A geometric approach to the construction of scientific impact indices

Marek Gągolewski
Systems Research Institute, Polish Academy of Sciences[1], and
Faculty of Mathematics and Information Science, Warsaw University of Technology[2],
E-mail: gagolews@ibspan.waw.pl. *(Corresponding author)*

Przemysław Grzegorzewski
Systems Research Institute, Polish Academy of Sciences[1], and
Faculty of Mathematics and Information Science, Warsaw University of Technology[2],
E-mail: pgrzeg@ibspan.waw.pl.

This is a pre-print version of the article published in *Scientometrics*. Please cite this paper as:

Gągolewski Marek and Grzegorzewski Przemysław. A geometric approach to the construction of scientific impact indices. *Scientometrics*, 81(3), 2009, 617–634. DOI: 10.1007/s11192-008-2253-y.

[1]Address: ul. Newelska 6, 01-447 Warsaw, Poland. Fax: +48 22 3810 105.
[2]Address: pl. Politechniki 1, 00-661 Warsaw, Poland.

Abstract

Two broad classes of scientific impact indices are proposed and their properties — both theoretical and practical — are discussed. These new classes were obtained as a geometric generalization of the well-known tools applied in scientometric, like Hirsch's h-index, Woeginger's w-index and the Kosmulski's *Maxprod*. It is shown how to apply the suggested indices for estimation of the shape of the citation function or the total number of citations of an individual. Additionally, a new efficient and simple $O(\log n)$ algorithm for computing the h-index is given.

Keywords: Hirsch's h-index, citation analysis, scientific impact indices.

1 Introduction

Recently we observe a gradually increasing interest in developing objective methods for characterizing productivity and quality of a scientific research. Scientometricians propose different numerical measures to quantify the research output and its impact both for individual scientists and for research teams, institutions etc. Such tools could be used in deciding upon grant allocation, employment, society membership or chair elections.

Traditional measures, such as the total number of papers or citations, the highest citation count of a paper, mean number of citations etc., have been broadly criticized. To compensate some of their drawbacks Jorge Hirsch [14] proposed an index to assess both the productivity and impact of a scientist. His so-called h-index is informally defined as follows: An individual has index h if h of his/her n papers have at least h citations each, and the other $n-h$ papers have no more than h citations each.

The h-index quickly received much attention in the academic community and became very popular. Many authors discussed its properties and restrictions, e.g. [1, 13, 21, 17, 5]. There were also attempts to apply the h-index not only for individuals' results but for entire journals or disciplines as well (see, e.g., [2, 25, 15, 6, 23, 8, 20]). Moreover, some systematic theoretical attempts to model the index behavior were also performed (e.g. [12, 3, 7, 22, 11, 27]).

Later many modifications of the h-index were proposed. For example, Egghe [9, 10] suggested the g coefficient, which is more sensitive to highly cited papers. On the other hand, the Kosmulski $h(2)$-index [18] correlates better with the maximal number of citations than the h-index and is more appropriate in the fields in which typical number of citations per article is high (e.g. in biology, chemistry and physics). Other indicators include

corrections for self-citations and co-authorship [4], effects of aging of papers [16, 24] and field-specific normalization [25].

In the present paper we observe that some of the well-known indices of scientific impact have a simple and clear geometric interpretation connected with the notion of metric in appropriate space. Therefore, using different metrics we obtain other indices. Moreover, through the appropriate choice of the metric we may construct an indicator which possesses desirable properties. It should be stressed that the suggested method for constructing indices of scientific impact is universal, as we do not make any assumptions on the distribution of citations.

The paper is organized as follows: first we establish some basic definitions which are used throughout the paper. In Sec. 3 we propose the class of one-parameter r_p-indices which might be perceived as generalizations of the h-index. Then (Sec. 4) we discuss their properties and give a formal method of determining their value. Sec. 5 deals with the two-parameter generalizations, called l_p-indices, which could be used to avoid some problems met for one-parameter indices.

2 Citation function

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.

Let us assume that an individual has published exactly $n \in \mathbb{N}$ papers. Each list of n published papers generates an ordered *citation sequence* $C = (c_1, c_2, \ldots, c_n)$ 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 total number of citations of the scientist equals $\zeta_C = \sum_{i=1}^{n} c_i$.

The following function, based on the citation sequence, will be useful.

Definition 1. A *citation function* based on a citation sequence C is a mapping $\pi_C : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ given by

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

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

The citation function has some obvious but interesting properties given in the lemma below.

Lemma 2. *Given a citation function π_C based on a citation sequence $C = (c_1, c_2, \ldots, c_n)$ the following holds:*

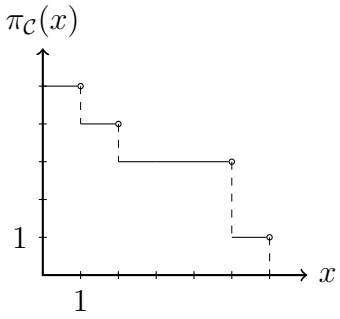

Figure 1: Citation function for $C = (5, 4, 3, 3, 3, 1)$.

(i) π_C *is nonincreasing.*

(ii) π_C *is a step function with steps located at* $x \in \mathbb{N}$.

(iii) $\forall x \in \mathbb{R}_0^+ \ \pi_C(x) \in \mathbb{N}_0$.

(iv) $\lim_{x\to\infty} \pi_C(x) = 0$.

(v) $\int_0^\infty \pi_C(x)\,dx = \sum_{i=1}^n c_i = \zeta_C < \infty$.

The proof is straightforward.

Definition 3. Consider a function $f : \mathbb{R}_0^+ \to \mathbb{R}_0^+$. We say that the citation function π_C *dominates* the function f (denoted $\pi_C \succeq f$) if $\pi_C(x) \geq f(x)$ 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_C will be denoted by L_{π_C}.

In next sections we consider some families of functions dominated by the citation function which seem to be useful both in characterizing some well-known citation indices and in defining another ones.

3 The r_p-indices

Recently Woeginger [27] noticed that "the h-index maximizes the volume of a scaled copied of an ℓ^∞ unit ball under the curve, while the w-index maximizes the volume of a scaled copied of an ℓ^1 unit ball under the curve". Let us formalize and generalize this interesting remark.

Denotation 4. 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

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

Moreover, for $p = \infty$ we have

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

Example of $s_{p,r}$ for $p = 1$, 2 and ∞ is shown in Fig. 2. It is easily seen that for $x \in [0, r)$ the graph $s_{p,r}(x)$ determines a part of an ℓ^p-sphere (circle) with radius r. This is the reason why $s_{p,r}$ is further on called the *p-sphere function* of *radius* r.

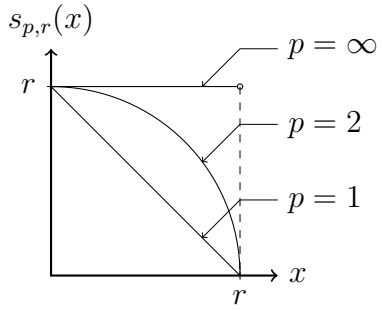

Figure 2: $s_{1,r}$, $s_{2,r}$, $s_{\infty,r}$.

One may easily see that the following lemma holds.

Lemma 5. *For any given $x \geq 0$ and fixed radius $r \geq 0$, a sphere function $s_{p,r}(x)$ is nondecreasing with respect to p.*

Let us now combine the p-sphere functions with the citation functions.

Definition 6. A *maximal p-radius* of a citation function π_C is the greatest number $r \geq 0$, for which π_C still dominates $s_{p,r}$, i.e.

$$r_p(\pi_C) := \max\{r : \ \pi_C \succeq s_{p,r}\}\,. \tag{4}$$

For abbreviation, the p-sphere function of maximal p-radius $s_{p,r_p(\pi_C)}$ will be denoted by S_{p,π_C} (or even S_p) and called the *maximal p-sphere function* for π_C.

Here are some properties of maximal p-radius and maximal p-sphere functions. These properties will be useful later on for calculating indices of scientific impact.

Lemma 7. *For any given $p \geq 1$ and a citation function π_C based on a citation sequence $C = (c_1, c_2, \ldots, c_n)$, the following properties hold:*

(i) $r_p(\pi_C) \leq n$.

(ii) $r_p(\pi_C) \leq c_1$.

(iii) If $p < \infty$ *then* $\{x < n : \ S_p(x) = \pi_C(x)\} \subset \mathbb{N}_0$.

(iv) If $p = \infty$ *then* $\{x < n : \ S_p(x) = \pi_C(x)\} \cap \mathbb{N}_0 \neq \emptyset$.

The proof is left to the reader.

Now we will show that the maximal p-radius is a generalization of Hirsch's h and Woeginger's w indices.

Proposition 8. *For any citation sequence $C = (c_1, c_2, \ldots, c_n)$*

$$r_\infty(\pi_C) = h, \tag{5}$$

where h is the h-index of an individual.

Proof. By the definition of the h-index

$$\begin{aligned} h &= \max\{k \in \mathbb{N} : \ c_i \geq k \text{ for } i = 1, 2, \ldots, k\} \\ &= \max\{k \in \mathbb{N} : \ c_k \geq k\}. \end{aligned} \tag{6}$$

Assume that $c_1 > 0$. Otherwise trivially $r_\infty(\pi_C) = 0 = h$.

Now we should notice that $r_\infty(\pi_C) \in \mathbb{N}$. Conversely, let $r_\infty(\pi_C) = r \notin \mathbb{N}_0$. But then either π_C has a step at r or $\pi_C(x) = r$ for some x, which contradicts Lemma 2, or we can find a number $r' > r$ such that $\pi_C \succeq s_{\infty,r'}$, which denies the definition of the maximal ∞-radius.

Hence by applying (4), (3) and (1) we get

$$\begin{aligned} r_\infty(\pi_C) &= \max\{k \in \mathbb{N} : \ \pi_C \succeq s_{\infty,k}\} \\ &= \max\left\{k \in \mathbb{N} : \ \pi_C(k^-) \geq k\right\} \\ &= \max\{k \in \mathbb{N} : \ c_k \geq k\} = h, \end{aligned}$$

which proves the proposition. □

Proposition 9. *For any citation sequence $C = (c_1, c_2, \ldots, c_n)$*

$$r_1(\pi_C) = w, \tag{7}$$

where w is the w-index of an individual.

Proof. By the definition of the w-index [27]

$$w = \max\{k \in \mathbb{N} : \ c_i \geq k - i + 1 \text{ for } i = 1, 2, \ldots, k\}. \tag{8}$$

Let us again assume that $c_1 > 0$. By Lemma 7 there exists $j \in \mathbb{N}_0$, $j < n$ for which $S_1(j) = \pi_C(j)$. $S_1(x) = r_1(\pi_C) - x$ for every $x \in [0, n)$ and $S_1(j) = c_{j+1} \in \mathbb{N}$. Thus we have $r_1(\pi_C) - j = c_{j+1}$. It implies that $r_1(\pi)$ is also an integer. Since $r_1(\pi_C)$ is the maximal 1-radius of π_C then

$$\begin{aligned} r_1(\pi_C) &= \max\{k \in \mathbb{N} : \ k - (i-1) \leq \pi_C(i-1) \text{ for } i = 1, 2, \ldots, k\} \\ &= \max\{k \in \mathbb{N} : \ k - i + 1 \leq c_i \text{ for } i = 1, 2, \ldots, k\} = w, \end{aligned}$$

and the proof is complete. □

As it was shown above, for $p = 1$ and $p = \infty$ the maximal p-radius $r_p(\pi_C)$ of a citation function π_C reduces to well-known measures of scientific impact. However $p = 1$ and $p = \infty$ are just boundaries of the set of all possible values that p can take. Therefore, for any other $1 < p < \infty$ we may also obtain candidates for scientific impact indices.

It can be shown that the following sine qua non conditions for being a scientific impact index hold. These conditions are similar to those proposed by Woeginger. Note that he defined the index as a function into $\mathbb{N}_0$. We have extended the class of acceptable functions.

Theorem 10. *For any $p \geq 1$, the maximal p-radius is a mapping from the set of all possible citation functions into $\mathbb{R}_0^+$ satisfying*

(i) If $\pi_C(x) = 0$ for every $x \in \mathbb{R}_0^+$ then $r_p(\pi_C) = 0$.

(ii) Given any two citation functions π_C and $\pi_{C'}$, if $\pi_C \succeq \pi_{C'}$ then $r_p(\pi_C) \geq r_p(\pi_{C'})$.

Thus through this simple construction we obtain an infinite class of scientific impact indices. Recalling our geometric intuitions, the r_p-index characterizes the radius of a maximal ℓ^p-sphere (circle) that is dominated by π_C.

One may ask a natural question about the relationship between indices corresponding to different ℓ^p-spheres. It is given by the following lemma.

Lemma 11. *For any $1 \leq p \leq q$ and for each citation function π_C we have*

$$r_q(\pi_C) \leq r_p(\pi_C) \leq 2r_q(\pi_C). \tag{9}$$

Sketch of the proof. The first inequality follows easily from Lemma 5. Woeginger [27] showed the right inequality for $p = 1$ and $q = \infty$. As a consequence, the inequality is valid for any $1 \leq p \leq q$. □

It is worth noting that $r_p(\pi_C)$ for $p < \infty$ satisfies the same Woeginger's axioms as the w-index, i.e. A2, B, C, D [27] and T1 [26]. However, as mentioned above, we define indices on the larger domain than Woeginger.

For practical applications, a procedure for determining the value of scientific impact index for any citation sequence should be computationally effective. Happily, the following proposition holds. Note the constructive proof.

Proposition 12. *Given an ordered citation sequence $C = (c_1, c_2, \ldots, c_n)$ the $r_p(\pi_C)$-index of scientific impact can be computed in linear time with respect to n.*

Proof. By Lemma 7 it follows immediately that one should consider only a finite number of points of π_C to determine $r_p(\pi_C)$. More precisely, only elements $c_1, c_2, \ldots, c_n$ of the citation sequence C are required for the computation. The pseudocode of the algorithm is given (see Fig. 3). Below we prove its correctness.

```
   Input: p ≥ 1; C = (c_1, c_2, ..., c_n), such that c_i ≥ c_j for 1 ≤ i < j ≤ n
   Result: r_p(π_C)

1  r := n;
2  for i = 1, 2, ..., n do
3      if s_{p,r}(i − 1) > c_i then
4          if p = ∞ then
5              r := max{i − 1, c_i};
6          else
7              r := ((i − 1)^p + c_i^p)^{1/p};
8  return r as the result;
```

Figure 3: Algorithm for computing $r_p(\pi_C)$.

The algorithm starts with the initial candidate for the output radius $r := r_0 = n$ (as $r_p(\pi_C) \leq n$ by property *(i)* of Lemma 7). Then it examines each consecutive element of the sequence C and reduces the value of r if necessary.

Consider the next candidate for the output radius $r := r_i$ at the i-th iteration. If $s_{p,r}(i-1) \leq c_i = \pi_C(i-1)$ then the radius r does not have

to be adjusted. Otherwise we find r' such that $s_{p,r'}(i-1) = c_i$. There are two cases: If $p = \infty$ then $s_{p,r}$ is not continuous and r' is chosen to be the maximal value of the pair $i-1$ and c_i, as π_C can be bounded here either by the number of papers or by the number of citations. If $p < \infty$ then r' is calculated by solving (2) for r. Since $r' < r$ and $c_j > c_i$ for $j < i$ hence $r := r'$ is a new candidate for the output radius.

After n iterations r is the maximal p-radius of π_C. Thus we obtain $r_p(\pi_C)$ in linear time with respect to n which is our assertion. □

It is worth noting that the value of r_∞ (i.e. the h-index) can be determined even faster. Please note again the constructive proof of the proposition.

Proposition 13. *Given an ordered citation sequence $C = (c_1, c_2, \ldots, c_n)$, $r_\infty(\pi_C)$ can be computed in logarithmic time with respect to n.*

Informal proof. As previously stated computation is based on the values $c_1, c_2, \ldots, c_n$. Here we additionally make use of the fact that $c_i \geq c_j$ for $1 \leq i < j \leq n$.

The pseudocode of the algorithm is given below (see Fig. 4). It is a modification of the binary search (binary chop) algorithm. We now prove its correctness.

```
    Input: C = (c_1, c_2, ..., c_n), such that c_i ≥ c_j for 1 ≤ i < j ≤ n
    Result: r_∞(π_C)

1   L := 1;
2   R := n;
3   repeat
4       d := ⌈(R−L)/2⌉;
5       M := L + d;
6       if c_M = M or L = R then
7           return M as the result;
8       else if c_M < M then
9           R := M − 1;
10      else
11          L := M;
```

Figure 4: Algorithm for computing $r_\infty(\pi_C)$ (the h-index).

The algorithm uses an interval of sequence indices $[L, R]$, which includes the value we look for. We start with the whole citation sequence, i.e. $[1, n]$. On every iteration we consider an element M lying in the middle of the interval. If $c_M = M$ or $L = R$, then we stop. Otherwise, we reduce the interval

either to $[L, M-1]$ or $[M, R]$ basing on the knowledge whether the result may be larger than M or not. This is because if $c_M < M$ then $r_\infty(\pi_C) < M$, while if $c_M \geq M$ then $r_\infty(\pi_C) \geq M$. The procedure is convergent, because the size of the interval is always reduced by ≥ 1.

As the bounding interval is halved in each iteration the running time of the algorithm is $O(\log_2 n)$, so the proposition holds. □

4 Discussion

Although Hirsch [14] states that, empirically, the total number of citations ζ_C is proportional to h^2, it is clear that without any assumptions on the citations distribution none of the r_p-indices of scientific impact can estimate ζ_C properly. Indeed, consider a trivial case when $C = (c_1)$. Then $(\forall p \geq 1)$ $r_p(\pi_C) = 1$ but $\zeta_C = c_1$ can be arbitrary.

Let us recall that by Lemma 2 the total number of citations is equal to the area below the citation function. Therefore, let us examine the relation between ζ_C and the area below the p-sphere function.

Lemma 14. *For any given $p \geq 1$ and a citation function π_C based on a citation sequence $C = (c_1, c_2, \ldots, c_n)$, the following properties hold.*

(i) $\int_0^\infty s_{p,r}(x)\,dx \leq \zeta_C$ *for any* $s_{p,r} \in L_{\pi_C}$.

(ii) $\int_0^\infty S_p(x)\,dx = \max\left\{\int_0^\infty s_{p,r}(x)\,dx : \ s_{p,r} \in L_{\pi_C}\right\}$.

(iii) If $p < \infty$ *then* $\int_0^\infty S_p(x)\,dx = r_p^2(\pi_C)\,\frac{1}{p}\,B\left(\frac{1}{p}, 1+\frac{1}{p}\right)$, *where* $B(\cdot,\cdot)$ *is the Euler beta function.*

(iv) $\lim_{p\to\infty} \int_0^\infty S_p(x)\,dx = r_\infty^2(\pi_C)$.

Proof. If $s_{p,r} \in L_{\pi_C}$ then, by definition, $\int_0^\infty s_{p,r}(x)dx \leq \int_0^\infty \pi_C(x)\,dx = \zeta_C$. Property *(ii)* follows directly from the definition of the maximal p-radius.

Now, let us consider the integral $\int_0^\infty S_p(x)\,dx$ for any $p < \infty$. Let $r = r_p(\pi_C)$. We get

$$\int_0^\infty S_p(x)\,dx = \int_0^r (r^p - x^p)^{\frac{1}{p}}\,dx.$$

By substituting $t = \left(\frac{x}{r}\right)^p$ we obtain

$$\begin{aligned}\int_0^r (r^p - x^p)^{\frac{1}{p}}\,dx &= \frac{r^2}{p}\int_0^1 t^{\frac{1}{p}-1}(1-t)^{\frac{1}{p}}dt \\ &= \frac{r_p^2(\pi_C)}{p}\,B\left(\frac{1}{p}, 1+\frac{1}{p}\right),\end{aligned}$$

which proves *(iii)*. Point *(iv)* follows from the limit properties of beta function. □

The lemma shows that the area below the maximal p-sphere function is the best estimate of ζ_C among all p-sphere functions dominated by the citation function under study. Moreover, this estimate is proportional to the square of the r_p-index, which coincides with the intuition of Hirsch. In particular, we get $r_\infty^2(\pi_C)$ for the h-index, $\frac{1}{2}r_1^2(\pi_C)$ for the w-index, $\frac{\pi}{4}r_2^2(\pi_C)$ for the index corresponding to ℓ^2-sphere, etc. Unfortunately, $r_p(\pi_C)$ gives only the lower bound of the total number of citations.

Generally, the problem which p is generally the best to estimate ζ_C is nontrivial. Our results suggest that $p \simeq 2$ performs well for many ad hoc citation sequences. We state that r_2-index is worth of deeper analysis in the future. Why? Consider the following intuition. One of the drawbacks of w (i.e. r_p for $p = 1$) is that it can include in its "core" too many publications with low number of citations. On the other hand, h (i.e. r_p for $p = \infty$) is very rigid and inflexible. A paper either has sufficient number of citations or it is completely left out of the "core". What is more, the actual value of citations for elements belonging to the "core" is unimportant. Hence it seems that by using $1 < p < \infty$ we can obtain an index which is more tolerant for paper close to the "core" (but not too tolerant) and index which pays more attention to papers with significantly high number of citations.

Another drawback of all one-dimensional indices is that they do not tell anything about the shape of the original citation function. They cannot answer such natural questions like: Has the citation function a long tail? Is it flat or peaked? To illustrate the problem better let us consider the following example.

Example
Let us consider three authors whose citation sequences are given in Table 1.

Id	citations
C_1	61, 59, 58, 51, 49, 49, 33, 32, 30, 30, 28, 28, 24, 24, 24, 23, 23, 23, 23, 23, 23, 22, 22, 17, 17, 17, 17, 17, 14, 14, 14, 14, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
C_2	29, 29, 28, 27, 27, 27, 26, 26, 26, 26, 25, 25, 24, 24, 24, 24, 23, 23, 23, 23, 22, 22, 22, 22, 22, 22, 21, 21, 21, 21, 21, 21, 21, 21, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 19, 18, 18, 18, 18, 16, 14, 14, 12, 12, 12, 12, 12, 12, 12, 12, 11, 11, 9, 9, 9, 9, 8, 8, 8, 8, 7, 7, 7, 7, 7, 6, 6, 5, 5, 3, 3, 3, 2, 1, 1, 1, 1
C_3	99, 96, 95, 81, 78, 75, 69, 63, 57, 56, 51, 48, 43, 43, 40, 37, 31, 27, 26, 24, 24, 23, 21, 20, 18, 18, 17, 16, 16, 10, 9, 4, 3, 2, 1, 1

Table 1: Citation sequences under study.

The corresponding citation functions are depicted in Fig. 5a.

Given citation sequences were chosen so that $h = 22$ for each author. However, it is easily seen that these citation functions differ significantly in shape. For C_2 we get a long tail and we could call him a "hard-worker" or a "big producer" (Nomenclature by [19] and [8]), because he published 100 articles, of which over than 60 gained about 20 citations. Hence it is seen that the h-index ignores more than half of his publications.

Quite different situation represents author C_3. He wrote not too many papers but some of them were very famous. This type is called a "genie" / "perfectionist" / "selective scientist".

On the other hand, C_1 represents a "typical" author.

The results of an analysis are shown in Table 2. It is seen that the considered $r_p(\pi_C)$-indices cannot effectively discriminate actual differences between these three authors. Similarly, an attempt to estimate the number of publications on the basis of those indices are not satisfactory. The number of citations is underestimated from 1.5 to 4 times. Fig. 6a shows estimated number of citations as a function of p.

Id	n	r_1	r_2	r_∞	ζ_{C_i}	$\int S_1$	$\int S_2$	$\int S_\infty$
C_1	60	35	26.83	22	957	612.5	565.5	484
C_2	100	29	26.68	22	1651	420.5	559.2	484
C_3	36	35	30	22	1342	612.5	706.9	484

Table 2: r_p-indices for citation sequences under study.

In next section we propose a wider class of indices that may eliminate or at least reduce some of the drawbacks of r_p-indices.

5 The l_p-indices

It is evident that some of the drawbacks of the r_p-indices described above are caused by symmetry of the p-sphere functions s_p. Therefore, here we define another class of functions and corresponding indices which have the ability of being much more adaptative to the citation function.

Denotation 15. Given an arbitrary $1 \leq p < \infty$, $a \geq 0$ and $b \geq 0$, let $e_{p,a,b} : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ denote a function

$$e_{p,a,b}(x) = \begin{cases} \left(b^p - \left(\frac{b}{a}x\right)^p\right)^{\frac{1}{p}} & \text{for} \quad x \in [0, a)\,, \\ 0 & \text{for} \quad x \geq a. \end{cases} \tag{10}$$

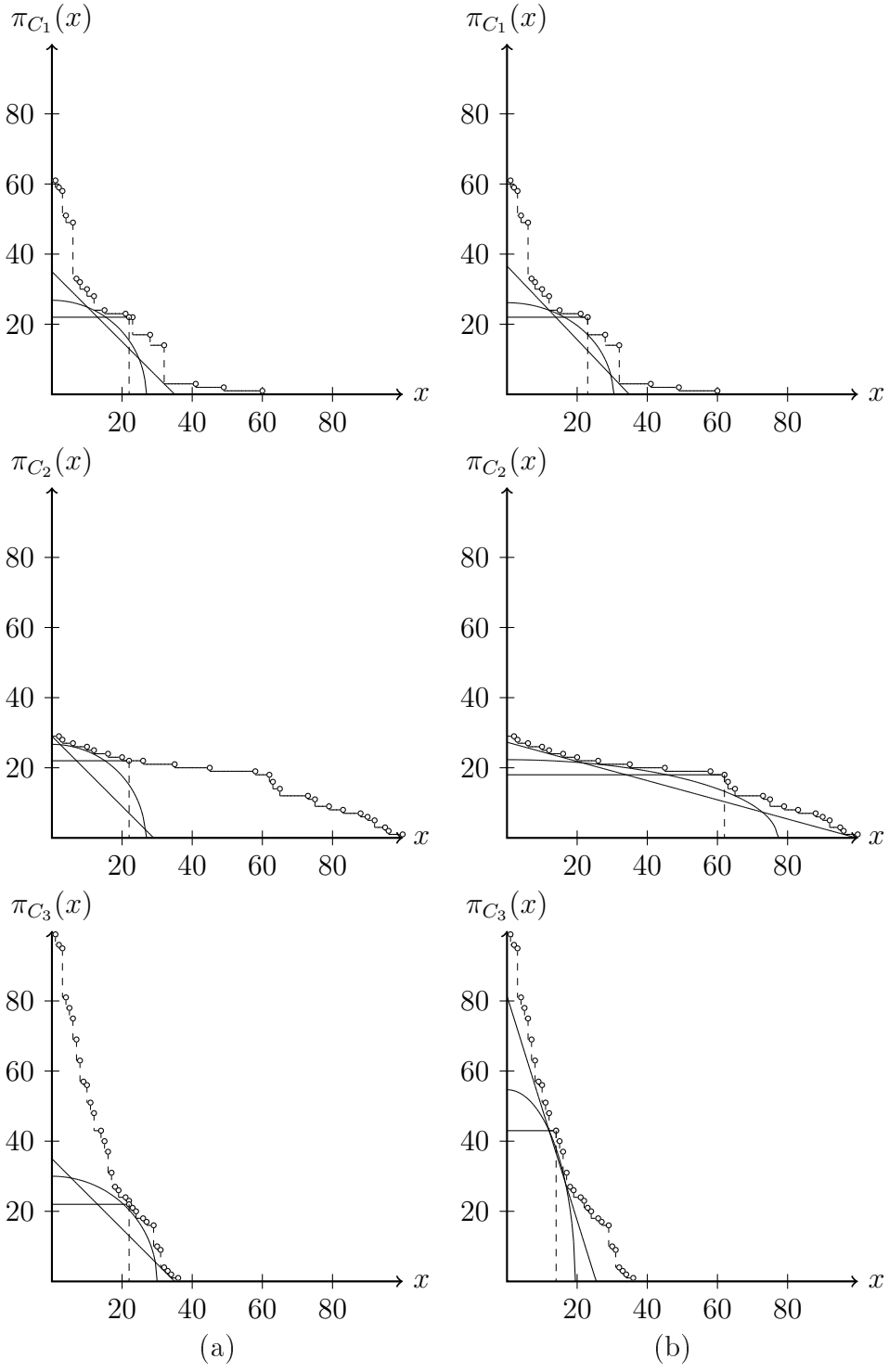

Figure 5: Citation functions for exemplary sequences from Table 1 and maximal (a) p-spheres S_1, S_2, S_∞ and (b) p-ellipses E_1, E_2, E_∞.

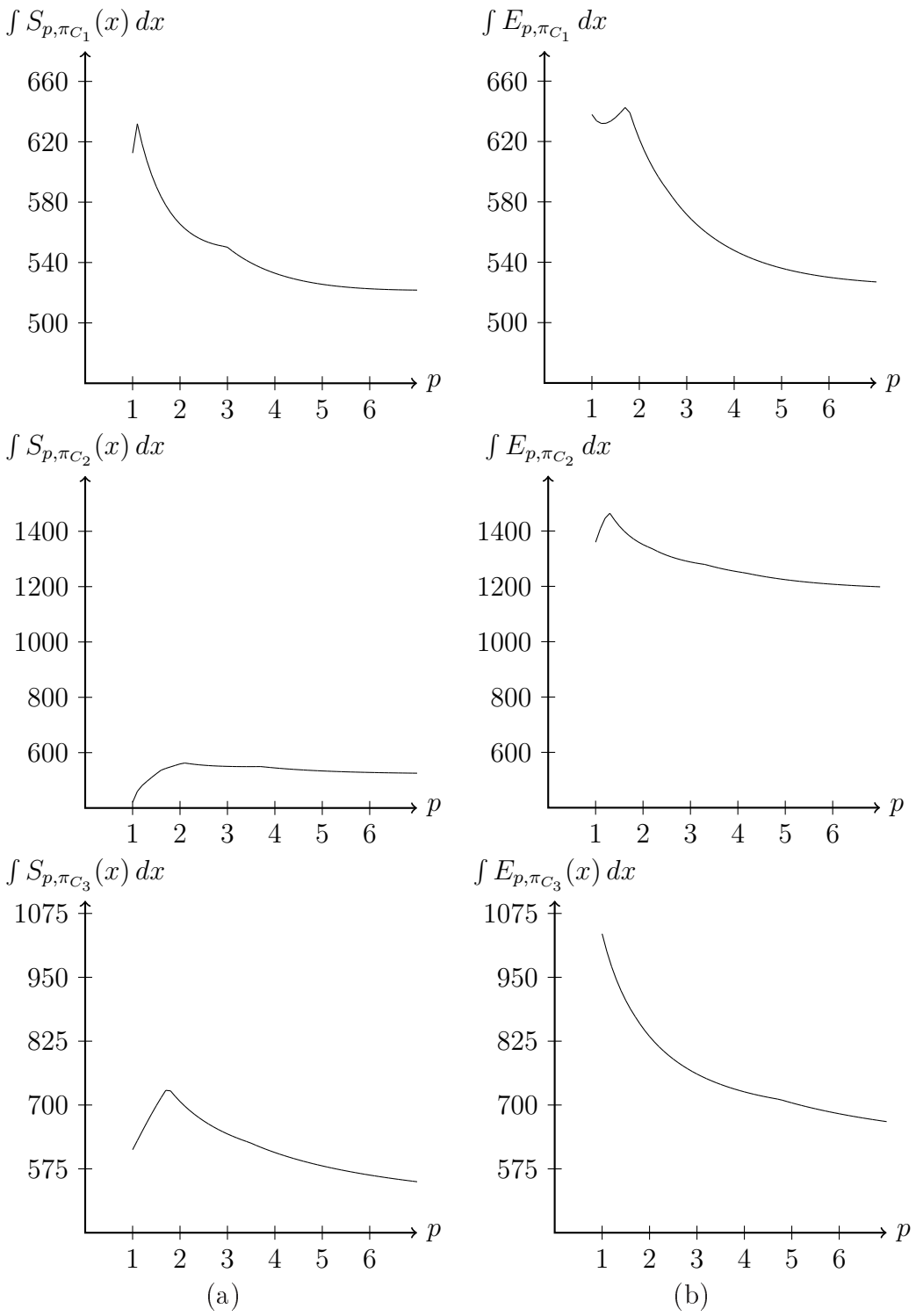

Figure 6: Estimated number of citations for exemplary sequences from Table 1 as a function of p.

Moreover, for $p = \infty$ we have

$$e_{\infty,a,b}(x) = \begin{cases} b & \text{for} \quad x \in [0, a)\,, \\ 0 & \text{for} \quad x \geq a. \end{cases} \tag{11}$$

We see that $e_{p,a,b}(x) = s_{p,b}(\frac{b}{a}x)$ for every x. Intuitively, for $x \in [0, a)$, the graph $e_{p,a,b}(x)$ determines a part of an ℓ^p-ellipse with semi-axes a and b. Therefore $e_{p,a,b}$ is further on called the *p-ellipse function* of *size* a and b.

Next we define a term analogous to the maximal p-radius $r_p(\pi_C)$.

Definition 16. A *maximal p-size* of a citation function π_C, denoted $l_p(\pi_C)$, is a pair (a, b), $a, b \geq 0$, which maximizes the area under $e_{p,a,b}$ and for which π_C still dominates $e_{p,a,b}$.

Please note that such definition is not strict, because there may exist $k > 1$ pairs satisfying the above condition. There are several possibilities to avoid this ambiguity. Selecting (a, b) such that $\frac{b}{a}$ is the closest to 1 tries to resemble the behavior of $r_p(\pi_C)$, as in this case equiaxial (or almost equiaxial) ellipse functions would be preferred. Another way is to choose the $\lceil \frac{k}{2} \rceil$-th pair in the sequence ordered by the first coordinate (the median). This case is more sensitive to shape of citation function. Please refer to the discussion below for further details.

For brevity, the p-ellipse function of maximal p-size $e_{p,l_p(\pi_C)}$ will be denoted by E_{p,π_C} (or even E_p) and called the *maximal p-ellipse function* for π_C.

Here are some basic properties of maximal p-size and maximal p-ellipse functions.

Lemma 17. *Given $p \geq 1$ and a citation function π_C based on a citation sequence $C = (c_1, c_2, \ldots, c_n)$, let $l_p(\pi_C) = (a, b)$. Then the following properties hold.*

(i) $a \leq n$.

(ii) $b \leq c_1$.

(iii) If $p < \infty$ then $\{x < n : \ E_p(x) = \pi_C(x)\} \subset \mathbb{N}_0$.

(iv) If $p = \infty$ then $\{x < n : \ E_p(x) = \pi_C(x)\} \cap \mathbb{N}_0 \neq \emptyset$.

The proof is left to the reader.

It is easy to see that for any citation function π_C

$$\int_0^\infty S_p(x)\,dx \leq \int_0^\infty E_p(x)\,dx \leq \int_0^\infty \pi_C(x)\,dx, \tag{12}$$

so by using l_p we get a better estimate of ζ_C than using r_p. Moreover, in this case, we are able to give an upper bound of the total number of citations.

Proposition 18. *For any given $p \geq 1$ and a citation function π_C based on the citation sequence $C = (c_1, c_2, \ldots, c_n)$, let $(a, b) = l_p(\pi_C)$. Then the following properties hold:*

(i) *If $p < \infty$ then $\int_0^\infty E_p(x)\,dx = a\,b\,\frac{1}{p}\,B\left(\frac{1}{p}, 1+\frac{1}{p}\right)$, where $B(\cdot,\cdot)$ is the Euler beta function.*

(ii) $\lim_{p\to\infty} \int_0^\infty E_p\,dx = a\,b.$

(iii) *If $p = \infty$ then*

$$\begin{aligned} \int_0^\infty \pi_C(x)\,dx &\leq (\ln n + 1) \int_0^\infty E_p(x)\,dx \\ &\leq (\ln a\,b + 1) \int_0^\infty E_p(x)\,dx. \end{aligned}$$

(iv) $\int_0^\infty E_p(x)\,dx \leq 2\int_0^\infty E_\infty(x)\,dx$ *for any* $1 \leq p < \infty$.

Proof. The proof of *(i)*, *(ii)* and *(iv)* is similar to that of Lemma 11 and 14.

Consider property *(iii)*. Of course, $\int_0^\infty E_\infty(x)\,dx = ab$. As (a, b) maximizes the area of $e_{\infty,a,b} \in L_{\pi_C}$ we have $c_1 \leq a\,b$. Indeed, if it does not hold we would have $\pi_C \succeq e_{\infty,1,c_1}$ and $\int_0^\infty e_{\infty,1,c_1}(x)\,dx > \int_0^\infty e_{\infty,a,b}(x)\,dx$ and we get a contradiction. Similarly we may show that

$$\begin{aligned} c_2 &\leq a\,b/2, \\ c_3 &\leq a\,b/3, \\ &\vdots \\ c_n &\leq a\,b/n. \end{aligned}$$

Then

$$\begin{aligned} \int_0^\infty \pi_C(x)\,dx &= \textstyle\sum_{i=1}^n c_i \\ &\leq a\,b \textstyle\sum_{i=1}^n \frac{1}{i} \\ &\leq a\,b\,(\ln n + 1)\,. \end{aligned}$$

Thus we have proved the first inequality. Since n can not be larger than ab, otherwise for $(n, 1)$ we could get a better dominating function (in sense of area), so $(\ln n + 1) \leq (\ln ab + 1)$ and the proposition follows. □

Considering computational aspects of l_p the problem is generally more difficult than in the preceding case. To determine the value of the maximal p-size we need two values (to define an unique ellipse). For $p = \infty$ the algorithm has linear complexity and is very easy to implement.

Unfortunately, for finite p the upper bound of the computational complexity is $O(n^3)$ ($O(n^2)$ pairs, $O(n)$ check for every pair). The routine is complicated and will not be included here. The possibility of its improvement is an interesting open problem.

Let us now go back for a moment to our example.

Example (cont.)
Let us consider again citation sequences given in Table 1. The corresponding citation functions together with p-ellipses are shown in Fig. 5b. The results of further analysis including p-sizes can be found in Table 3 and Fig. 6b.

We can see that l_p is a better estimate of ζ_C than r_p, especially for C_2. Even though l_∞ performance was the worst of the studied coefficients in this example, it has an important advantage: a clear and intuitive interpretation. As using the h-index one can state only that the second author has 22 papers with at least 22 citations each, applying l_∞ we can say that the most representative sample of his papers consists of 62 articles that received at least 18 citations each.

Please note, that the quotient b/a can be thought as a measure of type of the distribution. For $p = \infty$ in case of C_1 it equals about 0.96, for C_2 it is 0.29 and for C_3 3.07.

Id	n	l_1	l_2	l_∞	ζ_C	$\int E_1$	$\int E_2$	$\int E_\infty$
C_1	60	(34.9,36.6)	(30.3,26.1)	(23,22)	957	637.9	621.6	506
C_2	100	(99.7,27.3)	(77.1,22.3)	(62,18)	1651	1359.5	1350.9	1116
C_3	36	(25.4,81.4)	(19.4,54.7)	(14,43)	1342	1035.3	834.2	602

Table 3: Maximal p-sizes of citation sequences under study.

It is worth noting that $l_p(\pi_C)$ can be considered as a generalization of Kosmulski's *Maxprod* index.

Proposition 19. *For any citation sequence $C = (c_1, c_2, \ldots, c_n)$ let $(a, b) = l_\infty(\pi_C)$. Then*

$$a\,b = m, \tag{13}$$

where m is the, so called, individual's Maxprod-index [19].

Proof. By the definition of the *Maxprod*-index

$$m \;=\; \max\{i \cdot c_i :\; i = 1, 2, \ldots, n\}. \tag{14}$$

Assume $c_1 > 0$. Otherwise trivially $a\,b = 0 = m$. We have $\int_0^\infty e_{\infty,a,b}(x)\,dx = a\,b$, $e_{\infty,a,b} \in L_{\pi_C}$ and (a, b) maximizes the area under $e_{\infty,a,b}$. Then Lemma 17 and (11) gives immediately $a\,b = m$. □

As the maximal p-size is a two-dimensional measure, it cannot be used directly as an index of scientific impact. Although not recommended, due to loss of information, the maximal p-size can be projected into one dimension. We suggest here either using mean length of the axes, eg. $l_p^{(1)} := (a+b)/2$, or "normalized" diagonal length of the ellipse bounding rectangle $l_p^{(2)} := \sqrt{a^2+b^2}/\sqrt{2}$. Both measures in case $p=\infty$ give more credit to some (probably important and influential) contributions of "atypical" authors like those represented by C_2 and C_3 (see Table 4) than the h-index.

Id	r_∞	l_∞	$l_\infty^{(1)}$	$l_\infty^{(2)}$
C_1	22	(23,22)	22.5	22.5
C_2	22	(62,18)	40	45.7
C_3	22	(14,43)	28.5	32

Table 4: Maximal sizes of studied citation sequences projected to one-dimension.

It can be easily proved that the projected measures fulfill requirements for scientific impact indices given in Theorem 10.

Theorem 20. *For any given citation function π_C and an arbitrary $p \geq 1$, $l_p^{(1)}(\pi_C)$ and $l_p^{(2)}(\pi_C)$ are indices of scientific impact.*

6 Conclusions

In this paper we proposed a geometric approach to construction of scientific impact indices. Using suggested method we have obtained two families of indices described by one or two parameters, respectively. It was shown that some well-known indices, like Hirsch's h-index, Woeginger's w-index or Kosmulski's $Maxprod$-index are particular members of the proposed families.

The geometrical background of our indices might be useful in possible applications because one may choose a particular index basing on its clear geometric properties.

Our general method for constructing scientific impact indices was supplemented by practical effective algorithms for computing some of these indices.

References

1. Philip Ball. Index aims for fair ranking of scientists. *Nature*, 436:900, 2005.

2. Michael G. Banks. An extension of the Hirsch index: Indexing scientific topics and compounds. *Scientometrics*, 69(1):161–168, 2006.

3. Aparna Basu. A note on the connection between the Hirsch index and the Random Hierarchical model. *ISSI Newsletter*, 3(2):24–27, 2007.

4. Pablo D. Batista, Monica G. Campiteli, Osame Kinouchi, and Alexandre S. Martinez. Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68(1):179–189, 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. Quentin L. Burrell. Hirsch index of Hirsch rate? Some thoughts arising from Liang's data. *Scientometrics*, 73(1):19–28, 2007.

7. Quentin L. Burrell. Hirsch's h-index: A stochastic model. *Journal of Informetrics*, 1:16–25, 2007.

8. Rodrigo Costas and Maria Bordons. Is g-index better than h-index? An exploratory study at the individual level. *Scientometrics*, 2008. In press.

9. Leo Egghe. An Improvement of the H-index: the G-Index. *ISSI Newsletter*, 2(1):8–9, 2006.

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

11. Leo Egghe. Time-dependent Lotkaian informetrics incorporating growth of sources and items. *Mathematical and Computer Modelling*, 2008. In press.

12. 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.

13. Wolfgang Glänzel. On the opportunities and limitations of the H-index. *Science Focus*, 1(1):10–11, 2006.

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

15. Jorge E. Hirsch. Does the h-index have predictive power? *Proceedings of the National Academy of Sciences*, 104(49):19193–19198, 2007.

16. 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.

17. Clint D. Kelly and Michael D. Jennions. The h index and career assessment by numbers. *TRENDS in Ecology and Evolution*, 21(4):167–170, 2006.

18. 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.

19. Marek Kosmulski. MAXPROD — A new index for assessment of the scientific output of an individual, and a comparison with the h-index. *Cybermetrics*, 11(1), 2007.

20. Lokman I. Meho and Yvonne Rogers. Citation counting, citation ranking, and h-index of human-computer interaction researchers: A comparison between Scopus and Web of Science. *Journal of the American Society for Information Science and Technology*, 59(11):1711–1726, 2008.

21. Adam Proń and Halina Szatyłowicz. Habilitacja dodaje „skrzydeł"? *Forum Akademickie*, (3), 2006. (In Polish).

22. Ronald Rousseau. The influence of missing publications on the Hirsch index. *Journal of Informetrics*, 1:2–7, 2007.

23. Michael Schreiber. A case study of Hirsch index for 26 non-prominent physicists. *Ann. Phys.*, 16(9):640–652, 2007.

24. Antonis Sidiropoulos, Dimitrios Katsaros, and Yannis Manolopoulos. Generalized h-index for disclosing latent facts in citation networks. *Scientometrics*, 72(2):253–280, 2007.

25. Anthony F. J. van Raan. Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgement for 147 chemistry research groups. *Scientometrics*, 67(3):491–502, 2006.

26. Gerhard J. Woeginger. An axiomatic analysis of Egghe's g-index. *Journal of Informetrics*, 2(4):364–368, 2008.

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