Is it Possible to Measure Scientific Performance with the h-Index or with Another Variant from the Hirsch Index Zoo?

Michael Schreiber[1]
Institute of Physics, Technische Universität Chemnitz, D-09107 Chemnitz, Germany

Received 16.10.2013, accepted 17.12.2013, published 30.12.2013

The h-index proposed by Hirsch only 8 years ago is already frequently used to measure scientific performance. Nevertheless, several open questions are unsolved, e.g. what does the h-index really measure? Are there better variants available? How reliable is the determination of the h-index? Does it have predictive power?

1 Introduction

In recent years the attempts to measure scientific performance have been felt as a growing pressure on many scientists, exercised by administration, politics and even the general public. This is somewhat understandable because administrators, politicians and more general the citizens want to know whether the tax money is spent in a reasonable way. But it remains unclear not only what is reasonable, but also how to measure scientific performance. Various measures are in use, from counting PhD students (which has recently been strongly criticized due to plagiarism scandals) to counting allocated or spent third-party funding or counting publications and citations. It is not the purpose of the present paper to discuss or even weight the respective different measures. Rather, I will concentrate on the bibliometric issues, related to the h-index proposed by Hirsch[1] as a measure for the scientific performance in terms of citations. It is defined as the (largest) number h of a scientist's publications which have received at least h citations. Thus, it appears to be easy to determine it and therefore, the h-index has become famous among administrators but is also considered infamous by many scientists, even if they have obtained a relatively high index value.

The validity and the advantages and disadvantages of the h-index have been discussed ever since its introduction, and a plethora of variants has been introduced.[2] Nowadays, it is difficult to find a letter in the alphabet which has not yet been proposed at least once as a new bibliometric index in that context.[3] Many of the suggestions are driven by personal taste, some of them might also have been created due to the desire to improve the indicator value of the proposing author in comparison with competing colleagues. It is impossible to review all suggestions, even the reasonable proposals are too numerous. In the following, I will discuss only some of these bibliometric indicators what of course means a subjective choice according to my personal taste.

Let me add a caveat: I am not a scientometric expert but a physicist who has chosen this topic as a hobby horse (hobby deer) several years ago. In these years, I analyzed the citation records of several physicists. For the present paper, I have investigated the data of Prof. Dr. Kurt Binder from the Johannes Gutenberg-University Mainz, whom I got to know as an excellent scientist during my first professorship in the Institute of Physical Chemistry in Mainz three decades ago. He still is a very active researcher, publishing frequently and his papers have received and are still receiving very many citations. In the year 2001, he was distinguished by Thomson Reuters' Institute of Scientific Information (ISI) as one of the most cited authors in the world. There are only very few other German physicists with so many citations. Therefore, dear reader, while you may be impressed by the numbers below you should not be disappointed if your personal citation record is significantly lower.

2 The h-index and its determination

Probably the most simple way of measuring scientific performance in bibliometric terms is counting the number of publications. After this indicator came into use in particular in the US decades ago, some scientists adjusted their publication outcome to it. This led to the tactics of salami slicing, i.e. the apportionment of the obtained results into LPUs (least publishable units). As a countermeasure, people started counting the citations, what was made possible but cumbersome by the then regularly printed Science Citation Index.

Counting citations to help with economic decisions is nothing new. Already in 1927, the citation frequencies of 28 leading chemistry and physics journals in the previous 54(!) years were investigated[4] with the aim of determining for which journals the subscription should be continued or cancelled. Similarly, nowadays some people believe that citation records can be used for determining, whether to allocate research funds or whether to hire scientists or not.

The advance of large bibliometric databases has simplified the evaluation of citation statistics also for individual researchers significantly. For the present paper, I have downloaded K. Binder's citation record from the Web of Science (WoS) provided by Thomson Reuters (formerly ISI) on September 18, 2013. Then, 993 entries were found, but a careful check yielded only 884 publications that were written by the investigated person. This is the well known precision problem: up to now there is no reliable way of

[1] e-mail: schreiber@physik.tu-chemnitz.de

determining the citation record of the publication set of a person with high accuracy automatically. In some cases, like F. Wilczek, it is easy. In other cases, like M. L. Cohen, it is more difficult to eliminate homographs, i.e. to exclude papers which have not been written by the physicist Marvin L. Cohen.[5] In the other direction, for Pierre-Gilles deGennes one has to combine the WoS search results for deGennes PG, deGennes P, de Gennes PG, Gennes PG, and Gennes PGD. Similar problems occur for scientists whose names changed due to marriage. And the transliteration from other alphabets has changed over the years, which makes it sometimes nearly impossible to find all papers of Russian, Japanese, or Chinese scientists. For an accurate result, it is therefore indispensible to compare the papers in the downloaded citation record with the publication list provided by the author. Fortunately, such lists can usually be found in the WWW. In conclusion, the h-index value, which is automatically calculated in the WoS, is not reliable. One should also be aware that many conference proceedings are not included in the data base. Moreover, books have only recently started to be taken into account.

Sorting the papers according to the number c of citations allows an easy determination of the h-index, see Fig. 1: h is given by the largest rank r for which $c(r) \geq r$. In the present case, one gets $h = 95$. Graphically this means that one has to search for the largest r for which the data point in Fig. 1 lies on or above the diagonal $c(x) = x$. In order to avoid the unequality, it is often helpful to generalize the definition of h to the rational variant $\widetilde{h}$: if one uses a linear interpolation of the citation frequency $c(x)$ between r and $r+1$, then $\widetilde{h}$ can be determined from the equality $\widetilde{h} = c(\widetilde{h})$. Graphically, this means the intersection of the interpolating lines in Fig. 1 with the diagonal. The original h-index is obtained by rounding the rational version to the next lower integer value. In the present case, one gets $\widetilde{h} = 95.0 = h$.

3 Advantages and disadvantages

One advantage of the h-index was already stated by Hirsch in the original publication,[1] namely that it combines the dimension of quantity as expressed by the number of publications with the dimension of quality, assuming that the number of citations reflects the quality of a manuscript. This is certainly not obvious because sometimes faulty papers attract a considerable number of citations. It is an open question whether it is worthwhile to try and eliminate such incorrect publications. On the other hand, review articles are likely to be frequently cited, although they usually do not present new research results. It is another open question whether they should be included in the h-index or not.

The mentioned advantage, however, has been criticized from a methodological point of view because such a mixture of different dimensions into one indicator is questionable in principle. Moreover, only on first sight the mixture appears to be unique because the definition of h does not depend explicitly on any parameter. In fact, one can introduce a prefactor q and demand that h_q publications have obtained at least $q \cdot h_q$ citations each.[6] This arbitrariness allows one to define a generalized index h_q, or rather an infinite number of indices which are more or rather less useful. In particular, $q = 10$ has been suggested as a reasonable choice for highly cited researchers because then the results are much smaller so that the precision problem would be reduced.[7] In Fig. 1 the respective broken line yields $h_{10} = 24$. Already for more moderate values of q, the ranking of scientists can change considerably in comparison with the original h-index.[8] This underlines the problem that small differences in the index values should not be utilized for distinguishing the researchers. It would be an overinterpretation if differences of a few index points were taken as an indication that one scientist is better than the other.

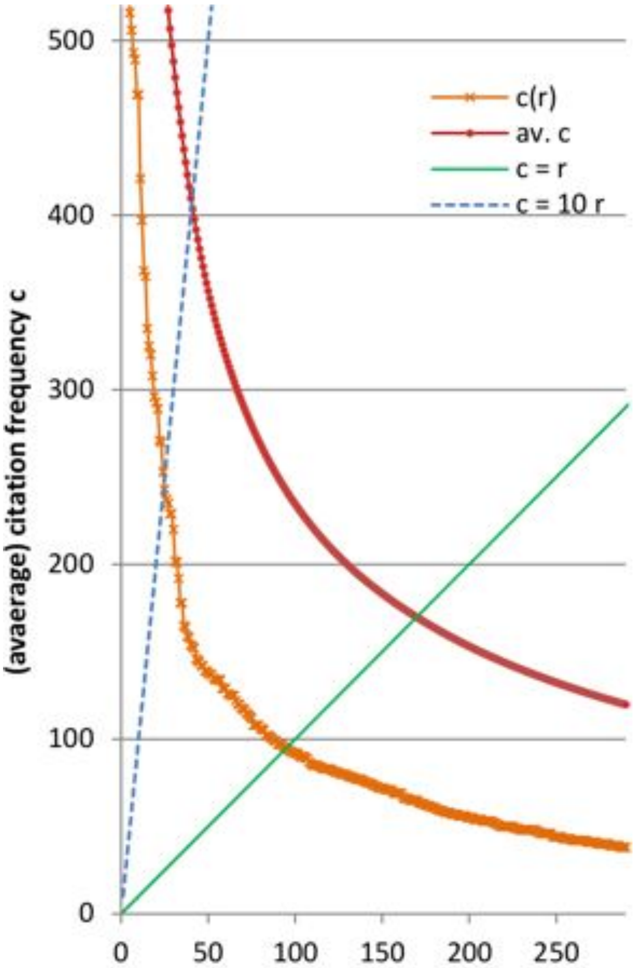

Figure 1: Citation record of Kurt Binder. The papers are ranked according to the number of citations (×). Also given is the average number of citations (●) up to rank r. The straight solid lines reflects the diagonal $c(r) = r$, the broken line indicates $c(r) = 10r$.

At first glance already, Fig. 1 shows that the citation curve is very skewed. This is usual and makes any use of average citation numbers questionable. Such averages have also been proposed for alternative indicators. The advantage is that they counterbalance the above mentioned salami slicing tactics. However, in my view, it is unfair to punish a productive author, even if some publications have had little impact. In any case, for pushing one's h-index the apportionment of the results to several papers is probably not helpful because possible citations are likely to be also likewise distributed, leading to lower citation frequencies for the smaller units.

The skewness of the citation record can be quantified: in the present case, the 5% most cited papers have attracted 40% of all the 42,531 citations. Excluding the highly cited review articles, the 5% most cited papers still have more than 34% of all the 35,457 citations. The distribution, thus, does not quite reach Pareto's principle, the 80-20 rule but comes close to that: not 80% but 69% of all citations are concentrated on the top 20% of papers. Consequently, one should not consider average values, because the use of averages for strongly skewed distributions violates most-basic scientific standards.

This is by the way a severe criticism which should be applied also against the impact factor which for a given year is calculated as the average number of citations within that year to the papers published in the two previous years. It was originally created to judge the quality of a journal in terms of citations. But as the citation distributions of journals are also strongly skewed, the impact factor is not a good measure.[9] And it should certainly never be used to judge the quality of a single paper, because there is a low correlation between the impact factor and the number of citations to an individual paper in a journal. As an example I note that according to Thomson Reuters, the impact factor in the year 2011 for Physical Review Letters was 7.37, but less than one third of the relevant papers (i.e. from 2009 and 2010) contributed 8 or more citations. The remaining two thirds had less citations and therefore effectively decreased the impact factor of the journal. It would have been better for Physical Review Letters if these papers had not been published in this journal.

4 Further variants of the h-index

One criticism against the h-index is based on the fact that additional citations to the papers in the h core, i.e. the set of h-defining papers, do not have any effect. Many people consider this to be unfair. This shortcoming was remidied in the g-index, defined originally as the largest number g of papers that together received g^2 or more citations.[10] I wonder whether the g-index has not become more popular, only because this definition appeared to be too difficult, when the sum had to be compared with a parabola. In fact, it is equivalent to the demand that the average number $\bar{c}$ of citations to the g most cited papers is larger or equal to g.[11] In this form it looks much more similar to the definition of the h-index. The respective average values are included in Fig. 1, yielding $g = 169$. Due to the averaging, the g data show a much smoother behavior than the h values in Fig. 1.

But in principle, the above reservations against averages apply. However, in this situation there is a way out: the average is only a mathematical formality with the aim of enhancing the index value in a reasonable way by taking into account the excess citations to the core papers,[10] i.e. the $c(r) - h$ citations of the r-th paper. E.g. in Fig. 1 this means $c(1) - h = 3148 - 95 = 3053$ citations to the first paper which are not relevant for the h-index but become relevant for the g-index.

Again, a linear interpolation $c(x)$ as above allows one to define an index $\widetilde{g} = \bar{c}(\widetilde{g})$ which in this case is a real number. Now, every additional citation to the papers in the core causes an albeit small increase of the index. I consider this to be an attractive feature.

Like the h-index, the definition of the g-index is not as unique as it looks. Similarly to the above, one can utilize a prefactor q and thus get an infinite number of generalized indices g_q.[6] Specifically, one gets $g_{10} = 40$ in Fig. 1.

There is another straightforward way of generalizing the g-index. Without explicitly mentioning it in the discussion above, the average was meant to be the arithmetic mean. But there are other means like the harmonic or the geometric mean. In general, one can use an exponent p to define the Hölder means also known as power means

$$\bar{c}_p(r) = \left(\frac{1}{r} \sum_{r'=1}^{r} \Big(c(r') \Big)^p \right)^{\frac{1}{p}}$$

and utilize these means in the same way as the average above.[12] For $p = 1$, one obtains the usual g-index. For $p = 0$ and $p = -1$, the geometric and the harmonic averages yield two generalized indices which have been labeled t and f previously.[13] Surprisingly, as far as I know, the quadratic average for $p = 2$ has not yet been exploited in the present context. In the limit $p \to -\infty$, one obtains the usual h-index. The other limit $p \to \infty$ yields the citation frequency $c(1)$ of the most cited paper which some people also consider a useful quality indicator. In summary, by varying the exponent p it is possible to give more or less preference to highly cited papers.

Other variants of the h-index are based on arithmetically averaged citation frequencies for different core sizes. Sometimes, the median is utilized. Further variants are based on the square root of the summed number of citations for different core sizes. More complicated definitions have been proposed, leading to rather exotic indices which are unlikely to be utilized because the calculation is too cumbersome for practical purposes. Several variants are discussed in short reviews but shall not be given undue credit here.[2,3]

5 Modifying the database

Up to now, the mentioned variants of the h-index have all been based on the original citation data. However, there are good reasons to modify these data. One problem concerns self-citations. Obviously, self-citations do not reflect the impact of a publication. In the WoS, total citation counts without self-citations can be obtained but here, only the direct self-citations, i.e. citations by the investigated scientist to his/her papers are taken into account. However, sometimes there is a co-author who is much more enthusiastically self-citing and of course these indirect self-citations should also be excluded from the citation record.

In his original publication,[1] Hirsch contended that self-citations would not have a big influence on the index. This conjecture is not true. I have shown that the exclusion of self-citations from the citation record can strongly change the ranking of scientists according to the so-called sharpened index h_s in comparison with the ranking according to h.[14] Unfortunately, as the exclusion of direct (let alone indirect) self-citations is at present not automatically done for each paper separately, I am afraid that this consideration will not be applied in many cases.

Another modification concerns the number of co-authors. It is an open question how multi-author papers can be treated in a fair way.[15] Usually, the contributions of the individual authors to a paper are not quantified. Nevertheless, suggestions have been made to give more weight to the first and/or last author of an author list for each publication. However, this is not very practical because there exist different traditions in different fields how the co-authors are arranged in the list. In conclusion, it remains an open question how to treat this problem. An imperfect way is to share the impact equally among the authors. One respective possibility would be to fractionalize the citation counts and attribute $c(r)/a$ citations to each of the a authors. But for the present purpose, this means that the papers have to be rearranged according to the fractionalized citation counts. This is not only impractical but it also appears unreasonable that highly cited papers with many authors are likely to drop out the core. A better way is to fractionalize the paper count, i.e. to attribute only a share of $1/a$ of the paper to each author. I have labeled the respectively modified index as h_m and shown that this modification can also have a strong effect on the ranking.[5]

Of course, one can combine the modification for multi-author papers with the index sharpened for self-citations and obtain the index h_{ms}.[16] Likewise, a modified sharpened index g_{ms} can be defined.

Another open question is how it is possible to compare the indices of scientists working in different fields. It is well known that there are different citation cultures, e.g. in mathematics and in engineering. Therefore a field normalization is required.[17] But even in one field like physics, it is doubtful whether the indices can be compared in a meaningful way.[18] For example, in mathematical physics the number of citations is usually considerably smaller than in biophysics. Therefore, a comparison without subfield normalization could be very unfair for mathematical physicists. But then, multidisciplinary papers become a problem because it is unclear which normalization should be applied.

Another difficulty occurs for large collaborations which are typical in high energy physics. If there, the paper counts are fractionalized, they are also marginalized which would be unfair. However, is it fair to take a paper with 1,000 authors fully into account 1,000 times?

6 The predictive power of the h-index

The h-index has been shown to have predictive power in the sense that there is a high correlation between the values after 12 years and after 24 years of the career of researchers.[19] This raises the question whether the h-index can be used profitably in academic appointment processes or for the allocation of research resources. However, I have shown that the evolution of the h-index with time is usually dominated for a long time by citations to previous publications rather than by new scientific achievements.[20] This is visualized in Fig. 2 where the time evolution of Binder's h-index is compared with the fictitious evolutions obtained under the assumption that he had stopped publishing in the selected years s. For example, for $s = 1988$ the index would have increased like h until 1993 and even in 2001 it would have been smaller only by 5 index points, less than 7%. If he had stopped publishing in 1997, there would have been no change compared to the actual h-index in the next 8 years and a change of no more than 2 index points until 2011, that is after 14 years! These observations should not been misinterpreted: The inertness of the h-index cannot be taken as an indication that recent publications had no impact. But it becomes more and more difficult for additional publications to contribute to the h-index when the index values are already high.

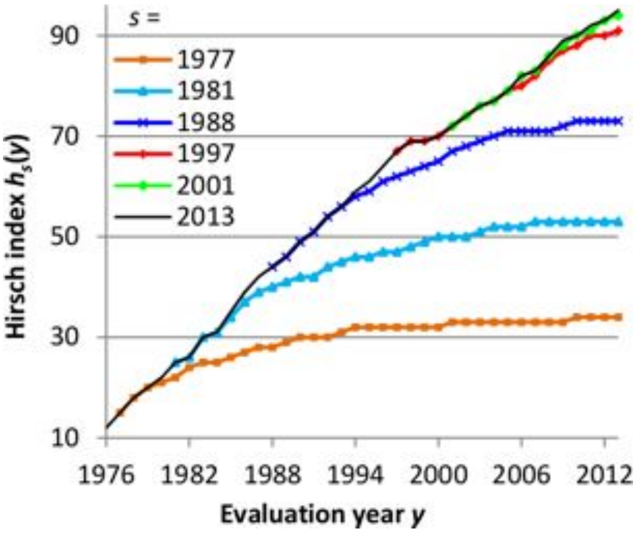

Figure 2: Time evolution of the h-index for Kurt Binder's publications (top black line). Additionally the evolution is displayed for selected years s (see legend) taking only publications up to the year s into account.

In conclusion, the h-index is a good predictor of itself due to its inertia, but it cannot predict future scientific performance. If a researcher goes to sleep in the year s, for example after getting tenure, the h-index is likely to increase anyway. Likewise, the past evolution of the h-index does not automatically mean that a candidate has performed well in recent years. The index would have increased more or less as it did, even if the candidate had gone to sleep several years ago. On the other hand, the h-index evolution does reflect the impact of the past achievements, although not so much of the very recent research. But of course, this earlier performance is also an aspect which one may want to consider in appointment processes or for allocation of research resources.

In a more general investigation of 10 citation-frequency based indicators,[21] the current annual number of citations was found to be the best predictor of future citations, but not surprisingly, none of the indicators was able to predict citations to future work well.

7 Discussion and conclusions

The h-index has become a popular measure for the scientific impact of a researcher's publications. Whether you or I like it or not, it is here to stay, not only but also because of its simplicity. But it remains an open question whether quality can be measured in this way. However, it is certainly better than just counting publications or citations. Nevertheless, it cannot replace peer review. And in scientometrics, it was recognized already in 1981 that "uncertainties make the concerted use of citation analysis and peer evaluation inevitable".[22] But of course, peer review means reading papers which is time consuming and requires thinking. And already one of the founding fathers of scientometrics warned that "citation analysis is not a shortcut to be used as a replacement for thinking".[23]

In any case, one should always be aware that it is impossible to judge the performance of a researcher by a single indicator. Even for the purpose of measuring the citation impact, more than one number would be better than just the h-index.

But if one really wants to condense the citation record into one indicator, there are some variants of the h-index which are more meaningful, while most versions are too complicated or exotic and will not have much impact. In my admittedly subjective view, the modified sharpened index g_{ms} would be the best variant.

Due to the many possibilities to select one of the index variants, any ranking based on a single indicator should be considered with reservation. Small differences in the index values should be interpreted with caution. One can easily find examples of prominent scientists with low index values. Therefore, it is reasonable to adhere to the principle of antidiagnostics, namely that "in scientometrics, numerical indicators can reliably suggest only eminence but never worthlessness".[24]

As usual, any indicator will lead to creative adjustment processes. Therefore one should be aware that the h-index can also be subject to possible manipulation: I have seen citation records which are surprisingly flat around the actual value of the h-index what can be achieved by the strategy of selfciting the respective papers with the aim of pushing them into the h core. A more clever because not so obvious way is enhancing the respective citation counts by a citation cartel. Indeed, I have already been asked by a colleague to cite specific papers for this purpose, because a minimum value of the h-index was demanded by the administration for the promotion to professorship in that country.

8 Prospect

Given the shortcomings of the h-index which can only be weakened but not remedied by the variants, the question arises whether there might be better alternatives to evaluate the scientific performance of an individual researcher in terms of citations. One possibility is the comparison of the specific citation record with reference sets which are accessable for example via the InCites database which is also provided by Thomson Reuters' ISI.[25-27] For this purpose, reference sets for different fields (or possibly subfields) and different publication years are utilized in order to determine a position of a publication within the reference distribution. For this aim, the papers in the reference distribution are sorted according to their citation frequency and median, quartile, decile or other percentiles are determined. Then age- and field-normalized impact scores can be calculated for each publication of an individual scientist by determining the respective percentile of the reference set to which the publication belongs. Such impact scores avoid most problems associated with the h-index and its variants, because they enable cross-field comparisons, avoid age-dependent discrimination of younger scientists, solve problems with the skewed citation distributions, and also make manipulations much more difficult.

As mentioned, for such an evaluation, a comparison of each publication record with the reference set has to be performed. This is of course much more tedious than the simple determination of the h-index. But given the importance of such evaluations with respect to allocating grant money or selecting candidates for an open position, simplicity and easy access to the data base should not play a decisive role. Of course, additional costs will occur for the determination of the reference distributions. But, compared to the costs of a miscast professorship or misplaced grants, the access to InCites is not so expensive that this means an impregnable hurdle.

On the other hand, the relatively large effort which is necessary for comparing the citation frequency of each publication might be the greater hurdle which is even increased by the necessity of selecting the appropriate reference set (or, rather, sets) especially for scientists which have worked in different (sub)fields or in cross-disciplinary research. Therefore, while the described evaluation in terms of normalized impact scores is certainly a much better way than the h-index comparisons, it appears doubtful whether these proceedings will be performed in many cases.

9 Summary

My answer to the question in the title of this paper is: In principle yes, but already the Greek philosopher Plato had realized that "a good decision is based on knowledge and not on numbers." What more is there to say?

References

[1] J. E. Hirsch, *P. Natl. Acad. Sci. USA* **2005**, *102*, 16569-16572.
[2] M. Schreiber, *Ann. Phys. (Berlin)* **2010**, *522*, 536-554.
[3] L. Bornmann, R. Mutz, S. E. Hug, H. D. Daniel, *J. Informetr.* **2011**, *5(3)*, 346-359.
[4] P. L. K. Gross, E. M. Gross, *Science* **1927**, *66*, 385-389.
[5] M. Schreiber, *New J. Phys.* **2008**, *10*, 040201–1-9.
[6] N. J. van Eck, L. Waltman, *J. Informetr.* **2008**, *2(4)*, 263-271.
[7] Q. Wu, *J. Assn. Inf. Sci. Technol.* **2010**, *61*, 609-614.
[8] M. Schreiber, *J. Informetr.* **2013**, *7(2)*, 379-387.
[9] B. Brembs, K. Button, M. Munafo, *Front. Hum. Neurosci.* **2013**, *7*, 291–1-12.
[10] L. Egghe, *Scientometrics* **2006**, *69(1)*, 131-152.
[11] M. Schreiber, *J. Assn. Inf. Sci. Technol.* **2010**, *61(1)*, 169-174.
[12] M. Schreiber, *J. Informetr.* **2010**, *4*, 647-651.
[13] R. S. J. Tol, *Scientometrics* **2009**, *80*, 317-324.
[14] M. Schreiber, *Ann. Phys. (Leipzig)* **2007**, *16(9)*, 640-652.
[15] J. E. Hirsch, *Scientometrics* **2010**, *85(3)*, 741-754.
[16] M. Schreiber, *Ann. Phys. (Berlin)* **2009**, *18*, 607-621.
[17] J. E. Iglesias, C. Pecharromán, *Scientometrics* **2007**, *73(3)*, 303-320.
[18] F. Radicchi, C. Castellano, *Phys. Rev. E* **2011**, *83(4)*, 046116–1-6 .
[19] J. E. Hirsch, *P. Natl. Acad. Sci. USA* **2007**, *104(49)*, 19193-19198.
[20] M. Schreiber, *J. Informetr.* **2013**, *7*, 325-329.
[21] A. Mazloumian, *PLoS ONE* **2012**, *7(11)*, e49246–1-9.
[22] G. Folly, B. Hajtman, J. I. Nagy, I. Ruff, *Scientometrics* **1981**, *3*, 135-147.
[23] E. Garfield, *Current Contents* **1983**, *45*, 5-14.
[24] T. Braun, A. Schubert, *Scientometrics* **1997**, *38*, 175-204.
[25] L. Bornmann, L. Leydesdorff, R. Mutz, *J. Informetr.* **2013**, *7(1)*, 158-165.
[26] L. Waltman, C. Calero-Medina, J. Kosten, E. C. M. Noyons, R. J. W. Tijssen, N. J. van Eck, T. N. van Leeuwen, A. F. J. van Raan, M. S. Visser, P. Wouters, *J. Assn. Inf. Sci. Technol.* **2012**, *63(12)*, 2419-2432.
[27] L. Waltman, M. Schreiber, *J. Assn. Inf. Sci. Technol.* **2013**, *64(2)*, 372-379.