

Comparative study of h-index sequences

Peer-reviewed author version

EGGHE, Leo (2009) Comparative study of h-index sequences. In:
SCIENTOMETRICS, 81(2). p. 311-320.

DOI: 10.1007/s11192-008-2170-0

Handle: <http://hdl.handle.net/1942/8491>

Comparative study of h-index sequences

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek,
Belgium¹

and

Universiteit Antwerpen (UA), Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium
leo.egghe@uhasselt.be

ABSTRACT

This paper studies four different h-index sequences (different in publication periods and/or citation periods). Lotkaian models for these h-index sequences are presented by mutual comparison of one sequence with another one.

We also give graphs of these h-sequences for this author on which a discussion is presented. The same is done for the g-index and the R-index.

¹ Permanent address

Key words and phrases: h-index sequence, Lotka, IPP, g-index sequence, R-index sequence.

Acknowledgement: The author is grateful to M. Goovaerts for the preparation of the raw citation data in WoS of this author and to A. Baeten for the construction of the graphs.

I. Introduction

At this moment it is not necessary anymore to give the definition of and references on the h-index since its introduction in Hirsch (2005) in 2005, the h-index has become a real hype and is even already published in the Web of Science (WoS) and Scopus.

One disadvantage of the h-index (in fact of any indicator) is that it is just one number, hereby reducing the evaluation of a researcher (or another source) to a one-dimensional scale. What we need is a sequence of h-indices, calculated for different publication periods and/or citation periods. The problem then, evidently, arises which publication and citation periods to take.

Liang Liming was the first to introduce h-index sequences (see Liang (2006)). Few papers have been written on h-index sequences: Burrell (2007), Rao and Rousseau (2008), Liu and Rousseau (2008) and Egghe (2008a). Liang Liming uses h-index sequences for authors but where the time goes backwards. This means that the k^{th} h-index ($k = 1, 2, \dots$) is calculated over the publication period (being also the citation period): last year - $(k - 1), \dots$, last year, where “last year” means the last year of a researcher’s career, i.e. the present year if the researcher is still active. This is rather strange: one is more interested in the real h-index sequence of a career, i.e. where the k^{th} h-index ($k = 1, 2, \dots$) is calculated over the publication period (being also the citation period): first year, \dots , first year + $(k - 1)$ where “first year” means the first year of a researcher’s career. That this is the “more logical” h-index sequence is also remarked in Burrell (2007). Furthermore, in Egghe (2008a) it is shown that, in general, both h-index sequences are different (even in their shapes) which implies that Liang’s h-index sequence cannot be used as a substitute for the more normal h-index sequence where time goes forward.

The reason for Liang’s use of the h-index sequences where time goes backwards is not given in Liang (2006) but we know that it is much more difficult to derive the h-index sequence of a researcher where time goes forward than to derive the h-index sequence of a researcher where time goes backwards (e.g. using WoS or Scopus). The reason is that the latter sequence can directly be determined from WoS or Scopus, since the periods all end in the present year (for

the former one, one is obliged to handle the raw citation data manually in order to be able to truncate the citation period in the past).

This paper will not deal with h-index sequences where time goes backward (for this, see Egghe (2008a)) but the above discussion indicates that more than one h-index sequence of a researcher can be defined. In fact, in Liu and Rousseau (2008), a list of 10 types of h-index sequences are presented. In Liu and Rousseau (2008), mathematical models for h-index sequences are lacking. In this paper we present 4 h-index sequences (3 of them also appearing in Liu and Rousseau (2008)) for which we give mathematical models. These sequences will be presented in the next section. The four h-index sequences are defined in an order where the periods of publication and/or the periods of citation are increasing. One model (the third) is the “classical” one, being the h-index sequence of the career of a researcher.

In the third section we present simple mathematical models for these h-index sequences and give some inequalities.

The fourth section presents an example of these sequences and similarly for the g-index and R-index, based on the citation data of this author (up to 2007) and the fifth section is a concluding section where also some open problems are presented.

II. Four types of h-index sequences

Let each h-index sequence depend on the time parameter $k = 1, 2, \dots, N$, where N is the length of the publication period (e.g. in years). For each k , let $T(k)$ denote the number of articles produced by an author in year k . For each $i, j = 1, 2, \dots, N$, $i \neq j$, let $C(i, j)$ denote the number of citations in year j to the $T(i)$ papers produced in year i .

II.1 Basic sequence for yearly publications

This model is lacking in Liu and Rousseau (2008). For each $k = 1, 2, \dots, N$, the h-index sequence $h_1(k)$ is calculated based on the $T(k)$ publications and the $C(k, k)$ citations in year k to these $T(k)$ publications. The sequence $h_1(k)$ is a basic (starting) sequence and is, in a

way, comparable with the immediacy index, published in the JCR of Thomson Scientific. So we have a scheme as in Fig. 1

k:	1	2	...	N
Pub:	T(1)	T(2)	...	T(N)
Cit:	C(1,1)	C(2,2)	...	C(N,N)

Fig. 1 Scheme determining the h-index sequence $h_1(k)$.

II.2 Basic sequence for cumulative publications

For each $k = 1, 2, \dots, N$ the h-index sequence $h_2(k)$ is calculated based on the $T(1) + \dots + T(k)$ publications and the $C(1,k) + \dots + C(k,k)$ citations. Hence, as in the previous sequence $h_1(k)$, the citation period is only one year and one checks the citations to all previous years $1, \dots, k$ (k included). So we have a scheme as in Fig. 2.

k:	1	2	...	N
Pub:	T(1)	T(1) + T(2)	...	T(1) + ... + T(N)
Cit:	C(1,1)	C(1,2) + C(2,2)	...	C(1,N) + ... + C(N,N)

Fig. 2. Scheme determining the h-index sequence $h_2(k)$.

It is already clear that, since for each $k = 1, \dots, N$, in sequence 2 one calculates (in year k) the citations to articles in a longer time period as in sequence 1, we have that

$$h_2(k) \geq h_1(k) \quad (1)$$

for all $k = 1, \dots, N$.

Also this second sequence is basic since the next sequence is built on the elements featuring in this second sequence.

II.3 Real career h-index sequence

For each $k = 1, 2, \dots, N$ the h-index sequence $h_3(k)$ is calculated based on the $T(1) + \dots + T(k)$ publications and the

$$\sum_{i=1}^k \sum_{j=i}^k C(i, j) \quad (2)$$

citations. Hence this h-index sequence is the real career h-index sequence of a researcher: for each $k = 1, 2, \dots, N$, the number $h_3(k)$ is the h-index of the researcher at time k of the career.

So we have a scheme as in Fig. 3.

k:	1	2	...	N
Pub:	$T(1)$	$T(1) + T(2)$...	$T(1) + \dots + T(N)$
Cit:	$C(1,1)$	$C(1,1) + C(1,2)$ $+ C(2,2)$...	$\sum_{i=1}^N \sum_{j=i}^N C(i, j)$

Fig. 3. Scheme determining the h-index sequence $h_3(k)$

Note that for this h-index sequence we do not need additional citation data, above the ones used in the calculation of the h-index sequence $h_2(k)$.

Since, for each $k = 1, \dots, N$, we calculate more citations than in sequence 2 to the same publication sets, it is clear that

$$h_3(k) \geq h_2(k) \quad (3)$$

for all $k = 1, \dots, N$.

II.4 Total career h-index sequence

For each $k = 1, 2, \dots, N$ the h-index sequence $h_4(k)$ is calculated based on the $T(1) + \dots + T(k)$ publications and the

$$\bigcirc_{i=1}^k \bigcirc_{j=i}^N C(i, j) \quad (4)$$

citations. More general, one can replace N in (16) by any $M > N$, allowing for even longer citing periods. Hence this h-index follows the publication career as in the previous case (real career h-index sequence) but counts citations to these publications over the whole career period (if we use N in (16)) or even over a period that is longer than the career period (if we replace N by $M > N$ in (16)). We have a scheme as in Fig. 4.

k:	1	2	...	N
Pub:	$T(1)$	$T(1) + T(2)$...	$T(1) + \dots + T(N)$
Cit:	$\bigcirc_{j=1}^M C(1, j)$	$\bigcirc_{j=1}^M C(1, j) + \bigcirc_{j=2}^M C(2, j)$...	$\bigcirc_{i=1}^N \bigcirc_{j=i}^M C(i, j)$

Fig. 4. Scheme determining the h-index sequence $h_4(k)$

Since for each $k = 1, \dots, N$ we calculate more citations than in sequence 3 to the same publication sets, it is clear that

$$h_4(k) \geq h_3(k) \quad (5)$$

for all $k = 1, \dots, N$.

Note: This are just 4 models of h-index sequences. Many other models are possible (cf. Liu and Rousseau (2008)). For instance, in Fig. 4, we could also have taken, for every k , $T(k)$

instead of $\sum_{i=1}^k T(i)$ as we did in the first model but then (17) fails (this h-index sequence is larger than $h_1(k)$ but we do not go into this further on).

III. Comparison of the h-index sequences in the Lotkaian framework

Here we suppose that the publication-citation system is Lotkaian:

$$f(j) = \frac{C}{j^\alpha} \quad (6)$$

($C > 0$, $\alpha > 1$), where f is the size-frequency function describing density of the papers with citation density $j \geq 1$. Suppose that there are T papers. Then we showed in Egghe and Rousseau (2006) that the h-index h of this system is given by

$$h = T^{\frac{1}{\alpha}} \quad (7)$$

If we apply this result to the first h-index sequence $h_1(k)$ and if we look at Fig. 1 we hence have

$$h_1(k) = T(k)^{\frac{1}{\alpha}} \quad (8)$$

for every $k = 1, \dots, N$, supposing that α applies for every k .

If we model $T(k)$ by

$$T(k) = T(1)k^\beta \quad (9)$$

where $\beta \geq 0$ then (8) leads to

$$h_1(k) = T(1)^{\frac{1}{\alpha}} k^{\frac{\beta}{\alpha}}$$

$$h_1(k) = h_1(1) k^{\frac{\beta}{\alpha}} \quad (10)$$

This h-index sequence always increases (or is constant in case $\beta = 0$, i.e. in case where the researcher publishes the same number of articles every year). Its shape is convex iff $\beta > \alpha$ and is concave iff $\beta < \alpha$. If $\beta = \alpha$ we have the linear increase. These results are readily seen.

The models of the next h-index sequences will be based on the ones of the previous h-index sequence. For $h_2(k)$ ($k = 1, \dots, N$) we assume we have another Lotka exponent, denoted $\gamma > 1$ (assumed to be valid for every k). A look at Fig. 2 now yields, applying again the general result (7)

$$h_2(k) = \sum_{i=1}^k T(i)^{\frac{1}{\gamma}} \quad (11)$$

for every $k = 1, \dots, N$ (since, for every k , we have $\sum_{i=1}^k T(i)$ publications).

We can express $h_2(k)$ in terms of $h_1(k)$ as given in the next proposition.

Proposition III.1: For every $k = 1, \dots, N$

$$h_2(k) = \sum_{i=1}^k h_1^\alpha(i)^{\frac{1}{\gamma}} \quad (12)$$

Proof: This is trivial:

$$h_2^\gamma(k) = \sum_{i=1}^k T(i)$$

$$= \prod_{i=1}^k h_1(i)^\alpha$$

using (8), from which the result follows. \square

It is clear that γ must be so that $h_2(k) \geq h_1(k)$ for every $k = 1, \dots, N$. This is e.g. the case if $\gamma \leq \alpha$.

In the special case of constant production (say $T(i) = T$ for every $i = 1, \dots, N$) we have by (8)

that $h_1(k) = T^{\frac{1}{\alpha}}$ is constant but $h_2(k)$ is concavely increasing: by (11)

$$h_2(k) = (kT)^{\frac{1}{\gamma}}$$

$$h_2(k) = k^{\frac{1}{\gamma}} T^{\frac{1}{\gamma}}$$

$$h_2(k) = k^{\frac{1}{\gamma}} \left(T^{\frac{1}{\alpha}} \right)^{\frac{\alpha}{\gamma}}$$

$$h_2(k) = k^{\frac{1}{\gamma}} h_1^{\frac{\alpha}{\gamma}} \tag{13}$$

which is concavely increasing in k since $\gamma > 1$.

Now we turn our attention to the comparison of the h-index sequences $h_2(k)$ and $h_3(k)$, $k = 1, \dots, N$. This will automatically lead to a comparison of $h_3(k)$ with $h_1(k)$, based on the above analysis. A look at Fig. 3 gives that, for every $k = 1, \dots, N$

$$h_3(k) = \prod_{i=1}^k T(i)^{\frac{1}{\gamma}} \tag{14}$$

for a certain Lotka exponent $\delta > 1$. Since $h_3(k)^3 \leq h_2(k)$ for every k we must have, by (11) and (14), that $\delta \leq \gamma$ (necessary and sufficient condition). The comparison of $h_3(k)$ with $h_2(k)$ and $h_1(k)$ is trivial: using (8) and (11) we have, by (14)

$$h_3(k) = h_2^{\frac{\gamma}{\delta}}(k) \quad (15)$$

$$h_3(k) = \sum_{i=1}^k a_i^{\frac{\gamma}{\delta}} h_1^{\frac{\alpha}{\delta}}(i) \quad (16)$$

Finally, for $h_4(k)$, an inspection of Fig. 4 gives, applying the general formula (7).

$$h_4(k) = \sum_{i=1}^k a_i^{\frac{\gamma}{\delta}} T(i)^{\frac{\alpha}{\delta}} \quad (17)$$

with $\varepsilon > 1$ the Lotka exponent of this system. Since $h_4(k)^3 \leq h_3(k)$ for every k we have, by (14) and (17) that (necessary and sufficient condition) $\varepsilon \leq \delta$. The comparison of $h_4(k)$ with $h_3(k)$, $h_2(k)$ and $h_1(k)$ is trivial:

$$h_4(k) = h_3^{\frac{\delta}{\varepsilon}}(k) \quad (18)$$

$$h_4(k) = h_2^{\frac{\gamma}{\varepsilon}}(k) \quad (19)$$

$$h_4(k) = \sum_{i=1}^k a_i^{\frac{\gamma}{\varepsilon}} h_1^{\frac{\alpha}{\varepsilon}}(i) \quad (20)$$

as is readily seen, using (8), (11), (14) and (17).

General note: Similar results as the ones obtained here for the h-index can be proved for the other h-type indices such as the g-index and the R-index cf., Egghe (2006), Jin, Liang, Rousseau and Egghe (2007). These results can be obtained from the ones presented here on

the h-index by taking into account that their values are obtained from the h-index and Lotka's parameter α :

$$g = \frac{\frac{\alpha}{\alpha-1} - \frac{1}{2}}{\frac{\alpha}{\alpha-1} - \frac{1}{2}} h \quad (21)$$

$$R = \frac{\frac{\alpha}{\alpha-1} - \frac{1}{2}}{\frac{\alpha}{\alpha-1} - \frac{1}{2}} h \quad (22)$$

as proved in the above mentioned references. We leave the details to the reader.

IV. Practical examples of $h_1(k)$, $h_2(k)$, $h_3(k)$ and $h_4(k)$ sequences and similarly for the g-index and the R-index

Fig. 5 presents, in one graph, the $h_1(k)$, $h_2(k)$, $h_3(k)$ and $h_4(k)$ sequences for this author (up to and including 2007 and starting in 1978, hence comprising a 30-year period). It is clear that for the $h_1(k)$ -sequence, for a single author, it is not easy to obtain h_1 -values above 0 or 1.

This is due to the well-known publication (and hence citation) delays. The sequence $h_1(k)$ could, however, be instructive for larger data sets such as for an institute or for a data set coming from a repository and based on downloads for which it is known that such delays are hardly existing, cf. Bollen, Van de Sompel, Smith and Luce (2005).

The h-index sequence $h_2(k)$ is interesting. It gives the “h-performance” of a career's publications based on a year by year citation count. The data show that $h_2(k)$ is fairly constant during the (larger) mid part of the career.

The third h-index sequence is the real career sequence and was already produced in Egghe (2008a) (in comparison of Liang's reverse time h-index sequence). The third h-index

sequence shows a more or less linear increase which can only be explained by a convexly increasing function of number of publications per year (see above or see also Egghe (2008a)).

The fourth h-index sequence is the total career h-index sequence (using the total career period (or even beyond) for citations) and, evidently, lies above the third sequence but, for the last year ($N = 30$) we have $h_3(N) = h_4(N)$. Note also that $h_1(1) = h_2(1) = h_3(1)$, by definition.

Remark that $h_1(k) \leq h_2(k) \leq h_3(k) \leq h_4(k)$, for all $k = 1, \dots, N$ as predicted.

Fig. 6 and Fig. 7 show the similar four graphs for the g-index and the R-index, with similar properties.

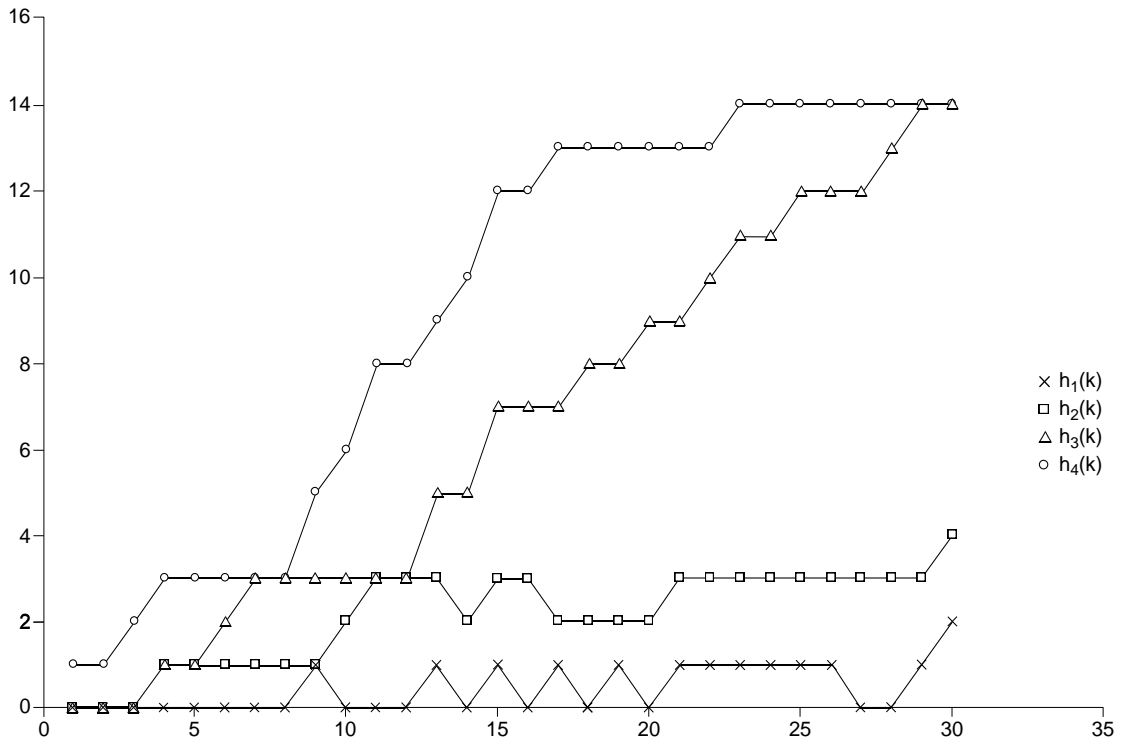


Fig. 5. The h-index sequences $h_1(k)$, $h_2(k)$, $h_3(k)$ and $h_4(k)$

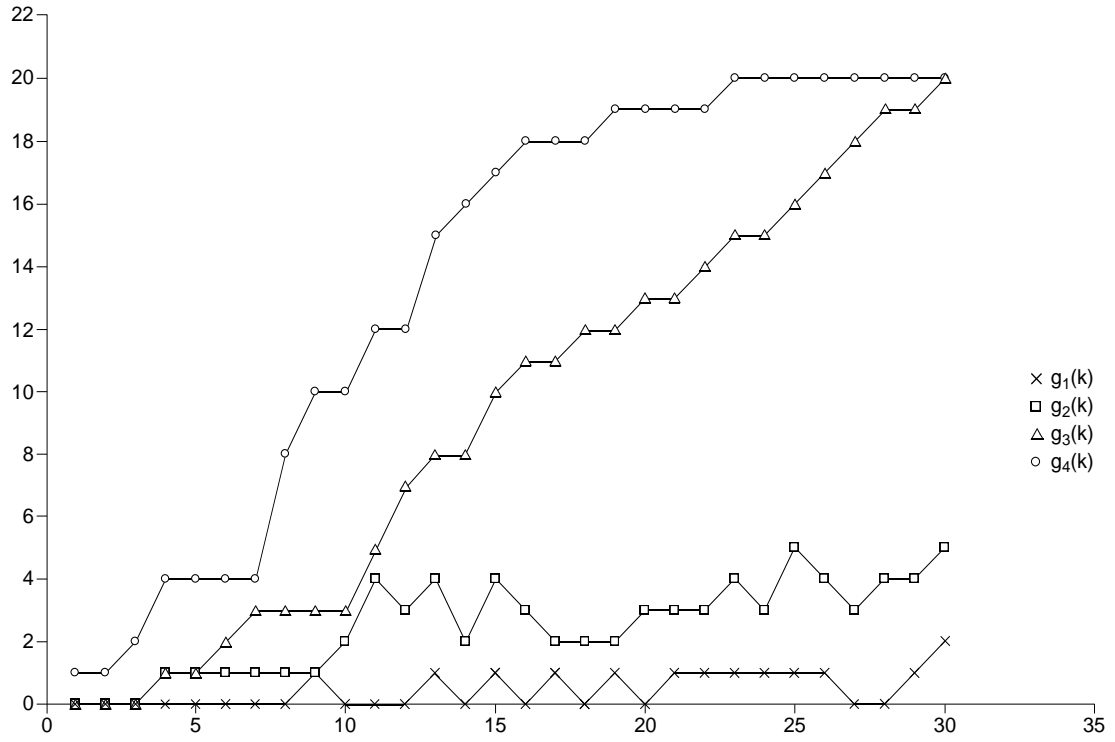


Fig. 6. The g-index sequences $g_1(k)$, $g_2(k)$, $g_3(k)$ and $g_4(k)$

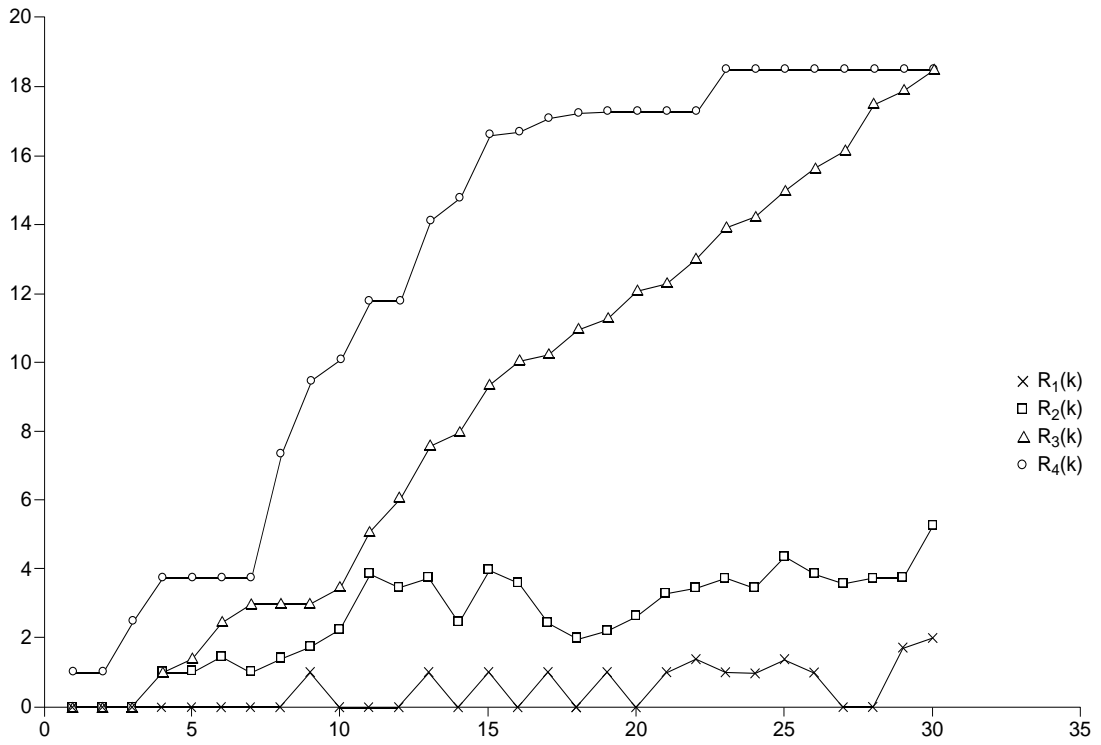


Fig. 7. The R-index sequences $R_1(k)$, $R_2(k)$, $R_3(k)$ and $R_4(k)$

V. Conclusions and open problems

In this paper we presented four examples of h-index sequences for which a Lotkaian model was given. We have that the consecutive h-index sequences have increasing values when compared to each other which is confirmed in the example given on these h-index sequences of this author.

It is not easy to get such type of data. In fact, any h-index sequence (of an author, institute (van Raan (2006)), journal (Braun, Glänzel and Schubert (2005, 2006)), topic (Banks (2006)),...) where citation periods are involved which end in the past must be truncated manually from the WoS data.

While the third h-index sequence – the real career h-index sequence – clearly is the most important one, it remains a challenge to further study the second h-index sequence, showing the yearly citation performance (in terms of the h-index) of a career (at each year). The fact that it is only slightly increasing (in the example) should be further studied and interpreted.

We encourage informetricians to generate similar data for larger sets of researchers (e.g. institutes,...) or journals or topics and to compile the h-index sequences (and similar for the g-index and R-index) as defined here. We also encourage informetricians to construct other h-index type sequences – other than the ones presented here – cf. Liu and Rousseau (2008) and also other ones than the ones given in the Liu and Rousseau article.

References

- M.G. Banks (2006). An extension of the Hirsch index: indexing scientific topics and compounds. *Scientometrics* 69(1), 161-168.
- J. Bollen, H. Van de Sompel, J. Smith and R. Luce (2005). Toward alternative metrics of journal impact. A comparison of usage and citation data. *Information Processing and Management* (Special issue on Informetrics, L. Egghe, ed.) 41(6), 1419-1440.

- T. Braun, W. Glänzel and A. Schubert (2005). A Hirsch-type index for journals. *The Scientist* 19(22), 8.
- T. Braun, W. Glänzel and A. Schubert (2006). A Hirsch-type index for journals. *Scientometrics* 69(1), 169-173.
- Q.L. Burrell (2007). Hirsch index or Hirsch rate ? Some thoughts arising from Liang's data. *Scientometrics* 73(1), 19-28.
- L. Egghe (2005). *Power Laws in the Information Production Process: Lotkaian Informetrics*. Elsevier, Oxford, UK.
- L. Egghe (2006). Theory and practise of the g-index. *Scientometrics* 69(1), 131-152.
- L. Egghe (2007). General evolutionary theory of IPPs and applications to the evolution of networks. *Journal of Informetrics* 1(2), 115-122.
- L. Egghe (2008a). Mathematical study of h-index sequences. Preprint.
- 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), to appear.
- L. Egghe (2008c). The influence of transformations on the h-index and the g-index. *Journal of the American Society for Information Science and Technology*, 59(7), 1-9.
- 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 USA* 102, 16569-16572.
- B. Jin, L. Liang, R. Rousseau and L. Egghe (2007). The R- and AR-indices : complementing the h-index. *Chinese Science Bulletin* 52(6), 855-863.
- L. Liang (2006). h-index sequence and h-index matrix: Constructions and applications. *Scientometrics* 69(1), 153-159.
- Y. Liu and R. Rousseau (2008). Definitions of time series in citation analysis with special attention to the h-index. Preprint.
- Y. Liu, I.K. Ravichandra Rao and R. Rousseau (2008) Empirical series of journal h-indices. *Scientometrics*, to appear.
- A.J. Lotka (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences* 16(12), 317-324.
- A.F.J. van Raan (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgement for 147 chemistry research groups. *Scientometrics* 67(3), 491-502.