On the correction of the h-index for career length

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), IBW, Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium

leo.egghe@uhasselt.be

ABSTRACT

We describe mathematically the age-independent version of the h-index, defined by Abt (Scientometrics 91(3), 863-868, 2012) and explain when this indicator is constant with age.

We compare this index with the one where not the h-index is divided by career length but where all citation numbers are divided by career length and where we then calculate the new h-index. Both mathematical models are compared.

A variant of this second method is by calculating the h-index of the citation data, divided by article age. Examples are given.

[1]Permanent address
Key words and phrases: age-independent, h-index (Hirsch-index)

Introduction

Let us have a researcher with T publications and let c_i $(i = 1,...,T)$ be the number of received citations of paper *i*. We suppose that the papers are arranged in decreasing order of number of received citations (i.e. $c_i \geq c_j$ if and only if $i \leq j$). Then the Hirsch-index (Hirsch (2005)) (or h-index) is the largest rank $r = h$ such that all papers on ranks $i = 1,...,h$ have at least h citations (i.e. the largest rank $r = h$ such that $c_h \geq h$ and hence $c_i \geq h$ for all $i = 1,...,h$).

It is clear that the h-index is age-dependent (i.e. is dependent of career length). Long careers usually have higher values of T (total number of publications) and of c_i $(i = 1,...,T)$ (number of citations received by paper *i*), when compared to shorter careers (e.g. younger researchers). This fact was already noted in the defining paper Hirsch (2005).

So, with the h-index, one should not compare researchers with different career length (as we should also not compare researchers from different fields, but that is the case for all citation-based indicators – in this paper we do not deal with this problem).

This has lead Abt (2012) to the following "age-independent" h-index (also implicit in Hirsch (2005)). Denote by $h(t)$ the h-index of a researcher at career length t (starting at the time of the researcher's first published paper). Then define

$$a(t) = \frac{h(t)}{t/10} \tag{1}$$

, i.e. the h-index (at time t) divided by the (fractional) number of decades since the first published paper. The factor 10 is only useful in practical examples; in theoretical models we might use

$$a^*(t) = \frac{h(t)}{t} \tag{2}$$

as well.

Abt (2012) claims that $a(t)$ is constant in t and gives practical evidence for it. Incidentally, a constant $a(t)$ implies that $h(t)$ increases linearly, as is trivial from (1) or (2). In this paper, based on the models developed in Egghe and Rousseau (2006) and Egghe (2009), we give a mathematical model for $a(t)$ and $a^*(t)$ and present necessary and sufficient conditions for $a(t)$ and $a^*(t)$ to be constant. Practical data on $h(t)$ and $a(t)$ for this author's career are given. This is done in the next section.

In the third section we describe a second way to reach an "age-independent" indicator: we do not divide $h(t)$ by t (or $t/10$) but we divide all citation numbers c_i by this career length and then we calculate the h-index of this transformed set of data. We denote by $h(t)$ this indicator (if we use $t/10$) and by $b^*(t)$ this indicator (if we use t). Also for these indicators we present a mathematical model based on Egghe and Rousseau (2006) and Egghe (2008a, b) and give necessary and sufficient conditions for $b(t)$ or $b^*(t)$ to be constant. An example from this author's career is given. Also in this section we remark that a third indicator can be constructed to reach "age-independence". In the previous two cases we divide always by career length t or $t/10$ where we do not take into account the different ages of the published papers. So, theoretically, it makes more sense to divide the number of citations c_i of the i^{th} paper by the age of the paper which is 2012 given by

$$2012 - \text{publication year} + 1 \tag{3}$$

and then calculate the h-index of this transformed set of data. If we use "age/10" we denote this indicator $c(t)$ and if we use "age" we denote this indicator $c^*(t)$ (as we did with the previous two indicators). We prove relations (inequalities) between these three indicators and present an example from this author's career.

The paper closes with some conclusions and suggestions for further research.

A model for Abt's "age-independent" index

In Egghe and Rousseau (2006) we assumed that the paper-citation system is Lotkaian (Egghe (2005)) with Lotka exponent $\alpha > 1$. We there showed that, if there are T papers in total, that the h-index of this system is given by

$$h = T^{\frac{1}{\alpha}} \tag{4}$$

In Egghe (2009) we assumed that we have a number (density) of publications per time unit (at t) equal to dt^{β}, where $d > 0$, $\beta \geq 0$ (d was denoted b in Egghe (2009) but we avoid this in order not to confuse with the second indicator $b(t)$, discussed in the introduction). Note that the case $\beta = 0$ is the case where we have a constant number of publications per year. The total number of publications T, dependent on t (denoted as $T(t)$) is given by

$$T(t) = \int_0^t dt'^{\beta}\, dt' \tag{5}$$

$$T(t) = \frac{d}{\beta+1} t^{\beta+1} \tag{6}$$

Combining (4) and (6) (supposing α to be independent of t) and denoting the time-dependent h-index h by $h(t)$, yields

$$h(t) = \left(\frac{d}{\beta+1}\right)^{\frac{1}{\alpha}} t^{\frac{\beta+1}{\alpha}} \tag{7}$$

Note that $h(t)$ is a concavely increasing function of t if and only if $\beta+1<\alpha$, is linear in t if and only if $\beta+1=\alpha$ and is a convexly increasing function of t if and only if $\beta+1>\alpha$.

By definition of Abt's indicator $a(t)$ we have, by (1) and (7)

$$a(t)=10\left(\frac{d}{\beta+1}\right)^{\frac{1}{\alpha}}t^{\frac{\beta+1-\alpha}{\alpha}} \tag{8}$$

(and the same for $a^{*}(t)$ with the factor 10 deleted).

We now have Proposition 1, yielding a necessary and sufficient condition for Abt's claim to be valid.

<u>Proposition 1:</u> The indicator $a(t)$ is constant if and only if

$$\alpha=\beta+1 \tag{9}$$

The same result is true for $a^{*}(t)$.

Proof: This is trivially following from (8) □

<u>Note:</u> The case $\alpha=\beta+1$ is a classical informetric case. The most classical Lotka exponent is $\alpha=2$ (see Egghe (2005)). This implies that $\beta=1$ and by (7) we have a linearly increasing $h(t)$ function, much in line with Fig. 1 ($h(t)$ for this author's career). So Proposition 1 indicates that a constant $a(t)$ function is classical and hence supports the finding in Abt (2012).

We further have Proposition 2.

<u>Proposition 2:</u>

(i) $a(t)$ is convexly decreasing if and only if

$$\beta + 1 < \alpha \tag{10}$$

(ii) $a(t)$ is increasing if and only if

$$\beta + 1 > \alpha \tag{11}$$

(iii) $a(t)$ is convexly increasing if and only if

$$\beta + 1 \geq 2\alpha \tag{12}$$

(iv) $a(t)$ is concavely increasing if and only if

$$\alpha < \beta + 1 \leq 2\alpha \tag{13}$$

(v) $a(t)$ is linearly increasing if and only if

$$\beta + 1 = 2\alpha \tag{14}$$

The same results are true for $a^*(t)$

Proof: All these results follow trivially from (8). □

These results are illustrated by this author's citation data yielding Fig. 1 $(h(t))$ and Fig. 2 $(a(t))$.

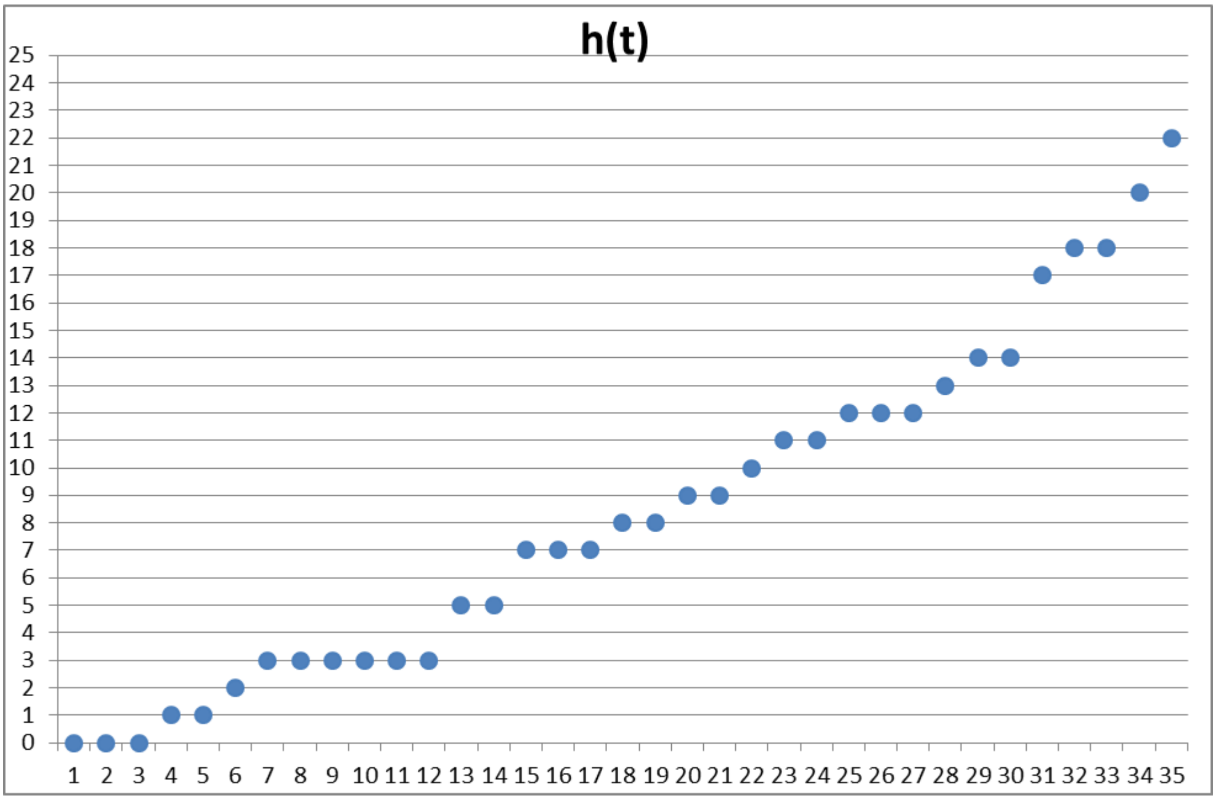

Fig. 1: h(t) sequence for this author's career

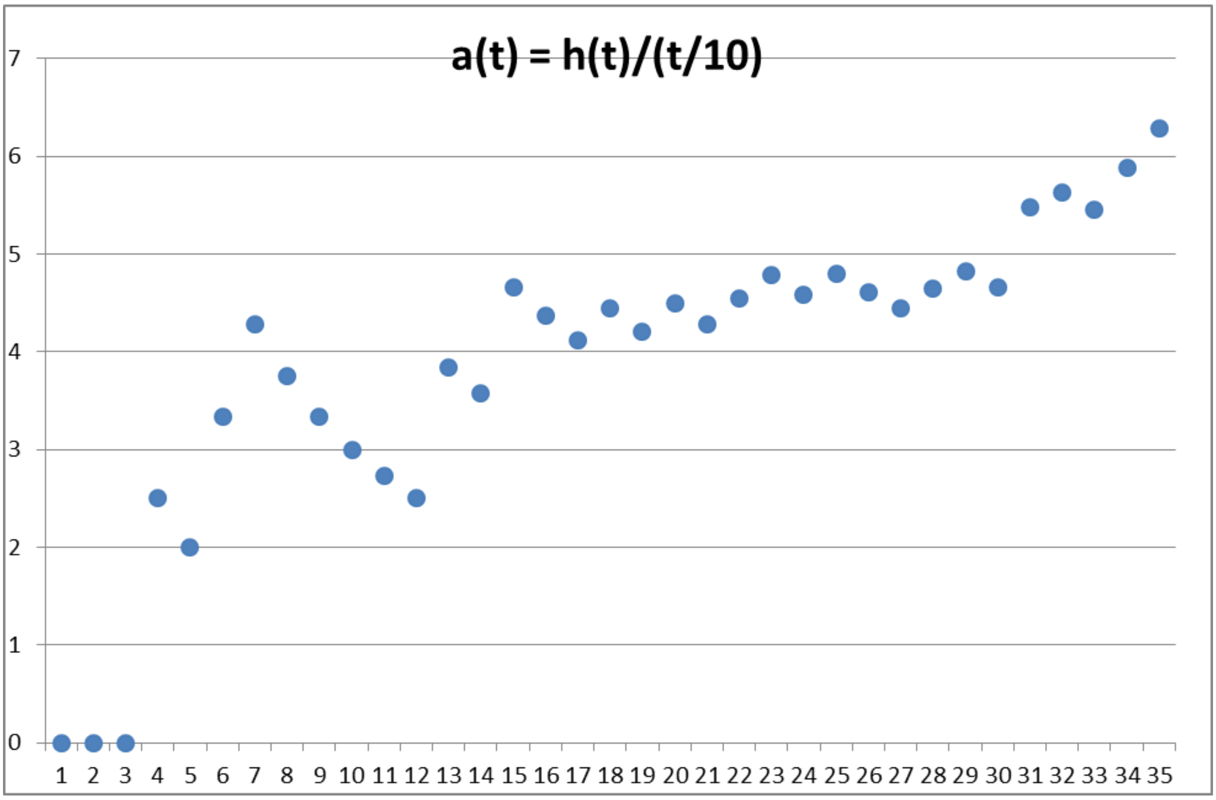

Fig. 2: a(t) sequence for this author's career

Fig. 1 is an updated version of Fig. 2 in Egghe (2009), updated to 35 career years ($t=1$ is 1978, the year of the first publication and $t=35$ is 2012, the present year). We can say that $h(t)$ is convexly increasing ($\beta+1>\alpha$) but is close to linear ($\beta+1\approx\alpha$) leading to an increasing $a(t)$ (see also Proposition 2 (ii)) but with a relatively constant middle part (see also Proposition 1).

An variant of Abt's "age-independent" index

There are several ways to correct the h-index for career length. One of them is Abt's index, discussed in the previous section: the simple idea there is by dividing the h-index by the career length t (or $t/10$).

A similar idea is as follows. We take the citation data c_i ($i=1,...,T$) and divide all these numbers by t (or $t/10$). For these "normalized" citation data we calculate the h-index. We denote this "age-independent" index by $b(t)$ (using $t/10$) and by $b^*(t)$ (using t). To model $b(t)$ and $b^*(t)$ we invoke a result from Egghe (2008 a, b) on transformations of the h-index.

<u>Proposition 3 (Egghe (2008 a, b)):</u> Let h be the h-index of the system c_i ($i=1,...,T$): the number of received citations for paper *i*, where there are T papers in total. If we do not change the total number T of papers but if we multiply each c_i by a positive number B, then the h-index h^* of this transformed system is given by

$$h^* = B^{\frac{\alpha-1}{\alpha}} h \tag{15}$$

where α is the Lotka exponent of the original system.

So we have the following formulae for $b(t)$ and $b^*(t)$.

$$b(t) = \left(\frac{1}{t/10}\right)^{\frac{\alpha-1}{\alpha}} h \tag{16}$$

$$b^*(t) = \left(\frac{1}{t}\right)^{\frac{\alpha-1}{\alpha}} h \tag{17}$$

Combining (16) with (7) yields

$$b(t) = 10^{\frac{\alpha-1}{\alpha}} \left(\frac{d}{\beta+1}\right)^{\frac{1}{\alpha}} t^{\frac{\beta-\alpha+2}{\alpha}} \tag{18}$$

and similarly for $b^*(t)$ (with $10^{\frac{\alpha-1}{\alpha}}$ deleted). From (8) and (18) we now see that

$$b(t) = 10^{-\frac{1}{\alpha}} a(t) t \tag{19}$$

and similarly

$$b^*(t) = a^*(t) t \tag{20}$$

Note that, since $t \geq 1$, it follows from (20) that

$$b^*(t) \geq a^*(t)$$

for all t.

This shows that $b(t)$ increases faster than $a(t)$ (and similar for $b^*(t)$ and $a^*(t)$). We have the following propositions, similar to Propositions 1 and 2.

<u>Proposition 4:</u> The indicator $b(t)$ is constant if and only if

$$\alpha = \beta + 2 \tag{21}$$

The same result is true for $b^*(t)$.

<u>Proposition 5:</u>

(i) $b(t)$ is convexly decreasing if and only if

$$\beta + 2 < \alpha \tag{22}$$

(ii) $b(t)$ is increasing if and only if

$$\beta + 2 > \alpha \tag{23}$$

(iii) $b(t)$ is convexly increasing if and only if

$$\beta + 2 \geq 2\alpha \tag{24}$$

(iv) $b(t)$ is concavely increasing if and only if

$$\alpha < \beta + 2 \leq 2\alpha \tag{25}$$

(v) $b(t)$ is linearly increasing if and only if

$$\beta + 2 = 2\alpha \tag{26}$$

The same results are true for $b^*(t)$.

The proofs of Propositions 4 and 5 follow trivially from (18) and the similar result for $b^*(t)$.

From Proposition 4 we see that, for the classical Lotka exponent $\alpha = 2$ we have that $b(t)$ is constant if $\beta = 0$. From (7) it follows that $h(t)$ is concavely increasing (since $\alpha > 1$).

The calculation of $b(t)$ is illustrated on this author's data for $t = 35$ (the year 2012). The citation data are as in Table 1. From this table we see that $h = h(35) = 22$. Hence, by (1)

$$a(35) = \frac{h(35)}{35/10} = 6.2857 \tag{27}$$

For $b(35)$ we have to divide the c_i -values by $35/10$. This is done in Table 2 from which it follows that

$$b(35) = 10 \tag{28}$$

Table 1. Citation data of this author, retrieved from the Web of Science on August 21, 2012 yielding $h = h(35) = 22$

i	c_i
1	258
2	137
3	106
4	61
5	60
6	52
7	48
8	41
9	41
10	35
11	31
12	31
13	30
14	29
15	26
16	26
17	24
18	24
19	24
20	24
21	23
22	23
23	20

Table 2. Citation data from Table 1, divided by 3.5, yielding $b(35) = 10$.

i	$\frac{c_i}{3.5}$
1	73.71
2	39.14
3	30.29
4	17.43
5	17.14
6	14.86
7	13.71
8	11.71
9	11.71
10	10
11	8.86

Both methods of normalizing the h-index use the career length t. There is a third normalizing method. It is the same as the one yielding $b(t)$ (or $b^*(t)$) but instead of dividing each c_i by $t/10$ (or t) we now divide each c_i by the (article age)/10 respectively by the article age and calculate the h-index of this new table. The new "age-independent" indices are denoted $c(t)$ and $c^*(t)$ respectively. Note that, by definition, $c(t) \geq b(t)$ and $c^*(t) \geq b^*(t)$ since article age $\leq t$.

The disadvantage of this third method is that it is complex to calculate: we have to check every article and the obtained new table is not decreasing anymore. A manual control of this author's citation data in the Web of Science on August 21, 2012 showed that $c(35) = 19$. Here $c(35) < h(35) = 22$ but the simple example in Table 3 shows that $c(t) > h(t)$ is possible. This occurs when articles are cited in a fast way. This is a good property.

Table 3. Example of $h(t)=1<c(t)=2$

i	c_i	age	$\frac{c_i}{age/10}$
1	10	1	100
2	1	1	10

Discussion on the three correction methods

As discussed above, we have selected three methods for correcting the h-index for career length. The first method is Abt's original proposal (Abt (2012)) by simply dividing the h-index of a researcher by the career length. A second method that is presented here is to divide all citation data by this career length and then calculate the h-index of this set of "normalized" data. A third and last method that is presented here is to divide each citation number by the age of the cited paper and then calculate the h-index of this set of "normalized" data.

Clearly the first method is the simplest, followed by the second method. The third method is the most logical one (since each article's citation number is divided by its age) but is the most intricate one since the order of the normalized citation data is different from the original one and hence one is obliged to consider all papers of the researcher. In this sense, the second method is an acceptable alternative for the third method since one divides each article's citation number by the career length. This is more logical than simply dividing the h-index by the career length as in the first method (Abt's method). The first method is not logical in this sense since this simple method indicates that a normalization is obtained by simply dividing the h-index by career length which would only be logical if the h-index is a linear function of time (career length) which is, by (7), not always the case.

From the above discussion one would be inclined to say that the second method is to be preferred since it is more logical than the first one and simpler than the third one. However, from (20):

$$a^*(t) = \frac{b^*(t)}{t} \qquad (29)$$

indicating that the first method is "equivalent" to the second method since it is obtained by dividing the normalized h-index $b^*(t)$ by the career length as indicated in (29). So the first method (although too simple in its definition) performs equally well as the second method, hence the first method should be preferred due to its simplicity. We also repeat that the first method yields a constant function (as indicated by Abt) in a classical informetrics case: $\alpha = \beta + 1$ (see Proposition 1) which is e.g. the case for $\alpha = 2$ (most classical Lotka exponent – see Egghe (2005)) and $\beta = 1$ (linear growth of the h-index).

Conclusions and suggestions for further research

This paper presented a mathematical model for the "age-independent" indicator of Abt. We give characterizations of the different shapes of this function of the career length t, amongst which a characterization of when this function is constant. We show that this happens in classical informetric cases giving evidence to Abt's claim that this indicator often is t-independent.

A second type of age-independent indicator is obtained by not dividing the h-index by t but by dividing each citation number c_i by t and then calculating the h-index of this transformed table. Also for this indicator, a mathematical model is presented and characterizations of the different shapes are given, amongst which a characterization of when this function is constant. Both models are also compared.

A third method of normalizing the h-index for career length t is as in the second method but, instead of dividing every citation number by t we divide by the age of each article. Although it is more difficult to calculate this third index it has the good property that, the faster articles are cited, the higher this index becomes.

We encourage the reader to conduct further experiments on these three types of "age-independent" indices and to define new variants of these three methods.

References

H.A. Abt (2012). A publication index that is independent of age. Scientometrics 91(3), 863-868.

L. Egghe (2005). Power Laws in the Information Production Process: Lotkaian Informetrics. Elsevier, Oxford, UK.

L. Egghe (2008a). The influence of transformations on the h-index and the g-index. Journal of the American Society for Information Science and Technology 59(8), 1304-1312.

L. Egghe (2008b). Examples of simple transformations of the h-index: Qualitative and quantitative conclusions and consequences for other indices. Journal of Informetrics 2(2), 136-148.

L. Egghe (2009). Mathematical study of h-index sequences. Information Processing and Management 45(2), 288-297.

L. Egghe and R. Rousseau (2006). An informetric model for the Hirsch-index. Scientometrics 69(1), 121-129.

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