https://doi.org/10.22306/asim.v10i2.109

Received: 27 Mar. 2024
Revised: 20 Apr. 2024
Accepted: 03 May 2024

The Hirsh index and its modifications

Waldemar Wozniak
University of Zielona Gora, Faculty of Mechanical Engineering, Institute of Materials and Biomedical Engineering, Department of Production Engineering and Transport, prof. Z. Szafrana 4, 65-516, Zielona Gora, Poland, EU, W.Wozniak@iimb.uz.zgora.pl (corresponding author)

Michal Sasiadek
University of Zielona Gora, Faculty of Mechanical Engineering, Institute of Mechanical Engineering, Department of Mechanics and Machine Design, prof. Z. Szafrana 4, 65-516, Zielona Gora, Poland, EU, M.Sasiadek@iim.uz.zgora.pl

Keywords: h-index, scientometry, nonadditive integrals.

Abstract: In recent years, scientometric indices, especially the Hirsch index, have gained significant attention from the scientific community due to some of their beneficial properties. The Hirsch h-index is widely used in various decision-making processes, such as evaluating scientific activity or allocating grants. Over time, many different indicators have been developed to address the shortcomings of the h-index. In this paper, we will focus on the Hirsch index, its strengths, and its weaknesses. Torra and Narukawa have demonstrated the relationship between the most commonly used scientometric indicators and fuzzy integrals, such as the Sugeno and Choquet integrals. This observation allows authors to define new indices using various fuzzy integrals and fuzzy measures.

1 Introduction

In 2005, the American theoretical physicist of Argentine origin, Jorge E. Hirsch, introduced the *h*-index, a new indicator for quantifying the results of scientists' research. It was designed as an alternative to other scientometric indicators, which until then primarily included the total number of publications, the average number of citations per paper, and the total number of citations received. The *h*-index is no longer used solely as a measure of scientific success for individual researchers but also for measuring the scientific output of research groups, scientific institutions, and even countries.

Let N_{p} represent the number of articles an author has published over n years. According to Hirsch's definition, a researcher has the h-index of h if h of their N_{p} articles have each been cited at least h times, while the remaining $(N_{\mathrm{p}} - h)$ articles have each been cited fewer than h times, cf [1]. Consider an author with a citation output of $\boldsymbol{x} = (6, 5, 3, 3, 1)$. One can see that they do not have the h-index of 4, as there are not 4 articles with at least 4 citations in the list. According to the definition, they should not have the h-index of 3, as well, because one of the remaining articles does not have less then 3 citations. This limitation is known as Hirsch’s threshold effect and the definition was improved by the author.

2 Mathematical background

In the whole text, we consider the set X of the input vectors $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ such that $x_1 \geq x_2 \geq \dots \geq x_n$ for $n \in \mathbb{N}$ and $x_k \in \mathbb{N} \cup \{0\}$ for all $k \in \{1, 2, \dots, n\}$. The vector $\boldsymbol{x} = (x_1, x_2, \dots, x_n) \in X$ represents a scientific record, where x_k is a number of citations of k-th publication. In the case of $x_k = 0$, the paper has no citations or does not exist. We will consider the scientific recocords for which $x_1 \geq 1$. In [2] Torra and Narukawa presented that the Hirsch index is the Sugeno integral with respect to the counting measure. Indeed, let $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ such that $x_1 \geq x_2 \geq \dots \geq x_n$ for $n \in \mathbb{N}$ and $x_k \in \mathbb{N} \cup \{0\}$ for all $k \in \{1, 2, \dots, n\}$ be the input vector of a scientist. Then the mathematical definition of the h-index reads as follows (1):

$$H(\boldsymbol{x}) = \max_{k \in \{1,2,\dots,n\}} \{\min\{k, x_k\}\} \tag{1}$$

To be this definition clearer, consider the scientific record $\boldsymbol{x} = (7, 6, 5, 4, 2)$. Then $H(\boldsymbol{x}) = \max\{\min\{1,7\}, \min\{2,6\}, \min\{3,5\}, \min\{4,4\}, \min\{5,2\}\} = \max\{1,2,3,4,2\} = 4$. Easier expression of the h-index should be (2)

$$H(\boldsymbol{x}) = \max_{k \in \{1,2,\dots,n\}} \{x_k \geq k\} \tag{2}$$

The graphical visualisation of the above example is demonstrated in the Figure 1.
One can see, that computing the h-index is finding the largest square under the number of citations.

Opinions on the usability of this index vary. The main discussion regarding the h-index concerns whether this index makes sense and whether it correctly reflects the level of influence of a particular researcher. As it turns out, everything has its strengths and weaknesses, which we will now briefly describe.

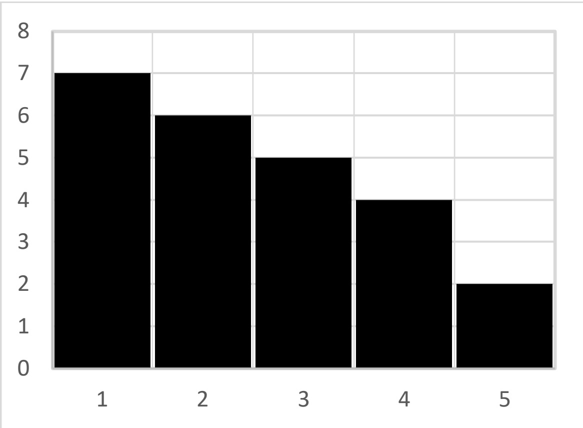

Figure 1 The Hirsch index

2.1 Pros and cons of h-index

2.1.1 Advantages

Hirsch's *h*-index offers several advantages that contribute to its widespread use in scientometric evaluations.

One of its key benefits is the simplicity of its calculation, making it an accessible metric for assessing scientific impact. It effectively combines both qualitative and quantitative indicators, as it takes into account both the number of citations and the number of publications. This dual nature allows for a more balanced evaluation of a researcher's contribution.

Furthermore, the *h*-index is highly versatile and can be applied not only to individual researchers but also to research groups, scientific institutions, and universities.

Another significant advantage is its immunity to uncited publications, ensuring that an author's overall impact is not skewed by a large number of works that have failed to gain recognition.

Additionally, the h-index provides a fair evaluation of scientists whose influence stems from a few highly cited papers that have made a significant impact on their field. At the same time, it supports researchers who consistently contribute to scientific progress through steady, ongoing work.

Lastly, the *h*-index remains independent of the author's prestige, focusing solely on the measurable impact of their research output rather than their reputation or institutional affiliation.

2.1.2 Disadvantages

Although the *h*-index is simple and effective, it has several weaknesses that limit its use in fairly evaluating scientific work. The aggregation of publications and citation frequency into a single representative value has been criticized by some authors because Hirsch, in his original work, assumed the equality of both inputs, which is an incorrect fundamental axiom, as demonstrated by the following example: Let author A have 10 papers with 10 citations each, and author B also have 10 papers with 10 citations each, but an additional 10 papers with 9 citations each. Obviously, we cannot consider these two authors equivalent, even though their *h*-index has the same value.
Many similar examples can be given because the *h*-index captures only a portion of publication and citation data. Despite its widespread use and advantages, the *h*-index has several limitations that have been identified over time, prompting researchers to propose various modifications and alternative indices.

One significant issue is that the *h*-index is affected by self-citations. Researchers who frequently cite their own work can artificially inflate their *h*-index, raising concerns about its accuracy. Although Hirsch originally claimed that self-citations have minimal impact, later studies, such as the one conducted by M. Schreiber in 2007, showed that self-citations can significantly influence the *h*-index, especially for early-career researchers with lower values. Since the *h*-index is commonly used for evaluating young researchers in promotions and competitive assessments, this poses a challenge. One possible solution is to exclude self-citations, though identifying and removing them can be difficult. Some scholars argue that even citations from co-authors should be disregarded to prevent bias.

Another drawback of the *h*-index is that it disadvantages researchers with shorter careers. Since the *h*-index is cumulative and never decreases, younger scientists are inherently at a disadvantage compared to senior researchers who have had more time to accumulate citations. To mitigate this issue, Hirsch proposed the m quotient, which normalizes the h-index by dividing it by the number of years since the researcher's first publication. This adjustment allows for a more fair comparison between scientists at different career stages.

Additionally, the *h*-index is insensitive to a large number of citations, meaning that an author with many highly cited papers may still have a relatively low *h*-index. For example, an author with four papers, each cited 200 times, has the same *h*-index as another author whose four papers have been cited only four times. This limitation led to the introduction of the *g*-index in 2006, which gives greater weight to highly cited papers. Another alternative is the *h(2)*-index, proposed by M. Kosmulski in 2006, which considers the number of citations squared for the most cited papers, providing a better representation of research impact.

A further limitation is that the *h*-index only takes integer values, meaning that a large group of researchers may share the same *h*-index, even if their actual citation impact differs significantly. Since it is rare for a scientist to achieve an *h*-index greater than 100 (which would require at least 10,000 citations), the index lacks fine granularity. To address this, some researchers have suggested mapping the *h*-index to the set of non-negative real numbers, improving its precision.

Another well-documented issue is that *h*-index values differ across scientific fields, making cross-disciplinary comparisons problematic. Researchers in some fields, such as physics, tend to receive more citations than those in fields like mathematics. Consequently, a mathematician with a lower h-index may have a greater impact in their field than a physicist with a higher h-index. This discrepancy arises due to differences in publication and citation practices among disciplines. One proposed solution is the constant ratio method, which normalizes citation counts within each field to enable fairer comparisons. However, in practice, this method provides only an approximate measure of influence.

Finally, the h-index can be affected by co-authorship. Researchers who collaborate extensively tend to produce more publications, potentially increasing their h-index. Additionally, papers with multiple authors may accumulate more citations, sometimes benefiting from self-citations among co-authors. To correct for this, several modified indices have been introduced. One approach involves adjusting the h-index based on the total number of authors contributing to an individual's publications. Another method, known as fractional citation counting, divides the number of citations for each paper by the number of authors before calculating the h-index, ensuring a more equitable measure of individual contribution.

These limitations highlight the need for alternative or complementary indices that address the shortcomings of the h-index, making scientific impact assessments more accurate and fair across different research disciplines and career stages.

3 Modifications of h-index

The first group of indices improving the h-index consists of the p-index, c-index and s-index. The p-index provides a number of publications with at least one citation whereas the c-index represents a number of citations of the most important paper. The p- and c-indices measure almost completely opposite aspects of the performance of a researcher. The p-index can be seen as a measure of productivity with focusing on productivity (i.e., number of papers) and paying almost no attention to impact (i.e., number of times a paper has been cited). On the other hand, the c-index can be seen as a measure of impact with focusing on impact and paying no attention at all to productivity. For instance, it prefers a single highly cited paper over a large number of slightly lower cited papers. Finally, the s-index defined by (3)

$$s(\boldsymbol{x}) = \sum_{k=1}^{n} x_k \tag{3}$$

Equals the total number of citations of all published papers. Thus the s-index takes into account all published papers of the scientist and not only the most cited.
Consider the output $\boldsymbol{x} = (14, 9, 9, 4, 2)$ of a scientist. Then the h-index $H(\boldsymbol{x}) = 4$, the p-index $p(\boldsymbol{x}) = 5$, the c-index $c(\boldsymbol{x}) = 14$ and the s-index $s(\boldsymbol{x}) = 48$.

To avoid some drawbacks of the h-index, there have been presented several other indices in the avaliable literature. One of them is the generalized Kosmulski index given by (4)

$$K_s(\boldsymbol{x}) = \max_{k\in\{1,2,\dots,n\}} \{x_k \geq s(k)\} \tag{4}$$

where $s: [0, \infty] \to [0, \infty]$ is arbitrary nondecreasing function. This index encompasses the Van Eck index [3] defined by (5)

$$h_\lambda(\boldsymbol{x}) = \max_{k\in\{1,2,\dots,n\}} \{x_k \geq \lambda. k\} \tag{5}$$

for arbitrary $\lambda > 0$ and *h(2)*-index (6) [4]

$$H2(\boldsymbol{x}) = \max_{k\in\{1,2,\dots,n\}} \{x_k \geq k^2\} \tag{6}$$

of Kosmulski or its extended version *h(m)*-index in the form (7)

$$HM(\boldsymbol{x}) = \max_{k\in\{1,2,\dots,n\}} \{x_k \geq k^m\} \text{ for } m = 3,4, \dots \tag{7}$$

Let us point out that the number λ indicates the degree of significance and for $\lambda = 1$ we have the h-index whereas *h(2)*-index takes into account the most cited publications.

Similar to the *h(2)*-index, there was introduced g-index [5] by L. Egghe to take into account influence of papers with high-citations number. It is defined by (8)

$$g(\boldsymbol{x}) = \max_{k\in\{1,2,\dots,n\}} \left\{\sum_{j=1}^{k} x_j \geq k^2\right\} \tag{8}$$

Considering the same scientist´s output $\boldsymbol{x} = (14, 9, 9, 4, 2)$. Then the Van Eck index $h_2(\boldsymbol{x}) = 3$, the Kosmulski index $H2(\boldsymbol{x}) = 3$ and the g-index $g(\boldsymbol{x}) = 6$.
The hg-index [6] is based on h- and g- indices as their geometrical mean, i.e. (9)

$$hg(\boldsymbol{x}) = \sqrt{H(\boldsymbol{x}). g(\boldsymbol{x})} \tag{9}$$

According to the definition, one can see that its geometrical interpretation is the square root of the rectangle with sides *H(**x**)* and *g(**x**)* while trying to maintain the advantages of both indices and minimize their disadvantages.

Another index containing the square root is the MAXPROD-index [7] which definition is (10)

$$MAXPROD(\boldsymbol{x}) = \sqrt{\max_{k\in\{1,2,\dots,n\}} \{k. x_k\}} \tag{10}$$

This index takes into account less cited publications with appropriate weights. For our scientis the value of hg-index is $hg(\boldsymbol{x}) = \sqrt{24}$ and the $MAXPROD(\boldsymbol{x}) = \sqrt{27}$.

Recently, in the work [8] authors proposed two indices which compensate some drawbacks of the h-index and they are called *the upper 2-h*-index H_2^u and *the lower 2-h*-index H_2^l. Their mathematical expressions are as follows (11)

$$H_2^u(\boldsymbol{x}) = \max_{k\in\{1,2,\dots,n\}}\{\min\{k+H(\boldsymbol{x}), x_k\}\},$$
$$H_2^l(\boldsymbol{x}) = \max_{k\in\{1,2,\dots,n\}}\{\min\{k, x_k+H(\boldsymbol{x})\}\} \quad (11)$$

In other words, we need to compute the classical *h*-index. By removing the number *H(x)* from the each citation, computing the *h*-index from the rest and adding this two numbers together, we get H_2^u. By removing the number *H(x)* from the number of papers, computing the *h*-index from the rest and adding this two numbers together, we get H_2^l. Let us point out that H_2^u compensates the *h*-index drawback about high-cited publications whereas H_2^l less- cited ones. We demonstrate the usage of this indices on the suitable example.

Consider the scientific record $\boldsymbol{x} = (6,6,4,3,1,1,1,0)$. Then the *h*-index $H(\boldsymbol{x}) = 3$. When removing the number of the *h*-index from the whole citations, we get the new vector $\boldsymbol{y} = (3,3,1,0,0,0,0,0)$ and the *h*-index of that vector $H(\boldsymbol{y}) = 2$. Thus $H_2^u = 5$. When removing the number of the *h*-index from the papers, we get the new vector $\boldsymbol{z} = (3,1,1,1,0)$ and the *h*-index of that vector $H(\boldsymbol{z}) = 1$. Thus $H_2^l = 4$. The situation is ilustrated in Figure 2.

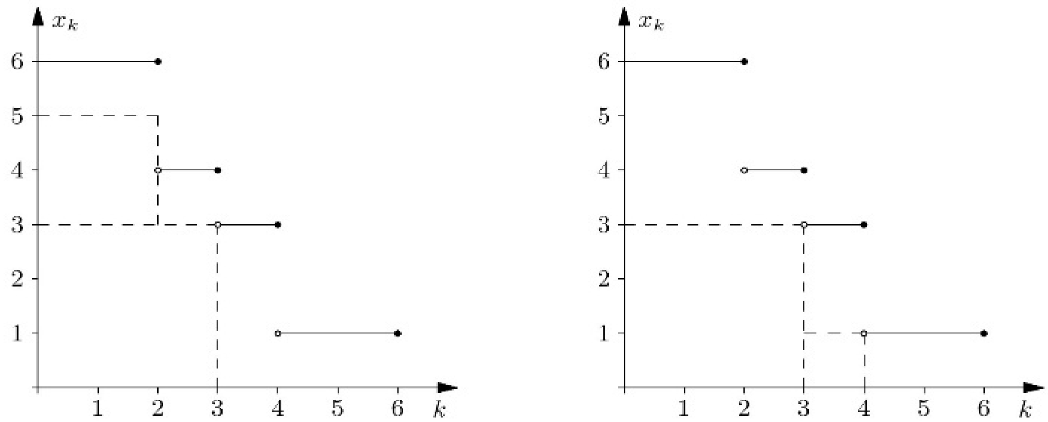

Figure 2 Comparison of H_2^u(left) and H_2^l(right)

4 Conclusion

The Hirsch index has gained significant attention from the scientific community due to its simple calculation and other favorable properties, such as its applicability not only to individuals but also to research groups and institutions. However, many scientists have pointed out the limitations of Hirsch's proposed index. Specifically, the *h*-index does not allow for fair comparisons between researchers from different disciplines. It is also highly sensitive to various parameters, such as career length. As a result, it is not suitable for comparing researchers from different age groups since the h-index is cumulative in nature. Another disadvantage is its insensitivity to citations of highly cited papers. The most frequently discussed issue is that the h-index can be influenced by self-citations, though this problem has been partially addressed in citation databases such as Scopus and Web of Science.

To mitigate these limitations, several alternative productivity indicators have been proposed. Many of them are merely variations of the h-index. Some of the more prominent alternatives include the *g*-index and *h(2)*-index, which address the issue of the *h*-index's insensitivity to highly cited papers. It has been shown that the *h*-index and certain other scientometric indices correspond to well-known aggregation functions, such as Sugeno, Choquet, and Shilkret integrals. These integrals belong to the category of fuzzy integrals, which have been extensively studied in recent years. This allows for the application of findings and knowledge from fuzzy integrals to citation indices. For this reason, in recent work [9] there has been presented the new overview of scientometric indices via iterated Sugeno integrals. The authors studied their mathematical properties in general which can be simply adopted to discrete version. Finally, in [10] there has been solved an open problem about full characterization of binary operation G satisfying the distributivity equation.

Acknowledgement
The state grant agency under the contract No. APVV-17-0258, APVV-19-0418, VEGA 1/0508/22, and KEGA 020TUKE-4/2023 supported this article.

References

[1] HIRSCH, E.: An index to quantify individual's scientific research output, *Proceedings of the National Academy of Sciences*, Vol. 102, No. 46, pp. 16569-16572, 2005. https://doi.org/10.1073/pnas.0507655102
[2] TORRA, V., NARUKAWA, Y.: The h-index and the number of citations: Two fuzzy integrals, *IEEE Transaction on Fuzzy Systems*, Vol. 16, No. 3, pp.795-797, 2008. https://doi.org/10.1109/TFUZZ.2007.896327

[3] VAN ECK, N.J., WALTMAN, L.: Generalizing the h- and g-indices, *Journal of Informetrics*, Vol. 2, No. 4, pp. 263-271, 2008. https://doi.org/10.1016/j.joi.2008.09.004
[4] KOSMULSKI, M.: A new Hirsch-type index saves time and works equally well as the original h-index, *ISSI Newsletter*, Vol. 2, No. 3, pp. 4-6, 2006.
[5] EGGHE, L.: Theory and practice of the g-index, *Scientometrics*, Vol. 69, No. 1, pp. 131-152, 2006. https://doi.org/10.1007/s11192-006-0144-7
[6] ALONSO, S., CABRERIZO, F., HERRERA-VIEDMA, E.: hg-index: A new index to characterize the scientific output of researches based on the h- and g- indices, *Scientometric*, Vol 82, No. 2, pp. 391-400, 2010. https://doi.org/10.1007/s11192-009-0047-5
[7] KOSMULSKI, M.: MAXPROD - a new index for assessment of the scientific output of an individual, and a comparison with the h-index, *Cybermetrics*, Vol. 11, No. 1, paper 5, 2007.
[8] MESIAR, R., GAGOLEWSKI, M.: H-index and other Sugeno integrals: Some defects and their compensation, *IEEE Transaction on Fuzzy Systems*, Vol. 24, No. 6, pp. 1668-1672, 2016. https://doi.org/10.1109/TFUZZ.2016.2516579
[9] BOCZEK, M., HOVANA, A., HUTNÍK, O: New monotone measure-based integrals inspired by scientific impact problem, *European Journal of Operational Research*, Vol 290, No. 1, pp. 234-357, 2021. https://doi.org/10.1016/j.ejor.2020.07.057
[10] BOCZEK, M., HOVANA, A., KALUSZKA, M.: On some distributivity equation related to minitive and maxitive homogeneity of the upper n-Sugeno integral, *Fuzzy Sets and Systems*, Vol. 430, pp. 102-113, 2022. https://doi.org/10.1016/j.fss.2021.02.016

Review process

Single-blind peer review process.