Modelling successive h-indices

Peer-reviewed author version

EGGHE, Leo (2008) Modelling successive h-indices. In: SCIENTOMETRICS, 77(3). p. 377-387.

DOI: 10.1007/s11192-007-1968-5
Handle: http://hdl.handle.net/1942/7917

Modelling successive h-indices

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium

leo.egghe@uhasselt.be

ABSTRACT

From a list of papers of an author, ranked in decreasing order of the number of citations to these papers one can calculate this author's Hirsch index (or h-index). If this is done for a group of authors (e.g. from the same institute) then we can again list these authors in decreasing order of their h-indices and from this, one can calculate the h-index of (part of) this institute. One can go even further by listing institutes in a country in decreasing order of their h-indices and calculate again the h-index as described above. Such h-indices are called by Schubert [Scientometrics 70(1), 201-205, 2007] "successive" h-indices.

[1] Permanent address

Key words and phrases: successive h-indices, distribution, global h-index
Acknowledgement. The author is grateful to M. Goovaerts for constructing the merged citation tables of Price awardees.

In this paper we present a model for such successive h-indices based on our existing theory on the distribution of the h-index in Lotkaian informetrics. We show that, each step, involves the multiplication of the exponent of the previous h-index by $\frac{1}{\alpha}$ where $\alpha > 1$ is a Lotka exponent. We explain why, in general, successive h-indices are decreasing.

We also introduce a global h-index for which tables of individuals (authors, institutes,…) are merged.

We calculate successive and global h-indices for the (still active) D. De Solla Price awardees.

I. Introduction

In a remarkable paper, Schubert (see Schubert (2007)) introduces the notion of "successive" h-indices. The principle is: calculate an h-index from a set of h-indexes. For the well-known definition of the h-index we refer to Hirsch (2005). A basic example uses the h-index of authors (as described above). If we have a group of authors (e.g. working in the same institute) and if we calculate the h-index for every author as described above (level 1), we can rank these authors in decreasing order of their h-indices. On this we can calculate a new h-index (same definition), now representing (whole or part of) the institute (level 2).

We can even go further: if authors are coming from different institutes (say of the same country) we can, in this way, calculate the h-indices of the different institutes (level 2), rank these institutes in decreasing order of their h-indices and then calculate the h-index of this table (level 3), i.e. of this (aspect of this) country. We can even rank countries according to their h-index and calculate the h-index (level 4) of this table.

We have here a chain as follows:

$$\text{countries} \rightarrow \text{institutes} \rightarrow \text{authors} \rightarrow \text{papers} \rightarrow \text{citations}$$

Another example is given in Schubert (2007):

countries → publishers → journals → articles → citations

(in each case “citations” refer to citations given to articles (papers) of authors/journals, respectively).

Without realizing it at the time of the writing of Egghe (2007), this paper is the basis for the modelling of such successive h-indices. This will be developed in the present paper. In Section II, we present concrete formulae for the calculation of successive h-indices in the case of Lotkaian informetrics at the different source-item levels. The use of Lotkaian informetrics is not controversial (see Egghe (2005), Chapter 1)) since the wide applicability, including networks (see the many references in Egghe (2005)) and even if Lotka’s law is not fully applicable, it serves as a first approximation of the real situation. In this model we show why the h-indices, at the different levels, in general, are decreasing when the level increases. We show that the h-index h_i on level i $(i = 1,2,3,...)$ equals

$$h_i = (T_{i-1})^{\frac{1}{\alpha_0 \alpha_2 ... \alpha_{i-1}}} \tag{1}$$

where T_{i-1} denotes the total number of objects at level $i-1$ and $\alpha_0, \alpha_2, ..., \alpha_{i-1}$ are the Lotka exponents for the Information Production Processes (IPPs) (source-item relations) (see Egghe (2005)) at the level $0,1,2,...,i-1$ (more details further on). We also explain why the lists of h-indices at level i show lower values than at level $i-1$ for all i. This is illustrated by the results (Table 1,2,3) in Schubert (2007).

These different aggregate levels gave us the idea of defining a global h-index, i.e. where all articles (no matter of which level they are generated) are counted as belonging to the same “meta author”. Also for this h-index (denoted h') we derive a model and we show that

$$h' \geq h_i \tag{2}$$

for all $i = 1,2,3,...$ and for all objects at the different aggregate levels, i.e. h' is larger than all h_1-indices of all authors, all h_2-indices of all institutes, all h_3-indices of all countries, and so on. This is done in Section III.

The different successive and global h-indices are then illustrated (in Section IV) on the group of (still active) D. De Solla Price award winners (data from Egghe (2006)).

The paper ends with conclusions and suggestions for further research.

II. Model for successive h-indices

Our model for successive h-indices is based on our theory on distributions of the h-index (Egghe (2007)) and on general formulae for the h-index in Lotkaian systems as proved in Egghe and Rousseau (2006). We give an overview of the results in the sequel.

Suppose that the articles-citation size-frequency function (e.g. of an author) is Lotkaian:

$$f(j) = \frac{C}{j^{\alpha}} \qquad (3)$$

where $f(j)$ is the article density with citation density j and we will use $j \geq 1$, $C > 0$, $\alpha > 1$ (see Egghe (2005)). Also from Egghe (2005) we have that T, the total number of articles, is given by

$$T = \frac{C}{\alpha - 1} \qquad (4)$$

In this case, it is shown in Egghe and Rousseau (2006) that the h-index is given by

$$h = T^{\frac{1}{\alpha}} \qquad (5)$$

Staying within the framework of Lotkaian informetrics we assume that the author-article IPP has a size-frequency function as follows

$$f_1(T) = \frac{D}{T^{\alpha_1}} \tag{6}$$

($T \geq 1$: only authors with at least one article are considered, and $D > 0$, $\alpha_1 > 1$): we are at level 1; from now on the article-citation relationship is considered at level 0 and, henceforth, we will denote $\alpha = \alpha_0$ in (3), (4) and (5) (and h in (5) will be denoted by h_1 to indicate level 1).

Then it is proved in Egghe (2007) that $\varphi_1(h_1)$ being the density of the number of authors with h-index h_1 equals

$$\varphi_1(h_1) = \frac{E}{h_1^{\alpha_0\alpha_1}} \tag{7}$$

where α_0, α_1 are as above and where

$$E = D\frac{\alpha_0\alpha_1 - 1}{\alpha_1 - 1} \tag{8}$$

Hence we refind Lotka's law with exponent $\alpha_0\alpha_1$.

This will be applied to calculate the h-index for higher aggregate levels 2,3,… .

The h-index h_2

The h-index h_2 is defined in Schubert (2007) as the h-index of the ranked list of authors in decreasing order of their h_1 values. Hence this is the h-index of the IPP whose frequency function is given by (7).

It is clear that $\alpha_0\alpha_1 > 1$ since $\alpha_0 > 1$ and $\alpha_1 > 1$. Then it is clear that (same argument as for (4) which is given in Egghe (2005)) the total number of authors (denoted S) is given by (use (6))

$$S = \frac{D}{\alpha_1 - 1} \tag{9}$$

and, because of (8)

$$S = \frac{E}{\alpha_0\alpha_1 - 1} \tag{10}$$

Since we have the law of Lotka on level 1 (formula (8)) we can apply (5) to level 1 in order to obtain the h-index for level 2 (h-index for a group of authors, e.g. an institute), denoted by h_2:

$$h_2 = (S)^{\frac{1}{\alpha_0\alpha_1}} \tag{11}$$

$$h_2 = \left(\frac{D}{\alpha_1 - 1}\right)^{\frac{1}{\alpha_0\alpha_1}} \tag{12}$$

So the h-index h_2 on level 2, successive to the h-index h_1 on level 1 follows in a remarkably simple way from the theory of distributions of the h-index, presented in Egghe (2007). It is nice to see that the theory is ready before the application to successive h-indices is made!

In the same way we can define the h-index at levels 3,4,… . What follows is level 3.

<u>The h-index h_3</u>

Now we replace (3) by (7) (density of the number of authors with h-index h_1) and introduce (13), replacing (6):

$$f_2(S) = \frac{F}{S^{\alpha_2}} \tag{13}$$

being the density of the institutes with author density $S \geq 1$ $(F > 0,\ \alpha_2 > 1)$.

We again apply Egghe (2007), obtaining that the density of the number of institutes with h-index h_2 is given by (using (11) and (13))

$$\varphi_2(h_2) = \frac{G}{h_2^{\alpha_0\alpha_1\alpha_2}} \tag{14}$$

where $\alpha_0, \alpha_1, \alpha_2$ are as above and where

$$G = F\frac{\alpha_0\alpha_1\alpha_2 - 1}{\alpha_2 - 1} \tag{15}$$

It follows from (13) (using again (4) at this level) that there are R institutes in total

$$R = \frac{F}{\alpha_2 - 1} \tag{16}$$

$$R = \frac{G}{\alpha_0\alpha_1\alpha_2 - 1} \tag{17}$$

(by (15)). Applying again (5), noting that (14) is Lotkaian, we have that the group of institutes (e.g. in a country) has a h-index (denoted h_3, indicating level 3) given by

$$h_3 = R^{\frac{1}{\alpha_0\alpha_1\alpha_2}} \tag{18}$$

It is now clear how to proceed to higher aggregation levels 4,… .

Evaluation of the successive h-indices

The formula (5) gives the h-index h_1 on the level 1 ($\alpha_0 =: \alpha$ and $h_1 =: h$): $h_1 = T^{\frac{1}{\alpha_0}}$ for an author with T papers. Formula (11) gives the h-index h_2 on the level 2: $h_2 = S^{\frac{1}{\alpha_0\alpha_1}}$ for an institution with S authors. Formula (18) gives the h-index h_3 on the level 3: $h_3 = R^{\frac{1}{\alpha_0\alpha_1\alpha_2}}$ for a country with R institutes.

In this sequence we can define the article-citation relation as being on level 0 and the h-index on level 0 can simply be defined $h_0 = C$ for a paper with C citations. We hence have a scheme as in Fig. 1

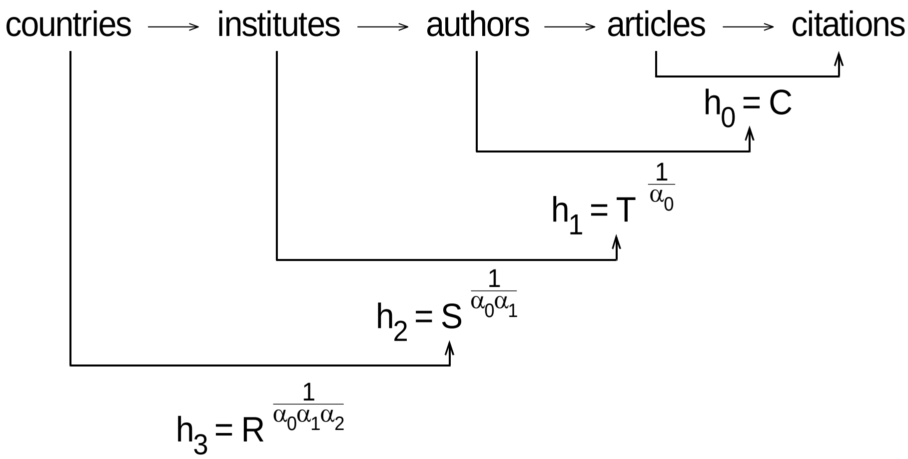

Fig. 1 Successive h-indices

In each case a list of h_i-values is produced per object (country, institute,…). Three such lists can be seen in Schubert (2007) for the levels: countries → publishers → journals (and, of course journals → articles → citations).

A look at these tables directly reveals lower h_i values for higher levels i, but a direct comparison is not possible since the ranks are for different objects (countries, publishers, journals) in the different tables. The models above give, however, a clear rationale for these lower h_i values the higher the level i.

1. If we assume that top countries have the top institutes and that the top institutes have the top authors then we will have, in most top cases:

$$R < S < T < C$$

(heuristic argument)

2. Since $\alpha_0,\ \alpha_1,\ \alpha_2 > 1$ we have

$$1 > \frac{1}{\alpha_0} > \frac{1}{\alpha_0\alpha_1} > \frac{1}{\alpha_0\alpha_1\alpha_2}$$

(exact argument). This even explains why, in general, $h_3 = h_2 = h_1$ since, for comparable R, S and T-values, h_{i+1} follows from h_i by taking a new root (e.g. square root for α_i » 2) ($i = 0,1,2$).

3. By the very definition, h_{i+1} is the h-index of the ranked h_i-table, hence h_{i+1} is a h_i-value somewhere inside the h_i-table and hence $h_{i+1} =$ top h_i-values which explains, heuristically, the decreasing values in consecutive tables.

<u>III. The global h-index</u>

Independent of the aggregation level: country or institute, for any group of authors, we can define a "global h-index" as the h-index of the "meta-author" which is composed of all the articles (and their citations) of the different authors in a group. This is similar (but in another context) with the definition of the global impact factor being the impact factor of a "meta-journal" composed of all the articles of the journals in a group: see Egghe and Rousseau (1996a,b).

In order to construct a model for this global impact factor, we first need two lemmas.

<u>Lemma III.1</u>: Suppose we have situation (3) and (4):

$$f(j) = \frac{T(\alpha - 1)}{j^{\alpha}} \tag{19}$$

with $\alpha > 1$, the article-citation size-frequency function, given that the author has T articles (densities). Then the overall article-citation size-frequency function, denoted $\Phi(j)$, is given by

$$\Phi(j) = A\frac{\alpha - 1}{j^{\alpha}}, \tag{20}$$

where A denotes the total number of articles. Hence we find the law of Lotka with the same exponent α.

Lemma III.2: The total number of articles A is given by, if $\alpha_1 > 2$:

$$A = \frac{D}{\alpha_1 - 2} \tag{21}$$

We omit the trivial proofs.

Note that it follows from (20) that also

$$A = \int_1^{\infty} \Phi(j)dj \tag{22}$$

as it should.

We now have the following theorem.

Theorem III.3: The global h-index h', being the h-index of the meta-author, is given by

$$h' = \left(\frac{D}{\alpha_1 - 2}\right)^{\frac{1}{\alpha}} \tag{23}$$

The proof follows readily from (5) applied to (20) and using (21).

We can prove the following property of h' in comparison with the successive h-indices h_1, h_2 and h_3:

$$h' > \max(h_1, h_2, h_3) \tag{24}$$

for all authors (h_1-values), institutes (h_2-values) or countries (h_3-values).

Indeed, for every author with T articles we have, by (5) that $h = h_1$ (first level) is given by

$$h_1 = T^{\frac{1}{\alpha}}$$

But A = total number of articles, hence $A > T$ (we assume that there is more than 1 author). Hence $h' >$ all h_1-values, by (22) and (23).

For every institute with S authors we have (second level) ($\alpha = \alpha_0$)

$$h_2 = S^{\frac{1}{\alpha_0 \alpha_1}}$$

$$< A^{\frac{1}{\alpha_0}} = h'$$

since α_0, $\alpha_1 > 1$ and since $S \leq A$. Analogously we prove that (third level):

$$h_3 = R^{\frac{1}{\alpha_0 \alpha_1 \alpha_2}}$$

$$< A^{\frac{1}{\alpha_0}} = h'. \qquad \square$$

That $h' >$ all h_2-values also follows from the following result:

$$h' = \mu^{\frac{1}{\alpha}} h_2^{\alpha_1} \tag{25}$$

with $\mu = \frac{A}{S}$, the average number of articles per author.

This follows readily from (23), (9), (21) and (12)

This ends our theory on the successive and global h-indices.

We close this paper by calculating an example.

IV. h′, h_1 and h_2 for Price awardees

In Egghe (2006), we presented tables for the (still active) D. De Solla Price awardees E. Garfield, F. Narin, T. Braun, A. van Raan, W. Glänzel, H. Moed, A. Schubert, H. Small, B. Martin, L. Egghe, P. Ingwersen, L. Leydesdorff, R. Rousseau and H. White (decreasing order of their $h = h_1$-indices. Their rankings and h_1-values can be read in Table 1.

Table 1. h-indices of Price awardees

r	h_1
1	27
2	27
3	25
4	19
5	18
6	18
7	18
8	18
9	16
10	13
11	13
12	13
13	13
14	12

From this it is clear that the h-index of this table is $h_2 = 13$. Note that for successive h-indexes it is quite possible that we have to go down to the bottom of the table (e.g. if the last 5 h_1-indexes are 14). This never occurs in the calculation of the h_1 indices themselves since citation tables always go down to low citation frequencies of articles.

For the global h-index h' of this group we have to merge the article-citation data of each author into one table in decreasing order of the number of citations to the articles. The result is given in Table 2 (truncated until we can calculate h'). It is clear that $h' = 60$.

Table 2 Merged article-citation data of Price awardees.

r	# citations	r	# citations
1	625	31	96
2	305	32	95
3	239	33	93
4	156	34	91
5	149	35	90
6	138	36	89
7	132	37	88
8	132	38	87
9	129	39	86
10	128	40	86
11	127	41	85
12	127	42	83
13	125	43	82
14	124	44	80
15	124	45	80
16	124	46	79
17	120	47	79
18	112	48	78
19	111	49	78
20	109	50	77
21	109	51	75
22	108	52	74
23	108	53	73
24	108	54	71
25	107	55	70
26	106	56	67
27	105	57	67
28	104	58	66
29	103	59	63
30	101	**60**	**63**
		61	59

Note that, although we do not have the complete article-citation list of the Price awardees, but only up to their $h = h_1$-index, it is possible to calculate h'. Indeed, by (24) we have that $h' >$ all h_1-values of the individual persons. Hence we only need to know each person's table up to the highest h_1 citation value.

Note that, for the global h-index $h' = 60$, your present author does not participate in the calculation of h' since (as can be seen in Egghe (2006)) his article with the highest number of citations (at that time) was 47.

It is clear that your present author does participate in the calculation of the impact factor of level 2: $h_2 = 13$.

General remark: There are no levels for the global h-index h' as we had with the successive ones. Indeed, it is the same if we start from the level “countries”, “institutes” or “authors” since the global list of articles and citations remains the same.

V. Conclusions and suggestions for further research

The successive h-indices, introduced in Schubert (2007), have been modelled in a Lotkaian framework. Formulae for h_2, h_3 (and extendable to higher levels) have been provided and a rationale for their overall decrease (with increasing level) has been given.

We introduced the global h-index being the h-index of the meta-author defined via the merging of the articles-citations lists of all the authors in the group. We show that the global h-index is larger than any successive h-index h_1, h_2, h_3 of all authors, institutes or countries.

It is clear that h' is an h-index which has different properties than the successive ones. We leave it as an open problem to further study the nature and applicability of all these h-indices.

However we are convinced that all these indices have good evaluation qualities in comparing different objects (authors, institutes, countries, or journals, publishers,…) as long as we use the same h-index. It would be interesting to find other applications for other general objects (e.g. research fields, databases,…).

References

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

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

Egghe, L. (2007). Distributions of the h-index and the g-index. Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics, Madrid (Spain) (D. Torres-Salinas and H.F. Moed, eds.), 245-253, CSIC, Madrid, Spain.

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

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, 16569-16572.

Schubert, A. (2007). Successive h-indices. Scientometrics 70(1), 201-205.