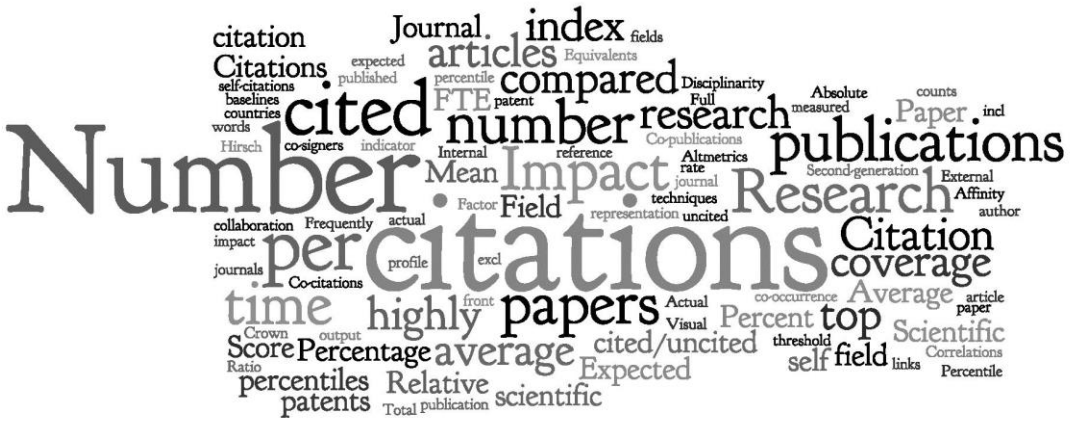
ΑΛΕΞΑΝΔΡΕΙΟ Τ.Ε.Ι. ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Bibliometric Methods for Researchers Evaluation

Βιβλιομετρικές Μέθοδοι Αξιολόγησης Ερευνητών



Του φοιτητή

Θρασύβουλου Καλούδη

Αρ. Μητρώου: 06/3105

Επιβλέπων καθηγητής

Αντώνης Σιδηρόπουλος

Θεσσαλονίκη 2016

“A politics of formalism rests on several things. First, a formal representation is an abstraction: It takes away properties from a particular situation. Second, it is a simplification: It reduces the complexity of real life situations in order to make them formally (usually, but not exclusively, mathematically) tractable. Third, and most important, every formal representation contains choices about what to keep in (what is important) and what to throw out. All such choices are political”.

Susan Leigh Star (Star S.L., 1995)

Prologue

The paper above deals with Bibliometric Methods for Researchers Evaluation. Its aim is to describe and analyze methods that evaluate scientific work, scientific journals, papers, websites and scientific libraries. It addresses to researchers who occupy with metrics about scientific work and scientists who want to develop their work and need to know the reliability and the impact of their sources.

Abstract

Bibliometric methods are too important to the scientific world due to their ability to evaluate not only individually scientist who have published their work but also journals, websites and digital libraries that publish scientific papers. The results of these metrics can be used by academic facilities for faculty recruitment, fair promotion practices and by individual scientists for comparison of their personal scientific merit with competitors. On this paper, firstly, described and analyzed metrics that evaluate journals and websites, such as *Impact Factor*. Secondly, described individual indices, such as *h-index* and thirdly described and analyzed the most important digital Libraries and Databases, such as *Google Scholar*. During the description of these metrics mentioned their advantages and disadvantages, furthermore, the reasons who constrained the experts to invent them. Finally, outlined the general benefits of bibliometric analysis.

Contents

Table of Figures

1. Introduction

The evaluation of the scientific work of a scientist has long attracted significant interest, due to the obvious benefits of obtaining unbiased and fair criteria that could give a brief evaluation of each paper and through it a ranking of the involved scientists' merits. Such a metric evaluation can be used by academic facilities for faculty recruitment, fair promotion practices and prize awarding, prompt funding allocation and by individual scientists for comparison of their personal scientific merit with competitors, etc. Similarly, the estimation of a publication's journal or conference quality is, in the era of the web, extremely important since it guides the scientists' decisions about where to publish their work, the researchers' preference in seeking for important articles, etc. Although, the issue of ranking a scientist or a journal / conference dates back to the 70's with the seminal work of Eugene Garfield (Garfield, 1972) and others (Holsapple, et al., 1994). Since the 00's this field has expanded due to the proliferation of digital libraries (Schwartz & Russo, 2004), whose huge amount of bibliographic data is practically impossible to process when seeking a paper for specific purposes and in the case of multiple papers extremely difficult to compare. Eugene Garfield (Garfield, 1972) made possible the widespread use of citation analysis through his creation of three citation indices: Science, Humanities and Social Science Citation Indices, which were combined and transformed into an electronic version called the Web of Science. These indices were based on the concept that a carefully selected subset of journals would produce the majority of important citing literature for any given article. Citation analysis has real world implications: for good or bad, citedness is considered in grants, hiring and tenure decisions. For many reasons professors and researchers may want to demonstrate the impact of their work and citation analysis is one way - a controversial one (Cheek J, Garnham B, Quan J., 2006) - to accomplish this. There have been two major popular ways for scoring scientific work and a third one, which is procured by combining the other two. In the first method, experts (appropriately assigned for the task) decide on the ranking. The second method relies on citation analysis, which involves examining the referring articles of an item (scientist / journal / conference) with the aim of validat-

ing them with the use of the appropriate indicators. A combination of the two which is favored quite recently is closer to the latter approach.

The first method which is based on an ad hoc approach works by collecting the opinion of different experts (or sometimes not, depending on the receiver of the evaluation) in a specific domain. E.g. instead of using a predetermined journal list, the respondents are asked to freely nominate their top-four research journals. This kind of work is very interesting, because a ranking according to readers' (and authors') perception is obtained, which is not always adequately expressed through citation analysis, but this method suffers from the fact that it is basically "manual" and sometimes (more times than not) biased, and not highly computerized (automated) and objective.

On the other hand, the second way of evaluating scientific work is by defining a mathematical function which calculates the output (a variable called "*score*") for each one of the important factors (called "*objects*") under evaluation, taking into account the graph mathematically created by a variable representing the *citations* (usually $c_{(i)}$) or a function of the ranking of the *citations* (usually $c_{(r)}$) among the published articles. Defining a quality and representative metric is not an easy task, since it should account for the productivity of a scientist and the impact of all of his/her work (analogously for journals/conferences) taking into account other important factors like the time of publication, the order of the names cited and many others. Most of the existing methods up-to-date are based on some form of mathematical function of the total number of authored papers, the average number of authored papers per year, the *total number of citations*, the average *number of citations* per paper, the average *number of citations* per year, etc.

Finally, in characteristic works in accordance with the hybrid approach, the scientists' *rankings* are realized by taking some averages upon the results obtained from the citation analysis and the experts' opinion, thus implementing a step incorporating both major approaches.

Although one cannot choose among citation analysis and experts' assessment

theoretically, given that both methods have their individual contributions, the former is usually the preferred method, because it can be performed in a fully automated and computerized manner and it is able to exploit the wealth of citation information available in digital libraries. All the metrics used so far in citation analysis, even the ones built on popular spectral techniques (Chakrabarti., et al., 1998), like HITS (Kleinberg, 1999) or *PageRank* (Page, et al., 1999) and its variations for bibliometrics, present the following disadvantages:

- They do not include the impact of the papers because the metrics are based solely on the total number of papers.
- They cannot evaluate productivity because the metrics are based on the average *number of citations* per paper.
- They are greatly influenced by a small number of extremely successful articles, which receive huge *number of citations*, whereas the rest of the articles may have negligible total impact because the metrics are based on the *total number of citations*.
- They present difficulty in setting administrative parameters, because the metrics are based on the number *x* of articles, which have received *y citations* each, which means that the metrics are actually based on the number *z* of the most cited articles.

To collectively overcome all these disadvantages of the indices, in 2005 J. E. Hirsch proposed the pioneering *h-index* (Hirsch J.E., 2005), which, in a short period, became extremely popular.

2. Indices

There is a multiplicity of bibliometric metrics, called indices, and their number is growing (probably a consequence of their increased use). All these indices are based on a quote from the source. They obviously inherit their source's issues. Indices provide quantitative data to make sound judgments about:

a. The parameters, which are set by an administrator and are not visible to the user initiating the action.
b. Productivity,
c. Specialization,
d. Collaboration,
e. Impact.

They provide adequate data to answer critical questions almost impossible or impractical to answer otherwise as the following:

- ✓ Which papers are most influential in a given field of research?
- ✓ Which authors are rising stars in their fields?
- ✓ How many articles has an institution produced in the past five years?
- ✓ How does that output compare to that of peer institutions?
- ✓ Has the research output of my country improved or declined in comparison with that of other countries?
- ✓ Where do the researchers who collaborate with researchers at an institution come from?
- ✓ Are researchers in a country performing better or worse than researchers in other countries publishing in the same journals?

2.1. Indices for journals

The evaluation of the scientific quality of an item is a delicate problem. A simple approach is to link the quality of a product to the quality of the medium in which it was published. Thus replaces an assessment of the support (usually a scientific journal) an individual assessment, which obviously greatly simplifies the work since there is much less media than articles. This approach is the initial model of the *Inter-Services*

Intelligence (ISI) which however has not changed under the pressure of client agencies.

2.1.1. The *Impact Factor* of a journal (*JIF*)

The *Journal Impact Factor* (*JIF*) is an index proposed by ISI in its *Journal Citation Reports* (JCR). It is calculated from the WoS[1]. This is initially a concept invented in the early sixties by Gene Garfield, founder of *ISI*. *JIF* of a journal to year n is defined as

"*the ratio between the number of citations received during the year n by articles published in years n-1 (C_{n-1}) and n-2 (C_{n-1}), and the total number of articles published in these two years ($P_{n-1} + P_{n-2}$)*" :

$$JIF_n = \frac{C_{n-1} + C_{n-2}}{P_{n-1} + P_{n-1}}$$

Equation 1: Journal impact factor *(JIF)*

The limitation to two years seems difficult to explain[2]. *JIF* is often considered an index of the quality of a journal and plays a significant role in the scientific world. The *ISI* however, indicates that *JIF* should not be used for different areas. In particular, the *JIF* in a field involving long research will automatically be lower than that of the journals of a rapidly changing field. It has been proven for example that the average molecular biology journals (area where an article is rapidly becoming obsolete) had a *JIF* much higher than the average journal of mathematics (Seglen P.O. , 1997). In 1999 the best *JIF* in mathematics corresponded to that of the 51st article in cell biology and the article of Andrew Wiles on Fermat's theorem contained only 4 of 84 references to publications that had been published in the two previous years (Sutherland W.J., 1999). Recent analyses show that this trend has not reversed. For example, the study of *JIF* of 181 journals in mathematics and 124 genetics journals based on the JCR 2005 showed that if the distribution were comparable, the average value

[1] Web Of Science

[2] ISI has recently proposed the instructions to calculate a *JIF* using a 5-year window.

of *JIF* varied by a factor of 10 between the two disciplines, in favor of genetics (Leydesdor L., 2007)[3].

Another logical argument against using the *JIF* for cross-domain comparisons is that of the journals with a low scientific content, which however, can have a domain *JIF* equal to that of another area with a high level of research since *JIF* depends only on *citations*. Conversely a very active area can have journals with a low *JIF* due to different citation practices or a reduced community. In this connection it is often said that the community size (measured in articles published) largely influences *JIF*. This is not true for all areas but many other factors can influence arbitrarily *JIF*: e.g. increasing the number of articles published in a journal in which the research is very active. *JIF* does not really measure the quality of an item or a specific author. Being based on WoS, *JIF* is often accused of an American bias even in the Anglo-Saxon community. This is due to the fact that one cites articles more easily in his own country and this bias was confirmed in special cases. Note that despite a lower coverage by *ISI* of life sciences they hold a special place for *JIF*: 12 of the 15 journals with a *JIF* above 10 fall into this discipline. The relatively general nature of life sciences in journals gives them an advantage in relation to good journals that are very rarely mentioned outside the community in examination.

2.1.2. The *Content Factor*

Impact Factor, the pre-eminent performance metric for medical journals, has been criticized for

- failing to capture the true impact of articles;
- favoring methodology papers;

[3] E. Garfield also draws attention to the absurdity of comparisons based on *JIF*: "*It is absurd to make invidious comparisons between specialist journals and multidisciplinary general journals like Nature and NEJM. To compare journals, you should stick to a particular category as is explained very carefully in the Guide to JCR*". (Garfield, 1998)

- being unduly influenced by statistical outliers; and
- examining a period of time too short to capture an article's long-term importance.

Also, in the era of search engines, where readers cannot and usually need not skim through journals to find information, the emphasis placed on citation efficiency through the calculation of the *Impact Factor* is probably redundant. A better metric should incorporate the effect of the *total number of citations* to all papers published by the journal - not just the recent ones - and not be influenced by the total number of papers published.

Bernstein and Gray (Bernstein & Gray, 2012) proposed a metric embodying the above principles called the "*Content Factor*". Thus, to remedy *Impact Factor*'s emphasis on recent *citations*, *Content Factor* considers the *total number of citations*, regardless of the year in which the cited paper was published. To correct for *Impact Factor*'s emphasis on efficiency, no denominator is employed. *The Content Factor is thus the total number of citations in a given year to all of the papers previously published in the journal*. The *Content Factor*, then, is simply the *total number of citations* in a given year to all of the papers the journal had published up to and including the year in question. The *Content Factor* is reported in kilo-cites (the *total number of citations* divided by 1000) to present units comparable in magnitude to those typically reported for *Impact Factor*. In a survey of 75 experienced orthopedic authors and a measurement of their perceptions of the "importance" of various orthopedic surgery journals conducted by the above, the *Content Factor* and the *Impact Factor* were found to be poorly correlated. The correlation between the "importance score" that the experts concluded and the *Impact Factor* was only 0.08 while the correlation between the "importance score" and *Content Factor* was 0.56. (Bernstein & Gray, 2012). Accordingly, the *Content Factor* reflects a journal's significance more accurately. In sum, while *Content Factor* cannot be accredited as the unique index of merit – meaning an easily obtained and intuitively appealing metric of the journal's knowledge contribution, not subject to gaming – it can be a useful adjunct, which with the appropriate contribution from other sources or indices can give an approximate rank of importance for a given paper.

2.1.3. The *immediacy index*

The *immediacy index* is the average number of times an article is cited in the year it is published. It shows how fast articles are cited following their publication. This index is also proposed by the ISI in its JCR[4]. It is defined as

"The ratio between the number of citations of articles published in year n (and only this one) and the number of articles published in the journal that year":

$$\boldsymbol{I_I = \frac{C_n}{P_n}}$$

Equation 2: The immediacy index

This index is often seen as a measure of the immediate impact of a journal. Because it is a per-article average, the *immediacy index* tends to discount the advantage of large journals over small ones. However, frequently issued journals may have an advantage because an article published early in the year has a better chance of being cited than one published later in the year. Many publications that publish infrequently or late in the year have low immediacy indices. For comparing journals specializing in cutting-edge research, the *immediacy index* can provide a useful perspective. It appears however that in many cases the journals presenting an elevated *immediacy index* obtain this figure due to a large *number of references* to editorials that do not appear in the denominator of this index.

2.1.4. *Cited half-life*

This index is also proposed by the *ISI* in its *JCR*. For the year *n* the *cited half-life* is the number of years *j* that 50% quotations from year *n* are previous *citations* to year *n - j* and 50% later. Thus, Nature Genetics Cited had a half-life of 4.7% in 2003 as it measured 46.38 % *citations* of 2003 from previous years dating back to 1999. This index provides information on the ongoing research in a given field. The indices such as *JIF* take into account only relatively recent *citations*; journals that have *cited half-life*

[4] Journal Citation Reports® offers a systematic, objective means to critically evaluate the world's leading journals, with quantifiable, statistical information based on citation data. By compiling articles' cited references, JCR helps to measure research influence and impact at the journal and category levels, and shows the relationship between citing and cited journals. Available in Science and Social Sciences editions. (Thomson Reuters, 2016)

rather small will have higher *JIF* mechanically than those with an important *cited half-life*.

2.1.5. The Scimago Journal Rank (*SJR*)

The Scimago Journal Rank (*SJR*) is based on the transfer of *prestige*[5] from a journal to another one; such *prestige* is transferred through the references that a journal does to the rest of the journals and to itself. The calculation of the final *prestige* of a journal is an iterative process, in which the *prestige* in the stage ***i*** of a journal depends on the *prestige* of the set of journals in stage *i-1*.

2.1.5.1. Calculation

The calculation of the *SJR* involves three stages:

- *Initial assignation of the SJR* : In this stage a default *prestige* is assigned to every journal. Having in mind that the *SJR* is calculated from an iterative process, which is based on the values, assigned in the previous step, it is necessary to have some initial values. The calculation of the *SJR* is a process that converges, so these initial values don't determine a final result, but just influence in the number of iterations needed.
- *Iteration process of calculation*: Starting from stage 1, the computation is iterated to calculate the *prestige* of each journal based on the *prestige* transferred by the rest.
- The process ends when the variation of the *SJR* between two iterations is less than a limit prefixed before the calculation process. The final result is the *SJR* of each journal.

[5] See Appendix A.4.2

$$SJR_i = \frac{(1-d-e)}{n} + e \cdot \frac{Art_i}{\sum_{j=1}^{N} Art_j} + d \cdot \sum_{j=1}^{N} \frac{C_{ji} \cdot SJR_i}{C_j} \cdot \frac{1 - (\sum_{k \in (Dangling-nodes)} SJR_k)}{\sum_{h=1}^{N} \sum_{k=1}^{N} \frac{C_{kh} \cdot SJR_k}{C_k}} + d \cdot \left[\sum_{k \in (Dangling-nodes)} SJR_k \right] \cdot \frac{Art_i}{\sum_{j=1}^{N} Art_j}$$

Equation 3: Calculation of SJR. The equation consists of 4 Addends. See 2.1.5.2. below

Where: C_{ji} - Citation from journal *j* to journal *i*.

C_j - *Number of citations* of journal *j*.

d - Constant (normally ≈ 0.85).

e - Constant (normally ≈ 0.10).

N - Number of Journals

2.1.5.2. Description of the equation

- *Addend 1*: It corresponds to the minimum *prestige* assigned to any considered journal, independently of any other factor (nº of articles, *citations*, etc.). It depends directly on the number of journals of the domain.
- *Addend 2*: *Prestige* that obtains a journal due to nº of articles published in the three-year window. It depends on both the number of articles published by the journal and the sum of all articles of all the journals of the domain[6].
- *Addend 4*: It is the *prestige* obtained by a journal which is represented by a dangling node. The amount of *prestige* to distribute is the sum of the *prestige* of all dangling nodes in the previous iteration. The *prestige* that a *concrete*

[6] Addends 1 and 2 are constant in all the iterations, their sum forming the minimum *prestige* that a journal receives.

node receives is directly proportional to the nº of articles published by the journal. A journal with more published articles receives more *prestige* from the dangling nodes than another with fewer *citations*.

- *Addend 3*: It is the *prestige* that a journal obtains from the journals that mention it. The percentage of the *prestige* that a journal *X* transfers to another one (*Y*) is constant in all the iterations, and depends on the ratio of nº connections of journal *X* to the total journal of connections of journal *Y*/nº of connections of journal *X*. The amount of *prestige* transmitted by a journal depends on that constant and the value of the JR in the previous iteration.
- *Computation of the prestige average per article (SJRQ)* : After stage 2 each journal has computed its *SJR*, an index of its global *prestige*. To obtain the *SJR*Q index we divide the *SJR* by the number of articles published in the citation window. The result is the *prestige* average per article, since logically the *prestige* obtained by a journal is the result of the *prestige* obtained by its articles. So, it could be compared to the *prestige* average per article without having in mind other factors like the frequency of each journal, the number of articles, etc.

$$SJRQ_i = \frac{SJR_i}{Art_i}$$

Equation 4: Calculation of SJRQ

Where : SJR_i : Scimago Journal Rank of the Journal *i*

Art_j : Number of Articles of journal *j*

2.1.5.3. Variables of each journal

1) Number of articles of a journal: the number of articles published in the citation window. The number of articles influences in: a) Determining the addend 2 in the stage 2[7]. b) Calculating the amount of *prestige* which is received from addend 4 in the stage 2. c) The Computation of *SJR*Q in stage 3.

2) Number of total references of a journal: The amount of *prestige* that a journal *X* transmits to another *Y* is defined by the division between *nº of references* from *X* to a *Y* by the *nº of total references* of *X*. Actually it is the *prestige* of a journal transmitted to another journal depending on both the *number of references* from *X* to *Y*, and on the *total number of references* of *X*. The *total number of citations* of a journal includes both those directed to journals of the domain considered and those directed to journals outside this domain.

3) Number of references received by a journal X: The *prestige* that any journal receives depends on the *number of citations* that it receives from the other journals; the bigger the *number of citations* to a journal the bigger will be the *prestige* of that journal.

[7] See Equation above.

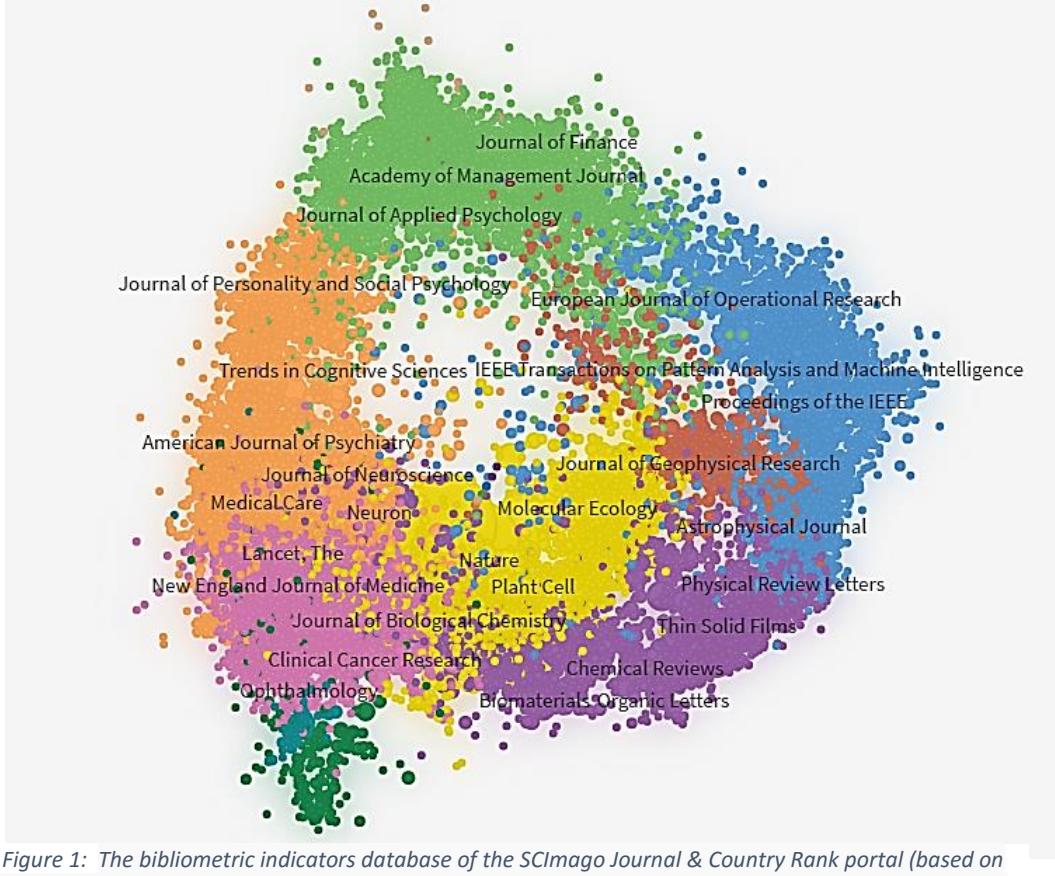


Figure 1: The bibliometric indicators database of the SCImago Journal & Country Rank portal (based on 2014 data) through the use of The Shape of Science, an information visualization project, which shows a very intuitive image of the interconnection of the different subject areas by the position of the journals. The individual profiles of the journals can be accessed from this interface. (Scimago Lab, 2016)

2.1.5.4. Global Values

1) *Limit of convergence criterion*: $|SJR_{i+1} - SJR_i| < Lim.$ When this criterion is fulfilled for all journals the calculation process is terminated.

2) *Number of Journals (N):* Its value corresponds with the total number of journals considered in the calculation; its value will be different if it varies the universe of journals considered. It determines addend 1, which actually is the minimum amount of *prestige* that each journal has in this domain.

3) *Global number of articles*: it is the sum of all articles of the journals considered in the calculation published in the three-year window. It influences addend 2 and 4.

4) *Constants d and e*: constants that determine the weight of the four addends of the *SJR* calculation equation.

5) *Dead-end (dangling) Nodes*: Certain journals of the domain that do not have references to any other journal of this domain, although they themselves might be cited or not. They constitute impasses in a graph since from them it is not possible to jump to other nodes. In order to assure that the iterative process is convergent, dead-end nodes virtually are connected to all those of the domain and their *prestige* is distributed between all the nodes (addend 4) proportionally to the number of articles of each one.

The *number of references* of a journal *X* to all the other journals of the domain is smaller than the *total number of references*, which means that part of its *prestige* is not distributed. As a consequence, the system does not converge. To solve this problem a *corrector* factor is determined in addend 3. This factor is common to all the journals that receive *citations*. It is used to distribute the *prestige* corresponding to the *citations* that go outside that domain between the mentioned journals of the domain. This *prestige* is distributed proportionally in accordance with the *citations*. (SCImago, 2007)

2.1.6. Scimago total cites

Total Cites (3years) is the *number of citations* received in the selected year by a journal to the documents published in the three previous years, --i.e. *citations* received in year *X* to documents published in years *X-1, X-2* and *X-3*. All types of documents are considered. The total number of times that a journal has been cited by all journals is included in the database in the JCR year. *Citations* to journals listed in JCR are compiled annually from the JCR years combined database, regardless of which JCR edition lists the journal and regardless of what kind of article was cited or when the cited article was published. Each unique article-to-article link is counted as a *citation*. *Citations* from a journal to an article previously published in the same journal are compiled in the total cites. However, some journals listed in JCR may be cited only journals, in which case self-cites are not included. (Thomson Reuters, 2015)

2.1.7. *Cites per doc*ument (impact)

Cites per doc is the average *number of citations* received per paper or the *total number of citations* (Times Cited) divided by the total number of Web of Science (WoS) *citations*. In a Global Comparisons Report[8], *cites per doc* limited to international collaboration counts is the *number of citations* received by papers with international collaboration divided by the total number of papers with international collaboration.

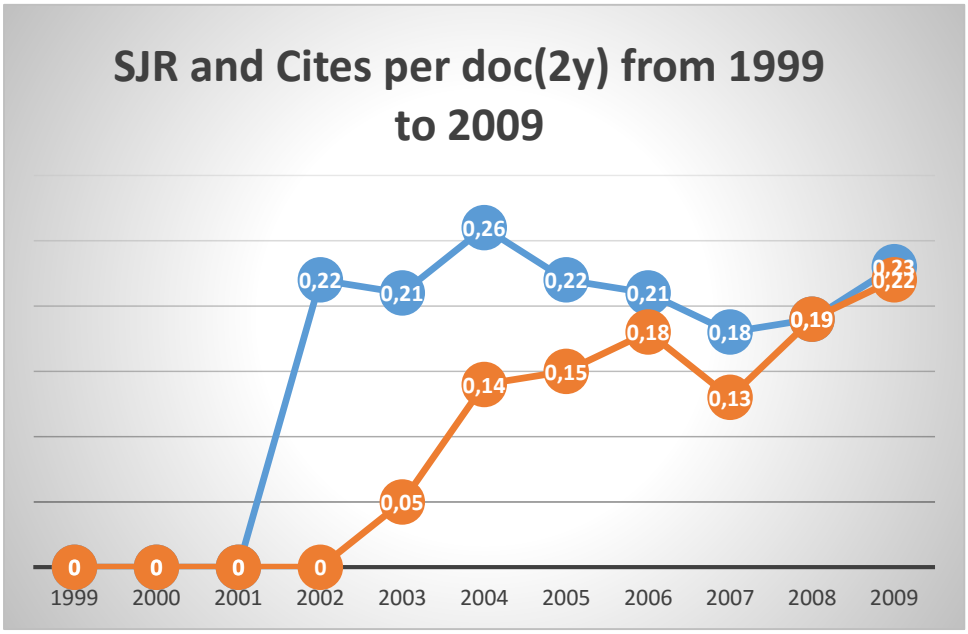


Figure 2: SJR -orange line- measures the scientific influence of the average article in a journal. Cites per Doc (2y) -blue line- measures the scientific impact of an average article published in the journal. For the PJP (Portuguese Journal of Pulmonology) it was 0.119 in 2007, increasing to 0.189. (Wincka & Morais, 2011)

[8] Global Comparisons metrics are calculated from Web of Science documents classified as Article, Note or Review. Proceedings papers are excluded unless they are also classified as Articles in Web of Science (some documents in Web of Science are assigned to more than one document type). The data used to calculate metrics--number of documents and Times Cited--are variables. They may change from year to year or once every few years or never.

2.1.8. *Centrality* indices

The prominence of a node in a weighted network is defined by two basic prominence classes[9]: *Centrality* and *Prestige*. The *centrality* of a node is calculated based on his/her volume of activity – how high his/her involvement is in many relations, regardless of their directionality (sending or receiving). There are three main *centrality* indices.

2.1.8.1. Degree *Centrality*

A node with a high degree *centrality* maintains numerous contacts with other network nodes. Nodes have higher *centrality* to the extent that they can gain access to and/or influence over others. A central node occupies a structural position (network location) that serves as a source or gate for larger volumes of information exchange and other resource transactions with other nodes. Central nodes are located at or near the center in a sociometric[10] network diagram. In contrast, a peripheral node maintains few or no relations and thus is located at the margins of a network diagram. *Actor-level*[11] *degree centrality* is simply each *actor*'s number of degrees in a non-directed graph:

$$\boldsymbol{CD_{(n_i)} = d_i(n_i)}$$

Equation 5: Calculation of degree centrality

where *i* is the *actor*'s *indegree*[12]

To standardize or normalize the *degree centrality* index, so that networks of different sizes (*g*) may be compared, divide it by the *maximum possible indegrees* (= *g-1*) nodes if everyone is directly connected to ***i***, and express the result as a proportion or percentage:

[9] See Appendix A.4.2. : Centrality & Prestige
[10] Sociometry is a quantitative method for measuring social relationships.
[11] The actor-network theory developed by Callon and his colleagues is an attempt to invent a vocabulary to deal with relationships among actors of a social network.
[12] See Appendix A.4.1.

$$C'D(n_i) = \frac{d_i(n_i)}{(g-1)}$$

Equation 6: Calculation of the normalized degree centrality index

where *g* is maximum possible *indegrees*.

2.1.8.2. Closeness Centrality

In the closeness concept, a *central actor* has minimum path distances from the *g-1* remaining nodes. An *actor* that is close to many others can quickly interact and communicate with them without going through many intermediaries. Thus, if two *actor*s are not directly tied, requiring only a small number of steps to reach one another, then, it is important to attain higher *closeness centrality*. *Actor closeness centrality* is the inverse of the sum of geodesic distances[13] from *actor i* to the *g-1* other *actor*s (i.e., the reciprocal of "how far" it lies):

$$C_C(n_i) = [\sum_{j=1}^{g} d(n_i, n_j)]^{-1}$$

Equation 7: Calculation of the closeness centrality of an actor

Where $d_{(ni,\ nj)}$ is the distance between the nodes n_i and n_j[14].

A *closeness index* can be standardized (normalized) by dividing it by the maximum possible distance expressed as a proportion or percentage.

2.1.8.3. Betweenness Centrality[15]

A central node occupies an intermediate position on the geodesics connecting many pairs of other *actor*s in the network. As a cutpoint[16] in the shortest path connecting two other nodes, a *between actor* might control the flow of information or the ex-

[13] The shortest path between two points according to the Riemannian metric is called a geodesic.
[14] In a directed graph, the geodesic distance between two actors may differ with the nodal order ($d_{(ni,nj)}$ may not equal $d_{(nj,ni)}$).
[15] See Appendix 0
[16] In topology, a cut-point is a point of a connected space such that its removal causes the resulting space to be disconnected.

change of resources, perhaps charging a fee or brokerage commission for transaction services rendered. If more than one geodesics link a pair of *actor*s, it is assumed that each of these shortest paths has an equal probability of being used. This probability is compensated by cut points i.e. between nodes. *Betweenness centrality* for *actor i* is the sum of the proportions, for all pairs of *actor*s *j* and *k*, in which *actor i* is between (i.e. involved in a pair's geodesics):

$$C_B(n_i) = \sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}}$$

Equation 8: Betweenness centralityfor actor i

As with the other *centrality* standardizations, the *betweenness centrality* is normalized by dividing it by the maximum possible *betweenness*, expressed as proportion or percentage.

2.2. Indices for websites - *PageRank*

2.2.1. The definition of *PageRank*

"*PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites."* (Page L. & Brin S., 2011)

Within the past few years, Google has become the far most utilized search engine worldwide. A decisive factor therefore was, besides high performance and ease of use, the superior quality of search results compared to other search engines. This quality of search results is based on *PageRank* which is a sophisticated method to rank web documents. It was devised by Google founders Lawrence Page and Sergey Brin from their time as graduate students at Stanford University. Nevertheless, much time has passed since the scientific work on *PageRank*, and within the past years most likely many changes, adjustments and modifications regarding the ranking methods of Google have taken place, but at least the fundamental concept behind *PageRank* still remains constitutive.

For the purpose of better search results and specially to make search engines resistant against automatically generated web pages based upon the analysis of content specific ranking criteria (doorway pages), the concept of link popularity was developed. Accordingly, the number of inbound links for a document measures its general importance. Hence, a web page is generally more important, if many other web pages link to it. The concept of link popularity often avoids good *rankings* for pages which are only created to deceive search engines and which don't have any significance within the web. So, within the *PageRank* concept, the rank of a document is given by the rank of those documents which link to it. Their rank again is given by the rank of documents which link to them. Hence, the *PageRank* of a document is always determined recursively by the *PageRank* of other documents.

2.2.2. The *PageRank* Algorithm

The original *PageRank* algorithm was described by Lawrence Page and Sergey Brin in several publications (Page & Brin, 1998). It is given by two equations (versions)

$$\boldsymbol{PR}\,(\boldsymbol{A}) = (\mathbf{1} - \boldsymbol{d}) + \boldsymbol{d} \cdot \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{i}=\boldsymbol{n}} \frac{\boldsymbol{PR}(\boldsymbol{T_i})}{\boldsymbol{C}(\boldsymbol{T_i})}$$

Equation 9: Calculation of PageRank (1st version) of a website

$$\boldsymbol{PR}\,(\boldsymbol{A}) = \frac{(\mathbf{1} - \boldsymbol{d})}{\boldsymbol{N}} + \boldsymbol{d} \cdot \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{i}=\boldsymbol{n}} \frac{\boldsymbol{PR}(\boldsymbol{T_i})}{\boldsymbol{C}(\boldsymbol{T_i})}$$

Equation 10: Calculation of PageRank (2nd version) of a website

Where

PR(A) is the *PageRank* of page *A*,

$PR(T_i)$ is the *PageRank* of pages T_i which link to page *A*,

$C(T_i)$ is the number of outbound links on page T_i ,

d is a damping factor which can be set between 0 and 1 (usually ≈ 0.85) and

N is the total number of all pages on the web (only 2nd version).

PageRank does not rank web sites as a whole, but is determined for each page individually. Further, the *PageRank* of page *A* is recursively defined by the *PageRank*s of those pages which link to page *A*. The *PageRank* of pages T_i which link to page *A* does not influence the *PageRank* of page *A* uniformly. Within the *PageRank* algorithm, the *PageRank* of a page *T* is always weighted by the number of outbound links *C(T)* to page *T*. This means that the more outbound links a page *T* has, the less will page *A* benefit from a link to it on page *T*. The weighted *PageRank* of pages T_i is then added up. The output of this is that an additional inbound link for page *A* will always increase page A's *PageRank*.

Finally, the sum of the weighted *PageRank*s of all pages T_i is multiplied with a damping factor d, which can be set between 0 and 1. Thereby, the extent of *PageRank* benefit for a page by another page linking to it is reduced. (Page, et al., 1999). The 2nd version of the algorithm does not differ fundamentally from the first one. Regarding the *Random Surfer Model*[17], the second version's *PageRank* of a page is the actual probability for a surfer reaching that page after clicking on many links. The *PageRank*s then form a probability distribution over web pages, so the sum of all pages' *PageRank*s will be one. On the contrary, in the first version of the algorithm the probability for the random surfer reaching a page is weighted by the total number of web pages. So, in this version *PageRank* is an expected value for the random surfer visiting a page, when he restarts this procedure as often as the web has pages. For example, if the web had 100 pages and a page had a *PageRank* value of 2, the random surfer would reach that page twice in average if he restarts 100 times. Because of the size of the actual web, the Google search engine uses an approximate, iterative computation of *PageRank* values. This means that each page is assigned an initial starting value and the *PageRank*s of all pages are then calculated in several computation circles based on the equations determined by the *PageRank* algorithm. By means of the iterative calculation, the sum of all pages' *PageRank*s still converges to the total number of web pages. So the average *PageRank* of a web page is 1. The

[17] A model of the behavior of a random surfer. The random surfer simply keeps clicking on successive links at random. However, if a real Web surfer ever gets into a small loop of web pages, it is unlikely that the surfer will continue in the loop forever. Instead, the surfer will jump to some other page (he will be bored). (Page & Brin, 1998)

minimum *PageRank* of a page is given by *(1-d)*. Therefore, there is a maximum *PageRank* for a page which is given by *d·N+(1-d)*, where N is total number of web pages. This maximum can theoretically occur, if all web pages solely link to one page, and this page also solely links to itself.

2.2.3. Influencing factors

The following potential influencing factors are included in the patent specifications for *PageRank*:

- Visibility of a link
- Position of a link within a document
- Distance between web pages
- Importance of a linking page
- Up-to-dateness of a linking page

$$\boldsymbol{PR}\ (\boldsymbol{A}) = (\mathbf{1}\ -\ \boldsymbol{d}) +\ \boldsymbol{d}\ \cdot \sum_{\boldsymbol{i=1}}^{\boldsymbol{i=n}} (\boldsymbol{PR}(\boldsymbol{T_i}) \times \boldsymbol{L}(\boldsymbol{T_i}, \boldsymbol{A})$$

Equation 11: PageRank modified to include $L(T_i,A)$ which represents the evaluation of each link with A

Where :

$L(T_i, A)$ represents the evaluation of a link which points from page T_i to page A. (Page, 1997)

2.2.3.1. The *Y-factor*

The *Y-factor* was introduced by Johan Bollen, Marko A. Rodriguez and Herbert Van de Sompel in 2006 (Bollen, et al., 2006). It is a simple combination of both the *ISI IF* (ISI *Impact Factor*) and the weighted *PageRank*. It was found that the *resulting journal rankings* correspond well to a general understanding of journal status. As articles cite one another, they define an article citation network in which each node represents an article and each directed edge represents a citation by that article to another. By grouping all articles published in the same journal under a single journal node, an article citation network can easily be transformed into a Journal Citation Network.

In that network, the directed edges between the journal nodes represent the collection of *citations* from one journal to another. This network can be formalized as a set of journals *V*, a set of directed edges $E \subseteq V^2$ that exist between the journals in *V*, and the function $W(v_i, v_j) \rightarrow N$ which maps each edge between the journal v_i and v_j to a positive, *integer citation frequency*. A range of journal status metrics can be applied to such a Journal Citation Network.

Having calculated the *ISI IF* (ISI *Impact Factor*) and the weighted *PageRank* respectfully, and in order to rank journals on the basis of both metrics combined, the product of the popularity-oriented *ISI IF* and the prestige-oriented Weighted *PageRank*, labeled *Y-factor* is defined as follows:

$$\boldsymbol{Y(v_j) = ISI\ IF(v_j) \times PRw(v_j)}$$

*Equation 12: Definition of the Y-*factor *metric*

where: *ISI IF* (v_j) is the ISI *Impact Factor* for v_j,

PRw(v_j) is the weighted *PageRank* for v_j.

In the assessment of scholarship, the ISI *Impact Factor* rules as the prime indicator or journal status. The *ISI IF* for a given journal is based on the *number of citations* it receives, and ignores the *prestige* of the citing journals. Therefore, it is an indicator of journal status that favors popularity over *prestige*. While the journal status metric which is obtained by computing Weighted *PageRank* for all journals in a Journal Citation Network strongly overlaps with the *ISI IF*, it also reveals significant and meaningful discrepancies. *PageRank* is a metric known to take the *prestige* factor *of status* into account. The fact that the widely used *PageRank* metric differs in a meaningful manner from the *ISI IF* is a substantial reason to contemplate the use of a variety of journal status metrics instead of just one. The simplistic definition of the *Y-factor rankings* may not be scientifically convincing, still the top scoring journals according to this ranking principle rather closely matched the perception of importance of eminent scholars. (Bollen, et al., 2006). (Senanayake, et al., 2015)

2.3. Downloads

The publishers' trend toward allowing online access leads to the use of another index: the number of downloads. This information has the advantage of being obtained in real time (Meho & Yang, 2007) and a correlation has been established between the number of downloads and the *number of citations*, though the degree of correlation varies significantly across disciplines. Under these conditions the number of downloads would provide an initial estimate of the future *number of citations* of articles. However, restrictions apply to this index:

- It is difficult to establish for a given author likely to publish in a variety journals because it would be very costly to examine all the journals of a database for each author.
- It does not take into account the new means of distribution used by researchers (personal pages, open archives).
- The practice of some publishers to mention the most downloaded items mechanically promotes these articles.
- Reliability is relative. These indices are calculated by the publishers themselves, which is an obvious interest conflict.

2.4. Quantitative indices

These are the easiest to establish indices from citation databases. Examples of such cases:

- *number of publications* and *citations* for a defined group of researchers,
- *number of publications* and *citations* per researcher for a defined group of researchers,
- percentage of world production,
- *number of publications* in ISI involved indices,
- *number of publications* in journals with high *JIF*

It is understood that the first three indices give no information on the quality of scientific work: they allow, at most, the evaluation of whether the group has an

"normal" publication activity, thus, making it feasible to compared with the average activity of other groups working in the same field. The next two have a validity correlated in that they are ISI indices for the same domain in consideration.

2.5. The individual indices

Organizations using the indices are demanding measures that would enable an individualized assessment of researchers, which is not the purpose of the indices of journals[18]. In this context, the journals' indices allow them an indirect view that does not lead to the quantified values they seek. Scientists also argue that an average of analyses based on journals' indices cannot reflect the quality of a particular item.

2.5.1. The h-index

2.5.1.1. General Information

J.E. Hirsch (Hirsch J.E., 2005) defined the h number of an author - the *h index* - of articles of the author, which have been cited at least h times each. It was proposed as an alternative to other indices (including the advantages and disadvantages listed in Hirsch's paper) as follows:

- *Total papers*: which measures productivity but not the impact.
- *Total number of citations*: It helps measure a form of total impact but may be strongly influenced by the number of co-authors and review articles (Royle, et al., 2013).

[18] E. Garfield on the purpose of the indices:" The *source of much anxiety about Journal Impact Factors comes from their misuse in evaluating individuals, e.g. during the Habilitation process. In many countries in Europe, I have found that in order to shortcut the work of looking up actual (real) citation counts for investigators the journal impact factor is used as a surrogate to estimate the count. I have always warned against this use."* (Garfield, 1998)

- *Citations per paper*: allows comparisons between scientists of different ages; but: it is difficult to estimate, it rewards low productivity and it penalizes high productivity.
- *Number of significant publications*: number of cited papers y times; it does not have the drawbacks of the previous indices but suffers from arbitrariness in the choice of y; we could also mention the difficulty of measurement.

The *h-index* is a measure of the number of highly impactful papers a scientist has published. The larger the number of important papers, the higher the *h-index,* regardless of where the work was published. The index was suggested in 2005 by Jorge E. Hirsch, a physicist at UCSD (University of California, San Diego), as a tool for determining theoretical physicists' relative quality and is sometimes called the *Hirsch-index* or *Hirsch number*. According to Hirsch*: "The h-index is defined by how many h of a researcher's publications (N_p) have at least h citations each."* To calculate it, only two pieces of information are required:

- The total number of papers published (N_p) and
- The *number of citations* (N_c) of each paper.

2.5.1.2. Calculation

Once a set of publications is identified, their *bibliographic metadata*, including *citations of each article*, is collected. After the collection of the metadata is completed, the records are put in order of their decreasing *citation counts* starting with the most frequently cited. The most frequently cited article is ranked in the first place and every record is ranked based on each *citation count*. To find the *h-index,* scroll down until the *number of citations* equals the number of the paper. A scientist has index h if h of his or her N_p papers have at least *h citations* each and the other (N_{p-h}) papers have *h or less citations* each (Hirch J.E., 2005).

2.5.1.3. Advantages of the h-index

The *h-index* can be very useful for comparative description of scientific topics (Banks, 2006) and most importantly for awarding scientific prizes (Glanzel & Persson, 2005).

The main advantage of the *h-index* is that it combines a measure of quantity (*publications*) and impact (*citations*) in a single indicator. More specifically, it relies on *citations* to the scientists' papers, not the journals, which is a truer measure of quality. Therefore, it is not increased by a large number of poorly cited papers, unlike total number of papers would be. Also, it performs better than other single-number criteria commonly used to evaluate the scientific output of a researcher such as: the *impact* factor*,* the *total number of documents* and the *total number of citations.* (Costas & Bordons, 2007)

2.5.1.4. Limitations

Although the *h-index* is a very useful tool for bibliometric analysis of scientists, it has its limitations.

Firstly, there are inter-field differences in typical h values due to differences among fields in productivity and citation practices (Hirsch J.E., 2005), so the *h-index* should not be used to compare scientists from different disciplines. Hirsch indicates that the measure of the *h-index* can be easily obtained from the Web of Science using the order Times Cited proposed by the ISI. But this requires good coverage of all areas of science (or at least a homogeneous coverage if we want to compare individuals) by the Web of Science, which is far from being the case. Only at the end of his paper he briefly addresses the problem of the value of the *h-index* across disciplines: *"h-indices in biological sciences tend to be higher than in physics….more research in understanding similarities and differences….in different field of science would be of interest"* (Hirsch J.E., 2005))

Secondly, another important problem Scientometrics[19] has to face, is that the use of the *h-index* could provoke changes in the publishing behavior of scientists, such an artificial increase in the number of self-*citations* distributed among the documents on the edge of the *h-index* (Van Raan, 2006). Self-*citations* can increase a scientist's h, but their effect on h is much smaller than on the total citation count since only

[19] Scientometrics is *the "quantitative study of science, communication in science, and science policy"* (Hess, D. J., 1997, p. 75)

self-*citations* with a *number of citations* just greater than h are relevant (Hirsch J.E., 2005). Among the many issues raised by this index include the reliability with which it can be measured. The ISI initially refused to provide this indicator, which led to the development of tools based on Google Scholar data, which besides the fact of having misidentified sources, it mismanages homonyms and self-*citations*. For authors with relatively common names the results obtained via Google Scholar are often fanciful (without doing an exhaustive search may well show that originally qualified authors of an *h-index* of 35, very high, saw this number reduced to 5 (average) as soon as the *citations* found in Google Scholar) were more closely examined. Meho (Meho & Yang, 2007) claims that the *h-index* is now readily calculable from the Web of Science (ISI in October 2006 changed its policy and proposed the *h-index* in its JCR), Scopus and Google Scholar. Nevertheless, the diversity of responses in these three databases requires some overlap and manual analysis of the validity of data before an approximately correct value of the index in the sense of its definition (without assuming its validity) can be obtained. (Costas & Bordons, 2007) In addition, the sources used mismanage references to books or book chapters, which strongly penalizes the references to book authors. One can thus find examples of authors who have written very few articles and whose books are known and have a very low h-index: *max quotes* is the only index that, if provided in addition to the h index, can then indicate the unconformity between the *h-index* and the real influence of the author, which then requires a very careful examination of *citations* themselves.

A critical analysis of the *h-index* (Roediger H.L., 2006) noted the following:

- the *h-index* is correlated with the *citation dates* (Egghe, 2007),
- the *h-index* can substantially increase even if the researcher is no longer active for a long time (Costas, et al., 2011),
- the *h-index* is underestimated for researchers with published books,
- the *h-index* does not highlight the very important contributions of an author,
- the fact that the *citations* are attributed to all the authors does not take into account the practical areas; the order of authors cited reflects the

importance of the contributions in most cases and this is not taken into account in its calculation,

- negative *citations* are not taken into account,
- it does not take into account the highly cited works and
- ignores the *total number of citations*,

As can be seen in below, most of the Nobel recipients in Physics achieved an *h-index* of 35-39. Is this a just representation of their influence?

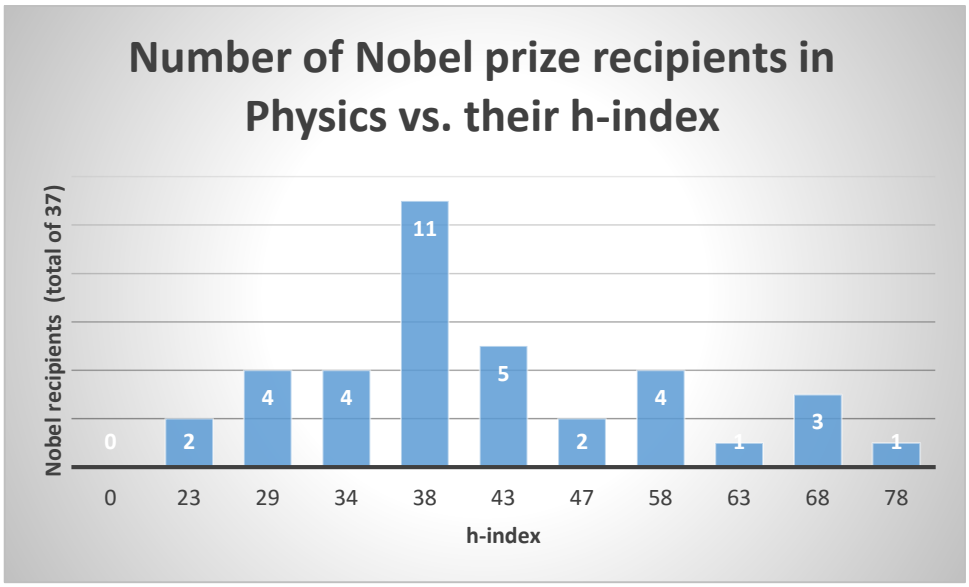


Figure 3: Histogram giving the number of Nobel Prize recipients in physics in the last 20 years versus their h-index. The peak is at the-h index between 35 and 39. (Hirsch, 2005)

As Hirsch puts it: *"There will be differences in typical h-values in different fields, determined in part by the average number of references in a paper in the field, the average number of papers produced by each scientist in the field, and also by the size (number of scientists) in the field … Scientists working in non-mainstream areas will not achieve the same very high h values as the top echelon of those working in highly topical areas".* He sums up with the statement: *"While I argue that a high h is a reliable indicator of high accomplishment, the converse is not necessarily always true."* (Hirsch, 2005, p. 4)

To overcome the limitations of the *h-index,* different modifications have been suggested that led to the creation of a big number of new *h-index* variants.

2.5.2. The *h-index* variants

Hirsch has encouraged the development of these alternative indices. However, we can estimate that the increased number of indices without a minimum of critical analysis casts doubt on the effectiveness of the relationship between scientific validity and indices. (Demaine, 2011-2012) (Bornmann, et al., 2008)

Podlubny (Podlubny, 2005) (Podlubny & Kassayova, 2006) showed that for nine broadly defined disciplines the average ratio of *total citations* to the *number of citations* in mathematics varied considerably (Mathematics: 1, Engineering/technology: 5, Biology: 8, Earth/space sciences: 9, Social/behavioral sciences: 13, Chemistry: 15, Physics: 19, Biomedical Research: 78, Clinical Medicine: 78).

Similarly, Iglesias & Pecharroman (Iglesias & Pecharroman, 2006) calculated the average *number of citations* per paper in the 21 different ISI fields and used this to design a normalization factor. Unfortunately, the discipline areas used in neither studies map closely enough onto the categories used by Google Scholar to use these normalization factors in Publish or Perish. However, they do show that comparisons of bibliometric data across fields are generally inappropriate.

Part of the differences between disciplines are caused by the fact that academics in the Natural Sciences typically publish more (and often shorter) articles and also publish with a large number of co-authors, while academics in the Social Sciences and Humanities typically published fewer (and longer) articles (or books) and publish with fewer co-authors.

However, differences in the number of co-authors also seem apparent within the same discipline. For instance, North American academics tend to publish articles with a larger number of co-authors than European academics. Since 1990, papers in the North-American Academy of Management Journal on average have 2.24 authors, papers in the British Journal of Management 2.01 authors, and papers in the Euro-

pean Management Journal 1.84 authors. Additional variations of the *h-index* have been proposed.

2.5.2.1. The Individual *h-index* (original)

The Individual *h-index* was proposed by Batista, Campiteli, Kinouchi, and Martinez (Batista, et al., 2006). It divides the standard *h-index* by the average number of authors in the articles that contribute to the *h-index,* in order to reduce the effects of co-authorship; the resulting *index* is called h_I.

$$\boldsymbol{h_I = \frac{h^2}{N_t}}$$

Equation 1: Calculation of the individual h-index

Where: N_t the number of authors considered in the *h* papers.

It was found that the distribution of the *h-index*, although it depends on the field, could be normalized by a simple rescaling factor. In fact, the normalization of the *h-index,* is carried out using a normalization of the *number of citations* of each article indexed, according to the equation:

$$\boldsymbol{S(i,t) = \frac{4}{(t - t_1 + 1)} \cdot C(i,t), t \geq t_1}$$

Equation 2: Calculation of the value of the citation for the i-th article at time t in which we calculate the number of citations (normalization of the h-index)

Where:

• *S (i, t)* is the value of the citation for the *i-th* article at time *t* in which we calculate the *number of citations*;

• *C (i, t)* is the *number of citations* detected from the data base at time *t* for the *i-th* article;

• t_1 is the year of publication of the article. (Harzing, 2013)

2.5.2.2. The contemporary h- index

The Contemporary *h-index* was proposed by Antonis Sidiropoulos, Dimitrios Katsaros, and Yannis Manolopoulos in their paper (Sidiropoulos, et al., 2007). It adds an age-related weighting to each cited article, giving (by default; this depends on the parameterization) less weight to older articles. The weighting is parameterized; the Publish or Perish implementation uses γ = 4 and δ =1, like the authors did for their experiments. This means that for an article published during the current year, its *citations* count four times. For an article published 4 years ago, its *citations* count only once (4/4). For an article published 6 years ago, its *citations* count 4/6 times, and so on. The contemporary *h- index* is defined in the following way: the normalization is made on each of the items by dividing the *number of citations* received by the number of years elapsed from the year of publication to the reference year of the data base, the total multiplied by 4 to obtain reasonable numerical (Sidiropoulos, et al., 2007). The following equation:

$$\boldsymbol{S^c(i) = \gamma \cdot (Y(now) - Y(i) + 1)^{-\delta} \cdot |C(i)|}$$

Equation 3: Calculation of the contemporary h-index

Where

Y(i) is the publication year of article *i* and

C(i) are the articles citing the article *i* .

If we set *δ*=1, then $S^c(i)$ is the *number of citations* that the article *i* has received, divided by the "age" of the article. The choice of using the "contemporary *h- index*" was dictated by the following considerations:

- It is an indicator well known in the literature and used in bibliometry;
- it includes a linear normalization for academic age of single item;
- it assigns a weight independent of the time during "the period of activity" of the article, and a decreasing weight in its "soft period" as the article gets older and does not accumulate more *citations*;
- it captures the concept of "active researcher", assigning a weight to the prevailing most recent publications;

- it performs well on a large sample.

The h_c*-index* corrects for the recentness of the *citations*, with recent *citations* carrying more weight (Sidiropoulos, et al., 2007).

A researcher has an index equal to h_c (h contemporary) if h_c of its publications have a citation indicator *S (i, t)* whose total is greater than h_c, and other publications have a citation indicator S^c*(i, t)* less than or equal to h_c. (Sidiropoulos, et al., 2007)

2.5.2.3. The $h_{I,norm}$ index (Publish or Perish (PoP) variation)

Publish or Perish also implements an alternative individual *h-index* called $\boldsymbol{h_{I,norm}}$ that takes a different approach: instead of dividing the total *h-index,* it first normalizes the *number of citations* for each paper by dividing the *number of citations* by the number of authors for that paper, then calculates the *h-index* of the normalized citation counts. This approach is much more fine-grained than Batista et al.'s. It accounts for any co-authorship effects that might be present more accurately and thus it is a better approximation of the per-author impact, which is what the original *h-index* has set out to determine.

2.5.2.4. The multi-authored h_m index

The third variation is due to Michael Schreiber (Schreiber, 2008). Schreiber's method uses fractional paper counts instead of reduced citation counts to account for shared authorship of papers, and then determines the multi-authored h_m index based on the resulting effective rank of the papers using undiluted citation counts. In scientometrics, the problem of how to count multi-authored publications has been discussed for a long time (Lindsey, 1980, Price, 1981), assigning credit proportionally to the number of authors which is usually called fractional counting or adjusted counting.

There have, however, evolved a number of different methods for accrediting publications for several authors, see e.g. Egghe et al. (2000). One difficulty is, that different scoring methods can lead to paradoxical effects and yield totally different *rankings* (van Hooydonk, 1997, Egghe et al., 2000) so that no unambiguous solution of the "multiple-author problem" (Harsanyi, 1993) exists. But fractional counting is

usually preferred since it does not increase the total weight of a single paper (Egghe et al., 2000). Egghe and Rousseau (1990) stated already "that the best way to handle multi-authored papers is to assign credit proportionally." Michael Schreiber proposed to modify the *h-index* by counting the papers fractionally according to (the inverse of) the number of authors, yielding the modified index hm (Schreiber, 2009). The same fractional counting of papers has been suggested by Egghe (2008) but the effect was relatively small, because of a large number of single-author papers in his data set. Schreiber's paper used obtained observations for data sets of more common average scientists.

The WoS allows an automatic arrangement of the publication lists in decreasing order according to the *number of citations* $c_{(r)}$, where *r* is the rank attributed to the paper. The *h-index* is readily available from this list as:

$$c_{(h)} \geq \boldsymbol{h} \geq c_{(h+1)}$$

according to Hirsch's original definition (Hirsch J.E., 2005).

2.5.2.5. The trend h-index

The original *h-index* does not take into account the year when an article acquired a particular citation, i.e., the "age" of each citation. Let's consider a researcher who contributed to the research community a number of really brilliant articles during the decade of 1960, which, say, got a lot of *citations*. This researcher will have a large *h-index* due to the works done in the past. If these articles are not cited anymore, it is an indication of an outdated topic or an outdated solution. On the other hand, if these articles continue to be cited, then we have the case of an influential mind, whose contributions continue to shape newer scientists' minds. There is also a second very important aspect in aging the *citations*. There is the potential of trendsetters, i.e., scientists whose work is considered pioneering and sets out a new line of research that currently is hot, thus this scientist's works are cited very frequently. To handle this, we take the opposite approach than the contemporary h-index's; instead of assigning to each scientist's article a decaying weight depending on its age,

we assign to each citation of an article an exponentially decaying weight, which is as a function of the "age" of the citation. This way, we aim at estimating the impact of a researcher's work in a particular time instance. It is of no interest how old the articles of a researcher are, but whether they still get *citations*. A researcher has an index equal to h_t (h trend) if h_t of its publications have a citation indicator *S (i, t)* whose total is greater than h_t, and other publications have a citation indicator *S (i, t)* less than or equal to h_t. We define an equation as follows:

$$\mathrm{S^t_{(i)}} = \boldsymbol{\gamma} \cdot \sum_{\forall \boldsymbol{x} \in \boldsymbol{C}(\boldsymbol{i})} (\mathrm{Y_{(now)}} - \boldsymbol{Y}_{(\boldsymbol{x})} + \mathbf{1})^{-\boldsymbol{\delta}}$$

Equation 4: Calculation of the citation indicator of the trend h-index

Where

Y(i) is the publication year of article *i* and

C(i) are the articles citing the article *i* .

If we set δ=1, then $S_t(i)$ is the *number of citations* that the article *i* has received, divided by the "age" of the article. Apparently, for $\gamma = \delta = 1$, the trend *h-index* coincides with the original *h-index*.

2.5.2.6. The tapered *h-index* (h_T)

An author's *h-index* cannot exceed his/her number of publications, and will usually be considerably less. Thus, in an unfair underestimation, the vast majority of even thousands of *citations* that accompany the most highly cited papers in reality contribute zero because each one of them only scores 1 towards the *h-index* score. Moreover, articles that have received many *citations* but sometimes fall just short of the number required to score for *h* (called "sleeping beauties" by Van Raan, (Raan, 2004)), also count for nothing; they are not reflected at all in the *h-index* score which in this case remains completely unaffected. Anderson, Hankin and Killworth (Anderson, et al., 2008), suggested that a bibliometric measure of publication output should be "strictly monotonic", which means that it should assign a positive score to each new citation as it occurs. At the very least, outstanding articles with numerous

citations should possess an accordingly increased index. The "tapered h-index" name was suggested by Prof. J.G. Shepherd for the new metric index.

Consider a scientist who has 5 publications which, sorted, have 6,4,4,2,1 *citations*. This publication output can be represented by a *Ferrers graph*[20], where each row represents a partition of the total 17 cites amongst papers (Figure below). The largest completed (filled in) square of points in the upper left hand corner of a *Ferrers graph* is called the *Durfee square*[21] (Andrews, 1984).

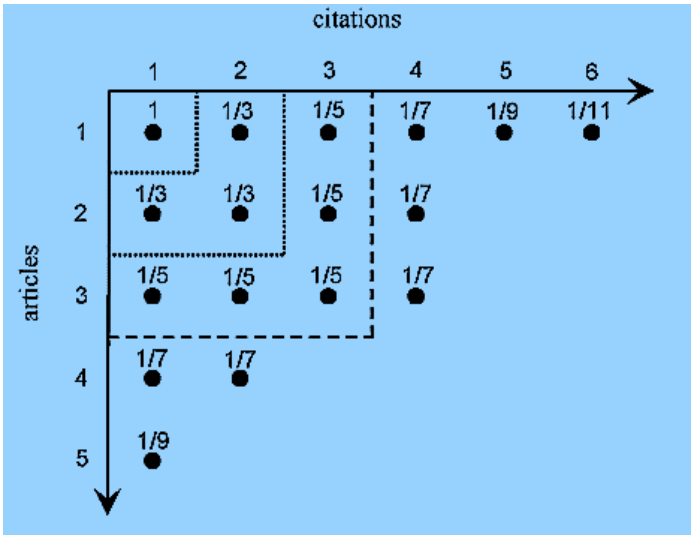


Figure 4: Example of a Ferrers diagram of an author's citations, in this case with 5 papers and a total of 6+4+4+2+1 = 17 citations, indicated in rows. The Durfee square is the 3-by-3 square indi-cated by a dashed line; this is the largest complete square in the Ferrers diagram. Citaon scores are shown according to the tapered h-index, h_T.

The *h-index* is equal to the length of the side of the Durfee square (in the case of Figure 4, h = 3), effectively assigning no credit (zero score) to all points that fall outside. Summing the relevant *citations*, scores of 1, 2, 3 are achieved for Durfee squares whose width is 1, 2, 3, matching the *h-index.* This notation immediately suggests a new index, h_T, which has the property that each additional citation increases the total score (the index has the property of "marginally increasing"), whether or not it lies within the *h-index* Durfee square. The score of any citation on a Ferrers graph is now given by

[20] See Appendix A.1.
[21] See Appendix A.2.

$$\boldsymbol{Citation\ score = \frac{1}{2L-1}}$$

Equation 5: Calculation of the citation score of the tapered h-index derived from the corresponding Ferrers diagram.

where *L* is the length of side of a Durfee square whose boundary includes the citation in question. The additional *citations* outside the Durfee square (of side 3) in Figure 4 above can now be scored, the five papers achieving scores of 1.88, 1.01, 0.74, 0.29 and 0.11, leading to a total score for h_T of 4.03.

The new bibliometric index, positively enumerates all *citations*, yet scoring them on an equitable basis with h. An advantage of this approach is that the scoring mechanism of h_T is on an equitable basis to that of h, permitting direct comparison of the two measures of output. The *h_T-index* is superior to *h-index*, both theoretically (it scores all *citations*), and because it shows smooth increases from year to year as compared with the irregular jumps seen in *h*. Conversely, the original *h-index* has the benefit of being conceptually easy to visualize. Qualitatively, the two indices show remarkable similarity (they are closely correlated), such that either can be applied with confidence. In mathematical terms, the most cited paper in a given list, with n_1 *citations*, generates a score, $h_T(1)$, of:

$$\boldsymbol{h_T(1) = \sum_{i=1}^{n_1} \frac{1}{2i-1} = \frac{ln(n_1)}{2} + o(1)}$$

Equation 6: Score, $h_T(1)$of the tapered h-index, for the most cited paper in a given list, with n_1 citations,

Where : $ln(n_1)$ is the natural logarithm of n_1[22]

$o(1)$ is a term for which applies the following:

$$\lim_{n_{1\to\infty}} o(1) = 0$$

The resulting $h_T(1)$ score as a function of n_1 (*citations* of most cited publication) is shown in Figure 5 below. If an author has *N* papers with associated *citations* n_1, n_2,

[22] Based on Euler's number (e ≈ 2.71828).

n_3, ..., n_N (ranked in descending order as in a Ferrers graph), the h_T score for any single paper ranked j in the list (with n_j *citations*), h_T(j), is:

$$(a)\ \boldsymbol{h_T(j) = \frac{n_j}{2j-1} \vdots n_j \leq j}, \qquad (b)\ \boldsymbol{h_T(j) = \frac{j}{2j-1} + \sum_{i=j-1}^{n_j} \frac{1}{2i-1} \vdots n_j > j}$$

Equation 7(a), 19(b): Calculation of the h_T score for any single paper ranked j in the list (with n_j citations).

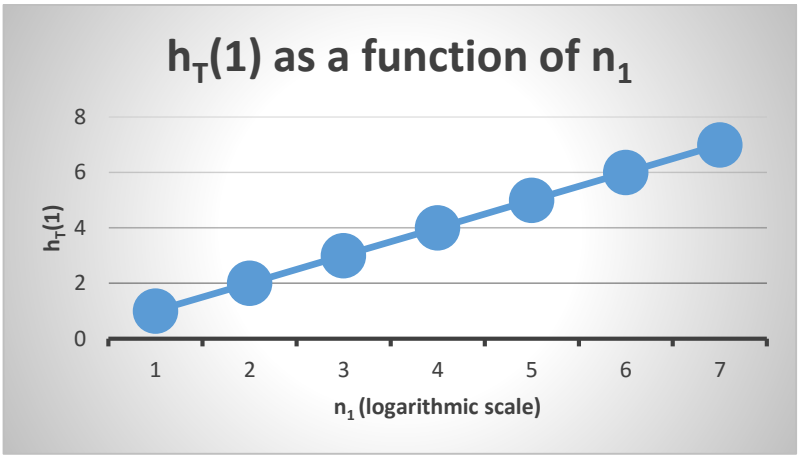


Figure 5: Tapered h-index of most cited paper hT(1) as a function of n1 in logarithmic scale i.e. 1=10, 2=100,.....6=1000000 etc.

2.5.2.7. The rational *h-index,* h_{rat}

Let *n* be the *number of citations* necessary for obtaining an *h-index* one higher. This number *n* is divided by the highest possible *n*, namely *2h+1*. Indeed, the lowest possible situation leading to a *h-index* of *h* consists of *h* articles with *h citations*, followed by an article without any citation. In order to get an *h-index* equal to *h+1* one more score for each of the first *h sources* is required, *h scores* in total, and *h+1 scores* for the last one: a total of *2h+1*. (Ruane & Toll, 2007)

For example, the ranking 3 – 3– 3 – 0 (articles ranked according to the *number of citations*) leads to *h* = 3. Its h_{rat} = 4 – 7/7 = 3. (Rousseau, 2008)

2.5.2.8. The dynamic h-index

Two scientists can have the same *h-index*, and even the same *R-index* but one's career can be on the rise (citation-wise), while the other one's is stagnating. A so-called dynamic index considering this aspect was proposed by Rousseau and Ye (Rousseau & Ye, 2008).

2.5.2.9. The $h_{I,annual}$

The individual, average annual increase of the *h-index* called $h_{I,annual}$ was proposed by Anne-Wil Harzing, Satu Alakangas and David Adams in their paper (Harzing, et al., 2014). The *average annual increase* in the individual *h-index* is useful for the following reasons:

- In common with the $h_{I,norm}$ index, it removes to a considerable extent any discipline-specific publication and citation patterns that otherwise distort the *h-index*.
- It also reduces the effect of career length and provides a fairer comparison between junior and senior researchers.

The $h_{I,annual}$ is meant as an indicator of an individual's average annual research impact, as opposed to the lifetime score that is given by the *h-index* or $h_{I,norm}$. (Harzing, 2013).

Discipline	Average h-index	Average # of authors per paper	Average academic age	Average hla-index
Humanities (n=19)	3.21[a]	1.90[a]	18.16[a]	0.14[a]
Social Sciences (n=24)	9.83[b]	2.62[ab]	19.54[a]	0.37[b]
Engineering (n=20)	12.50[b]	3.89[bc]	19.90[a]	0.34[b]
Sciences (n=44)	22.31[c]	4.66[cd]	29.36[b]	0.40[b]
Life Sciences (n=39)	23.95[c]	6.22[d]	25.69[b]	0.43[b]
F-statistic	33.894***	15.300***	10.427***	12.478***
Mean (SD)	16.92 (10.92)	4.33 (2.82)	23.86 (9.02)	0.36 (0.18)
Range	0-48	1.00-23.05	5-46	0.00-1.00
***Means with the same superscript are not significantly different at *p=0.05* (Tukey B-test, ***p<0.001**				

Table 1: The h-index compared with hI,annual index for different disciplines; source: (Harzing, et al., 2014)

2.5.3. The g-index

Proposed by Egghe (Egghe L., 2006), g is the number of articles whose total numbers of *citations* is at least g^2 (a *g-index* of 10 indicates that the author has written 10 papers which the sum of *citations* at least 100). Highly cited papers are important for the determination of the *h-index,* but once they are selected to belong to the top h papers, the *number of citations* they receive is rendered unimportant. This is a disadvantage of the *h-index, which* Egghe has tried to overcome through a new index, called *g-index*. "*Given a set of articles ranked in decreasing order of the number of citations that they received, the g-index is defined as the (unique) largest number such that the top g articles received (together) at least g2 citations*" (Egghe L., 2006). Another definition is given by Quesada (2010): "*The g-index is the maximum number g of papers by r such that the average number of citations of the g papers is at least g."* An easy way to determine the *g-index* is by calculating the *h-index* of the average citation count:

$$\boldsymbol{g} = \boldsymbol{h}_{(\boldsymbol{a}^{\boldsymbol{x}})}$$

Equation 8: Determination of the g-index.:

Where a^x is the vector of average *citations*.

By counting the average *citations* in the *h-core*, the *g-index* captures more of the impact of those highly-cited publications whose impact the *h-index* leaves out because they exceed h *citations*. It is therefore a variation of the *h-index.*

It can be calculated as follows:

$$\boldsymbol{g \geq \frac{1}{g} \cdot \sum_{i \leq g} c_i, \qquad g^2 \leq \sum_{i \leq g} c_i,}$$

Equation 9: Calculation of the g index.

Where c_i is a series of publications, denoted by their *number of citations*, in declining order.

2.5.3.1. Advantages & Disadvantages of the g-index

The *g-index* has its advantages and its disadvantages. One of the most important advantages of the *g-index* is that it accounts for the performance of the author's top articles and it helps to make the difference more apparent between the authors' respective impacts. The inflated values of the *g-index* help to give credit to lowly cited or non-cited papers while giving credit for highly-cited papers.

However, the *g-index* has been introduced in 2006 (one year after the introduction of the *h-index*) and it might not be as widely accepted as *h-index.* There's a lot of debate whether the *g-index* is superior to the *h-index* or not. Although, the *g-index* has a greater discriminatory power than the *h-index*, its discriminatory power is further enhanced by redefining the *g-index* as a rational (successive) number[23]. (Ruane & R.S.J. Tol, 2008)

Applying the *g-index* would also reveal that the *g-index* is more robust to the differences in domain size than is the *h-index.* (Vanclay, 2007)

[23] Rational or successive is a number that interpolates between *h* and *h + 1*. (Vanclay, 2007)

2.5.4. The e-index

The e-index, complementing the *h-index* for excess *citations* is the square root of the surplus of *citations* in the h-core beyond h^2. One of the aims of the *e-index* is to differentiate between scientists with identical h-indices but different *citations*. Another advantage of the *e-index* is that it can reflect the contributions of highly cited papers of an author, as usually ignored by the *h-index.* Zhang (Zhang, 2009) believes that the *e-index "…is a necessary h-index complement, especially for evaluating highly cited scientists or for precisely comparing the scientific output of a group of scientists having an identical h-index."*

2.5.4.1. Loss of citation information by the g-index

The *e-index proposed* here is aimed at considering the contributions of excess *citations*, which are mainly from highly cited papers. It is necessary to mention the *g-index*, which was proposed as being *"…sensitive to the level of the highly cited papers…"* (Egghe L , 2006). The *g-index* is defined as *"…the highest number of g of papers that together received g^2 or more citations.".* Although having some advantages, the *g-index* also suffers from the loss of citation information in many important cases, especially for distinguished scientists (most of whose papers are highly cited).

For instance, for any *k*, if

$$\sum_{j=1}^{k} \boldsymbol{c_j} > \boldsymbol{N^2} \text{ with } \boldsymbol{k} = \boldsymbol{1, 2, 3, .., N} \text{ (hypothesis 1),}$$

then the *g-index* has no definition. In fact, for any *N* conditions in hyp. 1, the *g-index* can have no definition.

Among the *N* conditions in hyp. 1, the strongest condition is: $\boldsymbol{c_1} > \boldsymbol{N^2}$, and the weakest condition is $\sum_{j=1}^{k} \boldsymbol{c_j} > \boldsymbol{N^2}$**.** (Zhang, 2009)

2.5.4.2. Definitions of the e-index

If all of a researcher's papers have at least *h citations* (Rousseau R, 2006), using the *h-index*, the only citation information that can be inferred is that at least h^2 *citations* have been received and additional *citations* for papers in the h-core are completely ignored. Here we define the *e-index* to complement the *h-index* for the ignored ex-

cess *citations*. The excess *citations* received by all papers in the h-core, denoted by e^2 are calculated as follows:

$$\boldsymbol{e^2 = \sum_{j=1}^{h} (c_j - h) = \sum_{j=1}^{h} c_j - h^2}$$

Equation 10: Calculation of the excess citations received by all papers in the h-core.

where c_j are the *citations* received by the *jth* paper and denotes the excess *citations* within the h-core. Then, we define d^2 as follows:

$$\boldsymbol{d^2 = \sum_{j=1}^{h} c_j}$$

Equation 11: Definition of d^2

Combining Eq. 22 and Eq. 23 we have: $\boldsymbol{d^2 = h^2 + e^2}$ and consequently we formulate:

$$\boldsymbol{e = \sqrt{d^2 - h^2}}$$

Equation 12: Calculation of the e index

Therefore, it can be seen that because of the loss of citation information, comparisons based on the *h-index* alone can be misleading when the ignored excess *citations* (e^2) are multiple times more compared to the h^2 *citations*. This means that for accurate and fair comparisons, it is necessary to use the *e-index* together with the *h-index.*

Other h-type indices like the *a-index* (Jin, 2006) and the *R-index* (Jin, et al., 2007) which are h-dependent, have information redundancy with *h*, and therefore, when used together with *h*, mask the real differences in excess *citations* of different researchers. Therefore, the *e-index* is a necessary *h-index* complement, especially for evaluating highly cited scientists or for precisely comparing the scientific output of a group of scientists having an identical *h-index* (Zhang, 2009) (Zhang, 2012).

2.5.5. The i10 index

The *i10-index* indicates the number of papers an author has written that have been cited at least ten times by other scholars. It was introduced by Google in 2011 as part of their work on *Google scholar*, a search tool that locates academic and related papers. Due to some of the problems with inaccurate counts, Google's *i10 index* has come under close scrutiny and criticism. It was created by *Google Scholar* and used in *Google's My Citations* feature.

i10-Index = the number of publications with at least 10 *citations*.

This very simple measure is only used by Google Scholar, and is another way to help gauge the productivity of a scholar.

Advantages of the *i10-Index*

- It is simple and straightforward to calculate;
- My *Citations* in Google Scholar is free and easy to use.

Its main disadvantage is that It is used exclusively only in Google scholar.

2.5.6. The hg-index

The *hg-index,* is a measure to characterize the scientific output of researchers which is based on both *h-index* and *g-index* to try to keep the advantages of both measures as well as to minimize their disadvantages. The *hg-index* fuses both measures in order to obtain a more balanced view of the scientific production of researchers and minimizes some of the problems that they present. (Alonso, et al., 2008)

The *hg-index* of a researcher is computed as the geometric mean of his *h*- and *g* indices, that is:

$$\boldsymbol{hg} = \sqrt{\boldsymbol{h} \cdot \boldsymbol{g}}$$

Equation 13: Calculation of the hg-index

The *hg-index* has some special features. It is, in some point, obvious that $h \leq hg \leq g$ and that $hg - h \leq g - hg$, indicating that the *hg-index* corresponds to a value nearer to *h* than to *g*. This property can be implied as a penalization of the *g-index* in the cases of a very low *h-index,* thus avoiding the problem of the big influence that a

very successful paper can introduce in the *g-index* (Alonso, 2010). It is interesting, that the *hg-index* can be interpreted in terms of geometry as the square root of the area of the rectangle with side lengths h and g. Ending, *hg-index* also introduces greater granularity than does the *h*- or *g-index* individually.

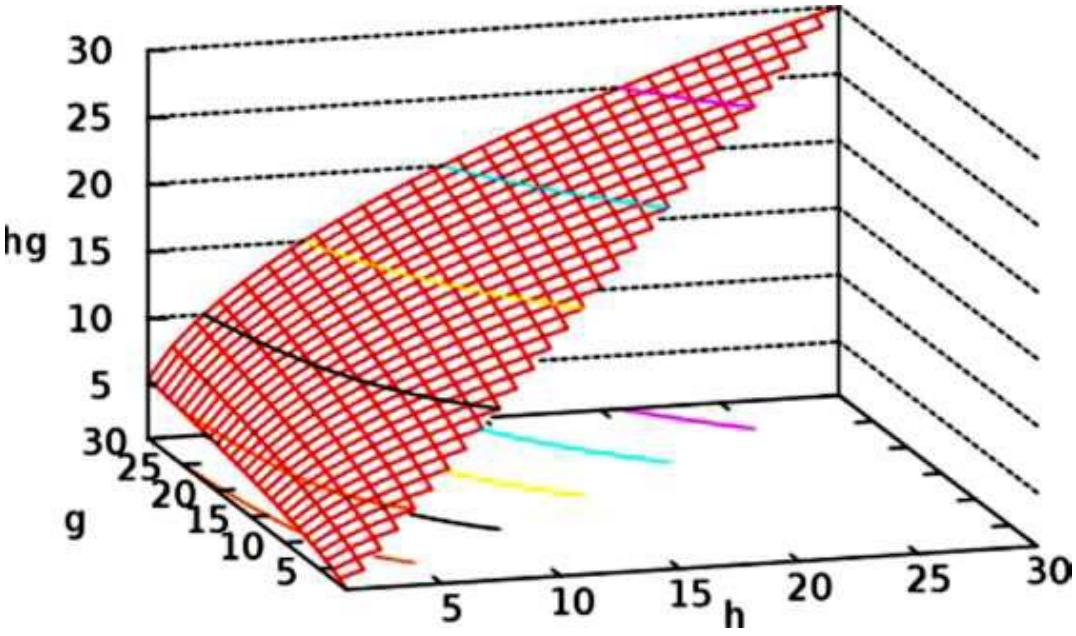


Figure 6: An example that shows the growth of hg index as function of h & g (Alonso, 2010)

2.5.7. Application of Pareto's Principle on *citations* of scientific papers

As first observed in 1965 by Price (Price, 1965), the numbers of *citations* received by scientific papers appear to have a power-law distribution. For example, the figure taken from the Science Citation Index shows the cumulative distribution of the *number of citations* received by a paper between its publication in 1981 and June 1997. (Redner, 1998).

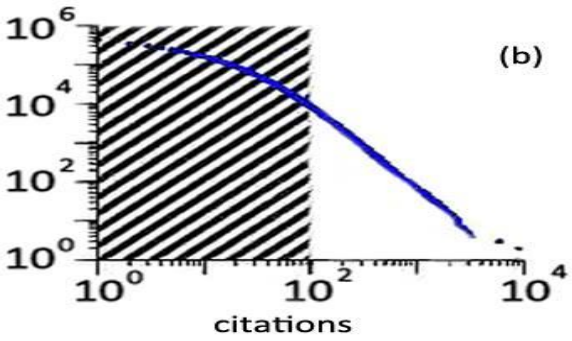


Figure 7: Cumulative distribution of the number of citations received by a paper between its publication in 1981 and June 1997. (Redner, 1998)

Suppose we have a system composed of a collection of objects, such as genera, cities, papers, web pages and so forth. Let us assume the objects to be papers of scientific value. New objects appear occasionally, as people publish new papers. Each object also has some property k associated with it, the *number of citations* to a paper, that is reputed to obey a power law, and it is this power law that we wish to explain.[24]

Newly appearing objects have some initial value of k which we will denote k_0. The value of k_0 can be zero in some cases: for instance, newly published papers usually have zero *citations*. In between the appearance of one object and the next, m new *citations* are added to the entire system. That is some papers will get new *citations*, but not necessarily all. In the simplest case, these are added to objects in proportion to the number that the object already has. Thus, *the probability of a paper getting a new citation is proportional to the number it already has*. In many cases, this seems

[24] See Appendix A.3.

like a natural process. For example, a paper that already has many *citations* is more likely to be discovered during a literature search and hence more likely to be cited again. Simon (Simon, 1955) named this type of process *the Gibrat principle*. It also appears with the names of *the Matthew effect* (Merton, 1968), *cumulative advantage* (Price, 1976), or *preferential attachment* (Barabási & Albert, 1999).

There is a problem however when $k_0 = 0$. If new papers appear with no *citations* and are supposed to collect *citations* in proportion to the number, they currently have, which is zero, and then the paper will never get any *citations*. To overcome this problem, new *citations* are assigned not in proportion simply to k, but to $k + c$, where c is a constant. Thus, there are three parameters k_0, c and m that control the behavior of the model. Real *citations* seem to have an exponent $\alpha \simeq 3$, so we should expect $c \simeq m$.

For *citations* of papers that have $k_0 = 0$ we must have $c > 0$ to get any *citations* or links at all. So

$$\boldsymbol{a = 2 + \frac{c}{m}} \quad (25)$$

Equation 14: Relationship among the three parameters that control the model of the Gibrat principle, leading to the formation of a Yule distribution

Many processes with different values of the three parameters have been proposed. Yule (Yule, 1925) and later Simon (Simon, 1955) showed mathematically that this mechanism produces what is now called *the Yule distribution*, which follows a power law[26] in its tail. The process proposed by Yule is a general mechanism that can explain a number of the power-law distributions observed in nature and can produce a wide range of exponents to match the observations by suitable adjustments of the parameters. Especially for *citations* it is now, the most widely accepted theory. (Newman, 2005).

[25] Remember that $\boldsymbol{a}$ must range between 2 and 3 and $\boldsymbol{c \simeq m}$.
[26] See Appendix A.3.

2.5.8. Other indices

More sophisticated indices, always based on *citations*, have been proposed to account for three possible biases:

- *Year of publication*: older publications are cited more.
- *Document type*: the *number of citations* varies considerably this type, for example, items of "review" are generally more cited as scientific articles.
- *The field*: publication practices differ significantly between global scientific fields.

Leading indices are based on a normalization process, which attempts to correct these biases. The best-known leading indices are:

- The field normalized citation score.
- *The crown index* that compares the average *number of citations* attributed to a "unit" (researcher, laboratory) with the *average number of citations* in international publications of the same year in the same field and the same type of document. E.g. a *crown index* of 0.9 indicates that the analyzed publications are cited 10% less than average.
- *The Top 5%* calculated for a group of authors, the share that is within 5% of the most cited papers in the world that year, in the same field for the same document type. A value greater than 1 indicates that the group has more publications in the group of 5% of the most cited publications than the world average.

The obvious desire to take greater account of the specific scientific fields to produce these leading indicators faces several difficulties:

- The definition of the area is prone to a lot of subjectivity (to which required level of granularity should one descend to truly reflect an individual field?)
- Data processing can only be manual (e.g. a newspaper can cover several areas), so it is necessary to sort the *citations*.

- The significant inaccuracy of citation sources occurs at two levels, on the calculation of the global average and the *citations* of the group considered.

The single use of *citations* to evaluate the scientific quality of an article and its impact raises many questions in the scientific community, the one that is obviously the most aware of the biases that this methodology can create. This question has prompted initiatives to propose other models. For example, in biology and medicine *the Faculty of 1000* service provided by *BioMed Central*, provides analysis of an article based on the cooperative reading of articles by a group of experts co-opted in a specific field.

Other averaging procedures have also been suggested, yielding the *t index* for the geometric mean and the *f index* for the harmonic mean (Tol, 2009), thus giving more weight to highly cited papers. Another complementary index, the *m-index* has been defined as the median of the *number of citations* in the h core (Bornmann, et al., 2008). Several Hirsch-type indices have been proposed based correspondingly on the square root of the *total number of citations* to the papers in the core. Again the size of the core could be *h* yielding once more a complementary index called *R* [27] (Jin, et al., 2007).

Using the g core reproduces the *g-index* according to Egghe's original definition (Egghe L., 2006). For the definition of the *h index* (Miller, 2006) the total number of publications is used as core; for the weighted *h-index,* h_w (Egghe & Rousseau, 2007) a subset of the h core is taken into account. The *e-index* [28] quantifies the square root of the excess *citations* to papers in the *h* core (Zhang, 2012).

Further variants, which are not based on one of the above classifications, have been suggested. E.g., the tapered *h-index,* h_T, [29] takes the *citations* to all publications into account in a complicated way (Anderson, et al., 2008). The *π-index* depends on the *total number of citations* to papers in the so-called elite set (Vinkler, 2009). Finally,

[27] See 2.5.4.2, p.51
[28] See The e-index
[29] See 2.5.2.6, p.42

the maxprod index was proposed to distinguish geniuses and hard workers from the typical researcher (Kosmulski, 2007).

When choosing to evaluate one of these indices, consideration must be given to the fact that the quality of the database is the most important criterion. The "*distinct author*"[30] is a significant feature now present in the Web of Science but it is practically almost impossible to establish the citation data of an individual scientist with high accuracy due to the huge amount of time and work it involves. In actual applications, precision is a formidable problem. Thus, the usefulness of these *rankings* is an ongoing controversial matter and scientists (who are the ones directly affected), bearing in mind the potential gain and profit associated, are strongly skeptical. (Schreiber, 2010)

2.6. General overview of selected variants and extensions of the h-index

In the table below we can see a comprehensive overview of the selected variants and extensions of the *h-index* and their basic details.

Type	Index name	Calculation	Author(s)	Notes
Variants	h	$h\,(f) = \max \min(f(i), i)$	Hirsch (2005)	Simple to calculate, objectively derived from popular DBs, robust to poorly-cited "tail". But ignores impact of important papers cited more frequently than *h* times.
	g	$g \leq \frac{1}{g}\sum_{i \leq g} Ci$	Egghe (2006)	Gives more weight to highly-cited papers but is not robust to influence of outliers.
	hg	$\sqrt{h \cdot g}$	Alonso et al.	The hg-index of a researcher is computed as the geometric mean

[30] 24 The Distinct Author Sets page lists sets of articles likely written by the author identified at the top of the page. All sets are created by the Distinct Author Identification System. Factors such as author name and citation data are combined to create clusters of articles likely written by an author. (Web of Science Help Distinct Author Sets: [Author Name])

				(2009)	of his h- and g-indices
	h_a		$h_a = \max(C_i \geq a \cdot i)$	Eck & Waltman (2008)	A generalized form of the *h-index.*
	A		$A = \frac{1}{h}\sum_{j=3}^{h} cit_j$	Jin (2006)	The a-index (as well as the m-index, r-index, and ar-index) includes in the calculation only papers that are in the Hirsch core.
	R		$R = \sqrt{\sum_{j=1}^{h} cit_j}$	Jin (2007)	The R- and AR-indices: Complementing the h-index
	m		$m = \frac{h}{y}$	Bomman et al. (2008)	The *m*-index is defined as *h*/*n*, where *n* is the number of years since the first published paper of the scientist
	h(2)		$h(2) \leq \frac{1}{h}\sum_{i \leq h} Ci$	Kosmulski (2006)	h(2)-index also gives more weight to highly cited articles.
	e		$e^2 = \sum_{j=1}^{h} cit_j - h^2$	Zhang (2009)	Independent yet complementary to the *h-index.* Useful for evaluating highly-cited researchers of differentiating between researchers with the same *h-index*.
	normalized h		$h^n = \frac{h}{N_p}$	Sidiropoulos et al. (2007)	Generalized Hirsch h-index for disclosing latent facts in citation
	tapered h		$H_{t(1)} = \sum_{i=1}^{n1} \frac{1}{2i-1} = \ln(n1)/2 + o(1)$	Anderson et al. (2008)	Takes all citations into account.
	rational h		$h_{rat} = (h+1) - \frac{n_c}{2 \cdot h + 1}$	Ruane & Toi (2008)	Increases in smaller step w than h: has more granularity.
Extensions	h_g		$h_g = \frac{n}{h\,(f)}$ n = citation impact	Egghe & Rousseau (2008)	The *h-value weighted by citation impact.*
	Temporal	m-quotient	$m = \frac{h}{y}$	Hirsch (2005)	Same as *m-index*.
		contemporary h	$S^c(i) = \gamma * (Y(now) - Y(i) + 1)^{-\delta} * \lvert C_{(i)} \rvert$	Sidiropoulos et	Older articles have less 'weight'. Identifies promising new re-

				al.	searchers
		trend h	$S^{t}(i) = \gamma * \sum_{\forall x \in C_{(i)}} (Y_{(now)} - Y_{(x)} + 1)^{-\delta}$	(2007)	Emphasizes recent citations, identifying researchers who are 'hot' now, even if their articles were are old.
		dynamic h-type	$R(T) * V_h(T)$	Rousseau & Ye (2008)	Tries to differentiate static vs. increasing *h-indexes*.
	Individuals	h_i	$h(f) = \frac{\max \min(f(i), i)}{n}$ n = the average number of authors	Batista et al. (2006); Imperial & Rodriguez-Navaro (2007)	Divides the standard *h-index* by the average number of authors in the articles that contribute to the *h-index,* in order to reduce the effects of co-authorship.
		h_m	$max_r(r \leq c(r))$	Schreiber (2008)	Uses fractional paper counts instead of reduced citations counts to account for shared authorship of papers. Determines the multi-authored h_m index based on the resulting effective rank of the papers using undiluted citations.
		fractional counting of papers	$max_r(r \leq c(r))$	Egghe (2008)	Same as h_m-index.

Table 2: Selected variants and extensions of the h-index (Demaine, 2011-2012)

3. Materials - Libraries and Databases

3.1. Medline

MEDLINE is the U.S. National Library of Medicine® (NLM) premier bibliographic database that contains more than 22 million references to journal articles in life sciences with a concentration on biomedicine. A distinctive feature of MEDLINE is that the records are indexed with NLM Medical Subject Headings (MeSH®). MEDLINE is the online counterpart to MEDLARS® (MEDical Literature Analysis and Retrieval System) that originated in 1964. The great majority of journals are selected for MEDLINE based on the recommendation of the Literature Selection Technical Review Committee (LSTRC), an NIH-chartered advisory committee of external experts analogous to the committees that review NIH grant applications. Some additional journals and newsletters are selected based on NLM-initiated reviews, e.g., history of medicine, health services research, AIDS, toxicology and environmental health, molecular biology, and complementary medicine, that are special priorities for NLM or other NIH components. These reviews generally also involve consultation with an array of NIH and outside experts or, in some cases, external organizations with which NLM has or had special collaborative arrangements. MEDLINE is the primary component of *PubMed*®, part of the *Entrez* series of databases provided by the NLM National Center for Biotechnology Information (NCBI).

3.1.1. Time coverage and Sources

Generally operating from 1946 to the present, with some older material. *Citations* from more than 5,600 worldwide journals in about 40 languages; about 60 languages for older journals. *Citations* for MEDLINE are created by the NLM, international partners, and collaborating organizations. (U.S. National Library of Medicine, 2015)

3.2. Google Scholar

Google Scholar uses the popular Google search engine to enable searches for scholarly materials such as peer-reviewed papers, theses, books, preprints, abstracts and technical reports from broad areas of research. It includes a variety of academic publishers, professional societies, preprint repositories and universities, as well as scholarly articles available across the web. Google Scholar includes full text and *citations*. Some links to full text ask for payment.

Google Scholar is a subset of the larger Google search index, consisting of full-text journal articles, technical reports, preprints, theses, books, and other documents, including selected Web pages that are deemed to be "*scholarly*." Although Google Scholar covers a great range of topical areas, it appears to be strongest in the sciences, particularly medicine, and secondarily in the social sciences. The company claims to have full-text content from all major publishers except *Elsevier* and the *American Chemical Society*, as well as hosting services such as *Highwire* and *Ingenta*.

Much of Google Scholar's index derives from a crawl[31] of full-text journal content provided by both commercial and open source publishers. Specialized bibliographic databases like *OCLC's Open WorldCat* and *the National Library of Medicine's PubMed* are also crawled. Since 2003, Google has entered into numerous individual agreements with publishers to index full-text content not otherwise accessible via the open Web. Google Scholar is fast and easy to search. It retrieves document or page matches based on the keywords searched and then organizes the results using a closely guarded relevance algorithm[32]. Because so much of the content of Google Scholar's index comes from licensed commercial journal content, most users will discover that clicking on a link in Google Scholar's search results may reveal only an abstract—not full text—accompanied by a pay-per-view option. Institutions can configure OpenURL link resolvers, such as *SFX*, to authenticate users to provide access to full-text content that is available through institutional subscriptions.

[31] Automatic searching and indexing by computer search engine software.
[32] A sorting algorithm based on the comparison of the relevance of the searched articles and the key words.

3.2.1. Disadvantages

The inadequacies of Google Scholar have already been well documented in reviews. These reviews focused on three major weaknesses of the tool:

- lack of sufficient advanced search features,
- lack of transparency of the database content, and
- uneven coverage of the database.

Henderson's review of Google Scholar demonstrated its significant limitations for clinician use (Henderson J., 2005). Tests conducted by Jacso (Jacso P., 2005) showed that Google Scholar typically crawled only a subset of the full available content of individual journals or databases. In February 2005 (Vine, 2006), Vine discovered that Google Scholar was almost a full year behind indexing PubMed records and concluded that "*no serious researcher interested in current medical information or practice excellence should rely on Google Scholar for up to date information"*. (Vine, 2006)

With a simple, basic search interface and only minimal advanced search features, Google Scholar lacks almost every important feature of MEDLINE.

- It does not map to Medical Subject Headings (MeSH)[33];
- does not permit nested Boolean searching;
- lacks essential features like explosions, subheadings, or publication-type limits; and
- Offers searchers no ability to benefit from the extraordinary indexing that the National Library of Medicine provides.

[33] MeSH (Medical Subject Headings) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed.

3.3. Scopus

The Scopus database provides access to STM[34] journal articles and the references included in those articles, allowing the searcher to search both forward and backward in time. The database can be used for collection development as well as for research. Scopus is an abstract and indexing database with full-text links that is produced by the Elsevier Co. The name, Scopus, was inspired by the bird, Hammerkop (Scopus umbretta), which reportedly has excellent navigation skills. The database, in development for two years, was developed working with 21 research institutions and more than 300 researchers and librarians. The verbal and behavioral feedback of these librarians and researches was analyzed and used to improve the product.

3.3.1. Content of Scopus

Scopus developers claim to index over 14,000 STM and social science titles from 4000 publishers, stating that it is the "*largest single abstract and indexing database ever built*" (Elsevier, 2015). The database claims 4600 health science titles are indexed including 100% MEDLINE coverage, 100% of EMBASE[35] coverage and 100% of Compendex[36] coverage. The list of titles indexed is selected based on user demand and market research. It contains 27 million abstracts with *citations* back to 1966. In addition to American journals, it includes European and Asia Pacific literature in both English and non-English. Indexing includes CAS[37] registry numbers, MeSH terms, EMTREE[38] terms and supplemental key terms added by indexers.

Some features of Scopus include:

- Links to both citing and cited documents, allowing the user to go both forwards and backwards in time.

[34] The International Association of Scientific, Technical and Medical Publishers (STM) is an international trade association organized and run for the benefit of scholarly, scientific, technical, medical and professional publishers

[35] Embase is a comprehensive biomedical database.

[36] Compendex is a comprehensive bibliographic database of engineering research, containing over 10 million records taken from over 5,000 engineering journals, conferences, and technical reports.

[37] A CAS Registry Number,[1] also referred to as CASRN or CAS Number, is a unique numerical identifier assigned by Chemical Abstracts Service (CAS) to every chemical substance described in the open scientific literature (currently including those described from at least 1957 through the present).

[38] Emtree is Elsevier's life science thesaurus, for full-text indexing of journal articles.

- Provides open access titles, which are included in the index.
- Indexes web pages and patents, with a claim to over 167 million relevant web pages.
- It is *OpenURL compliant* and works with any link resolver, using image-based linking.
- Runs an entitlement check prior to returning a full-text image if the article is available to the user.
- Can link to the publisher's web site to view the document.
- Developers claim *"citation accuracy is achieved by using state-of-the-art technology, with 99% of citing references and citing articles matched exactly.*"

(Burnham, 2006) (Chadegani, et al., 2013)

Scopus vs. Web of Science		
Features	**Scopus**	**Web of Science**
Noumber of Journals	18.000	12.000
Focus	Physical sciences, health sciences, life sciences, social sciences	Science, technology, social science, arts and humanities
Period Covered	1966-	1990-
Databases covered	100% Midline, Embase and more	Science Citation, Social Sciences Citations, Arts & Humanities Citation Indexes
Updated	Daily	weekly
Developer/Producer	Elsevier	Thomson Reuters
Citation Analysis	Yes	Yes
Controlled vocabulary	Yes-Index Terms field	No
Export feature	Yes	Yes
Alerts service	Yes	Yes
Strenghts	• More versatile search tool with advantages in functionality (default, refine, format of resultw of citation tracker and author identification • Covers 6256 unique journals, compared to WOS' 1467	• Greater time period of coverage • More options for citation analysis for institutions • Covers science and arts/humanities

	• Greater international coverage • Can use "first author" as a search field in Advanced Search • Can search with controlled vocabulary	
Weaknesses	Social science coverage, esp. sociology and prior to 1966	No controlled vocabulary

Table 3: Scopus vs. Web of Science (HLWIKI International, 2015)

3.4. The ISI Web of Knowledge

The ISI Web of Knowledge is an integrated, Web based platform designed to support all levels of scientific and scholarly research within academic, corporate, government or non-profit environments. It combines high quality, evaluated content with the tools needed to use, analyze and manage that content. The ISI Web of Knowledge platform provides a single, unified environment through which researchers can search and access different types of information, such as journal articles, proceedings papers, patents, chemical reactions and compounds, as well as web content. The content found within the platform is multidisciplinary, ensuring that scholars are not restricted along subject-specific lines. This becomes particularly important as research becomes more and more interdisciplinary in nature.

The *Thomson Reuters* core content covers over 16,000 international journals, books and proceedings in the sciences, social sciences and arts and humanities. The 10,400 international journals covered on an annual basis in the Web of Science are an important part of this data. Web of Science covers over 250 categories in every area of the sciences, social sciences and arts & humanities.

3.4.1. ISI Web of Knowledge Resources

3.4.1.1. Web of Science

The Web of Science provides access to current and retrospective multidisciplinary information from more than 10,400 of the most prestigious, high impact research journals in the world in the sciences, social sciences and arts and humanities – with

coverage back to 1900 (sciences), 1956 (social sciences) and 1975 (arts & humanities).

3.4.1.2. Biosis Previews / Biological Abstracts

BIOSIS Previews and Biological Abstracts contain bibliographic data & abstract text for research published in the life sciences. BIOSIS Previews and Biological Abstracts is the world's most comprehensive reference database for life science research covering over 5,500 journals. Additionally, it covers over 1500 conferences, 20,000 US Patents and over 10,000 Reviews and Monographs – all concentrated on the interdisciplinary life sciences.

3.4.1.1. ISI Proceedings

ISI Proceedings is available in two editions – Science Edition and Social Sciences and Arts & Humanities Edition. Both are multidisciplinary resources providing web access to bibliographic information, cited references and author abstracts from papers delivered at the top 2,600 international conferences, symposia, seminars, colloquia, workshops and conventions. Over 225,000 conference papers are indexed annually with data covering 1990–present.

3.4.1.2. Current Contents Connect (CCC)

Current Contents Connect is designed to be the quintessential current awareness product – an environment for the researcher whose primary goal is to be updated quickly and regularly on current research activity in a given field. The goal is to provide easy access to answer the question "What's new?" knowing that the researcher usually does n to spend much time in pursuit of the answer .CCC coverage includes:

- 8,000 international journals in seven different editions.
- 5,000+ evaluated scholarly and research-oriented websites brows able through Current Web Contents; subject specialist editors to ensure quality evaluate each website; annotations of each site are provided.
- 2,000+ books/books-in-series indexed at the chapter level.

3.4.1.3. Derwent Innovations Index

The Derwent Innovations Index opens the power of patent and citation searching combining value-added patent records from Derwent World Patents Index with patent citation information from the Derwent Patents Citation Index. It covers patents from over 42 international issuing authorities and includes descriptive titles and abstracts written by subject specialists. It includes data back to 1963 with over 16 million inventions and 33 million patent records in all technologies.

3.4.1.4. Web Citation Index

The Web Citation Index integrates scholarly web documents from over 500 evaluated institutional repositories. The content includes scholarly articles, preprints, theses, dissertations, proceedings, technical reports and other grey literatures. WCI combines cited reference indexing with automated indexing technology to produce a multidisciplinary index to web documents. All documents are full text searchable and enabled with full text links. (Center for Research Libraries, 2015)

3.5. Review of materials

Researchers turn to citation tracking to find the most influential articles for a particular topic and to see how often their own published papers are cited. For years, researchers looking for this type of information had only one resource to consult: the Web of Science from Thomson Scientific. In 2004 two competitors emerged – Scopus from Elsevier and Google Scholar from Google. Different scholarly publication coverage provided by the three search tools will lead to different citation counts from each. For many years Web of Science had a virtual monopoly on the provision of citedness tracking.

Late in 2004 two competitors to Web of Science emerged – Google Scholar and Scopus. The Internet search giant Google sponsored the creation of Google Scholar, a tool that attempts to give users a simple way to broadly search the scholarly literature. Google Scholar uses a matching algorithm to look for keyword search terms in the title, abstract or full text of an article from multiple publishers and web sites (Google Scholar does not share the specifics of how this algorithm works). The number of times a journal article, book chapter, or web site is cited also plays an im-

portant part in Google Scholar's ranking algorithm. Search results are displayed so that the more cited and highly relevant articles rise to the top of the set. This varies from the more traditional default "reverse chronological" order employed by most scholarly databases. Google Scholar neither lists the journal titles it includes, nor the dates of coverage; although they have indicated that they have agreements with most major publishers (except Elsevier). Another area of difference for Google Scholar is that unlike most scholarly research databases, it looks beyond journal literature to cover other modes of scholarly communication. Other sources covered in Google Scholar include preprint servers such as arXiv (physics), government, and academic Web sites. Google Scholar does not state how a Web site qualifies for inclusion in its searches.

At approximately the same time that Google Scholar was made public, Elsevier introduced Scopus, an indexing and abstracting service that contains its own citation-tracking tool. Scopus indexes a larger number of journals than Web of Science, and includes more international and open access journals. Citation coverage however only dates to 1996 (abstracts, but not citation coverage, are available back to 1966 for some journals.) Scopus includes its own Web search engine, *Scirus*. Scirus results are presented separately from other Scopus journal results. Also, material from

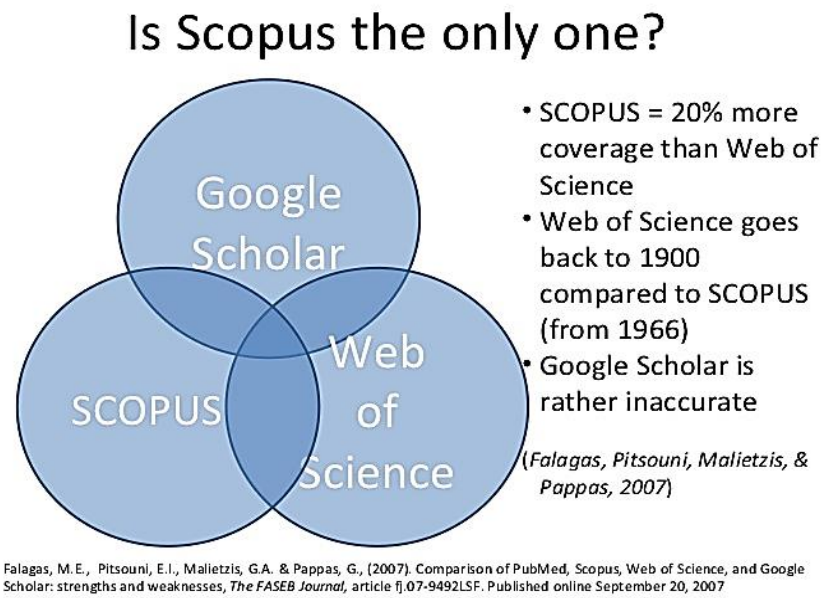


Figure 8: Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weakness (Falagas, et al., 2008)

Scirus does not figure into citation counts for Scopus journal records. (Bakkalbasi, et al., 2006) (Falagas, et al., 2008)

Each of these databases uses unique methods to record and count *citations*. The scope of these databases also (Falagas, et al., 2008) in that Web of Science and Scopus claim strong coverage of selected peer-reviewed journals, while Google Scholar might be better able to record *citations* from books and nontraditional sources, such as Web sites, dissertations, and open-access online journals. Any one of these three resources is not the unique answer to all citation tracking needs. Scopus shows strength in providing citing literature for current articles mainly, while Web of Science produces more citing material for older articles. All three tools provide some unique material. The question of which tool provides the most complete set of citing literature may depend on the subject and publication year of a given article.

Previous studies in some scientific fields, such as computing, biology, physics, and oncology (Harzing AWK, van der Wal R., 2008) (Kousha K, Thelwall M., 2008), have shown differences in citation counts among these databases. Differences in citation counts among the databases could have implications for citation analysis studies and in the use of citation counts for academic advancement decisions. If, however, the results across the databases are similar, then other features of the database, including cost and ease of use, may dictate preference. The Web of Science has long defined the standard for determining which *citations* are counted. The Web of Science claims as one of its strengths the selection process for only including certain journals in its content coverage. A description of the Web of Science Web site (Thomson Reuters., 2015) refers to *Bradford's Law*, first proposed in 1934, that states that the bulk of important scientific findings are reported in only a small number of journals. Therefore, the Web of Science emphasizes the quality of its content coverage, rather than the quantity. This scope of coverage, however, has been criticized for favoring North American–based, English-language journals (Meho, 2007) and for not fully covering other citation sources, such as books. Other citation databases offer alternative approaches to counting *citations*. Scopus, for example, covers more journals (approximately 15 000 peer-reviewed journals vs. 10 000 for Web of Science) with greater relative coverage of non–North American sources. Scopus claims that more

than half of its content originates from Europe, Latin America, and the Asia-Pacific region. Scopus also covers conference proceedings (which Web of Science also covers), trade publications, books, and several Web sources. Unlike Web of Science, however, whose content extends to1900, Scopus is limited in its coverage of older publications, especially those before 1996 (Falagas, et al., 2008). The automated, Web-based Google Scholar appears to include coverage of nontraditional online documents, including university theses and non–peer-reviewed Web sites. Google Scholar has been criticized, (Jacso P., 2005) in part for including *citations* from what many would consider non-scholarly sources, such as student handbooks and administrative notes. (Noruzi A. , 2005)

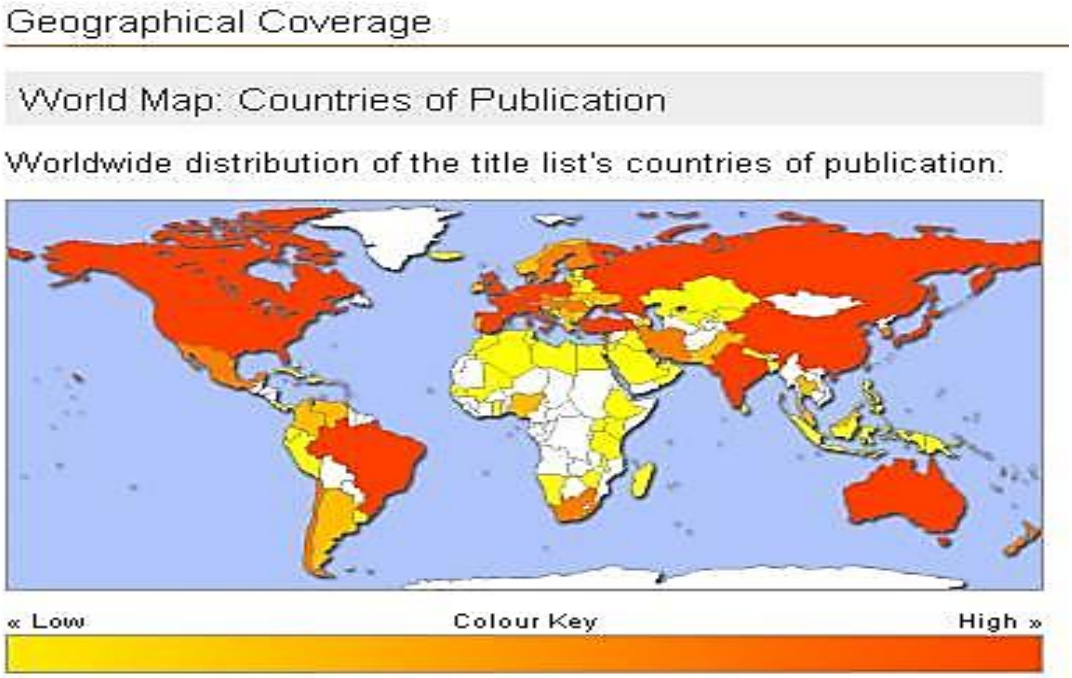


Figure 9: Scopus coverage map (Whitman, 2011)

4. Conclusion

The aim of bibliometric analysis is to record and provide reliable data, which, if positioned in a wider environment of indicators, are an important source of information for research. The assessment and interpretation of the indicators should take into account the constraints inherent in all bibliometric analyses.

An important fact that should be taken into account for the understanding and appreciation of the results is the number of publications as well as the systematic production of impact indicators such as coefficients of variation, the relevant impact indicators, the distribution and publications rates with specific characteristics etc.

The aim of bibliometric analysis is recording not only the overall trends but also scientific publications of exceptional performance, even if in some cases it is individual. In this direction, the study presents a wide range of indicators, through the combination of which a fuller picture of the research output is provided.

Bibliometric analysis involves the recording and processing of data related to scientific publications (e.g. the number of publications as well as the references to them, broken down by author, institution, scientific field, etc.) and thus:

- Reveals the characteristics of the research activity
- Identifies trends in research production to a body level
- Assesses the impact of scientific work
- Locates national and multinational networks between disciplines

it is important to emphasize that the indicators obtained from bibliometric databases should be listed in analogy. The indicators are based on a comparative approach; therefore, absolute values are by definition non-indicative, but employ their full significance only in comparison with those of other groups. Moreover, the analysis must incorporate the greatest amount of data possible in order to permit statistical compensation for any prejudice, which is able to affect every small undertaken venture that is considered separately. The data constraints used in Bibliometrics mainly stem from various media used by scientists to transfer information between them beyond the usual means of journals. Moreover, the verbal communication

among scientists is not limited to statistics, nor the internal reports between universities, laboratories and research groups and reports between the countries working together through committees, programs or workshops. There is also electronic communication between researchers, which is rapidly developing. All forms of communication, which are covered by the 'traditional' bibliometric methods, therefore consist of exchanges, which have been "formalized." Some unofficial or informal communication is not incorporated and probably will never be incorporated.

The traditional approach is even more restrictive in relation to anything including industrial research or research related to defense projects. There are long lags (time delays) in communication between science (primarily academic) and industry, due to the desire on the part of industry to protect its discoveries and the fact that the published results generally have an abbreviated form. Articles published by industrial laboratories deliberately provide a limited picture of the research objectives, which are, in general, creation of new products or procedures subject to commercial competition. Moreover, a large part of research related to defense (which is often associated with industrial research) is never included in the usual scientific communication, despite its technological importance and the fact that it tends to be in advance of basic research.

The studies based on bibliometric analysis are increasing in recent years in the international arena and are used to determine the characteristics and trends of research output at agency, country or wider set of countries level, the assessment of the scientific project impact, the evaluation of research activity and the emergence of national and multinational networks between scientists and disciplines. They are used to evaluate research systems or organizations and contribute to the development of national research policies at a global scale. The bibliometric indicators are an important but not unique part of a broader ecosystem of metrics of research activity. Disadvantages and limitations concerning their calculation and use are listed in literature, such as differences in practice publications and reports in the scientific fields (e.g. medical compared with humanities) that affect their impact indicators. Moreover, there are problems associated with the "cleansing" of the primary data and the identification of publications, the insufficiency of performance of other important

components of research activity etc. These concerns do not negate the importance of bibliometric indices as a valuable source of data and, as applies to the interpretation of most indicators, these shortcomings can be overcome when the bibliometric indicators are seen in the right context.

Bibliography

Alonso, S., Cabrerizo, F., Herrera-Viedma, E. & Herrera, F., 2009. h-index: A Review-Focused in its Variants, Computation and Standardization for Different Scientific Fields. Journal of Informetrics, 3(4), pp. 273-289.

Alonso, S., Cabrerizoy, F., Herrera-Viedmaz, E. & Herrera, F., 2008. hg-index: A New Index to Characterize the Scientific Output of Researchers Based on the h-and g- Indices,

Anderson, T., Hankin, R. & Killworth, P., 2008. Beyond the Durfee square: Enhancing the *h-index* to score total publication output. Scientometrics , Volume 76, pp. 577-588.

Anderson, T. R., Anderson, T. R., Hankin, R. K. S. & Killworth, P. D., 2008. Beyond the Durfee Square: Enhancing the *h-index* to score total publication output. Scientometrics, 76(3), pp. 577-588.

Andrews, G., 1984. The Theory of Partitions. Oxford: Cambridge University Press.

Bakkalbasi, N., Bauer, K., Glover, J. & Wang, L., 2006. Three options for citation tracking: Google Scholar, Scopus and Web of Science. Biomed Digit Libr. , 3(7)

Barabάsi, A.-L. & Albert, R., 1999. Emergence of scaling in random networks.*Science,* Volume 286, pp. 509-512.

Batista, P. D., Campiteli, M. G., Kinouchi, O. & Martinez, A. S., 2006. Is it possible to compare researchers with different scientific interests?. *Scientometrics,* 68(1), pp. 179-189.

Bernstein, J. & Gray, C. F., 2012. *Content Factor*: A Measure of a Journal's Contribution to Knowledge. *PLoS ONE ,* 7(7), p. e41554.

Bollen, J., Rodriguez, M. A. & Sompel, H. V. d., 2006. Journal Status. *Scientometrics,* 69(3), pp. 669-687.

Borgatti, S. P., Carley, K., Krackhardt, D., 2006. Robustness of *centrality* measures under conditions of imperfect data. *Social Networks,* 28(2), pp. 124-136.

Bornmann, L., Mutz, R. & Daniel, H., 2008. Are There Better Indices for Evaluation Purposes than the h Index? A Comparison of Nine Different Variants of the *h-index* Using Data from Biomedicine. *Journal of the American Society for Information Science and Technology,* 59(5), pp. 830-837.

Burnham, J. F., 2006. Scopus database: a review. *Biomed Digit Libr. ,* 3(1). Center for Research Libraries, 2015. *ISI Web of Knowledge.* [Online] Available at: http://adat.crl.edu/platforms/about/isi_web_of_knowledge

Chadegani, A. A., Salehi, H. & Yunus, M. M., 2013. A Comparison between Two Main Academic Literature Collections:. *Asian Social Science,* 9(5).

Chakrabarti., S. et al., 1998. *Spectral filtering for resource discovery,* s.l.: ACM SIGIR workshop on Hypertext Information Retrieval on the Web.

Cheek J, Garnham B, Quan J., 2006. What's in a number? issues in providing evidence of impact and quality of research(ers). *Qual Health Res.,* Volume 16, p. 423–435.

Chessor, 2016. *Scholarly Publishing: Author Metrics.* [Ηλεκτρονικό] Available at: http://libguides.library.nd.edu/c.php?g=221812&p=2281397 [Πρόσβαση 2016].

Costas, R. & Bordons, M., 2007. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics,* Volume 1, p. 193–203.

Costas, R. & Bordons, M., 2007. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics,* Volume 1, pp. 193-203.

Costas, R., Leeuwen, T. N. v. & Raan, A. F. J. v., 2011. The ‘‘Mendel syndrome’’ in science: durability of scientific literature and its effects on bibliometric analysis of individual scientists. *Scientometrics,* Volume 89, pp. 177-205.

Demaine, J., 2011-2012. h-index and Related Measures – Part 2 : Variants and Extension of the h-index, Vienna: European School of cientometrics.

Egghe L , 2006. An improvement of the h-index: the g-index.. *ISSI Newsletter.,* Volume 2, pp. 8-9.

Egghe L., 2006. Theory and practice of the g-index.. *Scientometrics,* 69(1), pp. 131-152.

Egghe, L., 2007. Dynamic h-index: The Hirsch*-index* in function of time. *Journal of the American Society for Information Science and Technology,* 58(3), pp. 452 - 454.

Egghe, L. & Rousseau, R., 2007. An *h-index* weighted by citation impact. *Information Processing & Management,* Volume 44, pp. 770-780. Elsevier, 2015. *http://www.elsevier.com.* [Online] Available at: http://www.elsevier.com/solutions/scopus

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A. & Pappas, G., 2008. Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. *The FASEB journal,* 22(2), pp. 338-342.

Freeman, L. C., 1978. *Centrality* in social networks: Conceptual clarification :Social Networks.

Garfield, E., 1972. Citation analysis as a tool in journal evaluation. *Science,* Volume 178, pp. 471-479.

Garfield, E., 1998. Letter to the editor. *Der Unfallchirurg,* 48(2), p. 413.

Haiman, M., 1994. On realization of Björner's "continuous partition lattice" by measurable partitions. *Trans. Amer. Math. Soc.,* 343(2), pp. 695-711.

Harzing AWK, van der Wal R., 2008. Google Scholar as a new source for citation analysis. *Ethics Sci Environ Polit.,* 8(1), pp. 62-73.

Harzing, A.-W., 2013. The publish or perish book : Your guide to effective and responsible citation analysis. 1st ed. Melbourne , Australia.: Tarma Software Research Pty Ltd,.

Harzing, A., Alakangas., S. & Adams, D., 2014. hIa: An individual annual *h-index* to accommodate disciplinary and career length differences. *Scientometrics,* 99(3), pp. 811-821.

Henderson J., 2005. Google Scholar: a source for clinicians?. *CMAJ,* 172(12), pp. 1549-1550.

Hess, D. J., 1997. *Science Studies An advanced introduction.* New York: New York University Press.

Hirsch J.E., 2005. An index to quantify an individual's scientic research output. *Proceedings of the National Academy of Sciences of the USA*, 102(46), p. 16569–16572.

HLWIKI International, 2015. *Scopus vs. Web of Science.* [Ηλεκτρονικό] Available at: http://hlwiki.slais.ubc.ca/index.php/Scopus_vs._Web_of_Science [Πρόσβαση 2016].

Holsapple, C. W., JOHNSON, H., MANAKYAN, L. E. & TANNER, J., 1994. Business computing research journals:A normalized citation analysis. *Journal of Management Information Systems,* 11(1), pp. 131-140.

Iglesias, J. & Pecharroman, C., 2006. Scaling the *h-index* for different scientific ISI fields. *Physics,* Τόμος 0607224.

Jacso P., 2005. *Google Scholar (redux).* [Online] Available at: http://www.gale.com/reference/archive/200506/google.html

Jin, B., 2006. h-index: an evaluation indicator proposed by scientist. *Science Focus,* Τόμος 1, pp. 8-9.

Jin, B., Liang, L., Rousseau, R. & Egghe, L., 2007. The R- and AR-indices: Complementing the *h-index. Chinese Science Bulletin,* Τόμος 52, pp. 855-863.

Kleinberg, J., 1999. Authoritative sources in a hyperlinked environment.. *Journal of the ACM,* 46(5), pp. 604-632.

Kosmulski, M., 2007. MAXPROD – A new index for assessment of the scientific output of an individual, and a comparison with the *h-index. Cybermetrics,* 11(1), p. paper 5.

Kousha K, Thelwall M., 2008. Sources of Google Scholar *citations* outside the science citation index: a comparison between four science disciplines. *Scientometrics,* 74(2), pp. 273-294.

Leydesdor L., 2007. Caveats for the use of citation indicators in research and journal evaluations,. s.l.:SIGMETRICS..

Meho, L., 2007. The rise and rise of citation analysis.. *Physics World.,* 20(1), pp. 32-36.

Meho, L. I. & Yang, K., 2007. Impact of data sources on citation counts and *rankings* of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the american society for information science and technology,* 58(13), pp. 2105-2125.

Merton, R. K., 1968. The Matthew effect in science.. *Science,* Volume 159, pp. 56-63.

Miller, C., 2006. Superiority of the *h-index* over the *Impact Factor* for Physics. *arXiv physics / 0608183.*

Newman, M. E. J., 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics 46, 323-351 (2005),* Volume 46, pp. 323-351.

Noruzi A. , 2005. Google Scholar: the new generation of citation indexes.. *LIBRI,*55(4), pp. 170-180.

Opsahl, T., Colizza, V., Panzarasa, P., Ramasco, J. J., 2008. Prominence and control : The weighted rich-club effect. *Physical Review Letters,* Volume 101, p. 168702

Opsahl, T., Agneessens, F. & Skvoretz, J., 2010. Node *Centrality* in Weighted Networks: Generalizing Degree and Shortest Paths. *Social Networks,* 32(3), pp. 245-251.

Page L. & Brin S., 2011. *Facts about Google and Competition,* s.l.: Google.

Page, L., 1997. Improved Text Searching in Hypertext Systems. USA, Patent No. 60/035,205.

Page, L. & Brin, S., 1998. *The anatomy of a large-scale hypertextual web search engine.* s.l., 7th International Web Conference.

Page, L., BRIN, S., MOTWANI, R. & WINOGRAD, T., 1999. *The PageRank citation ranking: Bringing order to the web,* Stanford: Stanford University.

Pareto, V., 1964 [1896]. *Cours d'Economie Politique.* Geneva: Droz.

Podlubny, I., 2005. Comparison of Scientific Impact Expressed by the *Number of citations* in Different Fields of Science. *Scientometrics,* 64(1), pp. 95-99.

Podlubny, I. & Kassayova, K., 2006. Towards a better list of citation superstars: compiling a multidisciplinary list of highly cited researchers. *Research Evaluation,* 15(3), pp. 154-162.

Price, D. J. d. S., 1965. Networks of scientific papers.. *Science,* Issue 149, pp. 510-515.

Price, D. J. d. S., 1976. A general theory of bibliometric and other cumulative advantage processes.. *J. Amer. Soc. Inform. Sci.,* Volume 27, pp. 292-306.

Raan, A. V., 2004. Sleeping beauties in science. *Scientometrics,* Volume 59, pp. 467-472.

Redner, S., 1998. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J. ,* B(4), pp. 131-134.

Roediger H.L., 2006. The *h-index* in science : a new measure of scholarly contribution.. *The Academic Observer,* 19(4).

Rousseau R, 2006. New developments related to the Hirsch index. Science Focus 1: 23–25.. *Science Focus,* Volume 1, pp. 23-25.

Rousseau, R., 2008. Reflections on recent developments of the *h-index* and h-type indices. *Collnet Journal of Scientometrics and Information Management,* 2(1)

Rousseau, R. & Ye, F., 2008. A proposal for a dynamic h-type index. *Journal of the*

American Society for Information Science and Technology, 59(10).

Royle, P., Kandala, N.-B., Barnard, K. & Waugh, N., 2013. Bibliometrics of systematic reviews: analysis of citation rates and journal *Impact Factor*s. *Systematic Reviews,* Volume 2, p. 74.

Ruane, F. & R.S.J. Tol, 2008. Rational (successive) h-indices: An application to economics in the Republic of Ireland.. *Scientometrics,* 75(2), pp. 395-405.

Ruane, F. & Toll, R., 2007. Rational (successive) H-Indices: economics in the Republic of Ireland. *Scientometrics,* Volume 75, pp. 395-405.

Schreiber, M., 2009. A Case Study of the Modified Hirsch-*index* hm accounting for multiple coauthors. *Journal of the American Society for Information Science and Technology,* 60(6), pp. 1274-1282.

Schreiber, M., 2010. Twenty Hirsch index variants and other indicators giving more or less preference to highly cited papers, Chemnitz, Germany: Institut für Physik, Technische Universität Chemnitz.

Schwartz, R. B. & Russo, M. C., 2004. How to quickly find articles in the top IS journals. *Communications of the ACM,* 47(2), pp. 98-101.

Scimago Lab, 2016. *SHAPE OF SCIENCE.* [Ηλεκτρονικό] Available at: http://www.scimagojr.com/

SCImago, 2007. *SJR* — SCImago Journal & Country Rank.. [Online].

Seglen P.O. , 1997. Why the *Impact Factor* of journals should not be used for evaluating research. *Bristish Medical Journal7,* 314(7079), pp. 458-502.

Senanayake, U., Piraveenan & Zomaya, A., 2015. The *PageRank*-Index: Going beyond Citation Counts in Quantifying Scientific Impact of Researchers. *PLoS ONE,* 10(8).

Sidiropoulos, A., Katsaros, D. & Manolopoulos, Y., 2007. Generalized Hirsch hindex for disclosing latent facts in citation networks. *Scientometrics,* 72(2), pp. 253-280.

Simon, H. A., 1955. On a class of skew distribution functions.. *Biometrika,* Volume 42, pp. 425-440.

Star S.L., 1995. The politics of formal representations: wizards,gurus and organizational complexity,. In: S. S.L., ed. *Ecologies of Knowledge. Work and Politics in Science and Technology.* Albany, NY: SUNY Press.

Sutherland W.J., 1999. *TREE,* 14(10), pp. 382-384.

Thomson Reuters, 2016. Journal Citation Reports. [Ηλεκτρονικό] Available at: http://thomsonreuters.com/en/products-services/scholarly-scientific-research/research-management-and-evaluation/journal-citation-reports.html

Thomson Reuters., 2015. *The Thomson Reuters journal selection process..* [Online] Available at: http://thomsonreuters.com/products_services/science/free/essays/journal_selection_process

Thomson Reuters, 2015. *http://thomsonreuters.com/en.html.* [Online] Available at: http://ipsciencehelp.thomsonreuters.com/incitesLive/glossaryAZgroup/g20/7411-TRS.html

Tol, R., 2009. The *h-index* and its alternatives: an application to the 100 most prolific economists. *Scientometrics ,* Volume 80, pp. 317-324.

U.S. National Library of Medicine, 2015. *http://www.nlm.nih.gov.* [Online] Available at: http://www.nlm.nih.gov/pubs/factsheets/medline.html

Van Raan AFJ, 2006. Comparison of the Hirsc*h-index* with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics ,* 67(3), pp. 491-502.

Vanclay, J., 2007. On the Robustness of the h-Index.. *Journal of the American Society for Information Science and Technology,* 58(10), pp. 1547 - 1550.

Vinkler, P., 2009. The π-index: a new indicator for assessing scientific impact. *Journal of Information Science 35, 602-612 (2009).,* Volume 35, pp. 602-612.

Vine, R., 2006. Google Scholar. J Med Libr Assoc., 94(1), pp. 97-99.

Whitman, K., 2011. Web of Science vs. Scopus: Which is better?. [Ηλεκτρονικό]

Available at: https://intellogist.wordpress.com/2011/03/21/web-of-science-vs-scopus-which-is-better-part-2/ [Πρόσβαση 2016].

Wasserman, S. & Faust, K., 1994. *Social Network Analysis: Methods and Applications.* Cambridge, ENG & New York, USA: Cambridge University Press.

Wincka, J. & Morais, A., 2011. *Impact Factor* and the portuguese journal of pulmonology: knockin’ on heaven's door?. Revista Portuguesa de Pneumologia, 17(4), pp. 151-152.

Yule, G. U., 1925. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis.. *Philos. Trans. R. Soc. London B,* Volume 213, pp. 21-87

Appendices

A.1. Ferrers diagram and conjugate partition

The Ferrers diagram, also called Young diagram, of a partition is a rectangular array of n boxes, or cells, with one row of length *j* for each part *j* of *λ*. The conjugate of a partition is the partition of *n* whose diagram you get by reflecting the diagram of *λ* about the diagonal so that rows become columns and columns become rows. We use the notation *λ** for the conjugate of *λ*.

A.2. The Durfee square

We define the Durfee square of λ to be the largest square array that fits in the upper left corner of the Ferrers diagram. If the Durfee square is *c* by *c*, we call c the Durfee number of *λ* The rest of the diagram of λ consists of two parts, which we call the arm (marked below with a 's) and the leg (marked with l 's). The arm and the leg are diagrams themselves. Obviously the arm can be any partition with at most c parts, and the leg any partition with parts at most c. (Haiman, 1994)

*	*	*	α	α	α
*	*	*	α	α	
*	*	*			
i					
i					

*Table 4: The Durfee square (marked with *'s); the arm and leg (marked with a's and l's respectively); (Haiman, 1994)*

A.3. Measuring Power Laws

When the probability of measuring a particular value of some quantity varies inversely as a power of that value, the quantity is said to follow a power law, also known variously as *Zipf's law* (Zipf, 1949) or the *Pareto distribution* (Pareto, 1964 [1896]). Cumulative distributions with a power-law form are sometimes said to follow Zipf 's law or a Pareto distribution. Instead of plotting a simple histogram of the data, we make a plot of the probability $P_{(x)}$ where *x'* has a value greater than or equal to *x*. Since power-law cumulative distributions imply a power-law form for $P_{(x)}$, "Zipf's law" and "Pareto distribution" are effectively synonymous with the power-law distribution. Zipf's law and the Pareto distribution differ from one another in the way the cumulative distribution is plotted[39]. The data depicted in the plots are of course identical. Identifying power-law behavior in either natural or man-made systems is quite demanding. The standard strategy makes use of a result we have already seen: a histogram of a quantity with a power-law distribution appears as a straight line when plotted on logarithmic scales. Just making a simple histogram, however, and plotting it on log scales to see if it looks straight is, in most cases, inadequate.

A.4. Node *Centrality* in Weighted Networks

The *centrality* of nodes, or the identification of which nodes are more "central" than others, has been a key issue in network analysis. Based on three features, Freeman (Freeman, L. C., 1978) formalized three different measures of node *centrality*: degree, closeness and betweenness.

- *Degree* is the number of nodes that a focal node is connected to, and measures the involvement of the node in the network. Its simplicity is an advantage: only the local structure around a node must be known for it to be calculated. However, there are limitations: the measure does not take into consideration the global structure of the network. For example, although a node might be connected to many others, it might not be in a

[39] Cumulative distributions like this are sometimes also called *rank/frequency plots.*

position to reach others quickly to access resources, such as information or knowledge (Borgatti et al.., 2006).

- *Closeness centrality* was defined as the inverse sum of shortest distances to all other nodes from a focal node aiming to accommodate the feature mentioned above. A main limitation of closeness is the lack of applicability to networks with disconnected components (e.g.dangling nodes).
- The last of the three measures, *betweenness*, assesses the degree to which a node lies on the shortest path between two other nodes, and are able to funnel the flow in the network. In so doing, a node can assert control over the flow. Although this measure considers the global network structure and can be applied to networks with disconnected components, it is not without limitations. For example, a great proportion of nodes in a network generally does not lie on a shortest path between any two other nodes, and therefore receives the same score of 0. The three measures have been generalized to weighted networks. These generalizations focused solely on tie weights and ignored the original feature of the measures: the number of ties. As such, a second set of generalizations was proposed by Opsahl et al. (Opsahl, et al., 2010) that incorporates both the number of ties and the tie weights by using a tuning parameter.

A.4.1. Degree

Degree is the simplest of the node *centrality* measures by using the local structure around nodes only. In a directed network, a node may have a different number of outgoing and incoming ties, and therefore, degree is split into out-degree and in-degree, respectively. Degree has generally been extended to the sum of weights when analyzing weighted networks (Opsahl, T., et al., 2008), and labeled node strength. Degree *centrality* of a node refers to the number of edges attached to the node. In order to know the standardized score, you need to divide each score by *n-1* (*n* = the number of nodes). For example if the graph has 7 nodes, 6 (7-1) is the denominator for this calculation.

A.4.2. Centrality & Prestige

A primary use of graph theory in social network analysis is to identify the "important" active nodes (*actor*s). *Centrality* and *prestige* concepts seek to quantify graph theoretic ideas about an individual *actor*'s prominence within a network by summarizing structural relations among the nodes. Group-level indexes of centralization and *prestige* assess the dispersion or inequality among all *actor*s' prominences. An *actor*'s prominence reflects its greater visibility to the other network *actor*s[40] (how big is the audience he/she attracts). An *actor*'s prominent location takes account of the direct sociometric choices made and choices received (outdegrees and *indegrees*), as well as the indirect ties with other *actor*s. The two basic prominence classes are:

- *Centrality*: The *actor* has high involvement in many relations, regardless of their send/receive directionality (volume of activity)
- *Prestige*: *Actor* receives many directed ties, but initiates few relations (his popularity exceeds his extensively)

(Wasserman & Faust, 1994)

[40] See Appendix A.4.