

A comparative assessment of additional bibliometric indexes in the case of a large excess of citations in relation to the h -index core

Ocena porównawcza dodatkowych wskaźników bibliometrycznych w przypadku dużego nadmiaru liczby cytowań w stosunku do wartości indeksu h

WIT GRZESIK*

DOI: <https://doi.org/10.17814/mechanik.2023.7.15>

A comparative assessment of citations based on additional bibliometric indicators and related to the excess of citations beyond that required for a given h -index is presented. The analysis was based on data from publications and the ranking in the category of scientific careers (“World’s TOP2% Most Influenced Scientists List”) published in 2022; two additional indicators were introduced: g and h_g . Numerous examples of the desirability and legitimacy of such an analysis for scientists from various scientific disciplines are given.

KEYWORDS: Hirsch index, World’s TOP2% Most Influenced Scientists, citations

Przedstawiono ocenę porównawczą cytowań przeprowadzoną na podstawie dodatkowych wskaźników bibliometrycznych i odniesioną do nadmiaru liczby cytowań poza wymaganą dla danego indeksu h . Analizę oparto na danych z publikacji oraz rankingu w kategorii kariery naukowej (“World’s TOP2% Most Influenced Scientists List”) opublikowanego w 2022 r.; wprowadzono dwa dodatkowe wskaźniki: g i h_g . Podano liczne przykłady celowości i zasadności takiej analizy dla naukowców z różnych dyscyplin naukowych.

SŁOWA KLUCZOWE: indeks Hirscha, World’s TOP2% Most Influenced Scientists, cytowania

Wprowadzenie

W poprzednich publikacjach autora [1–4] dokonano oceny stanu nauki polskiej, instytucji naukowych i dorobku naukowego indywidualnych autorów na podstawie dostępnych baz naukowych [5], najbardziej rozpowszechnionego indeksu Hirscha (h) oraz innych wskaźników bibliometrycznych wprowadzanych sukcesywnie w naukometrii [6–8]. Obecnie – pomimo wykazanych niedoskonałości i ograniczeń – wspomniane oceny są dokonywane prawie wyłącznie na podstawie bazowego indeksu cytowań h , który został wprowadzony na dużą skalę do naukometrii w 2005 r., a w 2009 r. został także zastosowany przez Schuberta do oceny publikacji indywidualnych autorów [8]. Chociaż indeks h łączy – jako jeden prosty wskaźnik – ocenę wydajności publikacyjnej na podstawie ilości publikacji autora i ocenę jakości na

podstawie liczby cytowań, to od początku pojawiają się liczne uwagi krytyczne co do jego stosowności jako podstawowej metryki oceny naukowej – o czym świadczy krytyczny artykuł opublikowany już w 2006 r. [2, 9]. Dotyczą one nie tylko braku uwzględnienia rzeczywistej liczby cytowań N_p dla danej publikacji poza indeksem h $N_h = (N_p - h)$, czyli **nadmiarowości cytowań**, wpływu narzędzi stosowanych do zarządzania danymi bibliometrycznymi czy silnej zależności od długości kariery naukowej, ale również występujących wyraźnych różnic w jego estymacji, gdy korzysta się z różnych baz danych, tj. Google’s Google Scholar (GS), Elsevier’s Scopus, Clarivate Analytics’ Web of Science (WoS) oraz akademickiego portalu społecznościowego (*academic social networking site* – ASNS) – ResearchGate (RG) [2, 4]. Kolejny artykuł w tym cyklu przybliży metodykę i efekty porównawcze wynikające z uwzględnienia dwóch dodatkowych wskaźników: g i h_g , które w sposób ilościowy i jakościowy charakteryzują występujący zwykle nadmiar cytowań w stosunku do definicji podstawowego indeksu h .

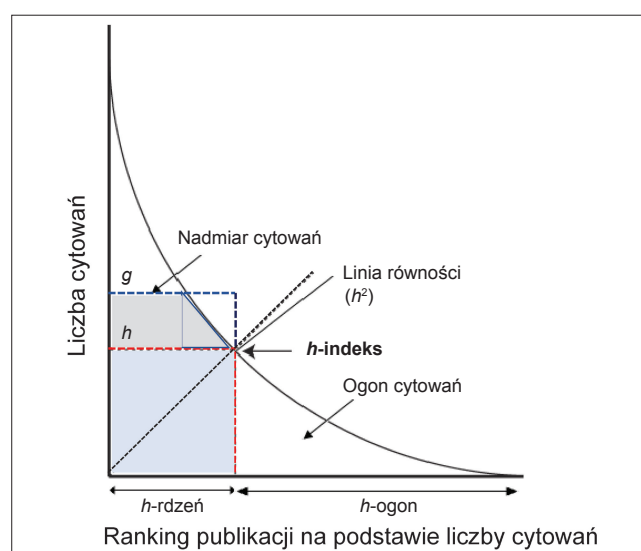


Fig. 1. Graphical illustration of the citation distribution based on the h -index definition [9, 10]

Rys. 1. Graficzna ilustracja rozkładu cytowań według definicji indeksu h [9, 10]

* Prof. dr hab. inż. Wit Grzesik – wit.grzesik@gmail.com, <https://orcid.org/0000-0003-3898-5119> – Opole, Polska

W niniejszym artykule jest rozważany ten pierwszy problem, przedstawiony graficznie na rys. 1; a ostatni problem został już częściowo ujęty we wcześniejszych publikacjach o charakterze ogólnym [3, 4]. Jeśli analizuje się dane uwzględnione w obydwóch kategoriach (kariery i cytowań za dany rok) w rankingu TOP2%, to zwykle się okazuje, że nadmiarowość liczby cytowań w stosunku do wymaganej liczby h^2 jest zjawiskiem powszechnym i również wymaga wnikliwej oceny bibliometrycznej. Występują bowiem liczne przypadki, gdy naukowcy mają wykazany mały indeks Hirscha, ale bardzo dużą liczbę cytowań jednego lub kilku artykułów (nadmiar ponad h -rdzeniem na rys. 1), nawet większą od naukowców z dużym indeksem h . Znakiem przykładem może być Albert Einstein, dla którego orientacyjnie wyznaczono indeks $h = 4(5)$, podczas gdy cytowania jego prac idą w setki tysięcy [11]. Z drugiej strony możliwa jest duża liczba artykułów o liczbie cytowań mniejszej niż h (h -ogon na rys. 1).

Hirsch [7, 12] ustalił doświadczalnie, że całkowita liczba cytowań N jest proporcjonalna do kwadratu indeksu h (h^2), tj. $h = \sqrt{\frac{N}{a}}$ ze współczynnikiem proporcjonalności $a = 3 \div 5$ (ale współczynnik $a = 1$ oznacza, że indeks h został uzyskany bez nadmiaru cytowań).

W pracy [9] podano dwa hipotetyczne przykłady, gdy jeden z naukowców opublikował 20 artykułów, z których tylko jeden był cytowany 500 razy (czyli uzyskał indeks $h = 1$), natomiast drugi opublikował też 20 artykułów, ale każdy z nich był cytowany po 20 razy (czyli wykazał wymagane minimum 400 cytowań do uzyskania indeksu $h = 20$). Jak w takim przypadku ocenić te dwa skrajne przypadki (500 cytowań vs. 400 cytowań) w kontekście wkładu naukowego i – co najważniejsze – na czyją korzyść?

Nawet dla osób z dużym indeksem h nadmiarowość cytowań w części rdzenia jest zwykle znaczna. Czy w takich przypadkach można jednoznacznie wskazywać na korzyść z dużego indeksu h ? Z analiz przeprowadzonych przez autora popularyzującego zasadne stosowanie wskaźników bibliometrycznych w światowych rankingach naukowych wynika, że w rankingu TOP2% są zamieszczeni naukowcy o takich strukturach cytowań (choć mniej liczni o dużym nadmiarze cytowań w zakresie h -rdzenia na rys. 1). Są to w większości naukowcy z długoletnim stażem naukowym, uwzględnieni w rankingu TOP2% w kategorii kariery naukowej [13], co z uwagi na liczne publikacje o najwyższej cytowalności może się przyczyniać do wzrostu nadmiaru cytowań dla tych publikacji.

W artykule dokonano analizy porównawczej tradycyjnego indeksu h oraz zaproponowanych w pracach [9, 10, 13, 14] indeksów g i h_g . Jak wskazano wcześniej, wysokie wartości indeksu g świadczą o dużych liczbach najczęściej cytowanych artykułów, które są odpowiednio większe od indeksu h . Z uwagi na ograniczoną dostępność do źródłowych baz danych w analizach wykorzystano wskaźniki określające całkowitą liczbę cytowań z autocytowaniami ($nc9621$) i wyznaczony na tej podstawie indeks h ($h21$) dla kryterium kariery naukowej w okresie 1996–2021, dostępny w rankingu TOP2% za 2021 r. [5].

Metodyka analiz i ocen porównawczych

Dla przypomnienia indeks h został zdefiniowany następująco [12]: *Badacz ma wykazany indeks h , jeśli opublikowanych P artykułów dostało co najmniej h cytowań każdy, a reszta ($P - h$) artykułów otrzymała nie więcej niż h cytowań.*

W formie matematycznej definicję tę można zapisać następująco [6]:

$$h = \max \{j: \text{cit}_j \geq j\}$$

gdzie: cit_j jest numerem cytowania j -tego artykułu. Artykuły uwzględnione w wyznaczeniu h -indeksu są zwykle oznaczone jako h -istotne (h -core papers), co zilustrowano na rys. 1.

Indeks g (g -indeks) zaproponowany w 2006 r. przez Egghe [9], a więc krótko po propozycji Hirscha [10], jest oparty na przekształceniu bazowego indeksu h i zdefiniowany jako liczba g najczęściej cytowanych publikacji, które uzyskały łącznie g^2 cytowań [2]. To również znaczy, że czołowe/topowe ($g + 1$) artykuły będą mieć mniej niż $(g + 1)^2$ cytowań. Przykładowo: autor, którego g -indeks = 10, opublikował 10 artykułów, których łączna liczba cytowań jest nie mniejsza niż 100. Wysokie wartości indeksu g świadczą o dużych liczbach najczęściej cytowanych artykułów i z reguły powodują nadmiarowość cytowań w zakresie h -rdzenia/trzonu (rys. 1). Przekłada się to na cenną zaletę w postaci ujęcia bieżących cytowań bez wykazania wzrostu indeksu h . To z kolei można traktować jako ograniczenie powodujące silną zależność indeksu g nawet od jednego artykułu cieszącego się największą popularnością. Można łatwo dowiedzieć, że $g \geq h$.

Indeks h_g jest definiowany jako średnia geometryczna [7, 10, 13, 14]:

$$h_g = \sqrt{h \times g}$$

Łatwo wykazać, że $h \leq h_g \leq g$ oraz że $h_g - h \leq g - h_g$, co oznacza, że h_g -indeks jest zbliżony co do wartości bardziej do h niż g . Ta cecha może być postrzegana jako ograniczenie dla indeksu g w przypadku bardzo małego indeksu h i unikania problemu wpływu artykułu o nadmiernej cytowalności. W aspekcie geometrycznym indeks h_g może być traktowany jako pierwiastek z zastępczego pola powierzchni prostokąta o wymiarach $h \times g$ w stosunku do pola powierzchni pod krzywą (hiperbolą) Hirscha (rys. 1), co przedstawiono na rys. 2. Indeks h_g dobrze koreluje z indeksem h , ponieważ współczynnik korelacji wyznaczony za pomocą rankingu Spearmana (*Spearman's rank correlation coefficient*) wynosi 0,992 [8], co też można dostrzec na rys. 4.

Jeśli porówna się dorobek publikacyjny dwóch autorów z uzyskaną identyczną liczbą cytowań, z których pierwszy opublikował 30 artykułów, ale tylko 1 z nich dostał 500 cytowań ($h = 1$), a drugi 50 artykułów i każdy z nich dostał 10 cytowań ($h = 10$) to indeksy g będą równe 22 ($22^2 = 484 < 500$ i $23^2 = 529 > 500$) i 10 ($10^2 = 100$ i $11^2 = 121 > 100$). Na tej podstawie

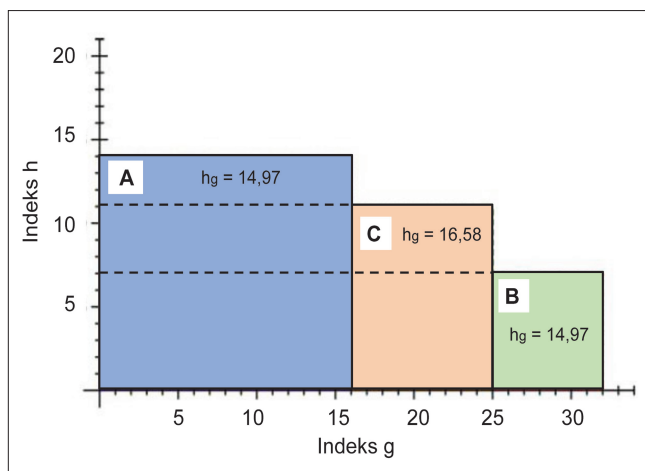


Fig. 2. Geometrical interpretation of the h_g -index [13]
Rys. 2. Geometryczna interpretacja indeksu h_g [13]

można na pierwszy rzut oka sądzić, że drugi autor jest lepiej postrzegany w środowisku naukowym ze względu na większą wydajność publikacyjną i większą dostrzegalność prac naukowych. Jednakże indeks g jest dużo mniejszy niż w przypadku pierwszego autora (10 vs. 22), którego 29 artykułów pozostało anonimowymi. Z tego względu uzasadnione jest wprowadzenie wskaźnika h_g , który bardziej bilansuje wkład naukowy obydwój autorów, tj. $h_{g1} = \sqrt{1} \times 22 = 4,7$ i $h_{g2} = \sqrt{10} \times 10 = 10$, a przez to minimalizuje wpływ tylko jednej wysoko cytowanej publikacji. Przedstawione relacje można łatwo odczytać z rys. 2, ponieważ jednakowy indeks $h_g = 14,97$ można uzyskać przy mniejszej wartości wskaźnika g i wyższym indeksie h (autor A: $h = 14, g = 16$), lub odwrotnie (autor B: $h = 7, g = 32$), co wynika z prostej relacji geometrycznej. Z kolei autor C ($h = 11, g = 25$) uzyskał wyższy indeks $h_g = 16,58$ z pośrednich wartości indeksów h i g od autorów A i B ($h_B < h_C < h_A$ ale $g_A > g_C > g_B$), co widać na rys. 3.

Na rys. 3 przedstawiono graficzną ilustrację warunków wzrostu indeksu h_g jako funkcji indeksów h i g . Można dostrzec, że indeks h_g kompensuje wpływ dużego indeksu g i małego indeksu h . Naniesiono też dane rzeczywiste z tablicy ($h = 10, g = 18, h_g = 13,42$).

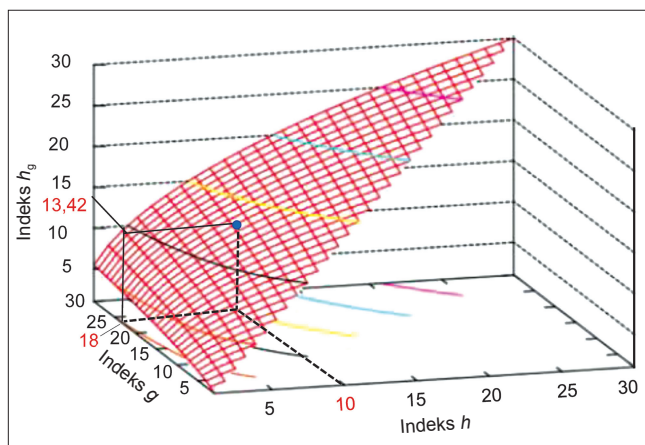


Fig. 3. Illustration of the growth in the h_g -index as a function of h and g indices [13]
Rys. 3. Ilustracja przyrostu indeksu h_g w funkcji indeksów h i g [13]

Porównanie osiągnięć naukowców w ujęciu wskaźników h, g i h_g

W tabelicy porównano wskaźniki bibliometryczne $h, h_m, g, g/h$ i h_g wyznaczone dla danych źródłowych dostępnych w bazach rankingu TOP2% [5]. Jest to uzupełnienie wcześniejszego opracowania, opublikowanego w artykule [2]. Przykłady dotyczą wybranych naukowców z Polski reprezentujących nauki inżyniero-techniczne.

TABLE. Authors ranking according to scientific career criterion in 2021 (Authors_career_2021)

TABLICA. Ranking autorów według kryterium kariery naukowej w 2021 r. (Authors_career_2021)

Nr	A $h_{21}(ns)$	B h_{21}	C h_m21	D g	E g/h_{21}	F h_g	G $nc9621$
1	73	81	50,8120	161	1,99	114,20	25 970
2	54	70	20,7059	132	1,89	96,12	17 476
3	41	52	16,2113	106	2,04	74,24	11 439
4	35	41	21,7077	80	1,95	52,27	6 538
5	29	31	22,1444	55	1,77	41,29	3 052
6	17	22	14,0833	38	1,29	28,91	1 498
7	13	17	12,7540	26	1,31	21,02	679
8	7	10	9,5000	18	1,43	13,42	329

Legenda:

A – $h_{21}(ns)$ – h -indeks na koniec 2021 r. (bez autocytowań)

B – h_{21} – h -indeks na koniec 2021 r. (z autocytowaniami)

C – h_m21 – h_m -indeks na koniec 2021 r.

D – indeks g

E – stosunek g/h_{21}

F – indeks h_g

G – $nc9621$ – całkowita liczba artykułów opublikowanych w latach 1996–2021

Źródło: "World ranking of Top2% scientists in 2021" (www.Elsevier.digitalcommonsdata.com).

Na rys. 4a i b przedstawiono związki między indeksami g i h dla ogółu naukowców z Polski (a) oraz indeksami g, h_g i h dla danych zebranych w tabelicy (b), które obrazują skalę nadmiaru cytowań występującą w rankingu TOP2% [5].

Wnioski:

- Potwierdza się charakter transformacji geometrycznej dla malejącego indeksu h , chociaż w świetle danych z rankingu TOP2% przekształcenia jak na rys. 2 mają raczej charakter teoretyczny.
- Nadmiar cytowań występuje dla całego zakresu indeksu h , jak na rys. 4a. Jednakże, widać wyraźnie, że dla dużych i bardzo dużych wartości indeksu h (orientacyjnie powyżej 30) występuje większy nadmiar cytowań.
- Relacja między wskaźnikami g i h dla dużego zbioru danych (rys. 4a) jest liniowa ($y = 2,2667 \times x - 7,3503$) ze współczynnikiem korelacji 0,93.
- Relacja między indeksami h i h_g wyrażona za pomocą współczynnika g/h zależy od wartości indeksu h . Dla wskaźnika h umownie powyżej 30 współczynnik g/h jest równy średnio 2, natomiast dla wskaźnika poniżej 20 maleje do wartości 1,3 (1,4).
- Dla rzeczywistych danych spełniony jest warunek matematyczny, że $h \leq h_g \leq g$.

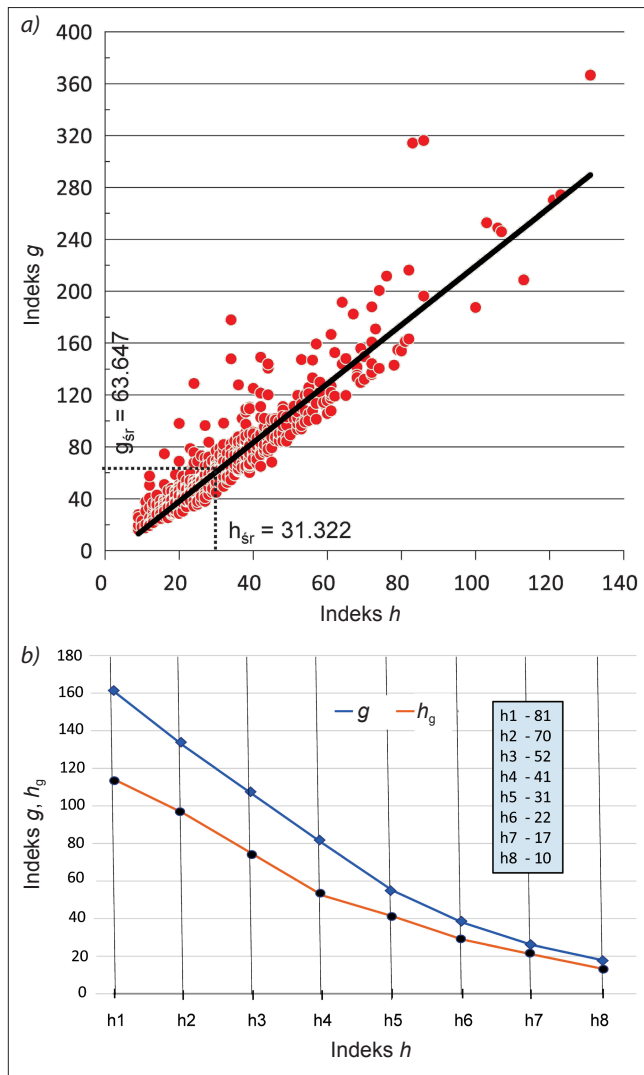


Fig. 4. Relationships between h - and g -indices and h_g -index based on data from TOP2% ranking for Poland [5] (a) and selected in the embedded table (b)

Rys. 4. Związki między indeksami h , g i h_g dla ogółu naukowców z Polski, oparte na danych z rankingu TOP2% [5] (a) i danych z tablicy (b)

• Na podstawie nowych wskaźników można stwierdzić, że znaczny nadmiar cytowań może pochodzić z autocytowań, jak w przypadku autorów nr 2 i 3 w tablicy.

Na rys. 5a–c porównano indeksy h (a), g (b) i h_g (c) dla 10 wybranych autorów według danych z artykułu [13]. Wartości indeksu h uszeregowano w kolejności malejącej od 27 do 12, czyli w zakresie odpowiadającym największej częstości w rankingu TOP2% [3, 4]. Można zauważyć, że jedynie dla autorów A1 i A2, którzy wykazali największe indeksy h równe 27, ustalono identyczną kolejność indeksów g i h_g . Rankingi dla pozostałych autorów są już przypadkowe ze względu na różne nadmiary cytowań i pojawiającą się z tego powodu dużą wrażliwość zmian.

Z przytoczonych danych (rys. 4 i 5) i przeprowadzonej analizy wynika, że indeksy g i h_g mogą stanowić cenne uzupełnienie oceny dorobku naukowego osób, które dysponują porównywalnym (np. trzech autorów na rys. 5a z indeksem $h = 18$) lub mniejszym indeksem h , ale mającym większy nadmiar cytowań w stosunku do wymaganej liczby $N_h = h^2$, czyli mini-

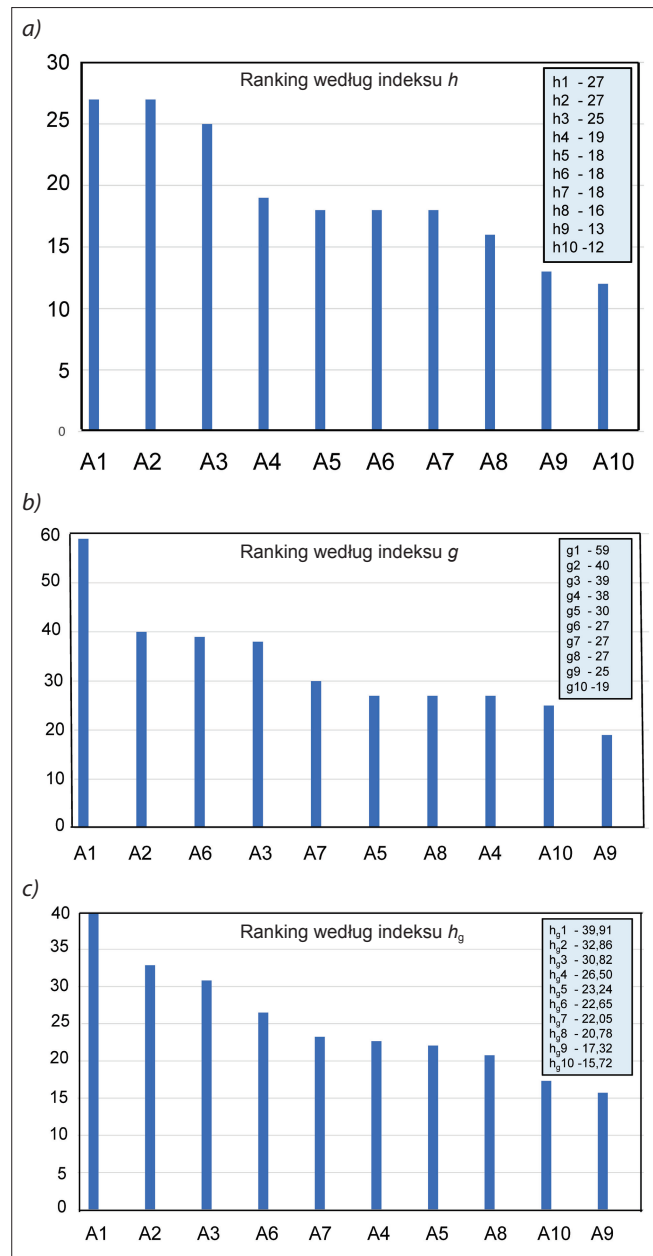


Fig. 5. Comparison of: a) h -index, b) g -index and c) h_g -index for 10 authors acc. to data from Ref. [13]

Rys. 5. Porównanie indeksów: a) h , b) g i c) h_g dla 10 autorów według danych z artykułu [13]

mum dla danego indeksu h . W grupie autorów A4–A8 indeks h wynosi 19, trzykrotnie 18 i 16, ale ponieważ autor A4 ($h = 19, g = 27$) nie dysponował dużym nadmiarem cytowań, zajął dopiero ósmą lokatę w rankingu według indeksu g (rys. 4b). Natomiast autor A6 z mniejszym indeksem h , ale większym nadmiarem cytowań ($h = 18, g = 39$) został sklasyfikowany w tym rankingu na trzeciej pozycji.

Więcej informacji o specyfice rozkładu indeksów h i h_g w różnych dyscyplinach naukowych, m.in. naukach ścisłych (chemii i fizyce), medycznych i inżynierskich, można znaleźć w artykułach analizujących publikacje autorów z różnych krajów [13, 14]. Na uwagę zasługuje fakt, że w tym przypadku największy nadmiar publikacji (ocenił na podstawie średniej wartości współczynnika g/h) mają naukowcy reprezentujący fizykę (1,85) i medycynę (1,60). W naukach inżynierskich jest to ok. 1,45. Natomiast średni współczynnik

nadmiaru cytowań dla wszystkich ujętych dyscyplin to ok. 1,5 [14]. Dla polskich autorów, zgodnie z modelem liniowym przedstawionym na rys. 4a, to ok. 2 ($h_{sr} = 31,32$; $g_{sr} = 63,65$). Jeśli natomiast uwzględnimy ok. 92% autorów z indeksem h do 50, to średnie indeksy $h_{sr} = 28$ i $g_{sr} = 56,13$.

LITERATURA

- [1] Grzesik W. „Stan nauk inżynierskich i technicznych w Polsce w świetle rankingu TOP2% Stanford University & Elsevier” [“Current state of engineering and technical sciences in Poland according to TOP2% Stanford University & Elsevier ranking”]. *Mechanik*. 8–9 [2022]: 28–30.
- [2] Grzesik W. “How one can properly quantify an individual scientific impact for multi-authored publications based on bibliometric data” [„Jak można właściwie ocenić indywidualny wkład naukowy w publikacjach wieloautorowych na podstawie danych bibliometrycznych”]. *Mechanik*. 3 (2023): 40–45, <https://doi.org/10.17814/mechanik.2023.3.6>.
- [3] Grzesik W. „Czy wzmocnienie pozycji polskiej nauki jest realne. Analizy w oparciu o ranking World’s TOP2% Scientists 2022”. *Forum Akademickie*. 4 (2023).
- [4] Grzesik W. “An attempt for the assessment of publications by Polish scientists in 2021 based on the world’s citation ranking” [„Próba oceny publikacji polskich naukowców w 2021 r. na podstawie światowego rankingu cytowań”]. *Mechanik*. 5–6 (2023): 48–51, <https://doi.org/10.17814/mechanik.2023.5-6.11>.
- [5] Bazy danych: www.Elsevier.digitalcommonsdata.com.
- [6] Todeschini R. “The j -index: a new bibliometric index and multivariate comparisons between other common indices”. *Scientometrics*. 87 (2011): 621–639, <https://doi.org/10.1007/s11192-011-0346-5>.
- [7] Alguliev R.M., Aliguliyev R.M., Fataliyev T.K., Hasanova R.Sh. “Weighted consensus index for assessment of the scientific performance of researchers”. *Collnet Journal of Scientometrics and Information Management*. 8, 2 (2014): 371–400, <https://doi.org/10.1080/09737766.2014.954864>.
- [8] Zhenbin Yan, Qiang Wu, Xingchen Li. “Do Hirsch-type indices behave the same in assessing single publications? An empirical study of 29 bibliometric indicators”. *Scientometrics*. 109 (2016): 1815–1833, <https://doi.org/10.1007/s11192-016-2147-3>.
- [9] Egghe L. “Theory and practise of the g -index”. *Scientometrics*. 69 (2006): 131–152, <https://doi.org/10.1007/s11192-006-0144-7>.
- [10] Bihari A., Tripathi S., Deepak A. “ h -index and its alternative: a review”. *Journal of Information Science*. (2021): <https://doi.org/10.1177/01655515211014478>.
- [11] https://pl.wikipedia.org/wiki/Wskaźnik_Hirscha.
- [12] Hirsch J.E. “An index to quantify an individual’s scientific research output”. *Proceedings of the National Academy of Science of USA*. 102, 46 (2005): 16569–16572, <https://doi.org/10.1073/pnas.0507655102>.
- [13] Alonso S., Cabrerizo F.J., Herrera-Viedma E., Herrera F. “ h_g -index: a new index to characterize the scientific output of researchers based on the h - and g -indices”. *Scientometrics*. 82 (2010): 391–400, <https://doi.org/10.1007/s11192-009-0047-5>.
- [14] Hadagali G.S., Kumbar B.D., Gourikeremath G.N., Hiremath R. “ g -index as an improvement of the h -index: A comparative study of prominent Indian scientists”. *International Journal of Information Dissemination and Technology*. 6(S1) (2016): S42–48. ■