Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 2–7
Journal of INFORMETRICS An International Journal
www.elsevier.com/locate/joi

The influence of missing publications on the Hirsch index

Ronald Rousseau [a,b,c]

[a] *KHBO (Association K.U.Leuven), Department of Industrial Sciences and Technology, Zeedijk 101, B-8400 Oostende, Belgium*
[b] *University of Antwerp (UA), IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium*
[c] *Hasselt University (UHasselt), Agoralaan, Building D, B-3590 Diepenbeek, Belgium*

Received 2 May 2006; received in revised form 30 May 2006; accepted 30 May 2006

Abstract

We show that usually the influence on the Hirsch index of missing highly cited articles is much smaller than the number of missing articles. This statement is shown by a combinatorial argument. We further show, by using a continuous power law model, that the influence of missing articles is largest when the total number of publications is small, and non-existing when the number of publications is very large. The same conclusion can be drawn for missing citations. Hence, the *h*-index is resilient to missing articles and to missing citations.

Keywords: Hirsch index; Influence of missing data; Power law model; Robustness

1. Introduction

Recently the Hirsch index, in short: *h*-index, has attracted a lot of attention in the scientific community (Bar-Ilan, 2006; Egghe, in press; Glänzel, 2006; Liang, 2006). This index, introduced by Hirsch (2005) is calculated as follows. Consider the list of publications co-authored by scientist S, ranked according to the number of citations each of these has received over a given period. Then scientist S' *h*-index is *h* if it is the largest natural number such that the first *h* publications received each at least *h* citations. Clearly, this definition can also be applied to some other source-item pairs, besides a scientist's publications and citations (Braun et al., 2005; Egghe & Rousseau, 2006; Rousseau, 2006).

In most applications citations have been taken into account only if the corresponding articles have been published in a journal covered by the Web of Knowledge (Thomson Scientific). Yet, it is also possible to collect citations from the Web via Google Scholar (Bar-Ilan, 2006), or from a local database such as the CSCD in China (Liu Zeyuan, personal communication). Expressed in a conglomerate framework this means that the used pool is essential (Rousseau, 2005). It is indeed quite feasible that a scientist's most cited works are published in conference proceedings, free web-journals or, generally, in sources not covered by the Web of Knowledge. What then is the influence of these highly cited articles on a scientist's *h*-index?

E-mail address: ronald.rousseau@khbo.be.

1751-1577/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2006.05.001

2. A simple discrete model

It is assumed that the number of missing articles, denoted as m, contains s highly cited ones, this is: articles above the level of the h-index. Secondly, it is assumed that in the initial situation the article at rank h receives exactly h citations. Finally, it is assumed that in the neighbourhood of the original h-index the difference between the numbers of citations received by consecutive articles in the ranking is a fixed number. None of these assumptions is crucial for the point we want to make, namely that the difference in ranking resulting from the missing articles is never equal to the number of missing articles, and usually much smaller. We note already that if s is zero the h-index remains the same. This happens when all missing articles receive a relatively low number of citations.

Case A. The h-index falls in a zone where there are many articles ($>s$) that received the same number of citations. In that case inclusion of the missing articles will result either in the same h-index or in an h-index that is one unit higher. Table 1 illustrates Case A.

Case B. The h-index falls in a zone (we assume that the length of that part of this zone before h is larger than or equal to s) where consecutive articles differ by exactly one citation received. In that case inclusion of the missing articles will result either in an increase of the h-index by $s/2$ (if s is even) or by $(s-1)/2$ (s is odd).

Indeed, assume that $s=2n$. Then after inclusion of s highly cited articles, the article originally at rank h is now ranked $h+2n$, and has still h citations. The article that occurred at rank $h+1$, now has rank $h+2n+1$, and has $h-1$ citations. In general, the article now at rank $h+2n+k$ receives $h-k$ citations. Hence, in order to determine the new h-index we solve: $h+2n+k=h-k$, yielding $k=-n$. So the new h-index is $h+n=h+s/2$. If s is odd, say $2n+1$, then we have to solve $h+2n+1+k=h-k$, yielding $k=-(2n+1)/2$. Because the standard Hirsch index is a natural number, also in this case the change is only equal to $n=(s-1)/2$, and not equal to s as perhaps expected. Table 2 gives an example for $s=3$.

Case C. Larger gaps

We assume that, in the original ranking, before the article at rank h there is, over the zone of interest, always a gap between the number of citations equal to G (>1). Clearly, Case B is the case $G=1$.

Table 1
An illustration of Case A for $s=3$

(a) h-index remains unchanged

Rank	Number of citations	Rank (including s new highly cited articles)	Number of citations
...	...	...	...
$h-3$	h	$h-3$	*
$h-2$	h	$h-2$	*
$h-1$	h	$h-1$	*
h	h	h	h
$h+1$	h	$h+1$	h

(b) h-index changes one unit

Rank	Number of citations	Rank (including s new highly cited articles)	Number of citations
...	...	...	...
$h-4$	$h+1$	$h-4$	*
$h-3$	$h+1$	$h-3$	*
$h-2$	$h+1$	$h-2$	*
$h-1$	h	$h-1$	$h+1$
h	h	h	$h+1$
$h+1$	h	$h+1$	$h+1$
$h+2$	h	$h+2$	h

* Exact value does not matter.

Table 2
An illustration of Case B for $s = 2n + 1 = 3$

Rank	Number of citations	Rank (including s new highly cited articles)	Number of citations
...	...	...	...
$h-4$	$h+4$	$h-4$	*
$h-3$	$h+3$	$h-3$	*
$h-2$	$h+2$	$h-2$	*
$h-1$	$h+1$	$h-1$	$h+4$
h	h	h	$h+3$
$h+1$	*	$h+1$	$h+2$
$h+2$	*	$h+2$	$h+1$

Here the new h-index is equal to $h+1$.
* Exact value does not matter.

If $s < G$ then the article formerly situated at rank h is now at rank $h+s$ and still receives h citations. The article formerly situated at rank $h-1$ moves to rank $h+s-1$ and has $h+G$ citations. Consequently, if $s \leq G$ then the new h-index is equal to $h+s-1$. Note that this means that h remains unchanged if $s = 1$. If $G < s \leq 2G$ then, one can easily see that the new h-index is $h+s-2$, and if $2G < s \leq 3G$ then the new h-index is $h+s-3$. This shows that even for relatively large gaps the influence of s missing highly cited articles is not equal to s, but smaller. Table 3 provides an example.

3. A first example: Citations follow a Zipf distribution

If citations follow a Zipf distribution this means that the number of citations of the source at rank r is equal to Z/r. In this case the h-index is found by solving the equation $h = Z/h$, hence $h = \sqrt{Z}$. Taking h equal to a natural number means that h is equal to the largest natural number smaller than or equal to $\sqrt{Z}$. This number is known as the floor function of $\sqrt{Z}$ denoted as $\lfloor \sqrt{Z} \rfloor$. In Table 4 the h-index is calculated for some values of Z, as well as the number of citations (rounded) of the sources at rank $h-1$. This allows us to assess if this situation corresponds (approximately) to Cases A, B or C. Recall that $h(1) = Z$, which corresponds to the source with the highest production (in this example: the article receiving the most citations).

Table 4 shows that for these realistically highest numbers of citations (Z-values) the gaps are rather small so that if, for example, four highly cited publications are missing, then this would result in a change of the h-index equal to one to three.

Table 3
An illustration of Case C for $G = 3$ and $s = 8$

Rank	Number of citations	Rank (including $s = 8$ new highly cited articles)	Number of citations
...	...	...	...
$h-4$	$h+12$	$h-4$	*
$h-3$	$h+9$	$h-3$	*
$h-2$	$h+6$	$h-2$	*
$h-1$	$h+3$	$h-1$	*
h	h	h	*
$h+1$	*	$h+1$	*
$h+2$	*	$h+2$	*
$h+3$	*	$h+3$	*
$h+4$	*	$h+4$	$h+12$
$h+5$	*	$h+5$	$h+9$
$h+6$	*	$h+6$	$h+6$
$h+7$		$h+7$	$h+3$
$h+8$	*	$h+8$	h

The new h-index is equal to $h+s-2 = h+6$.

Table 4
The Zipf model for the *h*-index

Z	*h*	# citations at rank $h-1$
2000	44	47
1000	31	33
500	22	24
200	14	15
100	10	11
50	7	8

4. A second example: Price awardees

Leo Egghe has recently introduced an alternative for the *h*-index (Egghe, 2006a, 2006b, 2006c). This is not the subject of this note, but we will use his tables of *h*-indices of Price medallists to study the influence of missing publications on a scientist's *h*-index. We will assume that for each of them five highly cited articles are missing and we will recalculate their *h*-index, based on the data in (Egghe, 2006c). Results are shown in Table 5.

This example shows that for this list of real *h*-indexes an additional five articles yields an increase in the *h*-index value between zero and four. On average the increase is 2.21 or less than half the number of missing publications.

5. An analytical model based on a power law

In this section we show that a power law, i.e. a Lotka model, as used in an earlier publication (Egghe & Rousseau, 2006) leads to the same conclusion as the combinatorial argument presented above.

In this earlier publication we proved that if citations (or in general: item frequencies) can be described by a negative power law with exponent $\alpha > 1$, and if the system has T sources, then the *h*-index (actually its real-valued version, because in this approach the *h*-index is not a natural number anymore) is equal to

$$h_1 = T^{1/\alpha} \tag{1}$$

Adding m missing articles (sources), and assuming that this addition does not alter the exponent α, then the *h*-index becomes:

$$h_2 = (T+m)^{1/\alpha} \tag{2}$$

Note that it is possible to keep the Lotka exponent α constant while T is replaced by $T+m$. This is explained in (Egghe & Rousseau, 2006). The argument is based on Egghe (2005, II.2.1).

Table 5
h-indices of Price awardees and recalculated *h*-indices (denoted as h'), based on the assumption that for each of them five highly cited articles are missing

Price awardees	*h*-index	h'	Difference
Braun T.	25	27	2
Egghe L.	13	15	2
Garfield E.	27	29	2
Glänzel W.	18	21	3
Ingwersen P.	13	16	3
Leydesdorff L.	13	15	2
Martin B.	16	19	3
Moed H.F.	18	20	2
Narin F.	27	28	1
Rousseau R.	13	13	0
Schubert A.	18	21	3
Small H.	18	22	4
Van Raan A.F.J.	19	20	1
White H.D.	12	15	3

We want to prove that $h_2 - h_1 \ll m$ or $\Delta h/\Delta T = (h_2 - h_1)/m \ll 1$. In a continuous framework this means that we have to show that $\mathrm{d}h/\mathrm{d}T \ll 1$.

Proposition. *Using the notation explained above we have*: $\mathrm{d}h/\mathrm{d}T \ll 1$.

Proof. $\mathrm{d}h/\mathrm{d}T = \mathrm{d}(T^{(1/\alpha)})/\mathrm{d}T = 1/\alpha T^{((1-\alpha)/\alpha)}$. As $\alpha > 1$, the exponent $(1-\alpha)/\alpha$ is always negative. As T is assumed to be relatively large, this means that $T^{((1-\alpha)/\alpha)} \ll 1$ ($<\alpha$). This shows that $\mathrm{d}h/\mathrm{d}T \ll 1$.

In (Egghe & Rousseau, 2006) we have already shown that the h-index is a concavely increasing function of T, T being the total number of sources (keeping the Lotka exponent α constant). Moreover, $\mathrm{d}h/\mathrm{d}T$ is convexly decreasing with $\lim_{T\to\infty}(\mathrm{d}h/\mathrm{d}T) = \lim_{T\to\infty}(1/\alpha)T^{(1-\alpha/\alpha)} = 0$. These results imply that the influence of missing articles is largest for small T, and tends to zero the larger the number of sources.

Note that in the discrete case we focused on the number of highly cited missing articles, as the other ones clearly have no influence on the value of the h-index. The argument used in the continuous case only uses missing articles, whether or not they are highly cited. Again, this is just a matter of convenience as missing articles that are not highly cited have no influence on the value of the h-index.

In our earlier article (Egghe & Rousseau, 2006) we also derived a formula for the h-index as a function of the total number of items (citations in this article). This relation is, for $\alpha > 2$, given by:

$$h = \left(\frac{\alpha - 2}{\alpha - 1}A\right)^{1/\alpha}$$

where A denotes the total number of items. Clearly, also the number of missed citations does not have a large influence on the value of the h-index. Indeed, $(\mathrm{d}h/\mathrm{d}A) = (1/\alpha)((\alpha-2)/(\alpha-1))^{1/\alpha}A^{((1-\alpha)/\alpha)} \ll 1$, showing that, under the assumption of a fixed α-value, the h-index is stable under the influence of missing citations. □

Remarks.

1. We do not claim that citations always follow a power law, or that adding new articles automatically leads to a new power law with the same α-value. We just say that within this model the analytical results confirm the combinatorial result. Moreover, in this model the influence of missing articles with respect to the number of sources (T) is exactly as one would expect: largest (but still smaller than the number of missed articles) for small T and non-existing for large T.
2. Similar conclusions can be made for the g-index, as introduced by Egghe (2006a, 2006b). Indeed, Egghe has shown that for the power law model $g = ((\alpha-1)/(\alpha-2))^{((\alpha-1)/\alpha)}T^{1/\alpha}$ (Egghe, 2006c). In this model, the difference between the h-index and the g-index is only a factor depending on α. Hence, the g-index, considered as a function of T behaves in the same way as the h-index.

6. Conclusion

Contrary to what one might intuitively expect, a relative small number of missing highly cited publications has only a small influence on the value of the h-index. This is usually the case as shown by the examples of citations following a Zipf distribution, and the h-indices of Price medallists. An analytical model reinforces our argument for missing publications, as well as for missing citations. We conclude that the h-index is resilient to missing articles and to missing citations.

Acknowledgements

Research for this note was performed while the author was a guest of WISE-Lab, Dalian University of Technology and of the National Library of Sciences of CAS (Beijing). He thanks Profs. Liu Zeyuan and Jin Bihui for their hospitality. The author further thanks Prof. Leo Egghe (Hasselt University) for a number of helpful suggestions, improving the obtained results. Research for this article was supported by NSFC Grant Nr. 70373055.

References

Bar-Ilan, J. (2006). *H*-index for Price medallists revisited. *ISSI Newsletter*, *2*(1), 3–5.
Braun, T., Glänzel, W., & Schubert, A. (2005). A Hirsch-type index for journals. *The Scientist*, *19*(22), 8.
Egghe, L. (2005). *Power Laws in the Information Production Process: Lotkaian Informetrics*. Oxford, UK: Elsevier.
Egghe, L. (2006a). How to improve the *h*-index. *The Scientist*, *20*(3), 14.
Egghe, L. (2006b). An improvement of the *H*-index: the *G*-index. *ISSI Newsletter*, *2*(1), 8–9.
Egghe, L. (2006c). Theory and practice of the g-index. *Scientometrics*, *69*, 131–152.
Egghe, L. Dynamic *h*-index: the Hirsch index in function of time. *Journal of the American Society for Information Science and Technology*, in press.
Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch index. *Scientometrics*, *69*, 121–129.
Glänzel, W. (2006). On the *h*-index—a mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, *67*, 315–321.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA*, *102*(46), 16569–16572.
Liang, L. (2006). *H*-index sequence and *h*-index matrix: constructions and applications. *Scientometrics*, *69*(1), 153–159.
Rousseau, R. (2005). Conglomerates as a general framework for informetric research. *Information Processing and Management*, *41*, 1360–1368.
Rousseau, R. (2006). A case study: evolution of JASIS' *h*-index. E-LIS: ID-code 5430.