E C O N O M I C S B U L L E T I N

On the Robustness of the h-index: a mathematical approach

Jean-Michel Courtault
CEPN (Centre d'Economie de l'Université Paris Nord), UMR 7115

Naïla Hayek
Université de Franche-Comté, and Laboratoire Marin Mersenne, Université de Paris I

Abstract

The h-index is an index recently proposed by Hirsch (2005) to measure scientific achievement by individual scholars. It is a compound measure of publications and citations. We show the robustness of this index. This means that h-index increases with both the number of publications and the number of citations only when these numbers are significant.

The authors thank the anonymous referee, together with Bertrand Crettez and Raphael Giraud for their useful comments.
Citation: Courtault, Jean-Michel and Naïla Hayek, (2008) "On the Robustness of the h-index: a mathematical approach." *Economics Bulletin*, Vol. 3, No. 78 pp. 1-9
Submitted: November 12, 2008. **Accepted:** December 20, 2008.
URL: http://economicsbulletin.vanderbilt.edu/2008/volume3/EB-08C00011A.pdf

1 Introduction

Recently a new measure of scientific achievement has been proposed by Hirsch (2005) that received a wide audience in all domains of the scientific community. Following Hirsch a scientist has an index h if h of his p papers have received at least h citations each and his other p-h papers have received at most h citations each. The h-index is a compound measure of productivity and quality as measured by the number of citations received by the published papers. The higher the index, the greater the number of significant papers published by an author and the higher the significance of the papers.

Many empirical papers have followed. Some were concerned with measurement of the impact of specific journals on their fields while many others were concerned with ranking of individual scientists in a particular field (economics as with Ursprung and Zimmer (2006)).

One of the main attractions of the index is its relative robustness. That is the h-index does not vary greatly if the number of documents included (e.g. if we exclude books or book chapters and consider articles only or if we exclude older papers) changes significantly. Neither the h-index increases significantly if the total number of citations increases. In particular, the h-index does not depend on the less interesting (i.e. quoted) papers an author has published and once a paper has reached h citations the extra number of citations does not increase the h-index. In particular, this means that the h-index does not give undue weight to review papers. If you have an h-index with value h then if you want to increase your index to h+1 you will often need to write more than one paper with h+1 citations (since you may not have had already h papers with h+1 citations). If one considers also that most papers ceased to be quoted anymore after a relatively short spell of time and that for any author the distribution of citations of his papers is very unequal, then a significant effort (both with respect to the quantity of papers published and their quality as measured by the number of citations received) has to be produced to increase one's h index.

This feature is particularly interesting for the Social Sciences. Indeed the value of a social scientist cannot be fully apprehended with a single result be it empirical (as in the medical science with the discovery of a remedy for a fatal desease) or theoretical (as in the mathematical science with the proof of a famous mathematical conjecture). Contrary to Nobel prize philosophy, very few Nobel prizes in economics were awarded for a single contribution (as the Black-Merton-Scholes formula for option pricing). Usually, they are awarded for several outstanding contributions, sometimes for the whole work as it is the case with scholars who have initiated a new subdiscipline. The h-index is also a useful characterization when we want to compare the contributions of many scholars since evaluation takes time. Researchers whose contributions are being evaluated earlier are not strongly disadvantaged relatively to researchers who are evaluated at the end of the investigation since the h-index does not depend significantly on the documents appearing after they were evaluated.

The object of this paper is to investigate the dependence of the h-index on the number

of papers included and the number of citations. We will first restate the definition of the h-index as a solution to a maximization problem (Glänzel's definition). We will then show explicitly the equivalence between Hirsch's definition (2005) of scientific achievement and Glänzel's definition (2006). We will then generalize Glänzel's definition in order to study the dependence of the h-index on the number of papers and citations. We will show several interesting properties as the increase of the h-index with set inclusion. Finally we will show the main result of the paper which is that the h-index has an upper limit .

2 Robustness of the h-index

Let p be the number of papers. Let X_j be the number of citations of the j-th most cited paper. $(X_j)_{1\leq j\leq p}$ is a decreasing sequence.

Definition
According to Hirsch a scientist has an index h if h of his p papers have received at least h citations each and his other p-h papers have received at most h citations each.

So he has an index h if he has h articles with more than h citations each while the remaining articles have less than or exactly h that is strictly less than h +1 citations. Thus the h-index satisfies $X_h \geq h$ and $X_{h+1} < h+1$.

Glänzel (2006) introduced the following alternative definition of the h-index as a solution to a maximization problem.

Definition
The h-index is defined as the solution of the following maximization problem

$$(P)\begin{cases} h = \text{Max } j \\ j \in \{1,...,p\} \\ X_j \geq j \end{cases}$$

If (P) has no solution we set $h = 0$.

Let us show that Glänzel's definition is equivalent to Hirsch's. First note that the h-index is uniquely defined by (P). If (P) admits no solution then h=0. If (P) admits a solution then this solution is necessarily unique.

Proposition 1
(i) $h = 0$ if and only if $X_1 = 0$
(ii) $h = p$ if and only if $X_p \geq p$.

(iii) If $0 < h < p$, then h is a solution of (P) if and only if $X_h \geq h$ and $X_{h+1} < h+1$.

Point (iii) shows the equivalence between Hirsch's definition and Glänzel's definition. It means that if h is a solution of (P) then the author has h papers with at least h citations each and the remaining p-h papers have at most (strictly less than h+1) h citations each and conversely.

Proof:
(i) If $X_1 = 0$ then for every $j > 1$, $X_j \leq X_1 = 0$ so for every j, $X_j = 0$ hence (P) has no solution so $h = 0$.
If $h = 0$ suppose that $X_1 > 0$ so $X_1 \geq 1$ so $h \geq 1$ which is a contradiction.
(ii) If $X_p \geq p$ then $h = p$ (since $p = Max\{1, ..., p\}$ and $X_p \geq p$).
If $h = p$ then by definition of h we have $X_p \geq p$.
(iii) If h is a solution of (P) then we have $X_h \geq h$. Suppose
$X_{h+1} \geq h+1$ then $h+1$ satisfies $X_j \geq j$ and $h+1 > h$ which contradicts that h is a solution.
So $X_{h+1} < h+1$.
If $X_h \geq h$ and $X_{h+1} < h+1$, suppose there exists $h' > h$ such that $X_{h'} \geq h'$.
$X_{h'} \leq X_{h+1} < h+1 \leq h'$. So $X_{h'} < h'$ which is a contradiction.∎

Remark
It is possible to give another characterization of the h-index. We construct a continuous decreasing function linking the decreasing sequence of citations. It has a fixed point and the h-index is the greatest integer less than or equal to that point.
So recall that $(X_j)_{1 \leq j \leq p}$ is a decreasing sequence. Let $X_1 \geq 1$ and $X_p < p$. Let f be a continuous decreasing function from $[1, p]$ to $\mathbb{R}$ such that $f(t) = X_t$ if $t = 1, .., p$.
Since $f(1) \geq 1$ and $f(p) < p$, we obtain that f has a fixed point $\hat{t}$.
$[\hat{t}]$ denotes the greatest integer less than or equal to $\hat{t}$. (Notice that $[\hat{t}] \leq \hat{t} < [\hat{t}] + 1$.)
We have $X_{[\hat{t}]} = f([\hat{t}]) \geq f(\hat{t}) = \hat{t} \geq [\hat{t}]$ and $X_{[\hat{t}]+1} = f([\hat{t}] + 1) \leq f(\hat{t}) = \hat{t} < [\hat{t}] + 1$.
Thus $h = [\hat{t}]$.

In order to study the dependence of the h-index on the number of papers and citations we generalize Glänzel's definition as follows.

Definition
Let $\mathcal{A}$ be the set of all papers. We define the function

$$h : \mathcal{P}(\mathcal{A}) \to \mathbb{N}$$

$$A \mapsto h(A)$$

$$\text{where } \begin{cases} h(A) = \text{Max } j \\ j \in \{1, ..., cardA\} \\ X_j \geq j \end{cases}$$

It is obvious that the only way so that a global increase of the total number of citations translates necessarily into an increase of the h-index, is that the number of citations of each paper does not decrease. This means that if an author has produced more papers than another author each of them being more quoted, then the first author cannot have a lower index than the second. However, he might not have a strictly greater h-index. For example, an author having 4 papers with $X_i = 4$ for i = 1, 2, 3 and $X_4 = 3$ will have the same h-index of 3 as an author having 3 papers each being quoted 3 times. It is also possible that an author with a greater h-index might have a smaller total number of citations. An author having published 100 papers each quoted 1 time will have an h-index of 1 with 100 citations whereas the author having published 3 papers each quoted 3 times will have an h-index of 3 with 9 citations.

Proposition 2
(i) If $X_i^A \leq X_i^B$ for every $i = 1, ..., cardA \leq cardB$, then $h(A) \leq h(B)$.
(ii) If $X_i^A \leq X_i^B$ for every $i = 1, ..., cardB < cardA$, and $h(A) < cardB$ then $h(A) \leq h(B)$.
(iii) If $cardB \leq h(A) \leq cardA$, then $h(A) \geq h(B)$.

Proof:
(i)-If $h(A) = 0$, then $h(A) \leq h(B)$. If $h(B) = 0$ then for every $i = 1, .., cardB$, $X_i^B = 0$. So for every $i = 1, .., cardA \leq cardB$, $X_i^A = 0$
hence $h(A) = 0$. Thus $h(A) \leq h(B)$.
-If $h(B) = cardB$ then $h(A) \leq cardA \leq cardB$.
-If $0 < h(B) < cardB$
If $h(A) = cardA \leq cardB$ then $X_{cardA}^A \geq cardA$ and $X_{cardA}^B \geq X_{cardA}^A \geq cardA$.
in this case if $X_{cardA+1}^B < cardA + 1$ then $h(B) = cardA$ so $h(A) \leq h(B)$
and if $X_{cardA+1}^B \geq cardA + 1$ then $h(B) \geq cardA + 1 > cardA = h(A)$, so $h(A) \leq h(B)$.

If $h(A) < cardA$ then suppose that $h(A) > h(B)$.
We have $X_{h(A)}^A \geq h(A)$ and $X_{h(A)+1}^A < h(A) + 1$. Since $h(A) \geq h(B) + 1$ we have $X_{h(A)}^A \leq X_{h(B)+1}^A \leq X_{h(B)+1}^B < h(B) + 1 \leq h(A)$ which is a contradiction, so $h(A) \leq h(B)$.

(ii) -If $h(A) = 0$, then $h(A) \leq h(B)$. If $h(B) = 0$ then for every $i = 1, .., cardB$, $X_i^B = 0$. So for every $i = 1, .., cardB$, $X_i^A = 0$, so for every $i = 1, .., cardA$, $X_i^A = 0$, hence $h(A) = 0$. Thus $h(A) \leq h(B)$.
-If $h(B) = cardB$ then $h(A) < cardB = h(B)$.
-If $0 < h(B) < cardB$ then suppose that $h(A) > h(B)$.
We have $X_{h(A)}^A \geq h(A)$ and $X_{h(A)+1}^A < h(A) + 1$. Since $h(A) \geq h(B) + 1$ we have $X_{h(A)}^A \leq X_{h(B)+1}^A \leq X_{h(B)+1}^B < h(B) + 1 \leq h(A)$ which is a contradiction, so $h(A) \leq h(B)$.

(iii) It is immediate to see that If $cardB \leq h(A) \leq cardA$, then $h(B) \leq cardB \leq h(A)$.■

Remark If we give a time interpretation to the set of papers of an author, then as time passes, the number of citations cannot decrease and so his h-index cannot decrease.

The h-index is not an increasing function of the number of papers. However, the h-index is increasing with respect to set inclusion. If a set of papers is included in another set of papers the h-index of the latter cannot be smaller than the h-index of the former. This is a good feature of a measure of scientific achievement, since an author cannot decrease his h-index as he increases his production.

Proposition 3
If $A \subset B$ then $h(A) \leq h(B)$

Proof: We first show that if $A \subset B$ then $X_i^A \leq X_i^B$ for every $i = 1, .., cardA$.
Indeed the elements of A are a finite subsequence of the sequence of elements of B. Thus there is a strictly increasing fonction q from $\{1, ..., cardA\}$ into $\{1, ..., cardB\}$ such that

$$X_i^A = X_{q(i)}^B$$

It is easy to show that if q is strictly increasing then $q(i) \geq i$ for every $i \in \{1, ..., cardA\}$.
Finally $X_i^A = X_{q(i)}^B \leq X_i^B$ since the sequence $(X_i^B)_i$ is decreasing.
The result follows using Proposition 2 (i).■

As an immediate consequence of proposition 3 we have:

Corollary
$Max(h(A), h(B)) \leq h(AUB)$.

Proof:
$A \subset AUB$ so $h(A) \leq h(AUB)$ using proposition 3
$B \subset AUB$ so $h(B) \leq h(AUB)$ using proposition 3
thus $Max(h(A), h(B)) \leq h(AUB)$.

Remark If we interpret A and B as the set of papers of different individuals then we have that the h-index of the group cannot be smaller than the h-index of each individual of this group.

Now comes the main result of the paper which shows where the robustness of the h-index lies. Robustness may have two interpretations. In one sense it means that a single paper cannot increase an index of scientific achievement by a big amount. Indeed a single paper cannot increase the h-index by more than 1 even if it has a lot of citations (even if the number of citations is greater than the initial h +1). On the contrary if we take the total number of citations as the index of scientific achievement it is often the case that a single paper can increase the index by a considerable amount. In a second sense robustness means that to increase an index it is necessary to add a significant number of papers significantly cited. A sufficient condition to increase the h-index is that the number of the new papers is at least as great as the initial h+1 and the number of citations of at least h+1 of them is cited h+1 times. If we take the number of publications as an index of scientific achievement

then it is possible to increase the index by the addition of a single paper.
The h-index has an upper limit. As an author increases the number of his scientific production the increase of his h-index is limited by the h-index of the new papers. That property is not shared by other indexes of scientific impact as the g-index of Egghe (2006). If we consider a set of 3 papers A with $X_1^A = 4$, $X_2^A = 3$ and $X_3^A = 1$ and the set of one paper B with $X_1^B = 8$ then $g(AUB) = 4 > g(A) + g(B) = 2 + 1$.
Notice that in the following proposition we do not assume that $A \cap B$ is empty.

Proposition 4
$h(AUB) \leq h(A) + h(B)$

Proof:
The elements ofAUB will be denoted Y_k.
The trivial cases:
If $A \subset B$ then $AUB = B$ and $h(AUB) = h(B) \leq h(A) + h(B)$
similarly if $B \subset A$ then $AUB = A$ and $h(AUB) = h(A) \leq h(A) + h(B)$).

So suppose A is not a subset of B and B not a subset of A:

• If $h(A) = 0$ or $h(B) = 0$.
Suppose $h(A) = 0$ then $\forall i = 1, ..., cardA$, $X_i^A = 0$, so $Y_k = X_k^B$ for $k = 1, ..., cardB$ and $Y_k = 0$ for $cardB < k \leq cardB + cardA$. Thus $h(AUB) = h(B) \leq h(A) + h(B)$
The case $h(B) = 0$ is similar.

• If $h(A) = cardA = a$ or $h(B) = cardB = b$
Suppose $h(A) = cardA = a$

-if $X_{h(B)}^B \leq X_a^A$
Let n be the numbers of common articles to A and B.
$Y_{a+h(B)-n} = X_{h(B)}^B$
In this case if $Y_{a+h(B)-n} < a + h(B) - n$ then $h(AUB) < a + h(B) - n \leq h(A) + h(B)$.
And if $Y_{a+h(B)-n} \geq a + h(B) - n$ then for $0 < h(B) < cardB$, since $Y_{a+h(B)-n+1} = X_{h(B)+1}^B < h(B) + 1 \leq h(B) + 1 + a - n$ (knowing $a - n \geq 0$), we obtain $h(AUB) = a + h(B) - n \leq h(A) + h(B)$.
And for $h(B) = cardB = b$, $Y_{a+b-n} \geq a + b - n = card(AUB)$, so $h(AUB) = a + b - n \leq h(A) + h(B)$.

-if $X_{h(B)}^B > X_a^A > X_{h(B)+1}^B$ where $0 < h(B) < cardB$
Let n be the number of common articles to A and B till the $h(B)$-th of B.
In this case if $Y_{a+h(B)-n} < a + h(B) - n$ then $h(AUB) < a + h(B) - n \leq h(A) + h(B)$.
And if $Y_{a+h(B)-n} \geq a + h(B) - n$ then, since $Y_{a+h(B)-n+1} = X_{h(B)+1}^B < h(B) + 1 \leq h(B) + 1 + a - n$ (since $a - n \geq 0$,) we obtain $h(AUB) = a + h(B) - n \leq h(A) + h(B)$.
-if $X_{h(B)}^B > X_a^A$, $h(B) = cardB = b$, $Y_{a+h(B)-n} = Y_{a+b-n} = X_a^A$. if $Y_{a+b-n} < a + b - n$ then $h(AUB) < a + b - n \leq a + b = h(A) + h(B)$,

and if $Y_{a+b-n} \geq a+b-n = card(AUB)$, so $h((AUB) = a+b-n \leq h(A)+h(B)$.

-if $X^B_{h(B)+1} \geq X^A_a$, where $h(B)+1 < cardB$
Let m be the smallest integer less than or equal to a such that $X^A_m \leq X^B_{h(B)+1}$,

-if $X^A_m \in]X^B_{h(B)+2}, X^B_{h(B)+1}]$, then let n be the number of common articles to A and B till the m-th.
$Y_{h(B)+1+m-n} = X^A_m < X^B_{h(B)+1} < h(B)+1$. In this case we cannot have $Y_{h(B)+1+m-n} \geq h(B)+1+m-n$, so $h(AUB) < h(B)+1+m-n \leq h(A)+h(B)$

-if $X^A_m \notin]X^B_{h(B)+2}, X^B_{h(B)+1}]$ then let n be the number of common articles to A and B till the $(h(B)+2)$-th.
$Y_{h(B)+2+m-1-n} = Y_{h(B)+1+m-n} = X^B_{h(B)+2} < h(B)+1$. In this case we cannot have $Y_{h(B)+1+m-n} \geq h(B)+1+m-n$, so $h(AUB) < h(B)+1+m-n \leq h(A)+h(B)$.

The case $h(B) = cardB = b$ is similar.

• If $0 < h(A) < cardA$ and $0 < h(B) < cardB$

-if $X^B_{h(B)} \in]X^A_{h(A)+1}, X^A_{h(A)}]$
Let n be the number of common articles to A and B till the $h(B)$-th of B. We have $Y_{h(A)+h(B)-n} = X^B_{h(B)}$. In this case:
if $Y_{h(A)+h(B)-n} < h(A)+h(B)-n$ then $h(AUB) < h(A)+h(B)-n \leq h(A)+h(B)$,
and if $Y_{h(A)+h(B)-n} \geq h(A)+h(B)-n$ then $Y_{h(A)+h(B)-n+1} = X^B_{h(B)+1} < h(B)+1 \leq h(B)+1+h(A)-n$, for $X^B_{h(B)+1} \in]X^A_{h(A)+1}, X^A_{h(A)}]$
and $Y_{h(A)+h(B)-n+1} = X^A_{h(A)+1} < h(A)+1 \leq h(A)+1+h(B)-n$ for $X^B_{h(B)+1} \notin]X^A_{h(A)+1}, X^A_{h(A)}]$,
so we obtain $h(AUB) = h(A)+h(B)-n \leq h(A)+h(B)$

-if $X^B_{h(B)} \leq X^A_{h(A)+1}$
Let m be the smallest integer less than or equal to $h(B)$ such that $X^B_m \leq X^A_{h(A)+1}$,

-if $X^B_m \in]X^A_{h(A)+2}, X^A_{h(A)+1}]$ then let n be the number of common articles to A and B till the m-th.
$Y_{h(A)+1+m-n} = X^B_m < X^A_{h(A)+1} < h(A)+1$. In this case we cannot have $Y_{h(A)+1+m-n} \geq h(A)+1+m-n$, so $h(AUB) < h(A)+1+m-n \leq h(A)+h(B)$.

-if $X^B_m \notin]X^A_{h(A)+2}, X^A_{h(A)+1}]$ then let n be the number of common articles to A and B till the $(h(A)+2)$-th.
$Y_{h(A)+2+m-1-n} = Y_{h(A)+1+m-n} = X^A_{h(A)+2} < h(A)+1$. In this case we cannot have $Y_{h(A)+1+m-n} \geq h(A)+1+m-n$, so $h(AUB) < h(A)+1+m-n \leq h(A)+h(B)$.

Finally the case $X^B_{h(B)} > X^A_{h(A)}$ is similar to the case $X^A_{h(A)} > X^B_{h(B)}$.■

Note, however that we do not have in general the stronger result $h(AUB) \leq h(A) + h(B) - h(A \cap B)$. Take for example the case of two authors the first one having published 3 papers with $X_1^A = 2$ and $X_2^A = 1 = X_3^A$ and the second one having published 2 papers with $X_1^B = 2$ and $X_2^B = 1$. One of the papers is written jointly by the two authors and has one citation. Then $h(AUB) = 2, h(A) = 1, h(B) = 1$ and $h(A \cap B) = 1$.

If $B \subseteq A$ then $h(AUB) = h(A)+h(B)-h(A \cap B)$ since $h(A \cap B) = h(B)$ and $h(AUB) = h(A)$. If $A \cap B = \varnothing$ (as in the case of new papers published by an author) then $h(AUB) \leq h(A) + h(B) - h(A \cap B)$ as $h(A \cap B) = 0$. However we may have strict inequality. Take for example the case of two sets of 3 different papers having exactly 3 citations each. Then the h-index of the union of the 6 papers is also 3 as the h-index of the two individual sets. If we have two sets of 3 different papers having exactly 6 citations each then the h-index of the union of the 6 papers is 6 whereas the h-index of the two individual sets is 3. For other popular indexes of scientific achievement as the total number of papers or the total number of citations, we always have equality.

3 Conclusion

The h-index is an interesting indicator of the scientific contribution of an individual or a group of individuals. It depends both on the quantity and the quality of the papers published as measured by the number of citations. We showed that this index is robust. A significant number of papers significantly cited must be published to increase the h-index. As an extension of the present work, a probabilistic approach could take into account other aspects of robustness. One of them is that after a certain number of years, old papers are less likely to be cited. Another one is that the distribution of citations is very unequal; it follows a Pareto type distribution in general (a very small number of papers receives most of the citations). So older papers cannot help an author to increase his h-index and among the new papers he will be writing only a very few of them will contribute to his h-index.

4 References

1. EGGHE Leo (2006), "Theory and practise of the g-index", Scientometrics, 69 (1), 131-152.
2. GLÄNZEL Wolfgang (2006), "On the H-index: A mathematical approach to a new measure of publication activity and citation impact", Scientometrics 67 (2), p. 315-321.
3. HIRSCH Jorge E. (2005), "An index to quantify an individual's scientific output", Proceedings of the National Academy of Science, 102 (46), p. 16569-16572.
4. URSPRUNG Heinrich W. and Markus ZIMMER (2006), "WHO is the "Platz-Hirsch" of the German Economics profession? A citation analysis", Working Paper, 39 p.