Scientometrics
DOI 10.1007/s11192-014-1373-9

Rescaling the h-index

András Schubert

Received: 15 May 2014

Abstract An estimation of the h-index is proposed for cases when the original variable underlying the distribution for which the h-index had been determined was rescaled. Within its validity limits, the approximation can be usefully applied for field normalization, change of time frames or other changes of measurement scales.

Keywords h-index · Journals · Subject fields

Introduction

The success of the h-index (Hirsch 2005) critically hinges on the fortunate fact that the publication count and the maximum citation rate of top scientists are generally in the same order of magnitude. This feature of the citation distribution grants a sensible "intersection" of the rank number and the citation count at a certain non-trivial h-value.

This requirement largely restricts the usability of the h-statistics concept outside the realm of publication/citation statistics. Another feasible example is the degree distribution of a general unweighted graph (Korn et al. 2009; Schubert et al. 2009), since the maximum degree of a node is automatically limited by the number of the nodes. Generalization for weighted graphs is more problematic (Zhao et al. 2011).

Let us consider the historical example of measuring cycling prowess by the "Jeffreys-index", j, j being the highest number of days on which one had cycled j or more miles (Edwards 2005). The index did work, i.e., after a certain time (say, a few years) it had a suitable distinguishing power according to cycling prowess. (Jeffreys himself has been told to have an index of 70. The presumable creator of the concept, Sir Arthur Eddington had an index of 87.) Obviously, were this index elaborated, say, in Berlin instead of Cambridge, a

A. Schubert (✉)
Department of Science Policy and Scientometrics, Library and Information Center of the Hungarian Academy of Sciences, Budapest, Hungary
e-mail: schuba@iif.hu

Published online: 16 July 2014

similar index based on kilometer scale would have worked equally well. The numerical values would have been different but in the same order of magnitude, the rankings would remain substantially unchanged, although some ties might have been resolved, others might have been created. It is important to note that there is no unambiguous rule to transform the index values from one scale to another.

A more radical change in the scale could completely ruin the index. Measuring the distance in feet (or yard or meter) would result in j-indices practically equal to the days when the cyclist sat in the seat at all; measuring in light-years would set all j-indices to zero.

Apparently, there is somewhere an optimal scale, where the h-index can be used most effectively. The closer is this scale to some kind of "natural scale", the more evidently the practical use of the index offers itself.

In this paper an attempt is made to find at least an approximate transformation rule for the h-index if the scale of the underlying distribution is changed. Some practical applications of such a transformation are proposed, as well.

Methodology and results

Glänzel (Glänzel 2006) suggested a simple relation among the h-index, h, the sample size (in the'classical' Hirsch representation, the number of publications, n) and the density (mean citation rate per paper, x):

$$h = c.n^{1/3}x^{2/3}, \tag{1}$$

where c is a positive constant of the order of 1. If the random variable underlying the distribution studied is multiplied by a constant factor, λ, obviously, the mean value will be

$$x' = \lambda.x,$$

and, using Glänzel's formula, a transformation factor, $\kappa = \lambda^{2/3}$, can be used for estimating the h-index of the transformed distribution:

$$h' = \lfloor \kappa.h \rfloor,$$

$\lfloor\rfloor$ stands for the integer value ("floor") function.

"Rescaling" of the random variable (multiplication by a constant) is not supposed to change the constant, c, and the sample size therefore, their actual values are cancelled in the transformation process. In what follows, it will be studied on "real-life" examples whether and how this approximation can be used in the practice.

A model experiment

A model experiment was made on the h-index values of 8,162 journals covered by the Science Citation Index in 2001. A three-year citation period of 2001–2003 was considered. The h-index has been determined in the usual way and also measured in 10 citation and 0.1 citations (so to say, in "*deka*citations" and "*deci*citations"), respectively. The empirical and calculated values are given in Table 1.

There is a striking similarity between the observed and calculated values in both directions, although the significance test shows statistically significant difference in the "decicitation" scale. A closer look at the data shows that the main reason of the deviation

Table 1 The empirical and calculated h-index values using original and transformed citation scales (mean values ± standard deviations)

Original citation scale	
Empirical h-index	5.527 ± 6.112
Measured in 10 citations	
Empirical h-index	0.734 ± 1.281, $n = 8{,}162$
Calculated h-index ($h' = \text{integer}(0.1^{2/3} \cdot h)$)	0.729 ± 1.310, $n = 8{,}162$
Student's t value, significance of difference	0.247, NO ($p = 0.8053$)
Measured in 0.1 citations	
Empirical h-index	22.067 ± 26.283, $n = 8{,}162$
Calculated h-index ($h' = \text{integer}(10^{2/3} \cdot h)$)	25.123 ± 28.378, $n = 8{,}162$
Student's t value, significance of difference	7.138, YES ($p < 0.0001$)
Measured in 0.1 citations (journals with at least 100 papers)	
Empirical h-index	46.446 ± 37.936, n = 2,249
Calculated h-index [h' = integer($10^{2/3}$·h)]	47.865 ± 41.027, n = 2,249
Student's t value, significance of difference	1.2043, NO (p = 0.2285)

is that the "upscaled" h-index values of smaller journals approach the total number of papers, that is, the majority of the papers belong to the "h-core". By restricting the study to journals with at least 100 papers (italicized area in Table 1), the difference becomes insignificant.

Anyway, this example clearly shows that the approximation has its natural limits (in this case about one order of magnitude in both directions). Measuring the citations in 100 and 0.01 citations ("*hekto*citations" and "*centi*citations"), respectively, not only the approximation but the usability of the h-index gets completely ruined.

Field normalization of the h-index

An example of rescaling of the h-index has been published earlier by Iglesias and Pecharromán (Iglesias and Pecharromán 2007). Starting from Glänzel's results (Glänzel 2006) they, actually, reached the same transformation formula: the normalization factor is the ratio of the mean citation rates on the power 2/3 (Equation 18 in the cited paper). They used the formula to normalize the Thomson-Reuters Essential Science Indicators (ESI) field to the field of Physics (http://thomsonreuters.com/essential-science-indicators/). The normalization factors were then used to compare the h-index values of highly cited Spanish scientists active on various fields. The exercise appeared to be successful, however, a systematic validation of the method have not been attempted ever since.

For a validation study we used all journals from the 2001 list mentioned above that could unambiguously be assigned to one of the 21 science fields used in ESI (a total of 6,929 titles). The full ESI journal coverage can be found at http://incites-help.isiknowledge.com/incitesLive/ESIGroup/overviewESI/scopeCoverageESI.html. Similarly to Ref. (Iglesias and Pecharromán 2007), the field "Multidisciplinary" was excluded from the study because of its inherent heterogeneity and extremely small size. The frequency distribution of h-index values of the journals was then determined for each field (see Fig. 1).

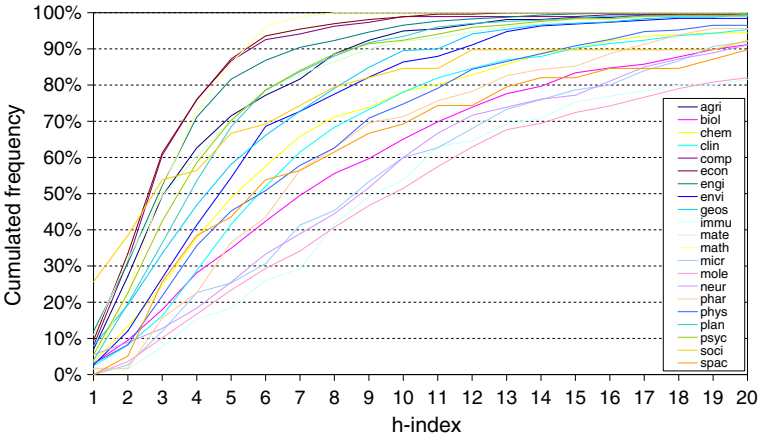

Fig. 1 The cumulated frequency distribution of journal h-index values for the 21 ESI fields

The presentation hardly makes possible to identify and study the single fields separately; the main message of the figure is that the curves fairly uniformly fill the chart area, i.e., the distributions are rather diverse.

In order to determine the normalization factor, λ, the mean citation rate of the total sample, x_{total}, and of each field, x_{field}, were calculated: $\kappa_{\text{field}} = (x_{\text{total}}/x_{\text{field}})^{2/3}$. The results are given in Table 2.

Figure 2 shows the frequency distribution of the normalized journal h-index values.

The normalized curves are spectacularly bunched together showing the efficiency of this simple normalization method and thereby strongly supporting the findings of Iglesias and Pecharromán (Iglesias and Pecharromán 2007). Apart from some errant sections in the curves of smaller fields stemming mainly from random fluctuations, the only substantially deviant field is physics (marked by a thick dashed line in Fig. 2). (In this sense, normalization to physics was not the most fortunate choice in Ref. (Iglesias and Pecharromán 2007).) A possible reason of this deviance and a correction option will be dealt with in the next section.

Sample size correction

According to Eq. (1), the h-index is depending also on the sample size, n, i.e. in the case of journal h-index, the number of papers in the journal. Conspicuously, the average annual number of papers in physics journals is extremely high: more than 300, while the overall average is about 120. Equation (1) suggests a simple way for correction. Since sample size, n, is on the 1/3 power in the formula, a correction factor similar to κ can be constructed as $\pi_{\text{field}} = (p_{\text{total}}/p_{\text{field}})^{1/3}$, where p is the average number of papers per journal. For physics, $\pi = 0.725$, and a sample-size corrected, normalized h-index can be calculated as $h'_{\text{corr}} = \pi \cdot h'$.

As can be seen in Fig. 3, the correction places physics back to the middle of the bunch. Since for most other fields the average annual number of papers per journals is substantially closer to the overall average, and the power of 1/3 strongly reduces the correction effect, the sample size correction is insignificant there.

Table 2 Mean citation rates and the normalization factor, κ, of the ESI fields

Field	Abbreviation	Mean citation rate	κ
Agricultural sciences	Agri	2.209	1.448
Biology & biochemistry	Biol	6.928	0.676
Chemistry	Chem	3.888	0.993
Clinical medicine	Clin	4.305	0.928
Computer science	Comp	1.416	1.948
Economics & business	Econ	1.264	2.101
Engineering	Engi	1.404	1.959
Environment/Ecology	Envi	2.952	1.194
Geosciences	Geos	3.090	1.158
Immunology	Immu	9.126	0.562
Materials science	Mate	2.136	1.481
Mathematics	Math	1.073	2.344
Microbiology	Micr	6.170	0.730
Molecular biology & genetics	Mole	11.351	0.486
Neuroscience & behavior	Neur	7.024	0.670
Pharmacology & toxicology	Phar	4.544	0.895
Physics	Phys	2.964	1.190
Plant & animal science	Plan	2.435	1.357
Psychiatry/Psychology	Psyc	3.085	1.159
Social sciences, general	Soci	1.314	2.047
Space science	Spac	7.331	0.651
All fields		3.850	

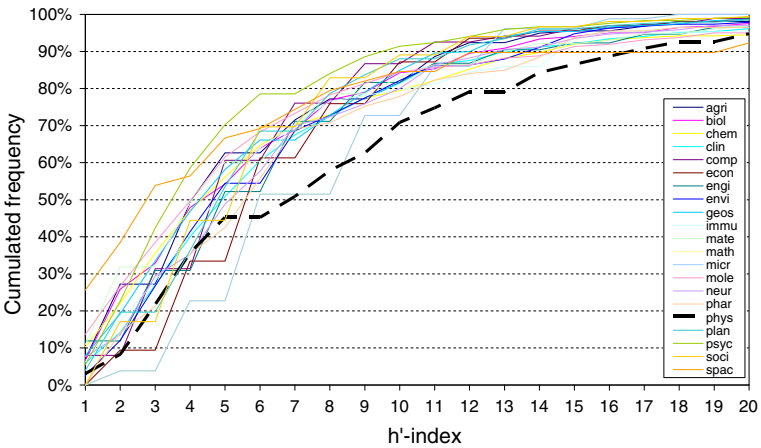

Fig. 2 The cumulated frequency distribution of the normalized journal h-index values (h') for the 21 ESI fields

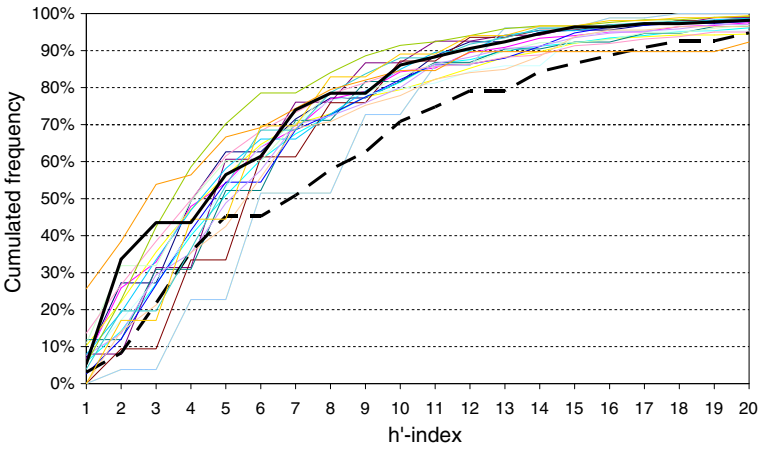

Fig. 3 The cumulated frequency distribution of the normalized journal h-index values (h') for the 21 ESI fields. Same as Fig. 2, with the sample-size corrected physics values added (marked by a *thick solid line*). For this latter, the horizontal axis is, obviously, h'_{corr}

Discussion and conclusions

It was shown that Glänzel's (Glänzel 2006) simple formula [Eq. (1)] is a suitable basis for estimating the h-index when the variable of the underlying distribution is rescaled. Multiplying the random variable by a constant factor, λ, the mean value will obviously be

$$x' = \lambda.x,$$

and a good approximation of the transformed h-index can be given as

$$h' = \lfloor \lambda^{2/3}.h \rfloor.$$

The validity of the estimation has its limits: within a multiplication or division factor of 10 the estimation seems to be effective, provided that the sample sizes and the maximum values of the variable remain commensurable. By setting it into a wider framework, the results support the field normalization method proposed by Iglesias and Pecharromán (Iglesias and Pecharromán 2007).

It should be stressed that the method cannot be applied for linear transformations with non-zero intercept (e.g., rescaling temperature from Celsius to Fahrenheit scale); it works, however, between the Celsius and Réaumur scales which are connected by a simple proportionality factor.

Equation (1) also suggests a method to take sample-size correction into account. This correction proved to be effective in normalizing the h-index values of physics journals.

The method may have real significance while comparing h-index values (whether of journals, individuals or other actors) between different fields, time frames, etc.

Acknowledgments This work was supported by the European Commission under the FP7 Grant No. 266588 (SISOB project) and the FP7 Grant No. 613202 (IMPACT-EV project).

References

Edwards, A. W. F. (2005). System to rank scientists was pedalled by Jeffreys. *Nature, 437*, 951.

Glänzel, W. (2006). On the h-index: A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics, 67*(2), 315–321.

Hirsch, J. E. (2005). An index to quantify an individual's scientific output. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 16569–16572.

Iglesias, J. E., & Pecharromán, C. (2007). Scaling the h-index for different scientific ISI fields. *Scientometrics, 73*(3), 303–320.

Korn, A., Schubert, A., & Telcs, A. (2009). Lobby index in networks. *Physica A, 388*, 2221–2226.

Schubert, A., Korn, A., & Telcs, A. (2009). Hirsch-type indices for characterizing networks. *Scientometrics, 78*(2), 375–382.

Zhao, S. X., Rousseau, R., & Ye, F. Y. (2011). h-Degree as a basic measure in weighted networks. *Journal of Informetrics, 5*, 668–677.