SOME NEW APPLICATIONS OF THE H-INDEX*

Wolfgang Glänzel[1,2]

[1]Steunpunt O&O Indicatoren
& K.U.Leuven, Faculty ETEW, MSI, Leuven (Belgium)
[2]Hungarian Academy of Sciences,IRPS, Budapest (Hungary)
wolfgang.glanzel@econ.kuleuven.be

In this note some new fields of application of Hirsch-related statistics are presented. Furthermore, so far unrevealed properties of the h-index are analysed in the context of rank-frequency and extreme-value statistics.

■ Introduction

Since its introduction in 2005, the h-index has mainly been used as a measure to quantify the research output of individual scientists. This is in line with Jorge E. Hirsch's intensions (Hirsch, 2005). Recent attempts to fine-tune or improve the indicator (e.g., Egghe, 2006 and Jin et al., 2007) or to extend its use to higher levels of aggregation (e.g., Braun et al., 2005) follow the original design. In what follows, we will show some new application possibilities of the h-index in the context of rank statistics. In particular, the properties of the characteristic extreme values of Pareto-type distributions provide the basis of the new statistics. The first application is actually found in the form of a composite indicator strongly related to the h-index. The second application relates the h-index with a generalised version of the Zipf-Mandelbrot law. While the first indicator can only be applied to distributions with finite expectation, that is $\alpha > 1$, the second application even works if $\alpha \leq 1$. Both applications are useful supplements in evaluative studies of research performance at the micro and meso level.

■ Theoretical background

In recent papers (Glänzel, 2006, Egghe and Rousseau, 2006, Burrell, 2007), attempts were made to interpret theoretically some properties of the *h*-index and to connect the results with traditional indicators of publication activity and citation impact (Schubert and Glänzel, 2007). The underlying citation distribution was assumed to be Paretian and on the basis of extreme-value statistics, important properties and regularities could be derived from the distribution. 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. In order to further elaborate these new approaches, we briefly summarise the mathematical rudiments.

Let X be a random variable. In the present 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 $G_k := 1 - F_k = P(X \geq k)$. Assume we have a sample $(\{X\}_{i=1,\ldots,n})$ of size n where all elements are independent and have the same distribution F. Gumbel's r-th characteristic extreme value (u_r) is then defined as **[1]**

$$u_r := G^{-1}(r/n) = \max\{k: G_k \geq r/n\},$$

where n is a given sample with distribution F (see *Gumbel*, 1958). The actual rank statistics $R(r) = X_r^*$ (where $X_1^* \geq X_2^* \geq \ldots \geq X_i^* \geq \ldots \geq X_n^*$

are the ordered/ranked elements of the sample $\{X\}_{i=1,\ldots,n}$) can be considered an estimator of the corresponding r-th characteristic extreme value u_r.

According to Glänzel (2006), the theoretical h-index (h) can be defined as **[2]**

$$h := \max\{r:\ u_r \geq r\} = \max\{r:\ \max\{k:\ G_k \geq r/n\} \geq r\}.$$

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

■ Methods and Results

For simplicity's sake we assume that the citation distribution under study can be approximated by a non-negative continuous distribution. In the case of continuous distributions we will write $F(x)$ and $G(x)$ instead of F_x and G_x, respectively. Furthermore, we assume that the underlying citation rates follow a Pareto distribution of the second kind. This general form of the Pareto distribution, also referred to as *Lomax* distribution, can be obtained from the infinite beta distribution if one of the parameters is chosen 1 (e.g., Johnson, Kotz, & Balakrishnan, 1994). In particular, we say that the non-negative random variable X has a Pareto distribution (of the second kind) if **[3]**

$$G(x) = P(X \geq x) = N^\alpha/(N + x)^\alpha\ ,\ \text{for all } x \geq 0$$

Clearly, if x is large ($x >> N$) we can neglect the parameter in the denominator and we have **[4]**

$$G(x) \sim N^\alpha/x^\alpha\ ,\ \text{for } x >> N.$$

Assuming a statistical sample with Lomax distribution and size n we obtain **[5]**

$$G(u_r) \sim N^\alpha/u_r^\alpha = r/n,\ \text{if}\ \ n >> r.$$

Consequently, we have $r \cdot u_r^\alpha = N^\alpha \cdot n$ and **[6]**

$$\zeta(r) := r^{1/(\alpha+1)} \cdot u_r^{\alpha/(\alpha+1)} = N^{\alpha/(\alpha+1)} \cdot n^{1/(\alpha+1)}.$$

Since the right-hand side does not depend on the particular rank r, the left-hand side must be a constant. Furthermore, we have $\zeta(h) = h$ by definition (namely $\zeta(h) = h^{1/(\alpha+1)} \cdot h^{\alpha/(\alpha+1)} = h$). Consequently, $\zeta(r) \equiv h$ for all $r << n$. This property yields the first important result, particularly, **[7]**

$$h = c(\alpha)^* \cdot E(X)^{\alpha/(\alpha+1)} \cdot n^{1/(\alpha+1)},\ \ \text{if}\ \alpha > 1,$$

where $E(X) = N/(\alpha-1)$ is the expected value of the underlying Lomax distribution and $c(\alpha)^* = (\alpha-1)^{\alpha/(\alpha+1)}$ is a positive real value which only depends on the parameter α. Taking into account that the continuous *Lomax distribution* model often rather poorly fits the empirical discrete, integer-valued distributions, one cannot expect a perfect correlation. Nonetheless, we have found a strong correlation between h and $m^{\alpha/(\alpha+1)} \cdot n^{1/(\alpha+1)}$ with m being the mean citation rate of scientific journals (Schubert and Glänzel, 2007). Solely the empirical $c(\alpha)^*$ value was usually somewhat lower that the theoretical one.

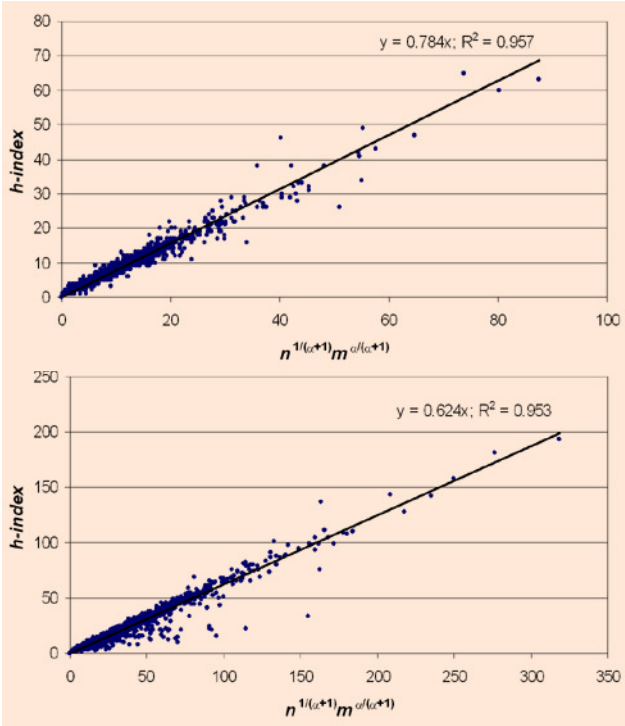

Figure 1 Correlation of the journal h-index with $n^{1/(\alpha+1)} m^{\alpha/(\alpha+1)}$ in all science fields combined top: citation window: 1980-1982 (α=2); bottom: citation window: 1980–2000 (α=1.5)

For largely different citation windows we have found solutions with different α values. For small windows comprising an initial period of about three years after publication, an α value around 2 has been found appropriate. For larger windows lower values yield an optimum solution. This change of exponent α with growing time intervals is in line with observations by Vlachý (1976) and Pao (1986).

Figure 1 shows the dependence of h on n and journal impact measures m for papers published in 1980 and indexed in the *Science Citation Index* of Thomson Scientific (Philadelphia, PA, USA). The impact measures have been calculated for a 3-year (top) and 21-year (bottom) citation window, respectively. In the first case $\alpha = 2$, for the longer citation period $\alpha = 1.5$ has been chosen.

Both theoretical considerations and empirical analysis lead to the conclusion that the h-index strongly correlates with $m^{\alpha/(\alpha+1)} \cdot n^{1/(\alpha+1)}$ which can be considered a composite indicator combining publication output and mean citation rate. Although this indicator has interesting properties, it is not designed to substitute the h-index. We just mention is passing that a similar indicator for journal impact was already suggested by *Lindsey* (1978) independently from the Hirsch index. For his *Corrected Quality Ratio* (CQR) we actually have $CQR^{0.4} = n^{0.4} \cdot m^{0.6}$, i.e., in this case we have $\alpha = 1.5$.A second important property can be observed when replacing the theoretical values in the left-hand side of Eq. (6) by the corresponding statistics. In particular, we obtain $z(r) = r^A \cdot R(r)^{(1-A)}$ with $A = 1/(\alpha+1)$, where $z(r)$ is expected to be an estimator of h for each $r << n$. In practice the deviation from $\zeta(r) \equiv h$ is quite large for the individual r values but the median M of the empirical $z(r) = r^A \cdot R(r)^{(1-A)}$ values proofed a strikingly robust estimator of h. Table 1 presents the corresponding statistics for selected *Price Awardees*. Both publications and citations have been counted from 1972 till May 2007 and no selection has been made for relevant literature. Thus papers in chemistry have been included for Schubert; the same applies for physics publication by van Raan and mathematical papers by Egghe, Rousseau and Glänzel. For the *Hirsch core* we obtained $M \sim h$ (with $\alpha = 2$). Except for Henry Small, z statistics provide robust approximation and the corresponding medians good estimators of the h-index. The reason for the poor fit for Henry Small lies in several highly cited papers of the 1970s on co-citation analysis and an extremely skewed citation distribution. For Henry Small an α value of about 1.0 resulting in constant z values around 20 would be more appropriate.

This new method for analysing the tail properties of Pareto-type distributions based on the z statistics works much better than the model described by Glänzel and Schubert (1988). The latter one was based on transformations of ordered statistics, namely on $r \cdot \ln(X_r^* / X_{r+1}^*) = r \cdot \ln[R(r)/R(r+1)]$ with $r << n$, which were extremely sensitive to ties. These statistics have an exponential distribution with parameter α, provided the underlying common distribution of X_r^* is Paretian with the same parameter (Glänzel et al., 1984). In practice, rank statistics of integer-valued discrete distributions often include ties (i.e. $R(r) = R(r+1)$ for some $r = 1, 2, \ldots$) resulting in $r \cdot \ln[R(r)/R(r+1)] = 0$. These ties can heavily distort the fit of the exponential distribution. By contrast, the new z statistics are more robust and much less sensitive to ties (see Table 1). One further important property is worth mentioning, namely that the z statistics can be considered a version of the Zipf-Mandelbrot law (cf. Yablonski, 1980, Egghe and Rousseau, 1990), where the constant value equals the h-index to the power $(\alpha+1)$, that is, $r \cdot R(r)^\alpha = \{z(r)\}^{\alpha+1} = h^{\alpha+1}$. Consequently, the case $\alpha = 1$ results in the following version of the classical Zipf's Law: $z(r) = \{r \cdot R(r)\}^{1/2} = h$ = constant, or equivalently, $r \cdot R(r) = h^2$.

(See table Table 1 for Hirsch-type indexes for selected bibliometricians who have been awarded the Price Medal (with $\alpha = 2$) on the next page!)

■ Conclusions

In this paper we have described two new applications of Hirsch-related indexes. The composite indicator, which expresses a multiplicative connection between derivatives of publication output and citation impact, proved surprisingly robust and works at both the meso and the micro level. Its strong correlation with the h-index is independent of the subject area (cf. Schubert and Glänzel, 2007). The z statistics, representing the second application, can be used to analyse the tail of citation distributions in the light of the h-index. At the same time, the h-index proved useful as truncation point for rank frequency analysis, for instance, by applying z and related statistics to the *Hirsch core* (e.g. Burrell, 2007) publications.

■ References

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

Burrell, Q. L. (2006), Hirsch's h-index: a stochastic model. *Journal of Informetrics*, 1 (1) 16-25

Burrell, Q. L. (2007), On the h-index, the size of the Hirsch core and Jin's A-index, *Journal of Informetrics*, 1 (2) 170-177.

Egghe, L., Rousseau, R. (1990). *Introduction to informetrics. Quantitative methods in library, documentation and information science*. Elsevier Science Publisher. Amsterdam.

Egghe, L., Rousseau, R. (2006). An informetric model for the h-index. Scientometrics, 69(1), 121-129.

Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics*, 69(1), 131-152.

	Egghe		Glänzel		Leydesdorff		Roussseau		Schubert		Small		van Raan	
r	*R(r)*	*z(r)*	*R(r)*	*z(r)*	*R(r)*	*z(r)*	*R(r)*	*z(r)*	*R(r)*	*z(r)*	*R(r)*	*z(r)*	*R(r)*	*z(r)*
1	53	14.1	131	25.8	116	23.8	33	10.3	131	25.8	335	48.2	113	23.4
2	42	15.2	54	18.0	40	14.7	26	11.1	128	32.0	249	49.9	56	18.4
3	40	16.9	40	16.9	32	14.5	19	10.3	86	28.1	135	38.0	56	21.1
4	36	17.3	37	17.6	29	15.0	18	10.9	61	24.6	114	37.3	41	18.9
5	22	13.4	36	18.6	23	13.8	17	11.3	60	26.2	88	33.8	40	20.0
6	19	12.9	35	19.4	22	14.3	16	11.5	40	21.3	82	34.3	34	19.1
7	17	12.6	34	20.1	21	14.6	16	12.1	34	20.1	82	36.1	32	19.3
8	17	13.2	34	21.0	21	15.2	16	12.7	34	21.0	79	36.8	32	20.2
9	16	13.2	34	21.8	20	15.3	15	12.7	28	19.2	78	38.0	31	20.5
10	16	13.7	28	19.9	19	15.3	15	13.1	27	19.4	51	29.6	26	18.9
11	15	13.5	27	20.0	18	15.3	15	13.5	27	20.0	46	28.6	25	19.0
12	15	13.9	27	20.6	17	15.1	15	13.9	27	20.6	41	27.2	25	19.6
13	14	13.7	27	21.2	15	14.3	15	14.3	26	20.6	30	22.7	25	20.1
14	14	14.0	26	21.2	14	14.0	14	14.0	23	19.5	28	22.2	25	20.6
15			24	20.5					23	19.9	25	21.1	24	20.5
16			23	20.4					22	19.8	23	20.4	24	21.0
17			22	20.2					22	20.2	23	20.8	23	20.8
18			22	20.6					19	18.7	18	18.0	22	20.6
19			22	21.0					19	19.0			21	20.3
20			21	20.7									21	20.7
21													21	21.0
h	**14**		**20**		**14**		**14**		**19**		**18**		**21**	
M	**13.7**		**20.4**		**14.9**		**12.4**		**20.2**		**31.7**		**19.6**	

Table 1 Hirsch-type indexes for selected bibliometricians who have been awarded the Price Medal (with $\alpha = 2$)

Glänzel, W., Schubert, A., Telcs, A. (1984), *Goodness of Fit Test for the Tail of Distributions*. Bolyai Colloquium on Goodness of fit (Debrecen, Hungary, June 25-28, 1984).

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.

Glänzel, W., Schubert, A. (1988), *Theoretical and Empirical Studies of the Tail of Scientometric Distributions*. In: L. Egghe, R. Rousseau (Eds.), Informetrics 87/88, Elsevier Science Publisher, 75-83.

Gumbel, E. J. (1958). *Statistics of extremes*. New York: Columbia University Press.

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).

Jin, B.H., Liang, L.M., Rousseau., R., Egghe, L. (2007), The R- and AR-indices: Complementing the hindex. *Chinese Science Bulletin*, 52 (6) 855-863.

Johnson, N. L., Kotz, S., Balakrishnan, N. (1994), *Continuous univariate distributions*. Volume 1, 2nd Edition, John Wiley & Sons, Ney York.

Schubert, A. Glänzel, W. (2007), A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics*, 1 (3), in press. (doi:10.1016/j.joi.2006.12.002)

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.

Yablonski, A. I. (1980), On fundamental regularities of the distribution of scientific productivity. *Scientometrics*, 2 (1) 3-34.

ISSI Newsletter is published by ISSI (http://www.issi-society.info/). Contributors to the newsletter should contact the editorial board **by email**.
Wolfgang Glänzel: wolfgang.glanzel@econ.kuleuven.be
Ronald Rousseau: ronald.rousseau@khbo.be
Liwen Vaughan: lvaughan@uwo.ca
Aparna Basu: basu.aparna@rediffmail.com
Balázs Schlemmer: balazs.schlemmer@econ.kuleuven.be
Accepted contributions are **moderated** by the board. **Guidelines** for contributors can be found at http://www.issi-society.info/editorial.html