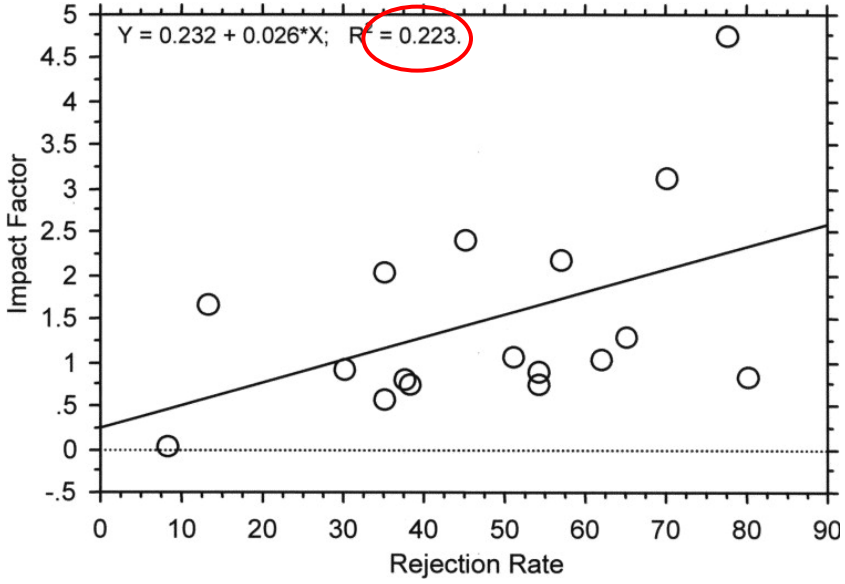
Some data - II

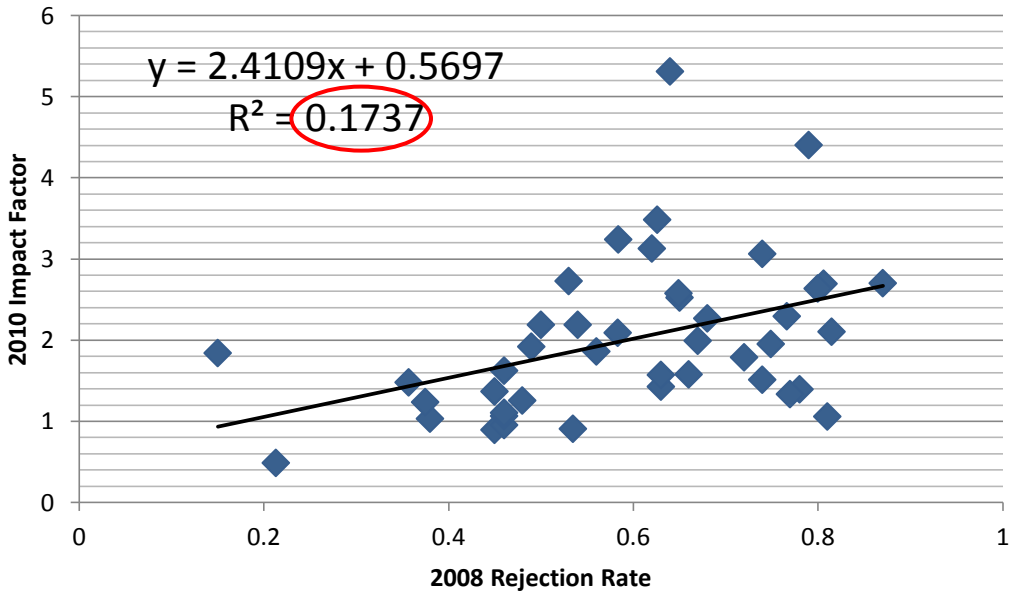
- **<u>Important</u>: regardless of IF, most papers in each journal are cited only a few times (if ever) and few papers are cited many times**
- Assuming that a randomly chosen paper in JSSC (IF=3.063) is better (has more citations) than one of TCAS-I (IF=2.240) **is wrong >36% of the time**
- Assuming that a randomly chosen paper in TIT (IF=2.612) is better than one of TIA (IF=1.672) **is wrong >43% of the time**

journal indicators are average quantities and give therefore **<u>no indication</u>** of the quality of any single published paper

Some data - III

- Indication of the **selectivity** of a journal: if the IF of a journal is large, is the review process "very strict"?
- This is **not supported by data** (at least if one assumes valid the equation "strict review process = high rejection rate"): the correlation coefficient **is on the order of 0.2**



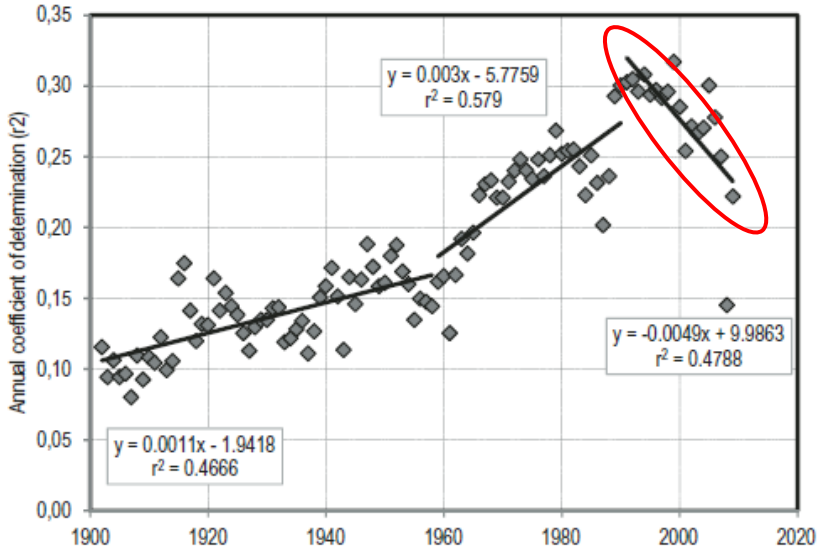


A. Kurmin, T. Krimis, "Exploring the Relationship Between Impact Factor and Manuscript Rejection Rates in Radiologic Journals, Acad Radiol 2006; 13:77–83

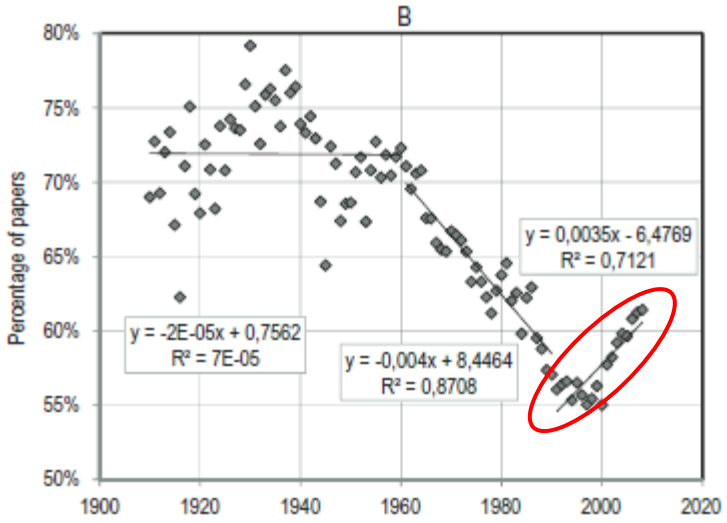
43 IEEE titles, Rejection Rate obtained by internal reports

Some data IV

- **Assumption**: the IF of a journal is large, papers published there are highly cited, if I publish there my paper has an higher probability to be highly cited
- This is **not supported by data** (neither in terms of correlation nor of probability) [G. A. Lozano *et al*., "The Weakening Relationship Between the Impact Factor and Papers' Citations in the Digital Age", J. American Society for Information Science and Technology, 63(11):2140–2145, 2012]



"Correlation coefficient" between IF in year of publication and citation rate in the following 2 years



Percentage of papers which are in the top 5% of the distribution citation in a given year which were NOT published in a journal in the top 5% of the IF ranking

Why this is happening?

- While the IF was historically created to help **librarians**, it is **misused** to evaluate **individual's research activity** (for hiring, tenure, promotion…)

The unintended use of the IF **made it the target and not the measure** and created incentives to its manipulation. From evaluation view point: "when a measure becomes a target, it ceases to be a good measure" - Goodhart's law

(from D. Arnold, K. Fowler, "Nefarious Numbers", Notices of the AMS, vol 58, n.3, pp 434-437)

According to the 2013 Nature article of Richard Van Noorden the EiCs of the 4 journals involved in a citation cartel created it because

"In Brazil, an agency in the education ministry, called CAPES, evaluates graduate programmes in part by the impact factors of the journals in which students publish research"

Several organizations toke positions against bibliometric misuse

Several other research agencies and professional organizations in the area of Physics, Medical Sciences, Biology toke positions againts bibliometrics misuse and abuse

SF Declaration on Research Assessment

San Francisco DORA
Declaration on Research Assessment

The San Francisco Declaration on Research Assessment (DORA), initiated by the **American Society for Cell Biology (ASCB)** together with a group of editors and publishers of scholarly journals, recognizes the need to improve the ways in which the outputs of scientific research are evaluated

Science 17 May 2013:
Vol. 340 no. 6134 p. 787
DOI: 10.1126/science.1240319

EDITORIAL

Impact Factor Distortions

Bruce Alberts

» Bruce Alberts is Editor-in-Chief of *Science*.

CREDIT: TOM KOCHEL

This Editorial coincides with the release of the *San Francisco declaration on research Assessment* (DORA), the outcome of a gathering of concerned scientists at the December 2012 meeting of the American Society for Cell Biology.* To correct distortions in the evaluation of scientific research, DORA aims to stop the use of the "journal impact factor" in judging an individual scientist's work. The *Declaration* states that the impact factor must not be used as "a surrogate measure of the quality of individual research articles, to assess an individual scientist's contributions, or in hiring, promotion, or funding decisions." DORA also provides a list of specific actions, targeted at improving the way scientific publications are assessed, to be taken by funding agencies, institutions, publishers, researchers, and the organizations that supply metrics. These recommendations have thus far been endorsed by more than 150 leading scientists and 75 scientific organizations, including the American Association for the Advancement of Science (the publisher of

IEEE statement on correct use of bibliometrics

IEEE position statement on correct use of bibliometrics (approved by BoD in 09/2013)

- IEEE joins several professional and scientific institutions (but none in the area of Engineering) to stress that **bibliometric indicators cannot be used (alone) to obtain an automatic evaluation of single researcher "scientific quality"**

- A web page was created to make the statement available to the IEEE community.

http://www.ieee.org/publications_standards/publications/rights/bibliometrics_statement.html

IEEE statement on correct use of bibliometrics

IEEE [illegible] (app[illegible]

- [illegible] in

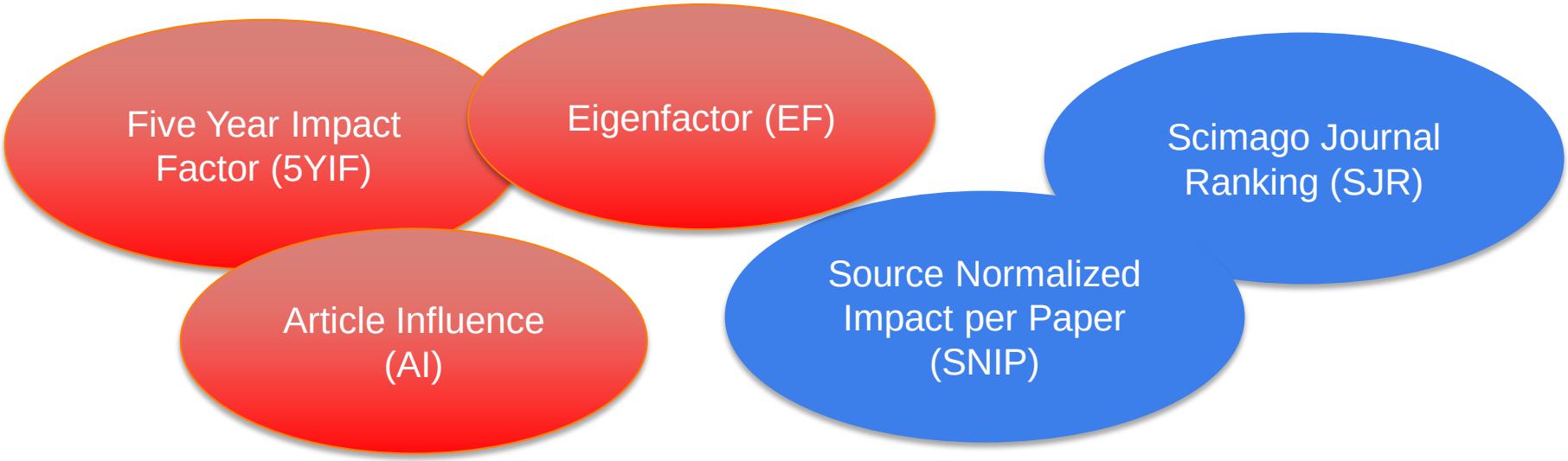
1. The use of **multiple complementary bibliometric indicators** is fundamentally important to offer an appropriate, comprehensive and balanced view of each journal in the space of scholarly publications.
2. **Any journal-based metric** is not designed to capture qualities of individual papers and **must therefore not be used as a proxy for single-article quality or to evaluate individual scientists**.
3. While bibliometrics may be employed as **a source of additional information** for quality assessment within a specific area of research, **the primary manner** for assessment of either the scientific quality of a research project or of an individual scientist **should be peer review**.
4. The IEEE **explicitly and firmly condemns** any practice aimed at influencing the number of citations to a specific journal with the sole purpose of artificially influencing the corresponding indices.

[illegible]EE

http://www.ieee.org/publications_standards/publications/rights/bibliometrics_statement.html

Other measures to solve IF issues for Journal evaluation

Several "successful" new indicators: 5 in either WoS or Scopus



- ***Increase the citation window*** : 3 or 5 years
- ***Introduce subject field normalization***: explicit (SNIP) or implicit (EF, AI, SJR)
- ***Exclude all (or most) self-cites***: eliminate the inflation issue (EF, AI, SJR)
- ***Only count "equivalent scientific" documents both at numerator and denominator***: eliminate another cause of inflation (EF, AI, SJR, SNIP)

Popularity vs Prestige

- An important distinction is between indicators measuring *popularity* or *prestige*

1. *Popularity indicators*: are based on an algebraic formula and count citations directly **independently of their source** (IF, 5YIF, SNIP)
2. *Prestige indicators*: are based on an recursive formula and weight the influence of citations **depending on their source** (EF, AI, SJR)

They evaluate **different aspects of Journal Impact**

At the very minimum, one needs to use **both** popularity (ex. IF, 5YIF) and prestige (ex. AI, SJR) indicators

How do we choose the Journal Bibliometrics?

Database: Thomson Reuters Web of Science (WoS) and Elsevier Scopus

- Author/Institution chooses the database for the evaluation

Classify with respect to more than one indicator:

- 5YIF and AI for WoS (5-year citation window; popularity & prestige) 5YIF has the advantage wrt IF to be somehow more stable in time and is less sensitive to "random" events;
- IPP and SJR for Scopus (3-year citation window; popularity & prestige, same definition as 5YIF and AI)

This procedure has the advantage to recognize that:

1. There is **no preferred database** for evaluation (i.e. the universe)
2. There is **no preferred indicator** for the evaluation
3. Less "emphasis" on bibliometric evaluation to avoid the "distorted perception" that **"IF is the article quality"**

How do we choose the Journal Bibliometrics?

Database: Thomson Reuters Web of Science (WoS) and Elsevier Scopus

- Author/Institution chooses the database for the evaluation

Classify with respect to more than one indicator:

- 5YIF and AI for WoS (5-year citation window; popularity & prestige) 5YIF has the advantage wrt IF to be somehow more stable in time and is less sensitive to "random" events;
- IPP and SJR for Scopus (3-year citation window; popularity & prestige, same definition as 5YIF and AI)
- Allow to compare the behavior wrt different indicators and call for GEV member opinion in case of large differences

- Author/Institution chooses the database for the evaluation

How do we choose the Journal Bibliometrics?

Database: Thomson Reuters Web of Science (WoS) and Elsevier Scopus

- Author/Institution chooses the database for the evaluation

Classify with respect to more than one indicator:

- 5YIF and AI for WoS (5-year citation window; popularity & prestige) 5YIF has the advantage wrt IF to be somehow more stable in time and is less sensitive to "random" events;
- IPP and SJR for Scopus (3-year citation window; popularity & prestige, same

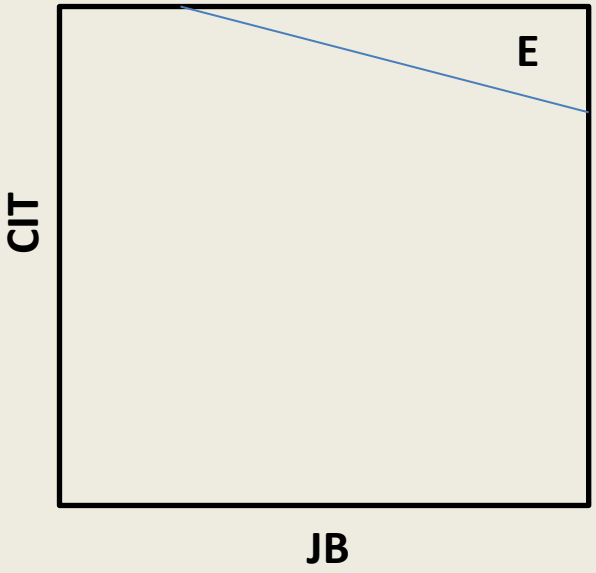
This procedure has the advantage to recognize that:

1. There is **no preferred database** for evaluation (i.e. the universe)
2. There is **no preferred indicator** for the evaluation
3. Less "emphasis" on bibliometric evaluation to avoid the "distorted perception" that **"IF is the article quality"**

Improved Bibliometric Algorithm: Slopes

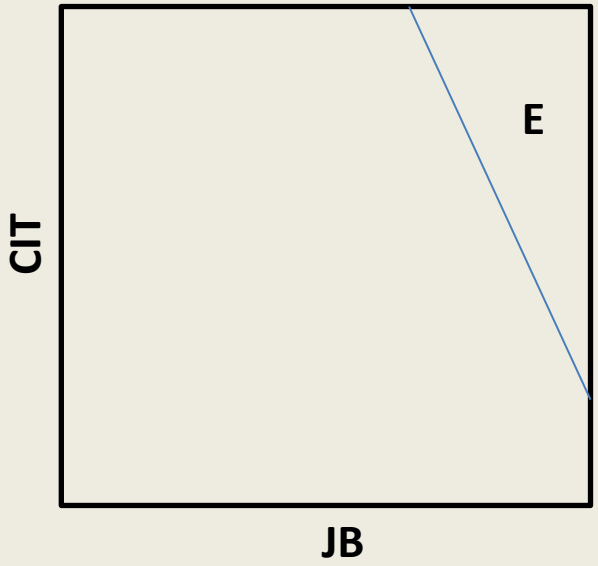
- The choice of A allow to give more weight to one of the 2 variables (JB,CIT), that is to give more importance to a classification of the product based on citation analysis or on the impact of the journal in which the paper was published

More weight to CIT



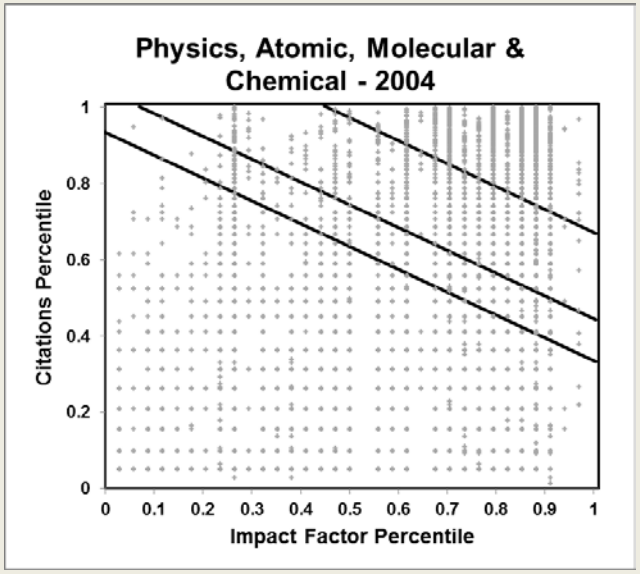
Older Papers

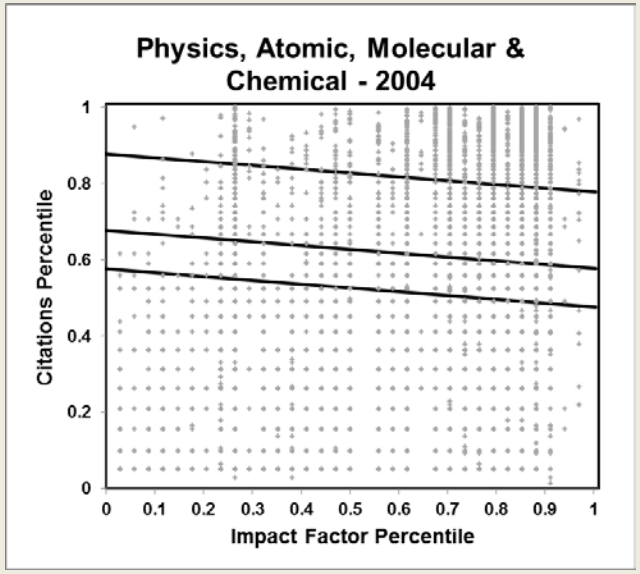
More weight to JB



Recent Papers

Improved Bibliometric Algorithm: Examples





- SC Physics, Atomic, Molecular and Chemical in 2004.
- The target distribution (in this case the 20-20-10-50 of VQR2004-2010) can be obtained using different values of A, that is, **the solution is not unique**

Improved Bibliometric Algorithm: Choices

GEV main task: on the basis of the knowledge of the editorial/citation practice of the community must choose:

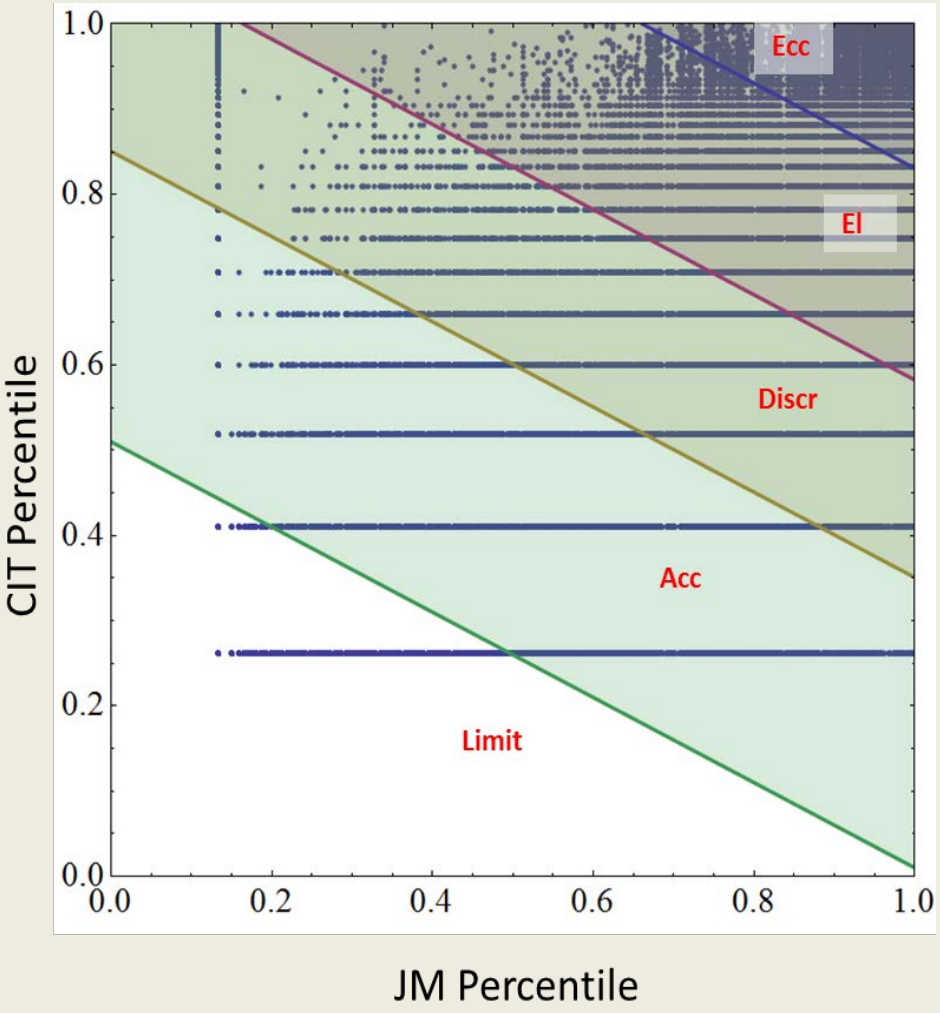
- ***Coarse grained calibration:*** GEV09 chose the slope for each year, that is the "degree of relative importance" of CIT vs JB
- Position and amplitude of the IR gray regions.
- In both cases there is an agreement among all bibliometric GEVs

ANVUR main task: on the basis of GEV citeria:

- ***Fine tuning:*** slopes A and constant terms B are chosen for all SC/ASJC to satisfy the desired global distribution of products;
- This guarantees that the **ex-ante probability** that an article has to fall in one of the 5 classes is the same (with a 1% tolerance) independently of the SC/ASJC of the journal in which it is published, of the year of publication, and of the GEV criteria.

This has the advantage to **make comparable the evaluation of the different areas** with respect to the corresponding international scientific community for all papers published in journals present in WoS or Scopus.

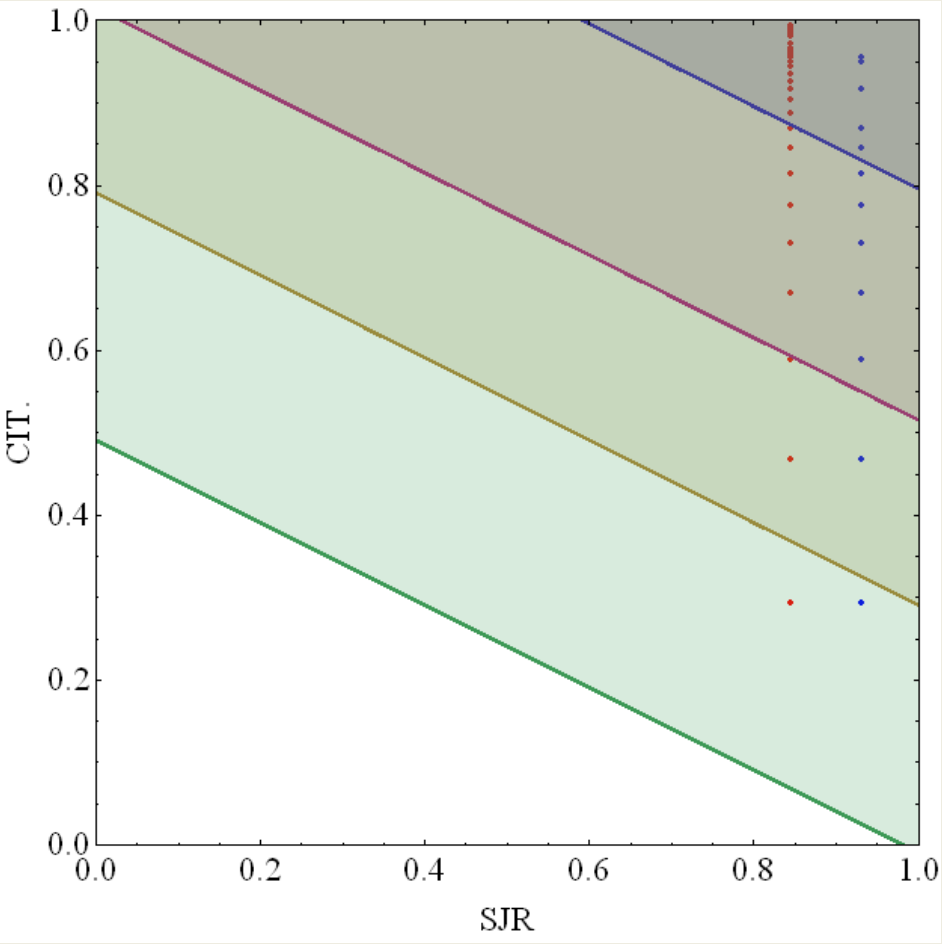
Example



IMPORTANT

- VQR classifies **products** and not journals
- It is possible that journals which are not in the top 10% of JB will have products in the top class
- Even the top journals may have products (and will have products in general) which are not in top class
- The number of citations to reach each class depends on JB, **but also on the behavior of all the other journals**

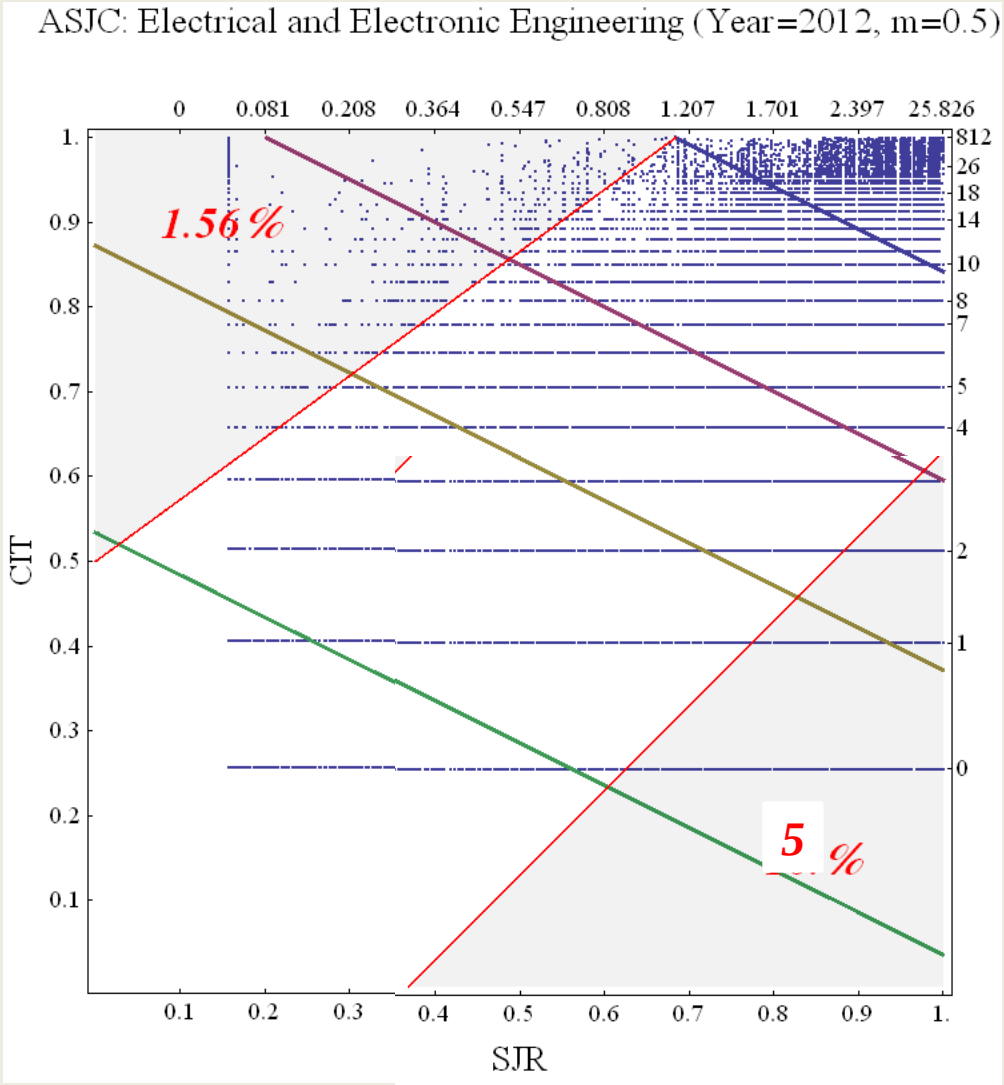
Example for two Journals



IMPORTANT

- Journal "red" is worse than journal "blue" but can have more products in zone 1
- Optimal choice of products is **not** easy and can be done only if one knows all the products in a given SC/ASJC

The IR Zones



IMPORTANT

- Highlight regions where the two information wrt JB and CIT have low correlation
- Isosceles triangles with 5%, 5% and 7% of the products in 2011, 2012, 2013 in the lower left corner
- Rectangular triangle with hipotenusa between (0,0.5) and the upper left corner of calss 1

Bibliometric Algorithm: Remark

- The final classification of the products evaluated using the bibliometric algorithm **is not automatically determined** by the algorithm **alone.**

- On the contrary **the final classification depends on the expert opinion of the GEV (members)** which will use all information at his/her disposal in addition to the bibliometric evaluation, such as his/her own knowledge of the area and the information contained in the document accompanying the product (scheda prodotto)

Peer review

What will be evaluated using peer review are:

- ***Journal Articles*** not indexed in WoS/Scopus
- ***Journal Articles*** for which peer review is requested by the authors (if adequately motivated) of which is requested by GEV members
- ***Journal Articles*** which are in the 10% set randomly determined for evaluation the correlation between the bibliometric evaluation methodology and peer review (this will NOT change the bibliometric evaluation)
- ***Journal Articles*** which ends in the IR regions
- ***Books*** with ISBN
- ***Chapters in Books/Articles in Conference Proceedings*** with ISBN
- ***Software anb Data-Bases*** if accompanied by a suitable description
- ***Patents:*** if awarded (not filed) between 1/1/2011 and 31/12/2014. Position and amplitude of the IR gray regions.

- Not all class of merit will be available for every product, as specified in the criteria

What do we do with self-citations?

1. Problem of disambiguation: we are not sure it may work for WoS;
2. It is not correct to eliminate them in the individual paper if we cannot do so in all papers to compute the thresholds;
3. May indicate activity of the authors (positive) or due to the fact that the community is not very large;
4. May be large in recent papers (due to the fact that "the authors know their work sooner than the readers")

Solution: indicate a warning if %self-citation > 50%. GEV members will evaluate the product and take the final classification decision. **Decision will consider all other information on the product** (citations to the corresponding conference paper, citations from patents, use in industrial products, awards, …)

Conclusions

Main differences of the bibliometrics algorithm classification in the VQR2011-2014 with respect to VQR2004-2010

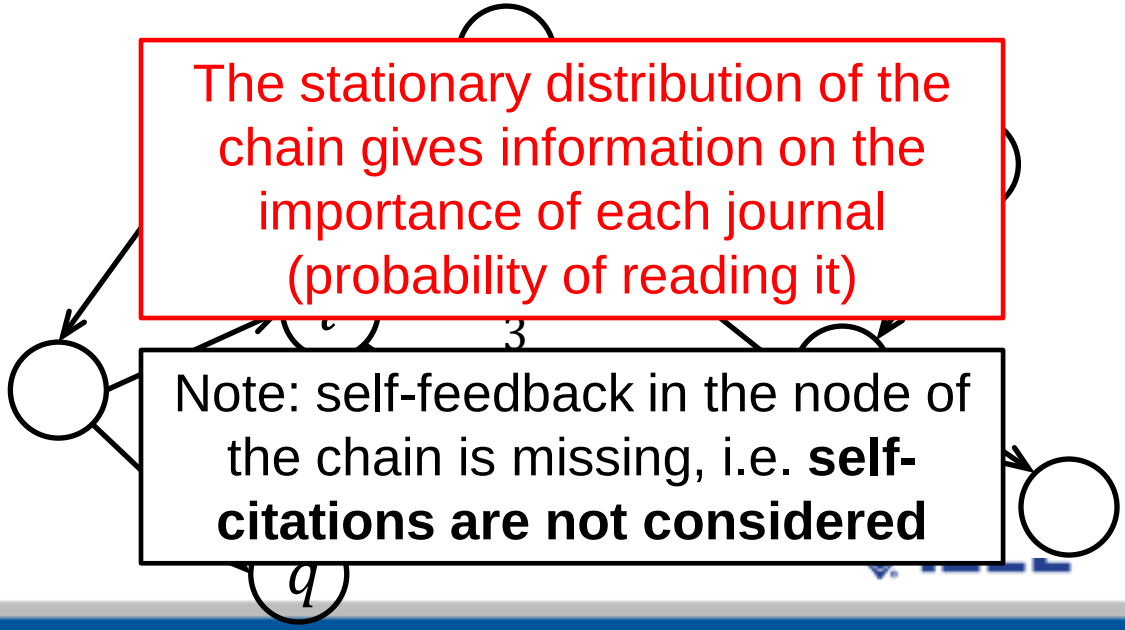
1. Use of different bibliometric indicators for taking into account the impact of the journal in the evaluation (to avoid to suggest that there is a "golden indicator")
2. (For most GEVs) Article level metrics (citations) influence (much) more the final evaluation with respect to the metric of the journal. The latter is used when the former is not (truly) reliable
3. The algorithm is calibrated to guarantee that the percentage of paper in each of the five classes (Excellent, High Quality, …) is respected for every SC/ASJC in each area. This will make results comparable among different areas (with respect to the relative international community)
4. Self-citations are a know data point in the final evaluation which may influence it (to reduce incentive towards self-citation inflation, but without "demonizing" self-citations practices)

Backup

Recursive (Pagerank) Prestige Measures - I

- Developed by Carl Bergstrom in 2007. The **EigenFactor** is computed by **Thomson** using the "same" algorithm used by Google to rank web pages
- Consider a collection of N journals. Each of them is represented by a node in a network
- Journal J_i gives in total 10 citations, 3 to J_j, 5 to J_p, and 2 to J_q. Of course 3/10 **represents the conditioned probability that a reader reads and article in J_j, assuming that he/she started reading J_i**

Markov Chain with transition matrix computed as the fraction of total citations given by 2 different journals



Recursive (Pagerank) Prestige Measures - II

- In formulas: how EF_i (for journal i) is "roughly" defined?:

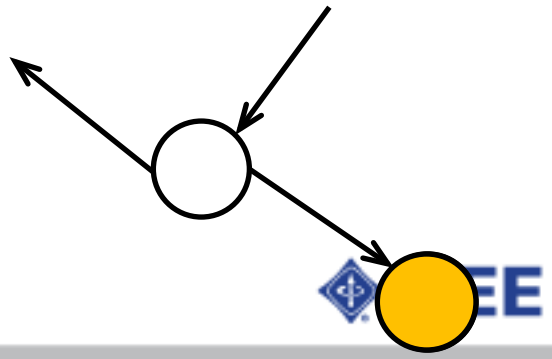
$$\pi_i[n] = \sum_{j=1, j\neq i}^{N} \frac{c_{ji}^{\Delta_2\Delta_1}}{\sum_{l=1, l\neq j}^{N} c_{jl}^{\Delta_2\Delta_1}} \pi_j[n-1]$$

$$\sum_{i=1}^{N} \pi_i[n] = 1$$

"self-citations" are not included (no self-feedback in the Markov Chain)

- It may happen that the reader at some point stop reading, than starts reading again picking a journal at random. He reads again exactly journal J_i with a probability given by the fraction of papers published in journal $J_i$with respect to the entire collection. (1-α) is the probability that the reader stop reading.

- Problems of the "dangling nodes", which represents journals that are only cited, but do not give any citations. Not considered in this version of the formula



Recursive (Pagerank) Prestige Measures - III

- One needs to compute

$$\pi_i = \lim_{n \mapsto \infty} \pi_i[n] \qquad \mathrm{EF}_i = 100 \sum_{j=1}^{N} \frac{\pi_j}{C_{j\rightarrow}^{\Delta_2 \Delta_1}} c_{ji}^{\Delta_2 \Delta_1}$$

- Remarks:
  1. The more "important" is the journal j (π_j is large) the more a citation from it to journal i increases EF_i
  2. Normalization by all citations given by journal *j* (citation potential)
  3. The EF_i represents the probability that a random reader picking journals at random and following citation will eventually read journal i
  4. It is a measure *per-journal* and not *per-paper* and therefore tends to be larger for journals publishing more papers (not necessarely a problem)

Recursive (Pagerank) Prestige Measures - IV

- The **Article Influence** is roughly the EF "normalized to" the number of papers published by each journal:
  - *Per-paper* measure (similar "physical meaning" w.r.t. the IF)
  - Further normalization to have AI=1 for the median journal

$$\mathrm{AI}_i = \beta \frac{\mathrm{EF}_i}{p_i^{\Delta_1}}$$

- **Pros (EF/AI):**
  1. Citations **are now weighted** depending on the source (a citation from *Science* is valued more than one from the "*Journal of Obscurity*")
  2. Time window for computing citations (Δ_1) **is 5 years**. This index is expected to exhibit fewer fluctuations over time
  3. Journal self-citations **are not** considered. The index is less prone to "external influence"
- **Cons (EF/AI):**
  1. More difficult to understand and compute
  2. **Not necessarily correct** to eliminate **all** self-cites.

SJR and SNIP - I

SNIP

- Introduced by Moed in 2010. Contained in Scopus

$$\mathrm{SNIP}_i = \frac{\mathrm{RIP}_i}{\mathrm{RDCP}_i} \qquad \mathrm{RIP}_i = \frac{C_{\rightarrow i}^{\Delta_2 \Delta_1}}{p_i^{\Delta_1}}$$

$$\mathrm{RDCP}_i = \theta \sum_{j \in \mathcal{I}} \frac{C_{j \rightarrow}^{\Delta_2 \Delta_1}}{p_{j \rightarrow i}^{\Delta_2 \Delta_1}}$$

Relative Database Citation Potential = average number of citations contained in any paper citing J_i in period Δ_1 normalized in such a way that the median journal in the database has $\mathrm{RDCP}_i = 1$

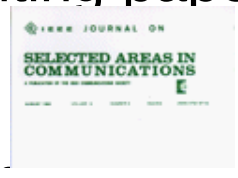
Raw Impact per paper = average number of citations per paper published in J_i in period $\Delta_1 = \{Y_{n-1}, Y_{n-2}, Y_{n-3}\}$ by papers published in all journals present in the data base in period $\Delta_2 = \{Y_n\}$

Same definition as the IF. The only difference is the rolling window of 3 years (instead of 2) to collect citations

SJR and SNIP - II

$$\mathrm{SNIP}_i = \frac{\mathrm{RIP}_i}{\mathrm{RDCP}_i} \qquad \mathrm{RDCP}_i = \theta \sum_{j \in \mathcal{I}} \frac{C_{j\rightarrow}^{\Delta_2 \Delta_1}}{p_{j\rightarrow i}^{\Delta_2 \Delta_1}}$$

- Assume that articles in journal J_i are cited by an article in journals J_k, J_l, and J_m



- Consi… citing paper i… itir… l
- Without consider… …Database… Potential for J_i is the av… …mber … …ned in every paper citing it, **which are in… …which … …ontained in the database**

SJR and SNIP - III

- **Pros (SNIP):**
  1. Time window for computing citations (Δ_1) **is 3 years**. This index is expected to exhibit fewer fluctuations over time
  2. An **explicit normalization to the citation potential for each journal** is considered which should make indicators for journals of different areas more comparable
  3. It is freely available from the Scopus homepage, without subscription to the database

Title	IF	SNIP	SC
IEEE Proceedings	6.81	5.97	Eng. E&E
PLOS Biology	11.45	1.94	Biology
Annual Review Biochemistry	34.32	8.27	Bioch&Mol. Bio.

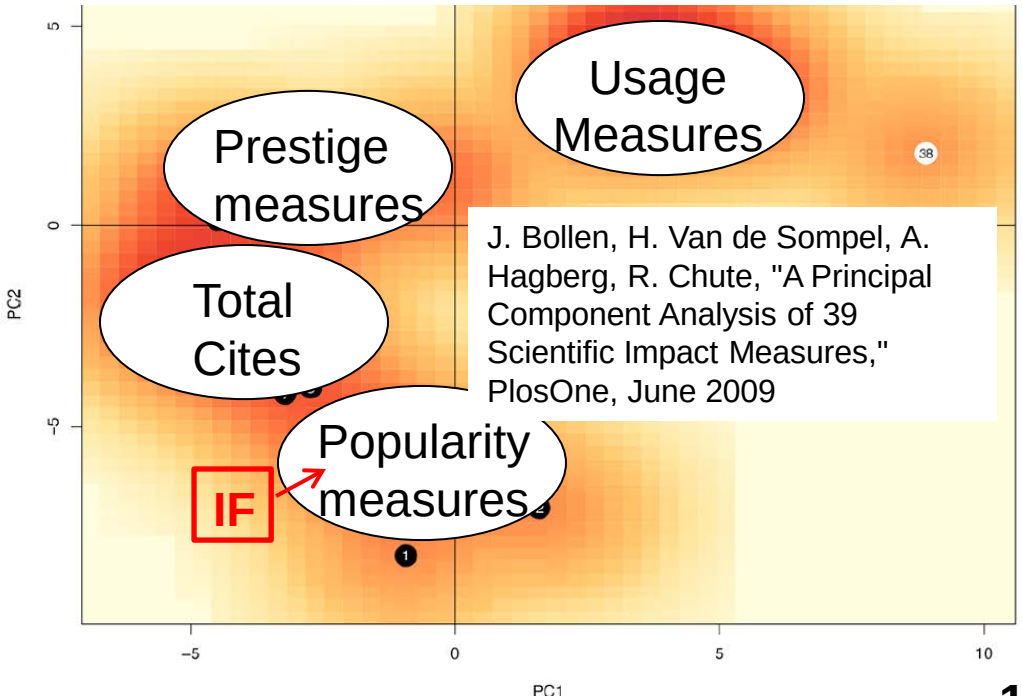
- **Cons (SNIP):**
  1. Citations **are not weighted** depending on the source
  2. More difficult to understand and compute (even if the definition is non-recursive)
  3. Self-cites are still considered

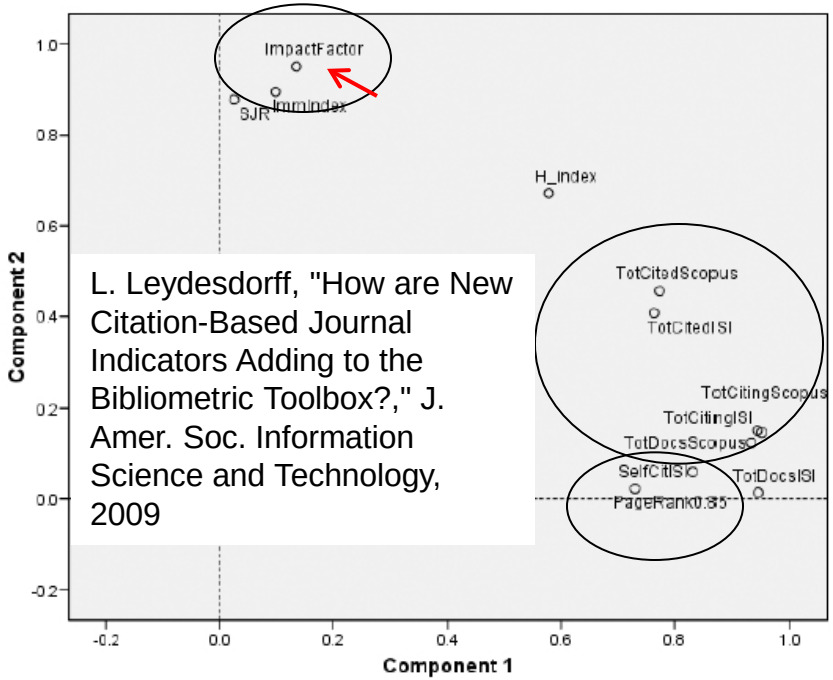
SJR and SNIP - IV

SJR

- Introduced by González-Pereira, Guerrero-Boteb, Moya-Anegónc in 2010. Contained in Scopus
- It has a definition similar to the Article Influence (i.e. it is a measure per-paper), bu consider self-citations up to 30%
- **Pros (SJR):**
  1. Citations **are now weighted** depending on the source (a citation from *Science* is valued more than one from the "*Journal of Obscurity*")
  2. Time window for computing citations (Δ_2) **is 3 years**. This index is expected to exhibit fewer fluctuations over time
  3. Journal self-citations **are** considered only partially. The index is less prone to "external influence"
  4. Freely available from Scopus homepage
- **Cons (SJR):**
  1. More difficult to understand and compute

Two PCA Analysis of bibliometric indicators





39x39 covariance matrix between indexes computed using Scimago, 2007 JCR and MESUR project for usage

13x13 covariance matrix between indexes computed using Scimago and 2007 JCR (no usage)

Compute the "principal components":

1. The problem is roughly 2-dimensional (83.4% cumulative variance)
2. Different clusters are present: prestige, popularity **measure different aspects of quality**
3. One **cannot use only one indicator** to "measure journal quality"

Making decisions based on multiple indicators (1/2)

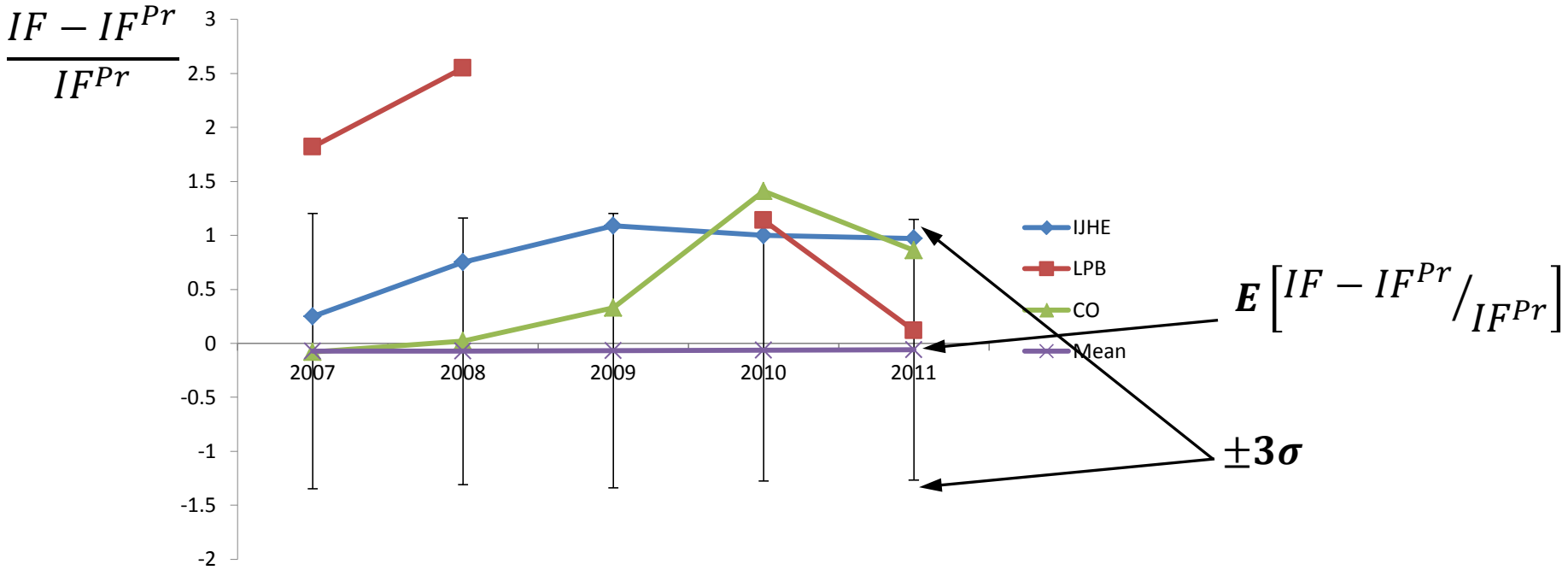
- EF, AI and IF measure journal quality, but IF uses self-cites while EF and AI **do not** use them
- If ranking wrt IF is **much greater** than wrt to EF and AI there may be a problem with self-cites
- Go back to the LPB vs Cortex issue. With respect to "SC per paper" *Cortex* in 2010 is worse than *Laser and Particles Beams* in 2008. Why was LPB removed from JCR and Cortex was not?

	Laser and Particles Beams (Physics, Applied)					Cortex (Behavioral Science)				
	2007	2008	2009	2010	2011	2007	2008	2009	2010	2011
Rk-IF	6	8	SUS	17	49	8	19	7	4	4
Rk-EF	32	54	SUS	45	57	9	15	12	15	10
Rk-AI	50	67	SUS	70	69	15	21	17	16	14

Difference in RK for LPB in 2008 is much larger than for CO in 2010

Making decisions based on multiple indicators (2/2)

- Similar information can be extracted using a **linear predictor*** $IF^{Pr} = F(EF, AI)$.
- A large relative difference may indicate problems



Using more than one indicator helps to obtain a better evaluation of the journal status

*Linear predictor on all Journals in 2007 to 2010 JCR which have all three IF,