



Robust h -index

Maurice Poirrier¹ · Sebastián Moreno¹  · Gonzalo Huerta-Cánepa¹

Received: 25 March 2020 / Accepted: 27 December 2020 / Published online: 5 February 2021
© Akadémiai Kiadó, Budapest, Hungary 2021

Abstract

The h -index is the most used measurement of impact for researchers. Sites such as Web of Science, Google Scholar, Microsoft Academic, and Scopus leverage it to show and compare the impact of authors. The h -index can be described in simple terms: it is the highest h for which an authors has h papers with the number of cites more or equal than h . Unfortunately, some researchers, in order to increase their productivity artificially, manipulate their h -index using different techniques such as self-citation. Even though it is relatively simple to discard self-citations, every day appears more sophisticated methods to artificially increase this index. One of these methods is collaborative citations, in which a researcher A cites indiscriminately another researcher B , with whom it has a previous collaboration, increasing her/his h -index. This work presents a new robust generalization of the h -index called rh -index that minimizes the impact of new collaborative citations, maintaining the importance of their citations previous to their collaborative work. To demonstrate the usefulness of the proposed index, we analyze its effect over 600 Chilean researchers. Our results show that, while some of the most cited researchers were barely affected, demonstrating their robustness, another group of authors show a substantial reduction in comparison to their original h -index.

Keywords h -Index · h -index manipulation · Robust h -index · Self-citation · Collaborative citation

Introduction

Several processes ranging from the ranking of applicants for academia/company positions, up to the assignment of millions of dollars on research grants, consider the measurement of a researcher's scientific production as one of their essential elements. The measurement of the scientific production of a researcher has evolved from the number of papers published in

✉ Sebastián Moreno
sebastian.moreno@uai.cl

Maurice Poirrier
mpoirrier@alumnos.uai.cl

Gonzalo Huerta-Cánepa
gonzalo.huerta.c@edu.uai.cl

¹ Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Viña del Mar, Chile

previous years to its current form, which considers not only the number of published work but other characteristics such as their citations (Egghe 2006; Jin et al. 2007; Cormode et al. 2013; Oberesch and Groppa 2017). Among these, the most important, in terms of usage, is the one known as h -index (Hirsch 2005), created in 2005 by Hirsch.

The popularity of the h -index is mainly due to its simple definition and calculation (Alonso et al. 2009). A definition of the index is as follows: for the highest h , if an author has h papers with more or equals cites than h , the index of the author is h . Given its simplicity, the h -index is used both on respected scientific production databases such as Web of Science, Google Scholar, Scopus, and academic institutions for tasks such as evaluation of academic departments (Meyers and Quan 2017). However, the h -index presents a problem: it can be manipulated through irregular citations to increase its value (Zaggl 2017; Bartneck and Kokkelmans 2011).

The most known irregular citation practice is through the use of self-citations (Seeber et al. 2019). Even though we consider that self-citations are essential to show the evolution of a research topic, the abuse of them is irregular. As an example, a study over three years of scientific production at Norway showed, on average, that 36% of citations were self-citations (Aksnes 2003). In preparation for this work, as an example, we found an author whose h -index decreases from 54 to 43 when self-citations are not considered. Previous researches propose several solutions to minimize irregular citations that range from not counting self-citations in the h -index calculation, to the creation of new indexes such as: $h(2)$ -index (Kosmulski 2006) and h_w -index (Egghe and Rousseau 2008). Even though most solutions have been focused on self-citations, there are other ways to increase the h -index, including anomalous citations among authors.

Irregular citations happen when co-authors of a paper agree to cite each other works in future works. For example, if authors A and B are co-authors in a couple of works, then author B indiscriminately cites work from A, with the sole purpose of increasing her/his h -index artificially, even when citations are barely related to the topic (Krampen et al. 2007).

In this paper, we propose a robust generalization of the h -index called robust h -index (rh_i -index), where i defines the minimum distance in the author collaboration network to consider a citation in the index's calculation. In this case, rh_0 -index and rh_1 -index are equivalent to h -index and h -index without self-citations, respectively. To show the validity of this work, we compare rh_2 , where co-authors' citations are not considered (citations previous to their collaboration are still counted), against rh_0 and rh_1 for 659 authors that work on any Chilean institution, and have an h -index greater than 10. Our results show that most of the researchers present a minimum variation between rh_2 and rh_1 , but some authors show significant differences, greater than 10 in a particular case.

The remainder of this paper is organized as follows. “***h*-Index**” section explains the current h -index with their deficiencies and some variations. “**Robust h -index (rh -index)**” section presents the new robust h -index. “**Analysis**” section shows the analysis of the rh_2 -index against rh_0 and rh_1 , describing data acquisition, preparation, results, and discussion. Finally, conclusions are presented in “**Conclusions**” section.

***h*-Index**

The Hirsch index, mostly known as h -index, was introduced by Hirsch in 2005 to facilitate quantification of scientific research output on two perspectives, impact and productivity. A simple definition of the Hirsch index is: for the highest h , if an author has h

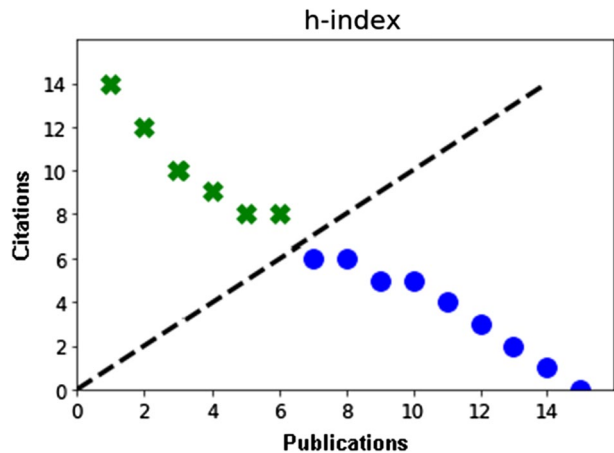
papers with more or equals cites than h , the index of the author is h . For instance, if an author has four papers with, 5, 4, 3, 1 cites each, she/he would have an h -index of 3 (three papers have three or more cites). Mathematically, the h -index is defined as follows: Let \mathcal{P}_h be the set of the h most cited publications of an author, and let N_i be the number of citations for the i -th article from \mathcal{P}_h . The h -index is the highest h value such that $\forall i \in \mathcal{P}_h, N_i \geq h$ is accomplished.

A simple way to calculate the h -index is to arrange the articles of an author by their amount of cites in decreasing order, and then look for the highest possible position where the number of cites of the paper is greater or equal than its position. Graphically, this can be observed in Fig. 1, where an author has an h -index of 6. Specifically, a cross or a dot represents the number of citations of the i -th most cited paper, and the articles are sorted based on their number of citations. The most cited paper has 14 citations, which is higher than its current position, i.e., first position. The second most cited paper also has 12 citations, which is also higher than its current position, i.e., second position. Eventually, we arrive at the sixth most-cited paper with eight citations, which is the highest position where the number of cites is greater or equal than the position, i.e., sixth position, obtaining an h -index = 6. For corroboration, the seventh most cited paper (or first dot) has only 5 citations, and since $N_7 = 5 < 7$, P_7 does not comply with $\forall i \in \mathcal{P}_7, N_i \geq 7$.

The h -index is the most popular index today, used on respected scientific production databases such as Web of Science, Google Scholar, and Scopus, mainly due to its simplicity (Alonso et al. 2009). The index is a quantification metric of the impact and the productivity of an author. This index facilitates the analysis of the impact of an author, and the comparison with other colleagues (Costas and Bordons 2007; Bornmann et al. 2008). Second, the index is consistent with the number of publications. Even though an author may have several publications, those articles with a low number of citations are not considered in the author’s h -index (Vanclay 2007; Glänzel 2006).

Besides the benefits, h -index also has some drawbacks, being the most important self-citations (Richard 2009; Schreiber 2007; Snyder and Bonzi 1998; Wilhite and Fong 2012; Livas and Delli 2018). Publications have estimated that self-cites could be on average a 36% of the total number of cites of an author (Aksnes 2003), and it could be strategically used to manipulate indexes (Zaggl 2017; Bartneck and Kokkermans 2011).

Fig. 1 h -index example for an author with 15 articles and h -index = 6. x -axis shows the i th article with most citations, while y -axis shows the number of citations for the i th article. Crosses represent the six most cited papers, where each of them has more citations than its current position. In contrast, dots represent other papers where their number of citations is lower than its actual position



Modifications

Apart from h -index's self-citations problem, other drawbacks have been found, leading to the creation of new indexes, which modify the main rule of the h -index (the top h articles have at least h citations) to solve its drawbacks. For example, the g -index (Egghe 2006) considers that the sum of citations for the top g papers should be at least g^2 citations, and the $h(2)$ -index (Kosmulski 2006) defines that the top h articles has at least h^2 citations. Even Hirsch implemented a new index called h_{bar} (Hirsch 2010), where, in an iterative process, papers belonging to \mathcal{P}_h are discarded according to the influence of co-authors. Other indexes weight the citation of a paper based on different characteristics. For example, h_w -index (Egghe and Rousseau 2008) is a version of the h -index weighted by citations impact. There are other indexes in which the weight is related to time. Most of these indexes take into account metrics about the productivity of authors in a period of time (Jin et al. 2007), the publication year (Burrell 2007), or author's contribution (Cormode et al. 2013). Another group considers, besides the number of citations, multiple other factors such as: the contribution of the author, the total number of papers and citations, age of the scientist, age of the publications, age of citations, co-authors of the publication, self-citation, and colleague citations (Abbasi et al. 2010; Hu et al. 2010; Mikhailov 2012; Chen and Wan 2016; Oberesch and Groppe 2017). Finally, other indexes give different importance to the citation based on the number of authors and co-authors of the citing papers: f -index (Katsaros et al. 2009), c -index (Bras-Amorós et al. 2011) and p -index (Horzyk 2014).

All these indexes solve different issues of the h -index, mainly avoiding self-citations or giving more dimensions to the index to be more accurate on the quantification. However, there is still a problem when two co-authors or more start to cite one another in a discriminatory way for the manipulation of the index. For example, a paper with 10 cites, where 6 of them are made from a previous co-author explicitly to increase the author h -index.

In the following section, we propose a generalization of the h -index, the robust h -index that overcomes this drawback.

Robust h -index (rh -index)

We propose a generalization of the h -index called robust h -index (rh_i -index). Similarly to the h -index, for the highest rh , an author with a rh -index of rh implies, that have rh papers with more or equals robust cites, where a robust cite is defined according to the i -robustness level. The i -robustness level represents the minimum distance among authors in a collaboration network to consider a citation, posterior to their collaboration, as valid; i.e., citations previous to their collaboration are always considered valid. Our approach is directly related to the collaboration network among authors, which can be visualized through a graph, where each node represents a single author, and edges indicate collaborations among authors. Based on this graph, the distance between the two authors is defined as the shortest path of the corresponding nodes. A visual example of this distance can be observed in Fig. 2, where the distance of an author with himself is considered to be 0, but the distance between authors A_1 and A_2 is 1, and authors A_1 and A_3 is 2. Similarly to the h -index, the rh_i -index of an author a is based on \mathcal{P}_{rh_i} , the set of rh_i most cited publications, and N_j , the number of citations for the j -th article from \mathcal{P}_{rh_i} . The construction of rh_i is based on valid citations, which are determined by: a citation where all authors have a distance of

at least i to author a , or, in case that the distance is less or equal to i , the time of the citation was previous to the collaboration with author a . Then, the rh_i -index rh_i is the highest value such that $\forall j \in \mathcal{P}_{rh_i}, N_j \geq rh_i$.

Considering these definitions, we defined the rh_i -index of an author a , for three i -robustness levels that will be used in this work:

- rh_0 the minimum distance for a valid citation is 0. All paper citations are valid, i.e., the original h -index.
- rh_1 the minimum distance for a valid citation is 1. All paper citations that do not include one of the authors a of the current document are considered valid, i.e., the h -index without self-citations.
- rh_2 the minimum distance for a valid citation is 2. All paper citations without author a and co-authors of a are considered valid (i.e., authors without a direct relation with a). Co-author citations previous to the collaboration with a are also valid.

To calculate \mathcal{P}_{rh_i} from scratch, we need to generate the collaboration network G_a based on publications citing author a . Let $G_a = (\mathbf{V}_a, \mathbf{E}_a)$ be the graph of collaboration authors. \mathbf{V} is a set of nodes including author a , authors of papers citing a 's, and collaborative authors of a , while \mathbf{E}_a is the set of edges, where $e_{jk} \in \mathbf{E}_a$ implies a direct collaboration between authors j and k . With G_a we can calculate the distance among all authors concerning a and generate $L_{a,i}$, a list of authors with at most i distance from the author a in G_a , and their time when they became collaborators. We state that, if a paper citing a work of a at time t , has at least one author belonging to $L_{a,i}$ before t , then the citation should not be considered when computing the author's index.

As an example, we will describe the process to calculate the rh_2 -index for a specific author a . First, we need to generate G_a , gathering all co-authors of a and the time of their first collaboration. In this case, $L_{a,2}$ are the co-authors from a and author a , and the time of their first collaboration (first publication together), but this set grows as r increases. Then, we analyze all cites for a to determine the valid cites. Specifically, if the author of a paper citing author a belongs to $L_{a,2}$, and the time of the cite is after their first collaboration, then the paper is not considered to calculate the rh_2 -index. Once all citing papers are analyzed, we proceed to calculate the rh_2 -index using only valid citations just as the h -index, allowing us to keep the simplicity and interpretability of the h -index.

Figure 3 shows the difference between rh_1 -index = 6 (h -index without self-citations) and rh_2 -index = 5 (our rh -index with a robustness level of 2). For both plots, x -axis

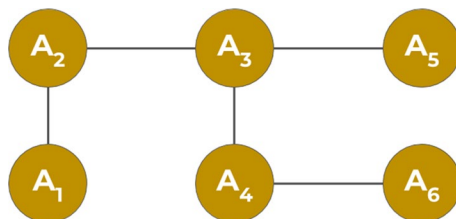


Fig. 2 Representation of a small collaboration network for a set of authors. Each node represents an author, and each edge represents at least one co-authorship paper between two authors. For example, A_3 has collaborated with A_2 , A_4 , and A_5 . The shortest path between the two authors gives the distance between them. For example, distance from A_1 to all nodes is given by $A_1 = 0, A_2 = 1, A_3 = 2, A_4 = 3, A_5 = 3,$ and $A_6 = 4$

represents the i th article with most citations, while y -axis represents the number of citations for the i th article. Crosses represent the rh_1 most cited articles from the researcher a that are considered to calculate rh_1 , circles are the least cited papers (not considered on the calculation of rh_1 -index), and squares are the i th article discarding the not valid cites based on a robustness level of 2. As can be observed in the left plot, several articles diminish their number of citations (the difference between crosses and dots versus boxes). Then, after reordering the most cited papers, in the right plot, we observe a decrease in the author's citation index. Besides this diminution, we observe an important change in the shape of the citation curve, indicating that most papers are not highly cited, as was believed in the left plot.

The advantages of our proposed metric are: a robust quality index of an author, generalization of the h -index, understandability of the robustness level, and ease incorporation on indexing databases. Robustness is achieved by considering only distant citations, showing the real influence of an author's work. Generalization comes from the fact that the proposed index is an extension of the well-known h -index based on a robustness level, meaning that the ease of computation, implementation, and understandability of the Hirsch index is also in our approach. Based on the rh -index, robust researchers should not have a big difference between their rh_0 -index, rh_1 -index, and rh_2 -index. A high value between any of these indexes shows possible irregular citations of the author. Lastly, our index is easy to implement on indexing databases such as Scopus, Google Scholar, or ISI Web of Science, where the collaboration network can be generated.

One of the main drawbacks of our approach is its time and space complexity when implemented from scratch. The h -index is faster than ours' proposal, as we need to generate the collaboration network. The implementation of a graph incurs in a higher space than the current version of the h -index. Both drawbacks, space, and time complexity are entirely expected as our approach takes into account much more information to generate better indexing of the author. However, these drawbacks are greatly diminished once a collaboration network is available. For example, indexing databases such as Scopus, Google Scholar, or ISI Web of Science already have all the necessary information to implement this index. As can be observed in those pages, they have a list of co-authors for any specific author. Then, it can be implemented directly, without much effort.

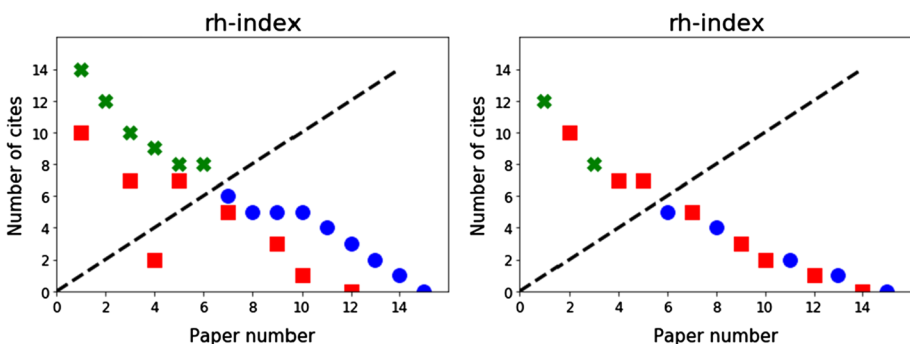


Fig. 3 rh_1 -index (h -index without self-citations) and rh_2 -index example of an author with 12 articles. Crosses represent most cited articles considered in rh_1 -index. Circles are the least cited articles not considered in rh_1 -index. Boxes represent the new number of valid citations for the i th paper for rh_2 -index. Left plot shows the current rh_1 -index (red boxes used only for visualization). Right plot shows the rh_2 -index, where the papers are reordered based on the red boxes. (Color figure online)

A second drawback could be related to collaboration disengagement. Given that new citations among co-authors will be considered invalid, it could lead to a reduction of collaboration among authors. However, positive effects such as the sharing of skills and knowledge between authors (Katz and Martin 1997), and the increase in the number and quality of publications (Presser 1980; Lee and Bozeman 2005) are significantly better retentions than unaccounted citations. Moreover, our studies show that the effect on recognized authors is minimum. On the opposite, possible fraudulent researchers are importantly affected.

As a summary, the proposed rh_t -index is a generalization of the h -index, adding new metrics to compare and quantify the impact and productivity of researchers. The main contribution of this new index is the non-inclusion of strategic collaborative cites that could be used to artificially increase the original h -index of an author, obtaining a robust quantification of a researcher's impact.

Analysis

In this section, we describe the data preparation and a specific case of our proposition the rh_2 -index. The data preparation subsection describes the data extraction and pre-processing phases, exposing procedures, faced issues, and solutions provided. After filtering and cleaning process, we obtained 659 authors. Furthermore, we compare different rh -indexes, against other indices, analyze the citation curves of two selected authors, and analyze a correlation plot among multiple indexes.

Data preparation

In 2016, Webometrics released a ranking with a list of 1009 researchers working at a Chilean institution. The list is ordered by the researchers' h -index, and includes other information such as their working institution, amount of citations, and a link to each author's Google Academic profile.

We leverage that list for the analysis performed in this work. To create a collaboration graph among authors, we enhance the previous data with information obtained using Microsoft's Knowledge API. Due to a difference in the way both APIs identify authors, we leverage the author's name to query Microsoft's API. For each author, we retrieved her/his publications. For each publication, we collected its metadata and all the articles that cited it. The metadata contains a unique ID, relevant keywords, subject, name, other ways to write the author's name (if available), among others.

We faced two issues when retrieving the authors' data. First, some researchers share the same name, causing that a publication is attributed to a researcher that did not write it. Second, there are authors with multiple IDs in the platform, which results in a split of her/his works as multiple authors. To solve both problems, we compare each author's Google Scholar profile against the one obtained from Microsoft's. For the previous comparison to work, a series of splitting and merging steps are necessary.

Authors may sign their articles in different ways: with an abbreviation of their names, with more than one last name, or different combinations of their first and last names. To match the names of authors, we measured their similarity with the Jaro–Winkler algorithm (Winkler 1999), set to a 0.8 threshold.

When similar authors are found, we compare their profiles by checking the list of articles on both platforms. If the names of articles are different on both platforms, we also leverage a Jaro–Winkler algorithm to measure the similarity between titles. For any pair of publication (a, b)—where a is from Google’s profile, and b is from Microsoft’s—if the similarity between the two titles is higher than 0.90, we assume that the two articles are the same. If any pair of articles is considered equal, we merge the authors’ IDs into one.

As a result of the cleaning process, we ended with 659 profiles, each one consisting of an author with their corresponding articles. For each article, we incorporate at least its authors, keywords, and the main topic.

Results

In this subsection, we compare our proposition, especially, rh_2 -index. First, we compare rh_0 -index (h -index), rh_1 -index (h -index without self-citations), rh_2 -index, and p -index. Second, we compare the ranking of the top authors with different indexes. Third, we select two authors and compare their curve of citations for each of the rh_1 -index. Finally, we compare the correlation among previous indices, adding to the comparison other two indexes: g -index and hc -index. We excluded these last two indexes from our first analyzes because of their high correlation with other indices.

We should note that of the analyzed indices, all the p -index, hc -index, and g -index, share the principles of rh -index. The p -index (popularity index) is an improvement of the f -index; instead of cites, each paper is measured based on the number of non-repeating authors that have cited it (Horzyk 2014). Then, an author with p -index = p implies that have at least p paper, where each paper has been cited by at least p different authors. The hc -index is a more robust version than rh_1 -index (h -index without self-citations), where all citations from any author of the paper are discarded (Schreiber 2007). As can be observed, this index is between rh_1 -index and rh_2 -index. Finally, the g -index measures the global citation performance of an author, where an author with g -index = g implies that the top g articles received (together) at least g^2 citations (Egghé 2006).

Left plot of Fig. 4 shows three levels of the rh -index values and p -index for each author in descending order (rh_0 -index). As can be seen, there are important differences among curves, showing clear differences according to the level of the rh -index. It is surprising, but expected, that some authors reduce their rh_2 -index by a value of 10 or greater (Right plot), meaning that several of their citations are related to previous co-author citations. As expected, p -index shows a completely different behavior than the other indices, as it related to authors (Left plot). However, we can observe that while some authors increase their p -index, their rh_2 -index decreases.

Figure 5 shows the differences in ranking for the top 10 authors according to their rh_0 -index. Several authors increase their ranking, for example, authors 2, 6, and 8. This implies that most of their citations are external, considering them as robust authors. In contrast, author 5, even though is fourth according to the rh_1 -index, drops to the ninth place with the rh_2 -index, which could imply some fraudulent citations. We want to highlight the main differences between the rh_2 and p indexes. First, author 6, considered a robust author, it has the first place according to the p -index, implying that most of their papers are cited from multiple authors. In contrast, author 8, even though is forth according to rh_2 -index, it drops to the tenth place according to the p -index. This could be explained by a reduced number of researchers in its area, where most authors cite each other, but they do not work among them. Finally, author 5, which is severally punished by rh_2 -index, has the second

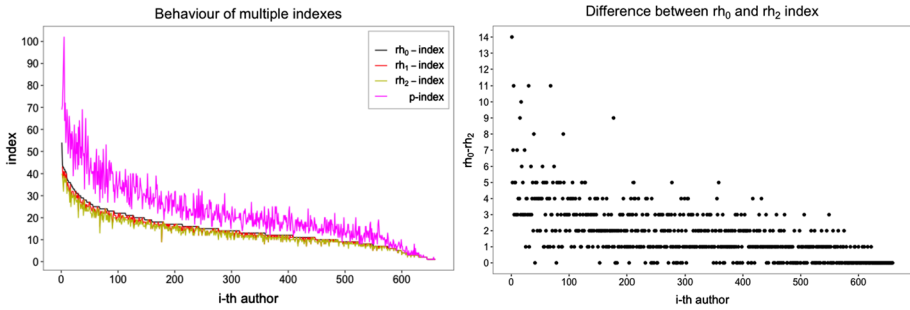


Fig. 4 Left: rh -index for all authors analyzed in the three different levels against the p -index. x -axis shows authors numbered and ordered by their rh_0 -index and y -axis shows the value of multiple indexes. The black line corresponds to rh_0 -index or h -index, the red line shows rh_1 -index, the citrus (dark yellow) line shows rh_2 -index, and magenta line shows p -index. Right: Highlight of the differences between rh_0 and rh_2 . (Color figure online)

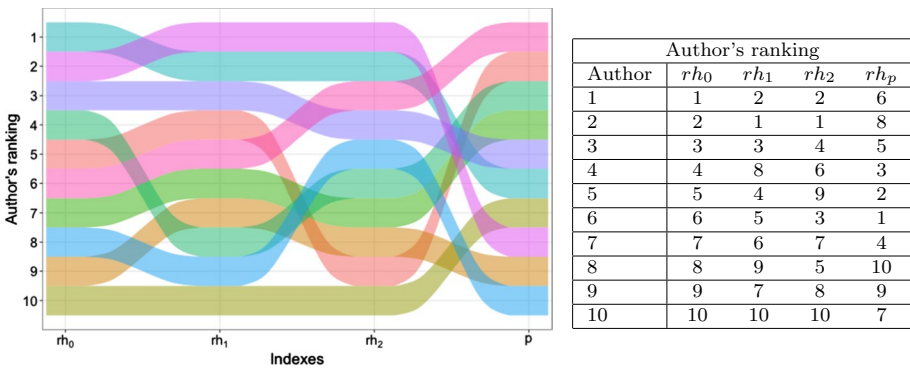


Fig. 5 Plot (left) and table (right) with the differences in their ranking among the top 10 authors according to the rh_0 -index. The first three columns show rh_0 , rh_1 , and rh_2 indexes, while the last column shows the p -index

place according to the p -index (recall, p -index considers authors). Then, even though several authors cited author 5, several of those citations correspond to close collaborators.

To continue with our analysis, in Fig. 6 we compare the citation curves for author 2 (A_2) and author 5 (A_5) as they show different behavior in their rankings. While A_2 has the highest ranking, A_5 decreases considerably. Each plot shows three different curves, one per index (solid: rh_0 -index, dash-dot: rh_1 -index, dash: rh_2 -index). Each plot shows the number of citations according to the rh -index for authors A_2 (left plot) and A_5 (right plot). As can be appreciated in the figure, we confirmed the robustness of author A_2 , as her/his curves show a minimum difference among them. In contrast, author A_5 shows significant differences between the curve for rh_2 against the curves defined by rh_1 and rh_0 . A posterior analysis determined that, from a total of 2105 collaborative citations for this author, a single co-author cited her/him over 300 times.

Our final analysis is a comparison among all indices of their Pearson correlation values, which measures the linear relation between two indices. The Pearson correlation varies between -1.0 and 1.0 , where values close -1.0 and 1.0 implies a strong linear relation between variables (negative or positive respectively), while values close to 0 shows

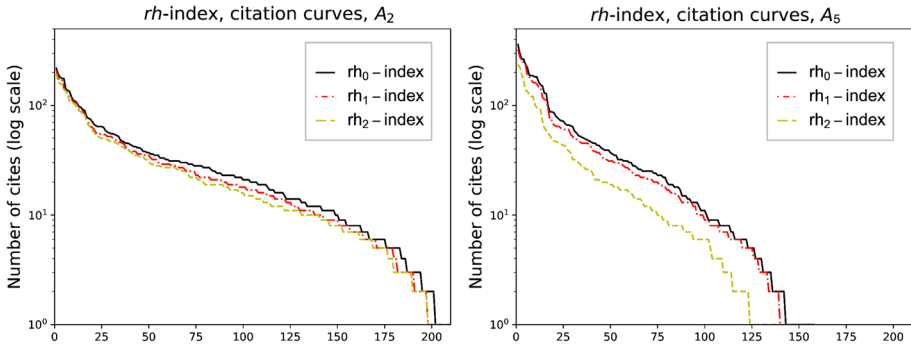


Fig. 6 rh_0 , rh_1 , and rh_2 citation curves for authors A_2 (left plot) and A_5 (right plot). Citations curves are represented by the solid line (rh_0), the dash-dot line (rh_1), and the dash line (rh_2). x -axis shows the j th most cited paper under the rh_r -index, while y -axis shows the total number of citations for each paper in the logarithmic scale

the lack of linear relation, but other types of relations could be observed. This analysis increased its importance as multiple indexes should be used to assess the importance of an author. For example, a recent article shows a strong correlation among several journal indices (Villasenor-Almaraz et al. 2019), being especially important the relation of the impact factor with the total number of citations per journal (Roldan-Valadez and Rios 2015; Diaz-Ruiz et al. 2018; Roldan-Valadez et al. 2018).

Figure 7 shows the Pearson correlation among all indices. The order of the variables was changed as the hc -index is related to rh_1 and rh_2 indices. All indices are highly correlated as all of them use similar ideas to the index calculation. As expected rh_0 , rh_1 , hc , and rh_2 show a high linear correlation among them. Recalls that these indices are similar as they diminish the number of valid citations, obtaining most of the information from the difference between these values. g -index also shows a high correlation, as it is a robust version of the h -index. Similarly, the high correlation of the p -index can be explained by the

Fig. 7 Pearson (linear) correlation among all indices. Pale red and blue pale colors show the highest and lowest values observed in the matrix. All indices are highly correlated as all of them use similar ideas to the index calculation. (Color figure online)

				p	
			g	0.858	
		rh₂	0.932	0.914	
	hc	0.995	0.934	0.916	
	rh₁	0.991	0.992	0.938	0.925
rh₀	0.989	0.982	0.981	0.925	0.943

relation of cites and the number of authors. As it is expected, the number citation should be related to the number of different authors, except in some specific cases, where areas are narrow.

Conclusions

In this work, we present a generalization of the h -index called robust h -index (rh_i -index). The i indicates the level of robustness required for a citation to be considered valid, based on the distance in the collaboration graph of an author, and the time of the citation. For instance, if we consider i to be one, then self-citations are not included. In particular, we analyze the effect of applying the index with a value i equals two (rh_2 -index), meaning that citations of co-authors, posterior to their collaboration, are not valid. In this analysis, we compute and compare the rh_0 -index (h -index), rh_1 -index (h -index without self-citations), the rh_2 -index, p -index, g -index, and hc-index for 695 authors belonging to Chilean institutions.

If we focus on the difference between ranking for the top 10 authors according to the h -index, some of them did not show a significant difference, indicating that most of their works are cited by researchers outside his nearby influence area. Nevertheless, a group of authors shows a considerable difference among their ranking, indicating that most of their works are cited among previous co-authors. Further analysis performed on one of these authors—with the highest difference between her/his rh_1 -index and rh_2 -index—reveal a hidden behavior: while rh_0 and rh_1 citation curves show that most of her/his papers are highly cited, the rh_2 citation curve shows that most papers are cited vaguely. Even more, we were able to identify up to 2105 collaborative citations for this author, where over 300 of them corresponds to a single co-author.

In summary, we state that based on our proposed index, it is possible to determine the robustness of an author, by determining if the researcher could be increasing her/his h -index in a fraudulent way. In the future, we plan to analyze the behavior of the rh_2 index for different scientific areas, given that some of them have a particular behavior. By doing so, we expect to create a model for a fair comparison of productivity among scientists in different areas. Besides, we will also analyze the inclusions of text analysis to determine a cite as valid, which could largely impact authors' indexes.

Acknowledgements Sebastian Moreno acknowledges the support of “Universidad Adolfo Ibáñez”.

Author Contributions Authors' contributions idea for the article: SM. Literature search: MP. Data analysis: MP, SM, GH-C. First draft: MP. Critical revision of the work: MP, SM, GH-C.

Data availability The original list of authors was obtained from Webometrics 2016. The final data was generated using data gathered from Microsoft's Knowledge .

References

- Abbasi, A., Altmann, J., & Hwang, J. (2010). Evaluating scholars based on their academic collaboration activities: Two indices, the re-index and the cc-index, for quantifying collaboration activities of researchers and scientific communities. *Scientometrics*, 83(1), 1–13.
- Aksnes, D. (2003). A macro study of self-citation. *Scientometrics*, 56(2), 235–246.

- Alonso, S., Cabrerizo, F., Herrera-Viedma, E., & Herrera, F. (2009). h-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, 3(4), 273–289.
- Bartneck, C., & Kokkelmans, S. (2011). Detecting h-index manipulation through self-citation analysis. *Scientometrics*, 87(1), 85–98.
- Bornmann, L., Mutz, R., & Daniel, H. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, 59(5), 830–837.
- Bras-Amorós, M., Domingo-Ferrer, J., & Torra, V. (2011). A bibliometric index based on the collaboration distance between cited and citing authors. *Journal of Informetrics*, 5(2), 248–264.
- Burrell, Q. (2007). Hirsch's h-index: A stochastic model. *Journal of Informetrics*, 1(1), 16–25.
- Chen, M., & Wan, Z. (2016). New nonlinear metrics model for information of individual research output and its applications. *Mathematical and Computational Applications*, 21(3), 26.
- Cormode, G., Ma, Q., Muthukrishnan, S., & Thompson, B. (2013). Socializing the h-index. *Journal of Informetrics*, 7(3), 718–721.
- Costas, R., & Bordons, M. (2007). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, 1(3), 193–203.
- Diaz-Ruiz, A., Orbe-Arteaga, U., Rios, C., & Roldan-Valadez, E. (2018). Alternative bibliometrics from the web of knowledge surpasses the impact factor in a 2-year ahead annual citation calculation: Linear mixed-design models' analysis of neuroscience journals. *Neurology India*, 66(1), 96–104.
- Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69(1), 131–152.
- Egghe, L., & Rousseau, R. (2008). An h-index weighted by citation impact. *Information Processing & Management*, 44(2), 770–780.
- Glänzel, W. (2006). On the opportunities and limitations of the h-index. *Science Focus*, 1(1), 10–11.
- Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569–16572.
- Hirsch, J. E. (2010). An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 85(3), 741–754.
- Horzyk, A. (2014). P-index-a fair alternative to h-index. Poland: Department of Automatics and Biomedical Engineering.
- Hu, X., Rousseau, R., & Chen, J. (2010). In those fields where multiple authorship is the rule, the h-index should be supplemented by role-based h-indices. *Journal of Information Science*, 36(1), 73–85.
- Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The r-and ar-indices: Complementing the h-index. *Chinese Science Bulletin*, 52(6), 855–863.
- Katsaros, D., Akritidis, L., & Bozanis, P. (2009). The f index: Quantifying the impact of coterminal citations on scientists' ranking. *Journal of the American Society for Information Science and Technology*, 60(5), 1051–1056.
- Katz, J., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26(1), 1–18.
- Kosmulski, M. (2006). A new Hirsch-type index saves time and works equally well as the original h-index. *ISSI Newsletter*, 2(3), 4–6.
- Krampen, G., Becker, R., Wahner, U., & Montada, L. (2007). On the validity of citation counting in science evaluation: Content analyses of references and citations in psychological publications. *Scientometrics*, 71(2), 191–202.
- Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, 35(5), 673–702.
- Livas, C., & Delli, K. (2018). Journal self-citation rates and impact factors in dentistry, oral surgery, and medicine: A 3-year bibliometric analysis. *Journal of Evidence Based Dental Practice*, 18(4), 269–274.
- Meyers, M., & Quan, H. (2017). The use of the h-index to evaluate and rank academic departments. *Journal of Materials Research and Technology*, 6(4), 304–311.
- Mikhailov, O. (2012). A new citation index for researchers. *Herald of the Russian Academy of Sciences*, 82(5), 403–405.
- Oberesch, E., & Gropppe, S. (2017). The mf-index: A citation-based multiple factor index to evaluate and compare the output of scientists. *Open Journal of Web Technologies*, 4(1), 1–32.
- Presser, S. (1980). Collaboration and the quality of research. *Social Studies of Science*, 10(1), 95–101.
- Richard, B. (2009). A simple method for excluding self-citation from the h-index: The b-index. *Online Information Review*, 33(6), 1129–1136.
- Roldan-Valadez, E., Orbe-Arteaga, U., & Rios, C. (2018). Eigenfactor score and alternative bibliometrics surpass the impact factor in a 2-years ahead annual-citation calculation: A linear mixed design model analysis of radiology, nuclear medicine and medical imaging journals. *La Radiologia Medica*, 123(7), 524–534.

- Roldan-Valadez, E., & Rios, C. (2015). Alternative bibliometrics from impact factor improved the esteem of a journal in a 2-year-ahead annual-citation calculation. *European Journal of Gastroenterology & Hepatology*, 27(2), 115–122.
- Schreiber, M. (2007). Self-citation corrections for the Hirsch index. *Europhysics Letters*, 78(3), 30002.
- Seeber, M., Cattaneo, M., Meoli, M., & Malighetti, P. (2019). Self-citations as strategic response to the use of metrics for career decisions. *Research Policy*, 48(2), 478–491.
- Snyder, H., & Bonzi, S. (1998). Patterns of self-citation across disciplines (1980–1989). *Journal of Information Science*, 24(6), 431–435.
- Vanclay, J. (2007). On the robustness of the h-index. *Journal of the American Society for Information Science and Technology*, 58(10), 1547–1550.
- Villasenor-Almaraz, M., Islas-Serrano, J., Murata, C., & Roldan-Valadez, E. (2019). Impact factor correlations with Scimago Journal Rank, Source Normalized Impact per Paper, Eigenfactor Score, and the Citescore in Radiology, Nuclear Medicine & Medical Imaging journals. *La Radiologia Medica*, 124(6), 495–504.
- Wilhite, A., & Fong, E. (2012). Coercive citation in academic publishing. *Science*, 335(6068), 542–543.
- Winkler, W. (1999). The state of record linkage and current research problems. Technical report statistical research report series RR99/04. Washington, DC: U.S. Bureau of the Census.
- Zaggl, M. (2017). Manipulation of explicit reputation in innovation and knowledge exchange communities: The example of referencing in science. *Research Policy*, 46(5), 970–983.