The mathematical relation between the impact factor and the uncitedness factor

EGGHE, Leo (2008) The mathematical relation between the impact factor and the uncitedness factor. In: SCIENTOMETRICS, 76(1). p. 117-123.

DOI: 10.1007/s11192-007-1902-x
Handle: http://hdl.handle.net/1942/7904

The mathematical relation between the impact factor and the uncitedness factor

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium
leo.egghe@uhasselt.be

ABSTRACT

In a general framework, given a set of articles and their received citations (time periods of publication or citation are not important here) one can define the impact factor (IF) as the total number of received citations divided by the total number of publications (articles). The uncitedness factor (UF) is defined as the fraction of the articles that received no citations.

It is intuitively clear that IF should be a decreasing function of UF. This is confirmed by the results in [T.N. van Leeuwen and H.F. Moed. Characteristics of journal impact factors: the effects of uncitedness and citation distribution on the understanding of journal impact factors. Scientometrics 63(2), 357-371, 2005] but all the given examples show a typical shape, seldom seen in informetrics: a horizontal S-shape (first convex then concave).

[1] Permanent address

Key words and phrases: impact factor, IF, uncitedness factor, UF.

Adopting a simple model for the publication-citation relation, we prove this horizontal S-shape in this paper, showing that such a general functional relationship can be generally explained.

I. Introduction

The Garfield-Sher impact factor (Garfield (1955, 2001)) is, perhaps, the indicator with the largest impact in informetrics. Essentially (and that is all we need in this paper) it is very simply

$$\mathrm{IF} = \frac{C}{P} \tag{1}$$

where P denotes the total number of publications (articles) and where C denotes the total number of citations to these articles. Of course one has to give information on the used publication period and citation period in order to be able to calculate (1) explicitely (for the Garfield-Sher impact factor, for each year x, the citation period is this year x and the publication period is determined by the years $x-1$ and $x-2$ (together) but, as said, this is of no importance here).

It is surprising that such a simple measure (IF being the average number of citations per article) – more than 40 years after its introduction – keeps on attracting papers devoted to the impact factor. We will not go into the different variants of the definitions of impact factors involving different citation and publication periods (we refer to Frandsen and Rousseau (2005) or Ingwersen, Larsen, Rousseau and Russell (2001) for this – see also references therein) but we will focus here on the understanding of journal impact factors in relation with other indicators. A good example of this is given in van Leeuwen and Moed (2005) where, e.g., the relation of I with the so-called “uncitedness factor” U is studied. The uncitedness factor is, simply, the fraction of uncited papers. Here we suppose, very generally, that we have a set of papers which are (or are not) cited in a certain fixed, but unspecified, time period. In short, we suppose to have a general information production process (IPP) where we have sources (articles) and items (citations to these articles) (see Egghe (2005)).

In van Leeuwen and Moed (2005), the functional relation between IF (as ordinate) and U (as abscissa) is found to be decreasing, a very logical fact. But there is more at stake here. Fig. 1 (reprinted with permission from Springer, Dordrecht, the Netherlands, to whom our sincerest thanks), (one of the $\mathrm{IF}(\mathrm{U})$ relations given in van Leeuwen and Moed (2005)) shows a very typical form, which is refound in all the graphs given in van Leeuwen and Moed (2005): first the curve is convexly decreasing followed by a concave decrease, hence a “horizontal” S-shape also implying a fast decrease around $\mathrm{U} = 0$ and $\mathrm{U} = 1$. That such a shape is refound in all the examples produced in van Leeuwen and Moed (2005) cannot be a coincidence and needs an explanation. This is the topic of this paper.

One of the referees wanted a comment on the importance of the problem. In itself, the relation U ® $\mathrm{IF}(\mathrm{U})$ is not so important in Scientometrics. It is clear that the relation should be decreasing. But, since in van Leeuwen and Moed (2005), all these relations have the typical S-shape as in Fig. 1, this cannot be a coincidence. Furthermore, as far as I am aware of, such an S-shape is only encountered here: first a convex decrease, followed by a fast concave decrease. Explaining such a shape is hence a logical (small) step in the development of this field. One cannot hide some mathematical explanations for scientometric regularities from the researchers in this field. Another possible interest in this regularity is that, in a certain field, the impact factor can be predicted by the uncitedness factor.

Using a very simple model for the size-frequency function $\mathrm{f}(\mathrm{n}) =$ the number of articles with $\mathrm{n} = 0,1,2,3,...$ citations, supposed to be decreasing in n, we indeed prove that the functional relation $\mathrm{IF} = \mathrm{IF}(\mathrm{U})$ is of the form described above and illustrated in Fig. 1.

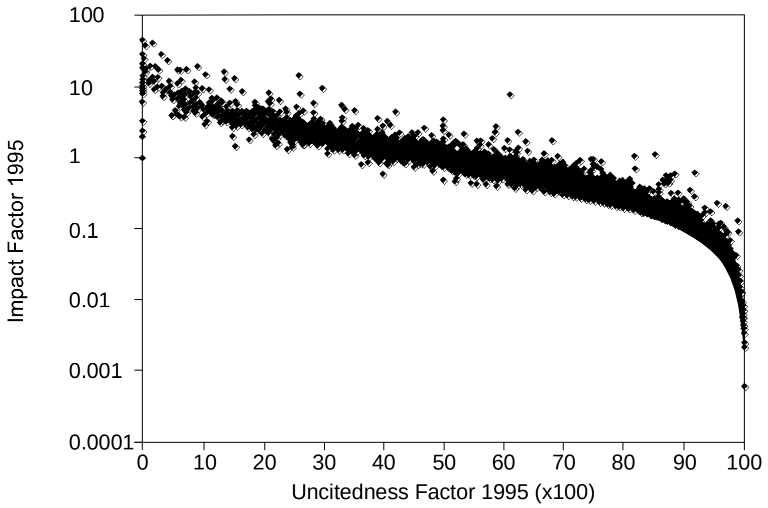

Fig. 1. Distribution of impact factor and uncitedness factor for all SCI fields, 1995, van Leeuwen and Moed (2005), reprinted with kind permission of Springer.

II. Proof of the relation between the impact factor IF and the uncitedness factor U in general systems

Let us have a general system in which we have articles that are (or are not) cited. Let P denote the total number of articles and let C denote the total number of citations. What we need is a size-frequency function $f(n)$, describing the number of articles with n citations, $n = 0,1,2,\ldots$. In this article, however, we will take n as a continuous variable, say n Î $[0,D]$.

The most classic function for f is Lotka's law (see Egghe (2005)): for α ³ 1

$$f(n)= \frac{C}{n^{\alpha}} \tag{2}$$

but it is clear that (2) does not cover the case $n = 0$ (which is crucial in this paper since we deal with the uncitedness factor).

One could use (as was also done in Rousseau (1997)), for $\alpha \geq 1$

$$f(n) = \frac{C}{(n+1)^{\alpha}} \tag{3}$$

for $n \geq 0$, but for this function I was not able (the calculations being too intricate) to complete the calculations for IF and U and hence I could not recover the $IF(U)$ functionality. In Redner (2005) and Burrell (2007), even more complicated models for $f(n)$ are presented which cannot be used, I assume, in the present context.

The least we can say is that these models are decreasing: $f(n)$ is decreasing in n. This is also what one can expect in most practical cases. The simplest decreasing function, that is positive on $n \in [0, D]$ is the linear one as in Equation (4):

$$f(n) = D - n \tag{4}$$

Although (4) will not be the right model for $f(n)$ I think that, if we can show, using (4), a graph as in Fig. 1 for the relation between IF and U, we then have a first explanation of this regularity and hence we have a mathematical understanding of the relation between the impact factor IF and the uncitedness factor U. This is what we will do in the sequel.

With (4) we have

$$P = \int_0^D (D - j)\,dj = \frac{D^2}{2} \tag{5}$$

$$C = \int_0^D j(D - j)\,dj = \frac{D^3}{6} \tag{6}$$

The number of articles that have at least one citation will be expressed by P_c:

$$P_c = \int_1^D (D - j)\,dj = \frac{D^2}{2} - D + \frac{1}{2} \tag{7}$$

Note that, in this model,

$$P - P_c = \int_0^1 (D - j)dj$$

stands for the number of articles with less than 1 citation, approximating (but not equalling) the number of uncited articles. Hence we approximate U by

$$U = 1 - \frac{P_c}{P} = 1 - \frac{\frac{D^2}{2} - D + \frac{1}{2}}{\frac{D^2}{2}}$$

$$U = \frac{2D - 1}{D^2} \tag{8}$$

By (1) we have, using also (5) and (6):

$$IF = \frac{C}{P} = \frac{D}{3} \tag{9}$$

Formula (9) in (8) yields

$$U = \frac{6(IF) - 1}{9(IF)^2}$$

or

$$IF = \frac{1 \pm \sqrt{1 - U}}{3U} \tag{10}$$

If we take the minus sign in (10) then $U \to 0$ implies $IF \to 0$ which cannot be the case. In fact, with the minus sign in (10) we find that IF is increasing in U (which is readily seen), which is impossible (formula (10) with the minus sign is an imported solution). Hence we only keep the plus sign in (10) yielding

$$IF = \frac{1 + \sqrt{1 - U}}{3U} \tag{11}$$

Now IF is a decreasing function of U:

$$\frac{d(IF)}{dU} = \frac{-\frac{3}{2}U - 3\left(\sqrt{1 - U} + (1 - U)\right)}{9U^2\sqrt{1 - U}} < 0 \tag{12}$$

since $0 \leq U \leq 1$. We have

$$\lim_{\substack{U \to 0 \\ >}} IF = +\infty \tag{13}$$

$$\lim_{\substack{U \to 1 \\ <}} IF = \frac{1}{3} \tag{14}$$

Result (14) might be surprising since one would expect that the lowest value of IF should be 0. But one should not forget that we approximated the uncitedness factor by the fraction U of "lowly cited" papers, i.e. for which $n \in [0,1]$ and this explains the positive number in (14).

Now we will check the double derivative of IF with respect to U in order to check the shape of Fig. 1. We have

$$\frac{d^2(IF)}{dU^2} = \frac{X}{3U^3(1 - U)^{\frac{3}{2}}} \tag{15}$$

where

$$X = -\frac{1}{2}U(1 - U) - \frac{1}{2}U + \frac{1}{4}U^2 + 2(1 - U)\sqrt{1 - U} + 2(1 - U) \tag{16}$$

The denominator of (15) is positive. Now, for $U = 0$, X equals $X = 4 > 0$ while for $U = 1$, X equals $X = -\frac{1}{4} < 0$. This shows that the graph $U \to IF(U)$ starts convexly and ends concavely:

this is so because (15) attains all values between the positive value in $U = 0$ and the negative value in $U = 1$ (using the fact that (15) is continuous). Of course, for a certain value $U \text{Î} \;]0,1[$ we have an osculation point (i.e. for which $\frac{d^2(IF)}{dU^2} = 0$). One can also see that, by (12),

$$\lim_{\substack{U \text{®} 0\\ >}} \frac{d(IF)}{dU} = \lim_{\substack{U \text{®} 1\\ <}} \frac{d(IF)}{dU} = - \text{¥} \tag{17}$$

, hence we have shown that the graph of the function $U \text{®} IF(U)$, i.e. the functional relation between the impact factor and the uncitedness factor is as in Fig. 2. This is clearly the same shape as the one of the cloud of points in Fig. 1, hence explaining this regularity.

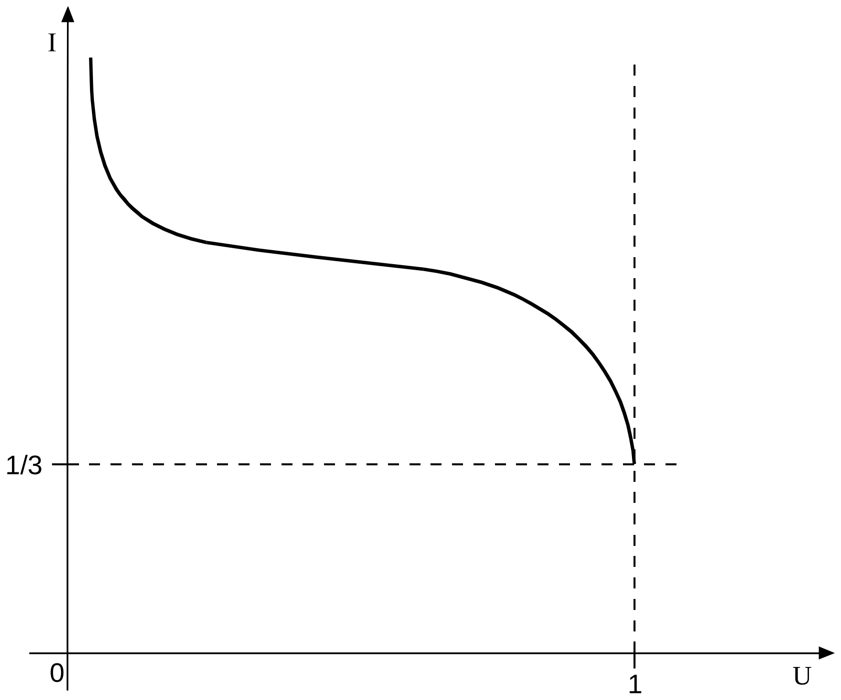

Fig. 2 The relation between IF and U

III. Open problems

It remains to explain the shape of Fig. 1 for a more realistic size-frequency function f. Also one needs to investigate this relation where U is exactly the uncitedness factor. Yet we think it is interesting to recover the shape of Fig. 1 for the relation between the impact factor and the fraction of “lowly cited” articles.

References

Burrell Q.L. (2007). Hirsch’s h-index: a stochastic model. Journal of Informetrics 1(1), 16-25, 2007.

Egghe L. (2005). Power Laws in the Information Production Process: Lotkaian Informetrics. Elsevier, Oxford, UK, 2005.

Frandsen T.F. and Rousseau R. (2005). Article impact calculated over arbitrary periods. Journal of the American Society for Information Science and Technology 56(1), 58-62, 2005.

Garfield E. (1955). Citation indexes to science: A new dimension in documentation through the association of ideas. Science 122, 108-111, 1955. Reprinted in: Essays of an Information Scientist 6, 468-471, 1983. ISI Press, Philadelphia, USA.

Garfield E. (2001). Recollections of Irving H. Sher 1924-1996: Polymath/Information Scientist *Extraordinaire*. Journal of the American Society for Information Science and Technology 52(14), 1197-1202, 2001.

Ingwersen P., Larsen B., Rousseau R. and Russell J. (2001). The publication-citation matrix and its derived quantities. Chinese Science Bulletin 46(6), 524-528, 2001.

Redner S. (2005). Citation statistics from 110 years of *Physical Review*. Physics Today 58(6), 49-54, June 2005.

Rousseau R. (1997). Sitations: an exploratory study. Cybermetrics 1(1), paper 1, 1997. http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

van Leeuwen T.N. and Moed H.F. (2005). Characteristics of journal impact factors: the effects of uncitedness and citation distribution on the understanding of journal impact factors. Scientometrics 63(2), 357-371, 2005.