Possible and Necessary h-indices

Marek Gągolewski[1,2] Przemysław Grzegorzewski[1,2]

1. Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
2. Faculty of Mathematics and Information Science, Warsaw University of Technology
pl. Politechniki 1, 00-661 Warsaw, Poland
Email: {gagolews,pgrzeg}@ibspan.waw.pl

This is a pre-print version of the article. Please cite this paper as:
Gągolewski Marek and Grzegorzewski Przemysław. Possible and Necessary h-indices. In: Proc. IFSA World Congress and Eusflat Conference IFSA/Eusflat 2009 (ISBN: 978-989-95079-6-8), Lisbon, Portugal, 20-24 July, 2009, pp. 1691-1695.

Abstract *— The problem of measuring scientific impact is considered. A class of so-called p-sphere indices, which generalize the well known Hirsch index, is used to construct a possibility measure of scientific impact. This measure might be treated as a starting point for prediction of future index values or for dealing with right-censored bibliometric data.*

Keywords — Hirsch's h-index, p-sphere indices, scientific impact, possibility theory, scientometrics.

1 Introduction

Fair and objective assessment methods of individual scientists had become the focus of scientometricians' attention since the very beginning of their discipline. A quantitative expression, i.e. measurement, of some publication-citation process characteristics is assumed to be a predictor of broadly conceived scientific competence.

Among the most popular scientific impact indicators is the h-index, proposed by J. Hirsch in 2005 [1]. It has been defined as follows. An author who had published n papers has the Hirsch index equal to H, if each of his H publications were cited at least H times, and each of the other $n - H$ items were cited no more than H times. This simple indicator quickly received much attention in the academic community [2, 3] and started to be a subject of intensive research. It was noted (see for example [4, 5]) that contrary to earlier approaches, i.e. publication count, citation count etc., the measure both concerns productivity and impact of an individual.

Many modifications of the h-index were later proposed, e.g. Egghe's g-index [6], Kosmulski's $h(2)$-index [7], Jin's R-index [8] or Schreiber's h_m-index [9]. It was also a matter of more formal studies (e.g. [10, 11, 12, 13, 14, 15]). It is worth noting that the h-index can be expressed as the Sugeno integral of some function with respect to a fuzzy counting measure [16]. For more details we refer the reader to the extensive scientometric literature.

2 Difficulties related to the h-index

From now on $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ denotes the set of all natural numbers and zero, while $\mathbb{R}_0^+ = \mathbb{R}^+ \cup \{0\}$ stands for the set of all nonnegative real numbers. For any sets X, Y and Z ($Z \subset X$) and $f : X \to Y$ by $f|_Z$ we mean a mapping satisfying $f|_Z(x) = \mathbf{1}_{x \in Z} f(x)$, where $\mathbf{1}$ is an indicator function.

Let us assume that an individual has published exactly $n \in \mathbb{N}$ papers. The list of papers generates an ordered *citation sequence* $\mathbf{C} = (c_1, c_2, \ldots, c_n)$ such that $c_i \geq c_j$ for $1 \leq i < j \leq n$, where $c_i \in \mathbb{N}_0$ is the number of unique citations received by the i-th article.

The problem with the h-index is twofold. Firstly, it assumes perfect knowledge of the author's citation sequence. In practice we gather bibliometric data from large online academic services, such as Thomson *Web of Science*, Elsevier *Scopus* or Google *Scholar*. The coverage of all digital libraries is limited, so in most cases we are dealing with right-censored data.

On the other hand, citing is a dynamic process. If you were applying for academic tenure and were asked to determine your h-value, would you be sure that the index is not about to increase in a while? That is because the Hirsch coefficient totally ignores the number of citations received by publications represented by $c_1, c_2, \ldots, c_H$ (we only know their citation counts are $\geq H$) and how close to H are $c_{H+1}, c_{H+2}, \ldots, c_n$.

As an illustration of the raised issues, consider the following citation sequences: $\mathbf{C}_1 = (4, 4, 4, 4, 0, 0, 0)$ and $\mathbf{C}_2 = (10, 9, 8, 7, 4, 4, 3, 1)$. Both have the h-index of 4, but in the latter case there is not much needed for $\mathbf{C}_2$'s h to increase even to the value of 6. Such property of the citation sequence could be called *saturation* or even *instability*.

In the next section we recall necessary information on a class of so-called "geometric" scientific impact indices [17]. In Section 4 we suggest how to evaluate the h-index stability by means of possibility theory. In our model a given, but usually uncertain, citation sequence will be treated as a source of information to estimate possible and necessary index values. Section 5 illustrates the construction of the suggested possibility measure for individual's h-index.

3 p-sphere indices

Let us recall some definitions from the paper [17]. The following function, based on the citation sequence, will be useful.

Definition 1. A *citation function* based on a citation sequence

$\mathbf{C}$ is a mapping $\pi_{\mathbf{C}} : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ given by

$$\pi_{\mathbf{C}}(x) = \begin{cases} c_i & \text{if} \quad x \in [i-1, i)\,, \quad i = 1, 2, \ldots, n, \\ 0 & \text{if} \quad x \geq n. \end{cases} \tag{1}$$

An exemplary citation function for sequence $\mathbf{C} = (5, 4, 3, 3, 3, 1)$ is depicted in Fig. 1.

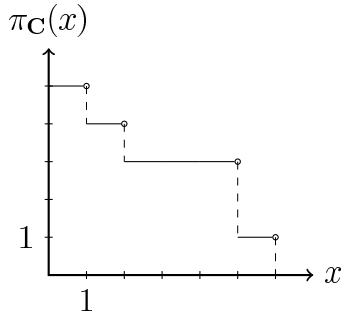

Figure 1: Citation function for $\mathbf{C} = (5, 4, 3, 3, 3, 1)$.

Definition 2. Consider a function $f : \mathbb{R}_0^+ \to \mathbb{R}_0^+$. We say that the citation function π *dominates* the function f (denoted $\pi \succeq f$) if $\pi(x) \geq f(x)$ for every $x \in \mathbb{R}_0^+$.

A set of all functions $f : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ dominated by the citation function π will be denoted by L_π.

Among the two new classes of scientific impact indices, the authors defined the p-sphere index.

Definition 3. Given an arbitrary real number $1 \leq p < \infty$ and any real number $r \geq 0$ let $s_{p,r} : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ denote a function

$$s_{p,r}(x) = \begin{cases} \left(r^p - x^p\right)^{\frac{1}{p}} & \text{for} \quad x \in [0, r)\,, \\ 0 & \text{for} \quad x \geq r. \end{cases} \tag{2}$$

Moreover, for $p = \infty$ we have

$$s_{\infty,r}(x) = \begin{cases} r & \text{for} \quad x \in [0, r)\,, \\ 0 & \text{for} \quad x \geq r. \end{cases} \tag{3}$$

Intuitively, for $x \in [0, r)$, the graph of $s_{p,r}(x)$ determines a part of an L^p-sphere (i.e. the boundary of an L^p-ball on a plane) of radius r (see Fig. 2) and centered at $(0, 0)$. Therefore $s_{p,r}$ is further on called the *p-sphere function* of *radius* r.

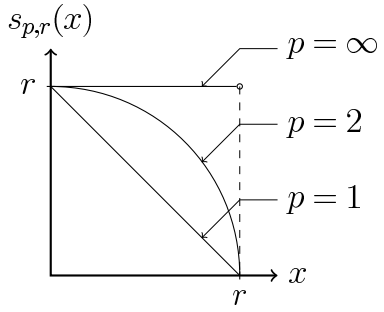

Figure 2: Exemplary p-spheres, for $p = 1, 2, \infty$.

Definition 4. The *p-sphere index* (or, originally, the *maximal p-radius*) of a citation function π is the greatest number $r \geq 0$, for which π still dominates $s_{p,r}$, i.e.

$$r_\pi(p) := \max\left\{r : \ \pi \succeq s_{p,r}\right\}. \tag{4}$$

The p-sphere index was primarily denoted $r_p(\pi)$. In this article we mainly consider the maximal radius to be a function of p, hence the change. For abbreviation, the p-sphere function of maximal p-radius $s_{p,r_\pi(p)}$ will be denoted by $S_{p,\pi}$ and called the *maximal p-sphere function* for π.

Here are some properties of the p-sphere index.

Lemma 5. *For any given $p \geq 1$ and a citation sequence $\mathbf{C} = (c_1, c_2, \ldots, c_n)$, let $\pi = \pi_{\mathbf{C}}$. Then the following properties hold:*

(i) $r_\pi(p) \leq n$.

(ii) $r_\pi(p) \leq c_1$.

(iii) If $p = \infty$, then $r_\pi(\infty) = H$, where H is the h-index of an individual (see [1]).

(iv) If $p = 1$, then $r_\pi(1) = W$, where W is the individual's w-index, as defined by Woeginger in [13].

(v) For any $q > p$, $r_\pi(q) \leq r_\pi(p) \leq 2r_\pi(q)$.

(ii) $r_\pi(p)$ is nonincreasing with respect to p.

Their proofs were given in [17].

4 Possible and necessary h-index

The theory of fuzzy measures and evidence has been well established. Therefore only definitions and properties that are necessary are recalled (for more details and further references the reader is referred, e.g., to [18]).

Further on we consider measures with respect to $\mathbb{R}_0^+$ and a family of all its subsets $\mathcal{P}(\mathbb{R}_0^+)$.

Definition 6. A function $\mu : \mathcal{P}(\mathbb{R}_0^+) \to [0, 1]$ is a *fuzzy measure* if it satisfies the following requirements:

M1. $\mu(\emptyset) = 0$ and $\mu(\mathbb{R}_0^+) = 1$ (boundary conditions),

M2. for all $A, B \in \mathcal{P}(\mathbb{R}_0^+)$, if $A \subseteq B$, then $\mu(A) \leq \mu(B)$ (monotonicity).

Definition 7. Let Pos denote a fuzzy measure. Then Pos is called a *possibility measure* iff for any family $\{A_k \in \mathcal{P}(\mathbb{R}_0^+) : k \in K\}$ and an arbitrary index set K,

$$\mathrm{Pos}\left(\bigcup_{k \in K} A_k\right) = \sup_{k \in K} \mathrm{Pos}(A_k). \tag{5}$$

For each possibility measure Pos we may associate another fuzzy measure Nec, called *necessity measure*, defined by

$$\mathrm{Nec}(A) := 1 - \mathrm{Pos}(\overline{A}), \tag{6}$$

where $A \in \mathcal{P}(\mathbb{R}_0^+)$.

It may be shown, that for every $A \in \mathcal{P}(\mathbb{R}_0^+)$ and for any possibility measure Pos and the associated necessity measure Nec following relations hold:

(i) $\mathrm{Nec}(A) > 0 \Rightarrow \mathrm{Pos}(A) = 1$,

(ii) $\mathrm{Pos}(A) < 1 \Rightarrow \mathrm{Nec}(A) = 0$.

Proposition 8. *Every possibility measure* Pos *may be uniquely determined by a* possibility distribution function *(abbreviated Pos.D.F.)* $R : \mathbb{R}_0^+ \to [0,1]$ *by the formula*

$$\mathrm{Pos}(A) = \sup_{x \in A} R(x), \tag{7}$$

where $A \in \mathcal{P}(\mathbb{R}_0^+)$.

Now we are ready to present a class of possibility measures for the h-index values.

Suppose we are given a citation sequence $\mathbf{C} = (c_1, c_2, \ldots, c_n)$ of some individual X. Some of the elements of $\mathbf{C}$ are either right-censored data or are potentially due to increase in a "short time". Let $H = h(\mathbf{C}) = \max\{i \in \mathbb{N}_0 : c_i \geq i\}$ be equal to X's Hirsch index. Thus we have sure evidence that the true value of X's Hirsch-index is *at least* H.

We propose some minimal requirements for a possibility distribution to describe hypothetical values of the h-index.

Definition 9. A *Pos.D.F. for the Hirsch index* is a mapping $R_h : \mathbb{R}_0^+ \to [0,1]$ which satisfies the following axioms:

H1. $R_h(x) = 0$ for $x < H$ or $x \notin \mathbb{N}_0$,

H2. $R_h(H) = 1$ (normalization),

H3. For any $x, x' \in \mathbb{N}_0$, if $H \leq x < x'$ then $R_h(x) > R_h(x')$ or $R_h(x') = 0$.

These axioms seem quite natural. Let us look at some properties of the fuzzy measures defined by such Pos.D.F. The following proposition might be easily proved.

Proposition 10. *Let* R_h *be an arbitrary Pos.D.F. for the Hirsch index. Furthermore, let* Pos *be the possibility measure determined by* R_h *and* Nec *be the associated necessity measure. Then for any* $[a,b] \in \mathcal{P}(\mathbb{R}_0^+)$ *the following properties are satisfied.*

- *If* $H \in [a,b]$ *then* $\mathrm{Pos}([a,b]) = 1$.
- *If* $H \notin [a,b]$ *then* $\mathrm{Nec}([a,b]) = 0$.
- *If* $H \in [a,b]$ *then for any* $b < b'$ *we get* $\mathrm{Nec}([a,b]) \leq \mathrm{Nec}([a,b'])$.

The postulated measures may give clues for questions such as: What is the (broadly conceived) possibility that the true value of X's h-index is really "equal to $H+1$" or "greater than H". Note that publishing-citing is an extremely complicated process and without drastic simplifications and idealizations it cannot be modeled using stochastic methods. Thus, in general, it is not reasonable to consider the results of individual scientometric measurements by means of the *probability* theory. Therefore, our problem is how to construct appropriate possibilistic measures. In the next section we propose a Pos.D.F. for the Hirsch index defined by means of the $r(p)$-indices.

5 Example

Lemma 11. *Assume we are given a citation function* $\mathbf{C} = (c_1, c_2, \ldots, c_n)$, $\pi = \pi_{\mathbf{C}}$ *and* $r_\pi(\infty) = H \in \mathbb{N}$. *Then the properties below are satisfied:*

(i) $c_i \geq H$ *for every* $i \leq H$,

(ii) $c_i \leq H$ *for every* $i > H$,

(iii) $r_\pi(p) \leq 2^{\frac{1}{p}} H$ *for any* $1 \leq p < \infty$.

Proof. Only (iii) is proved here. Let $1 \leq p < \infty$. By Lemma 5, $H \leq r_\pi(p) \leq 2H$. Let $a = \max\{c_1, 2H\}$. Consider a citation sequence $\mathbf{C}' = (c_1', c_2', \ldots, c_n')$, such that $c_i' = a$ for $i \leq H$, and $c_i' = H$ for $H < i \leq n$. Clearly, $r_{\pi_{\mathbf{C}'}}(\infty) = H$, and $c_i \leq c_i'$ for any i. For every $1 \leq p < \infty$, $S_{p,\pi_{\mathbf{C}'}}$ goes through (H, H), so $r_{\pi_{\mathbf{C}'}}(p)$ is the largest possible for fixed H. Solving (2) for r gives $r_{\pi_{\mathbf{C}'}}(p) = 2^{\frac{1}{p}} H$. Hence $r_{\pi_{\mathbf{C}}}(p) \leq 2^{\frac{1}{p}} H$ as stated. □

Let us define the inverse function of r_π. As r_π is not necessarily an injection, we need a special formula. In the sequel we assume that

$$r_\pi^{-1}(\varrho) := \max\{p : r_\pi(p) = \varrho\}, \tag{8}$$

for $\varrho \in [r_\pi(\infty), r_\pi(1)]$, and undefined otherwise. This definition is sensible, as it may be easily shown that for any $1 \leq p' < p''$ and $x \in \mathbb{R}_0^+$, $r_\pi(p') = r_\pi(p'')$ implies $S_{p',\pi}(x) \leq S_{p'',\pi}(x)$, hence the knowledge of maximal p satisfying $r_\pi(p) = \varrho$ is the most informative.

Now, let $R_\pi : \mathbb{R}_0^+ \to [0,1]$ be a function given by:

$$R_\pi(x) = \begin{cases} 0 & \text{for } x < r_\pi(\infty), \\ 2 - 2^{1/r_\pi^{-1}(x)} & \text{for } x \in [r_\pi(\infty), r_\pi(1)], \\ 0 & \text{for } x > r_\pi(1). \end{cases} \tag{9}$$

It is easily seen that for $x \in [r_\pi(\infty), r_\pi(1)]$, $R_\pi(x)$ is a strictly decreasing, but not necessarily continuous function.

For any $1 \leq p < \infty$, if $r_\pi(p) = 2^{\frac{1}{p}} H$, then $r_\pi^{-1}(x) = [\log_2(\frac{1}{H} x)]^{-1}$ and

$$R_\pi(x) = \begin{cases} 2 - \frac{1}{H} x & \text{for } x \in [H, 2H], \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Lemma 12. *Let* $\mathbf{C}$ *and* $\mathbf{C}'$ *be citation sequences for which* $\pi = \pi_{\mathbf{C}}$, $\pi' = \pi_{\mathbf{C}'}$, $r_\pi(\infty) = r_{\pi'}(\infty) = H$ *and for any* $1 \leq p < \infty$, $r_\pi(p) = 2^{\frac{1}{p}} H$.

(i) $R_{\pi'}(x) \leq R_\pi(x)$ *for any* x,

(i) If $R_{\pi'}$ *is given by* (10), *then* $r_{\pi'}(p) = 2^{\frac{1}{p}} H$ *for any* $1 \leq p < \infty$.

The proof is left to the reader.

It may be shown that $R_\pi|_{\mathbb{N}_0}$ is a Pos.D.F. for the Hirsch index for π. Thus the class of p-sphere indices may be used to construct a possibility measure having the properties of our interest. Obviously, one may find uncountably many transforms of r_π^{-1} to $[0,1]$, so that is just an example. Generating meaningful possibility measures by means of the p-sphere indices dependent on a type of the process affecting citation sequences is in scope of our future research.

Let us discuss the behavior of the proposed possible h-index on real-world data and see how it can be used to differentiate between individual's citation information. We consider the output of 3 Polish computer scientists, Prof. A having $n_A = 19$ publications, Prof. B with $n_B = 29$ publications and Prof. C with his $n_C = 18$ publications. Their citation sequences, according to *Scopus*[1], are given in Table 1. See Fig. 3 for the citation functions π_A, π_B, π_C and the maximal p-spheres for $p = 1, 2, \infty$. Each author's h-index equals $H = 7$, but they differ in values of other p-sphere indices, e.g. $r_{\pi_A}(1) = 12$, $r_{\pi_B}(1) = 14$ and $r_{\pi_C}(1) = 10$.

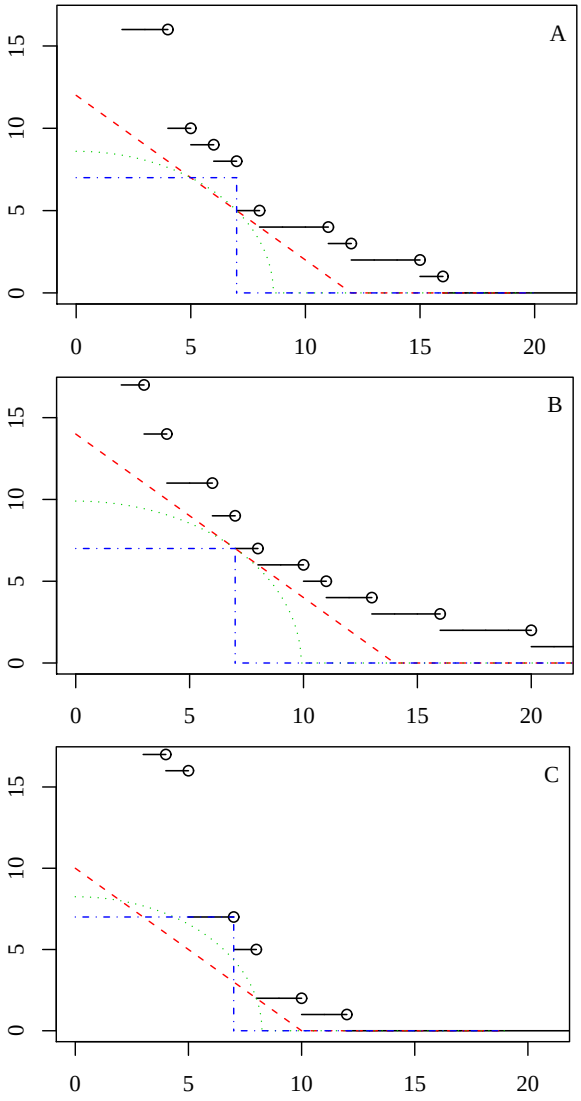

Figure 3: Citation functions of the 3 authors.

The author's R_π functions and resulting possibility distribution functions for the Hirsch-index are depicted in Fig. 4. Please note that R_{π_B} is of the form (10).

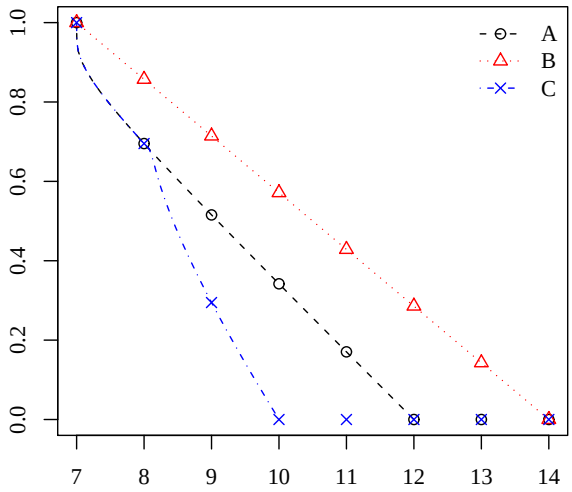

Figure 4: R_π and Pos.D.F. for the h-index of the 3 authors.

We see that Prof. B has the greatest possibility of increasing his h-index. On the contrary, some papers of Prof. C has not got a required number of citations so he can not have high expectations of a greater h-value. The proposed Poss.D.F. clearly differentiates between all the authors and can be used as a complement to the Hirsch index.

6 Conclusions

In the paper we discussed an important problem related to the Hirsch index: it assumes perfect knowledge of author's citation sequence and does not take into account a dynamic essence of the publication/citation process. Therefore we proposed a possibilistic approach to the indicator by setting several axioms for the possibility distribution function for the h-index.

In this model, given citation data are treated as an evidence for the minimal h-value and thus are just a starting point for speculation on its likely value in the case of perfect information.

We used a recently-proposed class of scientometric coefficients, the p-sphere indices, which is a generalization of Hirsch's index, to construct an exemplary possibility measure. A real-world example was presented for illustration.

As it was already mentioned, future work should definitely encompass the construction of different possibility measures, according to the type of a process affecting citation sequences.

[1]The publication data were gathered on January 20, 2009 and are limited to the field "Computer Science". Authors' names have been intentionally masked.

Table 1: Citation sequences of the 3 authors.

$\mathbf{C}_A$	(103, 20, 16, 16, 10, 9, 8, 5, 4, 4, 4, 3, 2, 2, 2, 1, 0, 0, 0),
$\mathbf{C}_B$	(56, 30, 17, 14, 11, 11, 9, 7, 6, 6, 5, 4, 4, 3, 3, 3, 2, 2, 2, 2, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0),
$\mathbf{C}_C$	(39, 34, 23, 17, 16, 7, 7, 5, 2, 2, 1, 1, 0, 0, 0, 0, 0, 0).

References

[1] Jorge E. Hirsch. An index to quantify individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005.

[2] Philip Ball. Index aims for fail ranking of scientists. *Nature*, 436:900, 2005.

[3] Lutz Bornmann and Hans-Dieter Daniel. The state of h index research. *EMBO Reports*, 10(1):2–5, 2009.

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

[5] Lutz Bornmann and Hans-Dieter Daniel. What do we know about the h index? *Journal of the American Society for Information Science and Technology*, 58(9):1381–1385, 2007.

[6] Leo Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[7] Marek Kosmulski. A new Hirsch-type index saves time and works equally well as the original h-index. *ISSI Newsletter*, 2(3):4–6, 2006.

[8] BiHui Jin, LiMing Liang, Ronald Rousseau, and Leo Egghe. The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, 52(6):855–863, 2007.

[9] Michael Schreiber. A modification of the h-index: The h_m-index accounts for multi-authored manuscripts. *Journal of Informetrics*, 2(3):211–216, 2008.

[10] Quentin L. Burrell. On the h-index, the size of the Hirsch core and Jin's A-index. *Journal of Informetrics*, 1:170–177, 2007.

[11] Leo Egghe. Modelling successive h-indices. *Scientometrics*, 77(3):377–387, 2008.

[12] Wolfgang Glänzel. On some new bibliometric applications of statistics related to the h-index. *Scientometrics*, 77(1):187–196, 2008.

[13] Gerhard J. Woeginger. An axiomatic characterization of the Hirsch-index. *Mathematical Social Sciences*, 56(2):224–232, 2008.

[14] Ronald Rousseau. Woeginger's axiomatisation of the h-index and its relation to the g-index, the $h(2)$-index and the r^2-index. *Journal of Informetrics*, 2(4):335–340, 2008.

[15] Antonio Quesada. Monotonicity and the Hirsch index. *Journal of Informetrics*, 3(2):158–160, 2009.

[16] Vicenç Torra and Yasuo Narukawa. The h-index and the number of citations: two fuzzy integrals. *IEEE Transactions on Fuzzy Systems*, 16(3):795–797, 2008.

[17] Marek Gągolewski and Przemysław Grzegorzewski. A geometric approach to the construction of scientific impact indices. *Scientometrics*, 2009. In press. DOI:10.1007/s11192-008-2253-y.

[18] George J. Klir and Bo Yuan. *Fuzzy sets and fuzzy logic. Theory and applications*. Prentice Hall PTR, New Jersey, 1995.