Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 179–184
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

Short communication

A systematic analysis of Hirsch-type indices for journals

András Schubert [a], Wolfgang Glänzel [a,b,*]

[a] *ISSRU/IRPS, Hungarian Academy of Sciences, Budapest, Hungary*
[b] *Steunpunt O&O Statistieken, K.U. Leuven, Dekenstraat 2, B-3000 Leuven, Belgium*

Received 13 November 2006; accepted 20 December 2006

Abstract

A theoretical model of the dependence of Hirsch-type indices on the number of publications and the average citation rate is tested successfully on empirical samples of journal *h*-indices.

Keywords: *h*-Index; Pareto distribution; Impact factor; Journals; Ranking

1. Introduction

Recently (Braun, Glänzel, & Schubert, 2005; Braun, Glänzel, & Schubert, 2006), it was suggested that the *h*-index, originally proposed by Hirsch (2005) "to quantify an individual's scientific output", can be usefully applied to the citation analysis of journals, as well. The point is not only to introduce a robust alternative indicator advantageously supplementing journal impact factors, but also to gain deeper insight into the properties of this interesting but unusual statistical measure. An amount of incomprehension that appears to surround Hirsch's concept (and that manifests itself both in unjustified refusal and in likewise unjustified glorification) stems from the lack of a solid statistical–theoretical background supported by a sufficient body of empirical evidences.

The citation analysis of journals seems to provide a particularly appropriate field for experimentation, since – more than in many other levels of investigation (e.g., individuals, groups or institutions) – there is a relative consensus (or, at least a tacitly accepted tradition) in some basic conceptual and methodological questions. As the definition of the journals themselves, the choice of the publication and citation windows, the categorization of journals into fields and subfields and similar questions are concerned, there is a vast literature to refer to; first of all, the 30 years of the Journal Citation Reports serves as a unique orientation pole—if not necessarily to be followed.

2. Theoretical background

In a recent paper (Glänzel, 2006), an attempt was made to interpret theoretically some properties of the *h*-index, given the underlying citation distribution, on the basis of extreme-value statistics. Specifically, the dependence of the *h*-index on the basic parameters of the distribution and on the sample size was discussed using Gumbel's characteristic extreme values.

* Corresponding author. Tel.: +32 16 32 57 13; fax: +32 16 32 57 99.
E-mail address: wolfgang.glanzel@econ.kuleuven.ac.be (W. Glänzel).

1751-1577/$ – see front matter
doi:10.1016/j.joi.2006.12.002

Let X be a random variable. In our case, X represents the citation rate of a paper. The probability distribution of X is denoted by $p_k = P(X = k)$ for every $k \geq 0$ and the cumulative distribution function is denoted by $F(k) = P(X < k)$. Put $G_k = G(k) := 1 - F(k) = P(X \geq k)$. Gumbel's rth characteristic extreme value (u_r) is then defined as

$$u_r := G^{-1}\left(\frac{r}{n}\right) = \max\left\{k : G(k) \geq \frac{r}{n}\right\}, \tag{1}$$

where n is a given sample with distribution F. The theoretical h-index (H) can consequently be defined as

$$H := \max\{r : u_r \geq r\} = \max\left\{r : \max\left\{k : G(k) \geq \frac{r}{n}\right\} \geq r\right\}. \tag{2}$$

If there exists such index r so that $u_r = r$ then we have obviously $H := r$ and we can write $H := u_H$.

Let us consider now an important special case, namely the discrete Paretian distributions with finite expectation. Most distributions used for modelling publication activity and citation processes belong to this category.

We say, that the distribution of a random variable X is Paretian if it asymptotically obeys Zipf's law, i.e, if $\lim_{k\to\infty} G_k k^{-\alpha} = \text{constant}$. Asymptotically Pareto distributed random variables obviously meet this definition since $p_k = P(X = k) \approx d\{N + k\}^{-(\alpha+1)}$ if $k \gg 1$; $\alpha > 1$, where N and d are positive constants. In what follows we will deal with this family of distributions. For $k \gg N$ we obtain $p_k = P(X = k) \approx dk^{-(\alpha+1)}$ and $G_k = P(X \geq k) \approx d_1 k^{-\alpha}$, where d_1 is a positive constant. Hence we have,

$$E(X) = \sum_{k=0}^{\infty} k p_k = \sum_{k=0}^{\infty} G_k < \infty, \quad \text{if}\, \alpha > 1. \tag{3}$$

By elementary manipulation of the cumulative distribution function we obtain the following approximation from the above definition of Gumbel's rth characteristic extremes.

$$u_r \approx c_1 \left(\frac{n}{r}\right)^{1/\alpha}, \tag{4}$$

where c_1 is a positive constant. Applying the Hirsch condition to this approximation results in the property

$$H = u_H \approx c_1 \left(\frac{n}{H}\right)^{1/\alpha}, \quad \text{if}\, n \gg 1. \tag{5}$$

Hence we have

$$H \approx c_2 n^{1/(\alpha+1)}, \quad \text{if}\, n \gg 1, \tag{6}$$

where $c_2 = c_1^{\alpha/(\alpha+1)}$ is a positive constant. In verbal terms, the h-index is approximately proportional to the $(\alpha + 1)$th root of the number of publications.

In the case of journals, the question is whether and how the h-index of a journal is determined by the parameters of its citation distribution (first of all, its expected value: the impact factor) and the "sample size": the number of papers published in the corresponding journal.

In case of a two-parameter Pareto distribution with parameters N, α, the expected value ("impact factor") is

$$\text{IF} = \frac{N}{\alpha - 1}, \tag{7}$$

while the constant in Eq. (4) is $c_1 = N$. Hence,

$$H = u_H \approx c_1 \left(\frac{n}{H}\right)^{1/\alpha} \approx N^{\alpha/(\alpha+1)} n^{1/(\alpha+1)}. \tag{8}$$

In the special case of $\alpha = 2$, which corresponds to a Lotka distribution with exponent 3, and is in line with common assumption in bibliometric models (e.g. Pao, 1986), Eq. (7) results in $\text{IF} = N$, hence we finally obtain

$$H = cn^{1/3}\, \text{IF}^{2/3}, \tag{9}$$

where c is a positive real value of order 1.

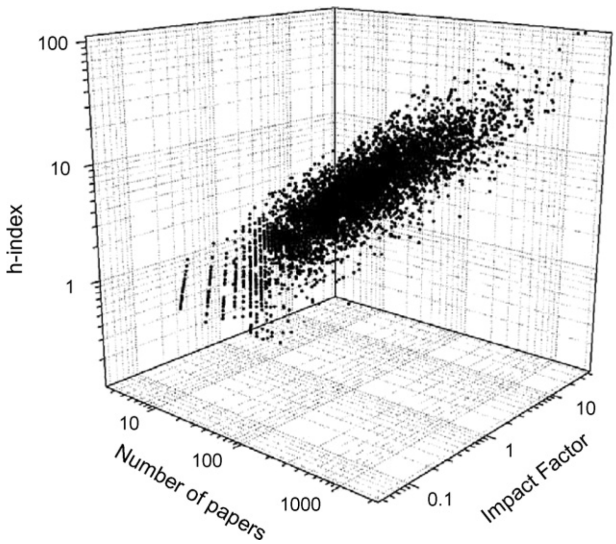

Fig. 1. Dependence of the journal h-index on the number of papers and the impact factor. Publication year: 2001; citation window: 2001–2003.

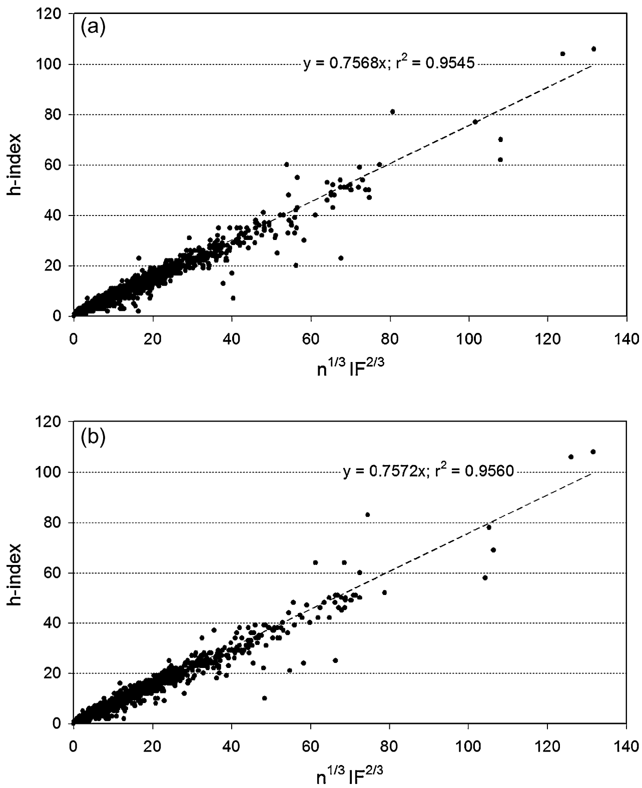

Fig. 2. Correlation of the journal h-index with $n^{1/3}\mathrm{IF}^{2/3}$. All science fields combined. (a) Publication year: 2001; citation window: 2001–2003. (b) Publication year: 2002; citation window: 2002–2004.

3. Methods

Journal citation data were collected from the Web of Science (WoS) database of Thomson Scientific. In the present study, we used one publication year (2001 and 2002) and a 3-year citation window (the year of publication plus the two succeeding years) to determine both the *h*-indices and the average citation rate (quasi-impact factor, IF) of the journals. Four document types: articles, letters, notes and reviews (as coded in the WoS database) were taken into consideration. For assigning journals to science fields, the classification scheme of Glänzel and Schubert (2003) was used.

4. Results

Number of publications (n), average citation rates (denoted, by tradition, as IF) and *h*-indices of 6406 journals in 2001 and 6481 journals in 2002 were determined.

Fig. 1 shows an overall view on the dependence of h on n and IF. On the triple logarithmic scale, the points representing the journals are apparently lying on a plane. This suggests a power function dependence of h both on n and IF.

In order to test systematically the validity of Eq. (9), H was plotted against the product $n^{1/3}\mathrm{IF}^{2/3}$ using data for all science fields combined, as well as for two science fields: biology and chemistry (according the classification scheme by Glänzel & Schubert, 2003). Data for two source years, 2001 and 2002 were analyzed. The test led to convincing positive results (see Figs. 2–4). It was not only proven that H showed a strong linear correlation with the product $n^{1/3}\mathrm{IF}^{2/3}$, but it could also be seen that the value of the constant c was remarkably independent of the science field and the source year, its value being around 0.75.

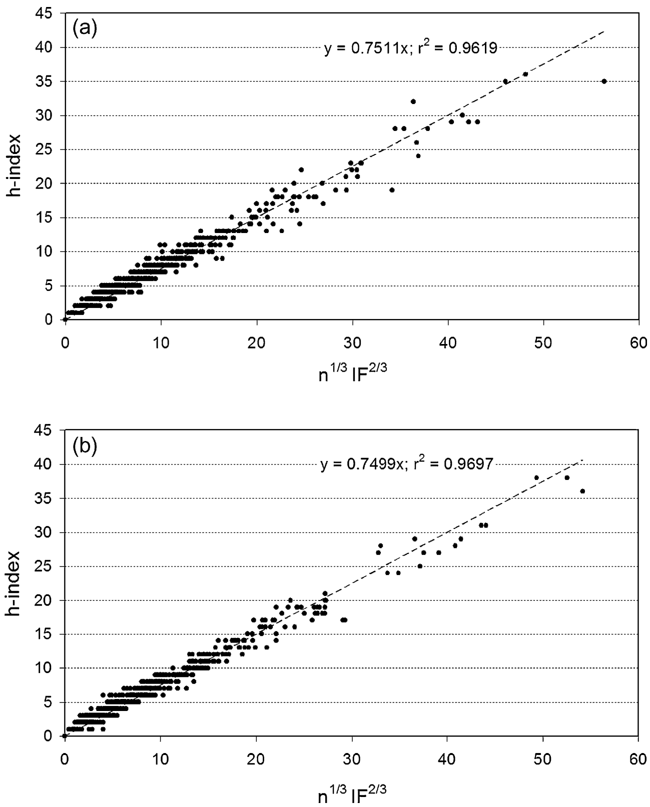

Fig. 3. Correlation of the journal *h*-index with $n^{1/3}\mathrm{IF}^{2/3}$. Biology. (a) Publication year: 2001; citation window: 2001–2003. (b) Publication year: 2002; citation window: 2002–2004.

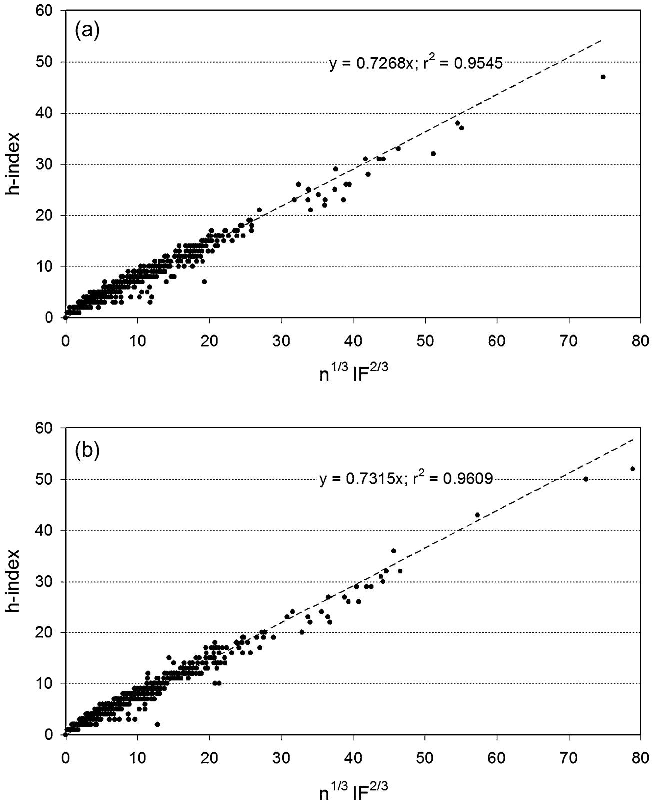

Fig. 4. Correlation of the journal h-index with $n^{1/3}\mathrm{IF}^{2/3}$. Chemistry. (a) Publication year: 2001; citation window: 2001–2003. (b) Publication year: 2002; citation window: 2002–2004.

5. Conclusions

Studying journal h-indices enabled us to submit a theoretical model of Hirsch-type indices to an empirical test. It was found that a simple theoretical relation (see Eq. (9)) among H, n and IF, derived from a Paretian model to the citation distribution within the journals, fits perfectly to the empirical data obtained from one publication year with a 3-year citation window.

Remarkably, the only free parameter of the model, c, proved to be practically independent of the science field. That means that no specific field dependence of the h-index should be accounted for beyond the well-known field dependence of publication productivity and citation rate. Eq. (9) in this sense allows for a kind of "similarity transformation" of h-indices between different fields.

There are a series of open questions waiting for future investigation. Whether Eq. (9) remains valid for other choices of publication and citation periods, as well, or does the exponent α change with growing time intervals as observed by Vlachý (1976) and Pao (1986) resulting in different forms of Eqs. (8) and (9)? How the constant c will behave with changing periods? Whether c will remain constant even in a wider range of science fields?

Hirsch's h-type indices will certainly challenge scientometrists for a while, and their use in the citation assessment of journals seems to have promising perspectives with a lot of systematic analysis and statistical background work to be done.

References

Braun, T., Glänzel, W., & Schubert, A. (2005). A Hirsch-type index for journals. *The Scientist*, *19*(22), 8.

Braun, T., Glänzel, W., & Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics*, *69*(1), 169–173.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for bibliometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.

Glänzel, W. (2006). On the *h*-index—A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, *67*(2), 315–321.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572 (also available at: arXiv: physics/0508025, accessible via http://arxiv.org/abs/physics/0508025)

Pao, M. L. (1986). An empirical examination of Lotka's law. *Journal of the American Society for Information Science*, *37*(1), 26–33.

Vlachý, J. (1976). Time factor in Lotka's law. *Probleme de Informare si Documentare*, *10*(2), 44–87.