Journal of Informetrics 9 (2015) 762–776

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

On a formula for the h-index

Lucio Bertoli-Barsotti[a,*], Tommaso Lando[b]

[a] Department of Management, Economics and Quantitative Methods, University of Bergamo, via dei Caniana 2, 24127 Bergamo, Italy
[b] Department of Finance, VŠB—TU Ostrava, Sokolskà 33, 70121 Ostrava, Czech Republic

ARTICLE INFO

Article history:
Received 24 May 2015
Received in revised form 25 June 2015
Accepted 13 July 2015
Available online 1 September 2015

Keywords:
h-Index
Citation data
Citation statistic
Geometric distribution
Lambert W function

ABSTRACT

The h-index is a celebrated indicator widely used to assess the quality of researchers and organizations. Empirical studies support the fact that the h-index is well correlated with other simple bibliometric indicators, such as the total number of publications N and the total number of citations C. In this paper we introduce a new formula $\tilde{h}_w = \tilde{h}_w(N, C, c_{\text{MAX}})$, as a representative predictive formula that relates functionally h to these aggregate indicators, N, C and the highest citation count c_{MAX}. The formula is based on the 'specific' assumption of geometrically distributed citations, but provides a good estimate of the h-index for the general case. To empirically evaluate the adequacy of the fit of the proposed formula $\tilde{h}_w$, an empirical study with 131 datasets (13,347 papers; 288,972 citations) was carried out. The overall fit (defined as the capacity of $\tilde{h}_w$ to reproduce the true value of h, for each single scientist) was remarkably accurate. The predicted value was within one of the actual value h for more than 60% of the datasets. We found, in approximately three cases out of four, an absolute error less than or equal to 2, and an average absolute error of only 1.9, for the whole sample of datasets.

1. Introduction

The h-index, h, is a widely recognized representative measure of individual scientific achievement, so that nowadays it is computed by default in specialized databases (such as Scopus or Web of Science—WoS). As well known, the h-index is statistically related to other simple standard bibliometric indicators, such as the total number of publications N and the total number of citations C. Indeed, on the basis of empirical research data, h has been found to be significantly positively correlated with C as well as with N (van Raan, 2006). It is also well known that mathematically the h-index cannot exceed the number of publications (cited at least once) N and, symmetrically, it cannot exceed the highest citation count, c_{MAX}. Moreover it cannot exceed $\left[\sqrt{C}\right]$, which is the integer part of the square root of the total number of citations C. Then, in symbols $h \le \min\left\{N, \left[\sqrt{C}\right], c_{\text{MAX}}\right\}$ (Bertoli-Barsotti, 2013). Interestingly, in this paper we make use of these three simple indicators (C, N and c_{MAX}) not exclusively for determining an upper bound for the Hirsch index, but also for estimating its value. For this reason, uncited publications are omitted in the present analysis. In what follows, we shall use the following notations:

T: total number of publications
N: total number of publications cited at least once

* Corresponding author.
E-mail address: lucio.bertoli-barsotti@unibg.it (L. Bertoli-Barsotti).

http://dx.doi.org/10.1016/j.joi.2015.07.004
1751-1577/

C: total number of citations
c_{MAX}: citation count of the most cited publication
$m = C/N$: mean number of citations per publication

Generally, the h-index may be interpreted as a function of both N and C because it combines, in a loose sense, both productivity, expressed as the total number of papers N, and quality, expressed as a mean number of citations per paper $m = C/N$ (Prathap, 2010a), in one single measure. On the other hand, increasing publications *alone*, or the total number of citations *alone*, or C/N *alone*, does not have an immediate effect on the h-index. According to Adler, Ewing, and Taylor (2009) the h-index captures only "a small amount of information about the distribution of a scientist's citations". Put otherwise, the h-index is relatively insensitive to moderate variations of the 'type' of the citation distribution, and this may be an advantage if, as in the present paper, attention is restricted to finding an estimate of this index.

In fact, the aim of this paper is to present a new mathematically representative predictive model for h. More precisely, we introduce a formula that relates h functionally to N, C and c_{MAX}, say $\tilde{h}_W = \tilde{h}_W(N, C, c_{MAX})$, that is, equivalently, $\tilde{h}_W(N, m, c_{MAX})$. To do so, we will assume that citations are geometrically distributed. The formula is of interest because it makes it possible, at least theoretically, to determine how h changes as a function of the number of publications and the number of citations. We note that the idea is not new, in that this approach has already been successfully employed by Burrell (2013a) (for an in-depth analysis of the probabilistic mechanism that governs the citation process, see also Burrell, 2007)—but without giving an explicit formula for the h-index.

But before proceeding with this task, in the next section we briefly survey the methods best known in the literature for obtaining mathematical models (that is, mathematical estimators) for the h-index. Then, in the subsequent sections we will describe our formula in detail. We will also present a case study demonstrating the ability of the formula to produce good estimates of the 'true' h-index, for single authors.

2. Mathematical models for the *h*-index

Several alternative mathematical models for the h-index have been proposed in the literature. These models essentially depend on the assumption of a specific citation distribution function, say $n(x)$, representing the number of papers which have been cited a total of x times.

Regardless of the fact that a single (simple) probability model is perhaps unable to describe citation distributions over the whole range of citations (Redner, 1998; van Raan, 2001) – unless a relatively large number of parameters is used – examples of models of citation distributions (sometimes in terms of a rank-size formulation, and sometimes as size-frequency distribution) are: (a) the exponential distribution (Lancho-Barrantes, Guerrero-Bote, & Moya-Anegon, 2010); (b) the Weibull distribution, after Weibull, 1951; see also Johnson, Kotz, & Balakrishnan, 1994, p. 628), also referred to as 'stretched exponential distribution' (Bletsas & Sahalos, 2009; Laherrère & Sornette, 1998; Iglesias & Pecharroman, 2007); (c) the Tsallis distribution, also known as q-exponential distribution (Tsallis, 1988; Tsallis & de Albuquerque, 2000; Burrell, 2008; Anastasiadis, deAlbuquerque, deAlbuquerque, & Mussi, 2010; Wallace, Larivière, & Gingras, 2009); (d) the so-called 'log-normal' distribution (Redner, 2005; Perc, 2010; Stringer, Sales-Pardo, & Amaral, 2008; Radicchi, Fortunato, & Castellano, 2008); (e) the discrete generalized beta distribution (Martinez-Mekler et al., 2009; Campanario, 2010; Petersen, Stanley, & Succi, 2011; Mansilla, Köppen, Cocho, & Miramontes, 2007); (f) the Yule distribution (de Solla Price, 1976); (g) the logarithmic distribution (Bertoli-Barsotti and Lando, 2015); (h) the negative binomial, or Pascal distribution (Mingers & Burrell, 2006); (i) the Price distribution (Glänzel, 2006); to cite only some. In passing, note that some of these (a–d) are continuous random variables, while others (e–i) are discrete random variables. The formers here cited assume, typically, a real non negative support, while the latters range over positive integers (e–g), or non-negative integers (h and i). All these distributions may potentially define, correspondingly, a theoretical model for the h-index, but this may not be easy to find, depending principally on the existence of a cumulative distribution function in analytically closed form. Moreover, and more importantly, this possible theoretical model for h may not depend in a simple way on a few basic standard indicators, such as the total number of papers published, the total number of citations, or the mean number of citations per paper. In this sense, two Pareto-type citation models of special interest in bibliometrics and in citation analysis constitute well-known (positive) exceptions.

(1) The power-law/Pareto citation distribution, also known as 'inverse' power-law (Burrell, 2008), or Lotkaian informetric distribution or Lotka's law (Rousseau & Rousseau, 2000; Egghe, 2005a,b; Egghe & Rousseau, 2006; Lafouge, 2007), is probably the distribution most known and used in Informetrics. According to this probability model, the citation distribution function $n(x)$ (or size-frequency function) is equal to x^{-a} up to a normalizing factor, namely

$$n(x) \propto x^{-a}, \quad x \geq 1, \quad a > 1. \tag{1}$$

To be noted is that in our context the number x of citations is a discrete random variable. Accordingly, this probability model should only be viewed as a rough approximation of the Riemann zeta distribution (also known as discrete Pareto distribution, or Zipf distribution) $n(x) \propto x^{-a}$, $x = 1, 2, 3, \ldots$,, which is clearly more appropriate, even if more difficult to handle analytically (Nicholls, 1987).

More specifically, from (1) one obtains

$$n(x) = N(\alpha - 1)x^{-a}, \qquad x \geq 1, \quad \alpha > 1. \tag{2}$$

where N is the total number of published papers (receiving at least one citation). This law coincides, up to a constant, with a special case (i.e. with support $x \geq 1$) of a *Pareto distribution of the first kind P*(*I*)(1, α), where $\alpha > 1$ is a shape parameter (Arnold, 1983; Johnson et al., 1994, p. 573). In order to warrant the existence of its expectation, $\mu = (\alpha - 1)/(\alpha - 2)$, the condition $\alpha > 2$ must be assumed (unless one considers a truncated version of the same distribution). This model may be represented by a linear dependence in a double logarithmic axis plot (log–log plot) of the observed frequency n versus the number of citations x.

Adopting this model and assuming $\alpha > 2$, Egghe and Rousseau (2006) obtained the following formula for the h-index:

$$h = N^{1/\alpha} \tag{3}$$

By reparameterization, this can be rewritten as (Egghe, Guns, & Rousseau, 2011)

$$h = N^{(\mu-1)/(2\mu-1)} \tag{4}$$

This expression depends on unknown parameter values, but a simple estimate may be obtained by substituting the expected value μ with its observed counterpart, that is, the average number of citations per publication $m = C/N$, yielding the formula

$$h = N^{(m-1)/(2m-1)} \tag{5}$$

Alternatively, by taking the 'default' value of $\alpha = 2$ (that, strictly speaking, is correct only for an infinitely high value of m), the alternative simple formula

$$h = \sqrt{N} \tag{6}$$

may also be deduced (Ye, 2009), but this assumption differs from the conclusion reached by Redner (1998), who analyzed approximately 800,000 papers and found a typical value of about 3 for the parameter α—at least for the large-citation tail of the citation distribution.

Note that the latter formula can be rewritten as $h = m^{-0.5}\sqrt{C}$. In partial agreement with this, in a case study van Raan, 2006 found a good correlation between the h-index and the function $0.42C^{0.45} \cong 5.7^{-0.5}C^{0.45}$. Besides, Hirsch (2005, 2007) himself suggested the possible rule $h = r^{-0.5}\sqrt{C}$, where r is a constant ranging between 3 and 5.

(2) A similar but different approach (sometimes confused with the one above) has been considered by Glänzel (2006) (see also Schubert & Glänzel, 2007; Glänzel, 2007, 2008). This time, starting from a *Pareto distribution of the second kind P*(*II*)(0, σ, θ), also known as *Lomax distribution* (Johnson et al., 1994, p. 575), or Tsallis distribution (Shalizi, 2007), one has

$$n(x) \propto (x + \sigma)^{-\theta-1}, \qquad x \geq 0, \quad \theta > 0, \tag{7}$$

where $\sigma > 0$ is a scale parameter, and θ a shape parameter (Arnold, 1983, p.44). More specifically, one obtains $n(x) = T\theta\sigma^{\theta}(x + \sigma)^{-\theta-1}$ (see Shalizi, 2007, Eq. (4)). Here, in order to warrant the existence of the expectation $\mu = \sigma/\left(\theta - 1\right)$ of the distribution, the condition $\theta > 1$ must be assumed.

Adopting this model (and assuming $\theta > 1$), Glänzel (2006) obtained an *approximate* formula (valid only for $x \gg \sigma$) for the h-index, namely:

$$h \approx \sigma^{\theta/(\theta+1)}T^{1/(\theta+1)}$$

In this case also, by taking the 'default' value of $\theta = 2$ (incidentally, note that Glänzel, 2007, found that the most relevant range for this parameter is between 2 and 3.5), the formula simplifies to

$$h = c\sigma^{2/3}T^{1/3}$$

where c is a positive real value 'of order 1' (Schubert & Glänzel, 2007), and where it is intended that the expectation becomes $\mu = \sigma$. A simple way to estimate h is to substitute the expected value μ with its observed counterpart, $m_0 = C/T$, yielding

$$h = cm_0^{2/3}T^{1/3} \tag{8}$$

though still remaining to be identified and interpreted is the parameter c. A value of c around 0.75 was found applicable by Schubert and Glänzel (2007), in a study applied to the citation analysis of journals, while Iglesias and Pecharroman (2007) suggested the value $c = \sqrt[3]{1/4 = 0.63}$ (see also Vinkler, 2009). In words, this rule states that the h-index can be approximated by the product of a power function of the sample size and a power function the sample mean. Prathap

(2010a) interpreted $h = m_0^{2/3}T^{1/3}$ as a substitute or mock h-index, and renamed it the 'p-index' (Prathap, 2010b). Empirical applications of this formula, with possible small variants, are numerous: see for example, Glänzel (2008), Bletsas and Sahalos (2009), Csajbók et al. (2007), Vinkler (2009) and Schubert et al. (2009).

A similar approach, starting from a shifted Pareto distribution of the first kind,

$$n(x) = T(\alpha - 1)(x + 1)^{-\alpha}, \qquad x \geq 0, \quad \alpha > 1, \tag{9}$$

was proposed by Egghe and Rousseau (2012). It is immediate to see that this model is equivalent to a Pareto distribution of the second kind $P(II)(0, \sigma, \theta)$, by taking $\sigma = 1$ and substituting $\theta = \alpha - 1$ They easily obtained the equation $h(h+1)^{\alpha-1} = T$. Then, after substituting the expected value, $\mu = (\alpha - 2)^{-1}$, with its observed counterpart, $m_0 = C/T$, they deduced the equation

$$h(h + 1)^{(m_0+1)/m_0} = T,$$

which can be solved for h, but unfortunately not in explicit form.

For empirical comparative studies on some of the above formulas for the h-index see, for example, Abbas (2012), Ye (2009, 2011), Burrell (2013b) and Malesios (2015). Summarizing, according to these formulas, the h-index mainly depends on two factors: productivity, as the numbers of published papers, and quality/impact, as the average number of citations per publication—also called 'citedness' (Vinkler, 2010).

3. The main result

3.1. Power series distributions and the geometric distribution

Under the assumption of geometrically distributed data, the frequency-size function

$$n(x) = Nq^{x-1}p, \quad x = 1, 2, \ldots \quad (0 < q < 1, \quad p = 1 - q) \tag{10}$$

expresses the number of articles with exactly x citations (e.g. Np represents the number of papers with exactly one citation). Note that this model of citation distribution is based on a *shifted* geometric distribution, because its support does not contain the value $x = 0$. As said above, our declared goal is to express the h index as a function of N, the number of publications cited at least once. Then, since the primary interest of this work is the prediction of the value of the h index (and not to fit the whole citation distribution), we decided to exclude uncited papers from the analysis. Indeed, by definition, the derivation of the h-index does not depend on these publications.

For an interesting theoretical justification of the proposed geometric distribution, the reader is referred to Burrell, 2007, 2013a, 2014). Besides, this model can also be "formally" motivated by arguing that this distribution is nothing but the discrete version of the logarithmic transformation of the Pareto distribution of the first kind, $P(I)(1, \alpha)$. In particular, it is easy to see that the logarithmic transformation of a $P(I)(1, \alpha)$ is an exponential distribution. In symbols, under this assumption, the citation distribution function is

$$n(x) \propto e^{-\eta x}$$

Unlike the Pareto-type citation models, the exponential random variable has finite moments of all orders for every value of its parameter. The n versus x plot on a semilog scale approximates a straight line of slope $-\eta$, since $\log n = a - \eta x$; thus semilog plots can be easily used to check this model. By substituting $\theta = e^{-\eta}$, we can equivalently write $n(x) \propto \theta^x$. In its discrete version, the model can be regarded as a special case of a *power series distribution* (PSD, Johnson et al., 2005). Membership of the class confers a number of special properties. A PSD follows the probability mass function of the type $c^{-1}a_x\theta^x$, for $x = 0, 1, 2, \ldots$, where $a_x \geq 0$, θ ($\theta > 0$) is the so-called *power parameter*, and $c = \sum_{i=0}^{\infty} a_i\theta^i$ is the *series function*. Then, the geometric probability function $c^{-1}q^x$, $x = 1, 2, \ldots$, is an instance of a PSD, with q as power parameter, $a_0 = 0$, $a_x = 1$ for every $x = 1, 2, \ldots$, and $c = q/p$, $p = 1 - q$. The distribution is simply qualified by a straight line $\log n = a + bx$, where a represents the *logit* of p, $\log\left(p/q\right)$, and $b = \log q$, when plotting $\log n$ as a function of the number of citations x (semilog plot).

As can be seen, the citation distribution (10) has two parameters, one for normalization (N), and one that characterizes the shape of the citation distribution. The parameter p, or, equivalently, its complement to one: that is, the power parameter $q = 1 - p$, quantifies the 'fatness' of the tail; the smaller the value of p (the higher the value of q), the fatter the tail. The expectation is $\mu = 1/p$. The role of p can also be interpreted in the light of the level of concentration of the citations (in few papers).

3.2. A formula for the h-index

The assumption of geometrically distributed data enables estimates to be made of the expected theoretical value of h. Now, the value $N\sum_{x=1}^{k} q^{x-1}p = N\left(1 - q^k\right)$ provides an estimate of the number of papers with a number of citations less than

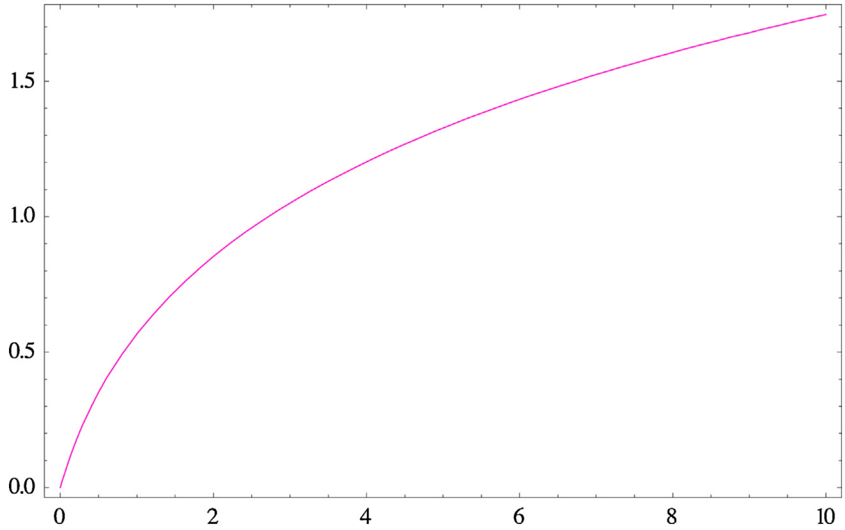

Fig. 1. The Lambert *W* function for values of its argument in the range (0,10).

or equal to k (i.e. the number of papers receiving at most k citations). Then, the complementary cumulative distribution function

$$R(k) = N - N\left(1 - q^k\right) = Nq^k$$

provides an estimate of the number of papers receiving at least $k+1$ citations. Hence, the h-index is determined by the equality

$$R(h-1) = h$$

As mentioned above, this equation was firstly proposed by Burrell (2013a) (but without giving an explicit solution), with the only slight difference that he considered a non-shifted version of the geometric distribution.

This equation can be solved as follows. First of all, recall that the *Lambert W* function (Wolfram Research Inc., 2013; see Fig. 1) is the inverse function $w(y)$ of the function

$$y = we^w.$$

The equation $R(h-1) = h$ is equivalent to $q^k = kN^{-1} + N^{-1}$, where $h = k + 1$. By substituting in the above equation $k = -t-1$, we obtain

$$tq^t = -Nq^{-1},$$

that is equivalent

$$(\log q)\, t \exp(t \log q) = -(\log q)\, Nq^{-1}.$$

Then, by substituting in the above equation $z = t\log q$, we obtain

$$ze^z = -(\log q)\, Nq^{-1}.$$

Hence, by definition, we have $z = W\left(-(\log q)\, Nq^{-1}\right)$, which yields the final solution

$$h = k + 1 = -t = -\frac{1}{\log q} W\left(-(\log q)\, Nq^{-1}\right) \tag{11}$$

To illustrate, in Fig. 2a and bwe represent the value of that solution $W\left(q^{-1}N \cdot \log\left(q^{-1}\right)\right) / \log\left(q^{-1}\right)$ as a function of q and N. As can be seen, even if N grows, this does not imply that h increases. Indeed, the number of publications should increase for at least an equal value of the mean of the number of citations. Note that similar graphs have been obtained by Bletsas and Sahalos (2009), but adopting other models, i.e. Tsallis and Weibull distributions.

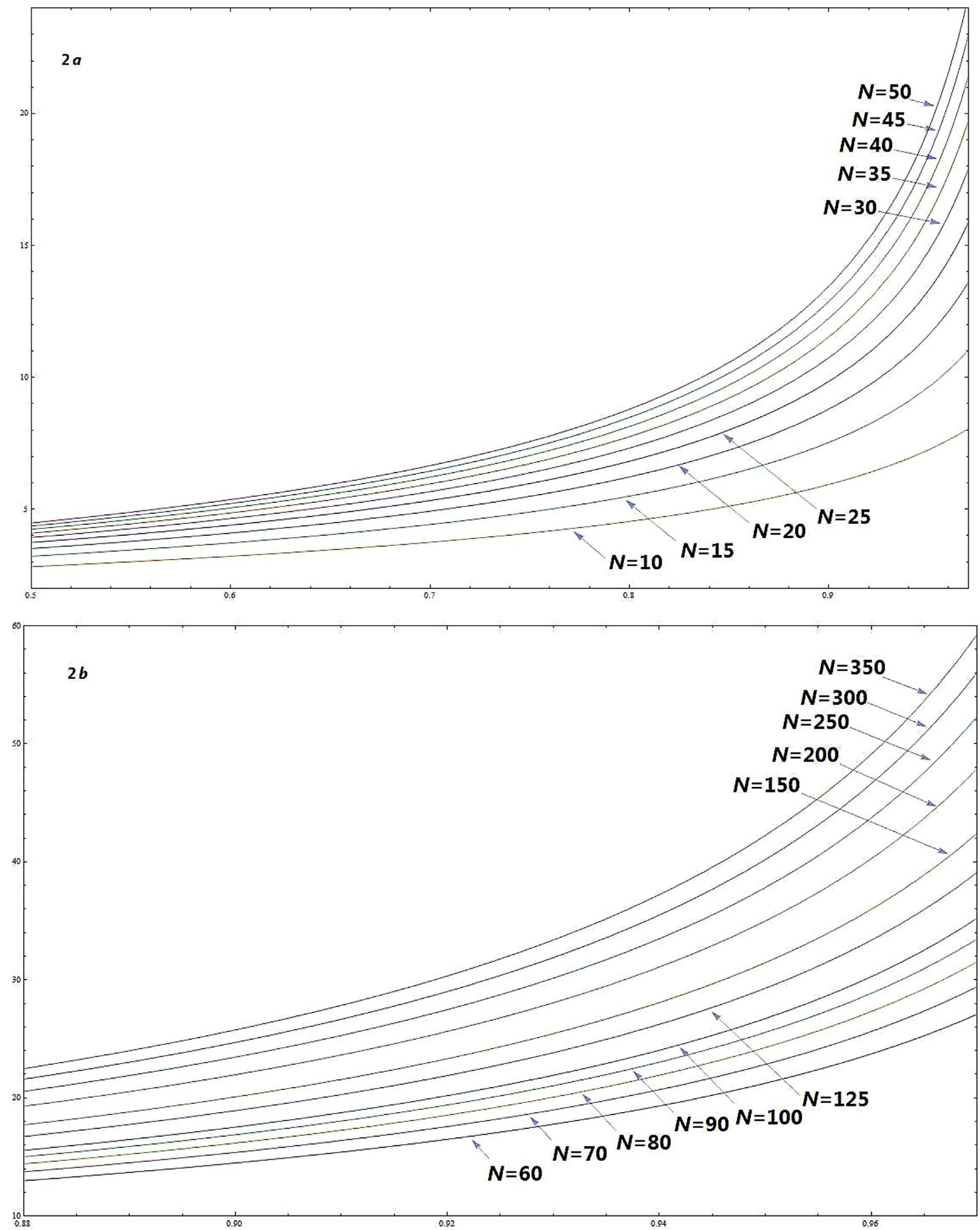

Fig. 2. The different curves represent the theoretical h-index as a function of the power parameter $q = 1 - N/C$ for different values of N (number of publications cited at least once): for N ranging from 10 to 50, in steps of 5 (a), and for N equal to 60, 70, 80, 90, 100, 125, 150, 200, 250, 300 and 350 (b). For fixed N, h_W is limited from above by N.

For the reader's information and convenience, in Appendix A the function $W(y)$ is briefly tabulated for values of y from 0.5 to 10, in steps of 0.05. Since the h-index is modeled as a non-negative integer, we will take the integer part of that solution, which we shall denote with h_w,

$$h_W = \left[\frac{W\left(q^{-1}N \cdot \log\left(q^{-1}\right)\right)}{\log\left(q^{-1}\right)} \right] \tag{12}$$

where $[z]$ denotes the integer part of z. In passing, note that, as q tends to 1 for fixed N, h_w tends to N (from below). Indeed, we have

$$\lim_{q\to 1} h_W = \lim_{q\to 1} \frac{W(Ny)}{y}$$

which is an indeterminate form of the type 0/0. But, since $W'(0) = 1$, applying the De l'Hopital's theorem we find

$$\lim_{y\to 0} \frac{W(Ny)}{y} = \lim_{z\to 0} \frac{W(z)}{z} N = N \lim_{z\to 0} \frac{W'(z)}{1} = N$$

Then since, for fixed N, h_w is an increasing function of q, it is always limited from above by N.

3.3. A formula for the estimation of h_w

It is important to distinguish between h_w and its empirical counterpart, i.e. its estimate. Estimation of the parameter of the geometric distribution is particularly straightforward. Because it is a PSD, the maximum likelihood estimation and the method of moments (by considering the first order moment equation) lead to the same estimate, for this random variable. It is easy to see that $\hat{p} = 1/m$. Then, a simple estimate of h_w is obtainable by substituting, in its expression, the unknown parameter q with its maximum likelihood estimate (MLE), $\hat{q} = 1 - m^{-1}$, where $m = C/N$. We then obtain the formula:

$$\hat{h}_W = \left[W \frac{\left(\left(CN/(C-N)\right) \cdot \log\left(C/(C-N)\right)\right)}{\log\left(C/(C-N)\right)} \right] \tag{13}$$

(Remember that, because of the invariance property of the MLE, the MLE of $h_W = h_W(q)$ is $\hat{h}_W = h_W\left(\hat{q}\right)$, where $\hat{q}$ is the MLE of q.)

The problem of how a single or few outliers can disproportionately inflate the statistic C is well known (Hirsch, 2005). Due to the highly skewed nature of the typical distribution of citations, it is often the case that the presence of individual highly cited papers tends to *overestimate* C, and consequently h_w, in comparison to the h-index (h is notoriously insensitive to a single 'big hit', outstandingly highly cited, paper). Elsewhere the term 'king effect' has been coined (Laherrère, 1996; Laherrère & Sornette, 1998; Malacarne, Mendes, & Lenzi, 2001) to indicate the case of a *single* high-value outlier—that is, the 'record value'. From a bibliometric point of view, the informative role of the most cited paper is controversial; for instance, according to Anderegg et al. (2010), "a single, highly cited paper does not establish a highly credible reputation but might instead reflect the controversial nature of that paper (often called the single-paper effect)". In conclusion, to contrast the tendency of $m = C/N$ to give an estimate almost systematically biased upward, this value should be conservatively substituted by a *trimmed* mean:

$$\tilde{m} = \frac{\tilde{C}}{(N-1)}$$

where we write $\tilde{C} = C - c_{MAX}$ for short. This trimmed mean is calculated by averaging all but the largest observation, c_{MAX}. The same adjustment was proposed by Burrell (2013a, pp. 779–780), but only for the most extreme outliers (chosen ex-post, on subjective basis). Differently, in our formula we include c_{MAX} as a "systematic" bias-reducing adjustment term. Then, our final formula reads as follows:

$$\tilde{h}_W = \tilde{h}_W(N, C, c_{MAX}) = \left[\frac{W\left(\left((N-1)/\left(1-\tilde{m}^{-1}\right)\right) \cdot \log\left(1/\left(1-\tilde{m}^{-1}\right)\right)\right)}{\log\left(1/\left(1-\tilde{m}^{-1}\right)\right)} \right] \tag{14}$$

Note that we can equivalently write $\tilde{h}_W = \tilde{h}_W(N, m, c_{MAX})$. This means that $\tilde{h}_W$ can also be interpreted as a function of quantity/productivity, here represented by N, and quality/impact, represented by m (Prathap, 2014). Moreover, note the 'subtractive' role in our formula of c_{MAX}, which is only used here to reduce the upward bias induced by c_{MAX} itself on the estimate of the 'true' mean m. Technically, our formula $\tilde{h}_W$ can then be interpreted as a trimmed MLE of h_w; more precisely, a MLE with a bias-reducing adjustment.

4. A case study

4.1. Sample database

In this section, we describe a case study that we carried out to evaluate empirically the adequacy of the fit of the proposed formula $\tilde{h}_W$ to the 'real' h-index, h, calculated from the full publication list of an author. For this case study we used a databese containing the publications of applicants to the so called "Abilitazione Scientifica Nazionale" (ASN), a nation-wide evaluation based on scientific qualification criteria for the recruitment of academic staff in Italy. These data were also considered elsewhere for a comparative study concerning 13 different bibliometric indices (Lando & Bertoli-Barsotti, 2014). The ASN involved tens of thousands of candidates. Here we focus on its first edition, year 2012 (for candidates, the deadline for applications was November 20, 2012), so-called ASN 2012. The evaluation relied completely on applicants' research productivity (and it did not require any personal interaction between evaluators and candidates).

For our study, we considered a cohort of 131 physicists (from the original sample of 149 applicants, 18 scientists were discarded from the analyses due to insufficient citation data – e.g. an h-index less than 2 – or difficulties in identifying the single scientist) who were applicants in the ASN 2012 for a full professorship. The whole sample can be considered as highly homogeneous, in that information regarding individual publications were collected from a single well-defined area within Physics, i.e. Condensed Matter Physics, and all candidates had a similar level of scientific maturity and similar academic qualifications. The publication and citation data were retrieved from Scopus, in January 2014.

4.2. Statistical analysis

Prior to their applications to the ASN, the applicants had published a total of T = 13,347 papers (in scholarly refereed journals), N = 11,079 of which cited at least once. The total number of citations was C = 288,972. We did not remove self-citations. The average percentage of uncited paper was 17%. Table 1 includes selected summary statistics from our database. We identify authors through a progressive number according to the alphabetical order (names not reported), in the first column. The following columns show respectively: the total number of publications, T; the total number of publications cited at least once, N; the total number of citations, C; the citation count of the most cited publication, c_{MAX}; the percentage of citations of the most cited publication, $c_{\text{MAX}}\% = \left(c_{\text{MAX}}/C\right) 100\%$; the mean of the number of citations per publication, m; the trimmed mean of the number of citations per publication, $\tilde{m}$; the trimmed MLE of the power parameter, $\tilde{q}$; the Hirsch index, h; the trimmed MLE of h_w; the absolute error, $\text{AE} = \left|\tilde{h}_W - h\right|$; the absolute relative error, $\text{ARE} = \left|\tilde{h}_W - h\right|/h$. As can be seen, the publications (cited at least once) received an average of 25 citations each (median = 23). The applicants' h-index values were on average 21.6 (median = 22), and ranged from a minimum of 2 to a maximum of 53. We found an average h-index of 21.6. The maximum observed value for h was 53. In contrast, only 13% of the scientists had an h-index smaller than 10. The average percentage of citations of the most cited publication was 16%. 77% of the authors received at least 1000 citations, and approximately 44% of the authors had at least 100 publications cited at least once. The most prolific author published 405 papers.

To study closeness of the estimated values to the exact ones, we computed the percentage errors in the theoretical values of h-index, $\tilde{h}_W$, as given by formula (5), with respect to the exact values of h. More precisely, a comparison between $\tilde{h}_W$ and the 'true' value of h was performed by computing the AE and the ARE. To characterize the overall quality of the results, the mean (mean absolute error, MAE, and mean absolute relative error, MARE) and the quartiles of these two types of errors were also computed. We found a very good fit, provided that, for the whole sample of 131 researchers considered, the MARE resulted less than 0.09, and the median of the ARE was 0.056. The observed median of the AE was equal to 1. More precisely, approximately two-thirds (63%) of all researchers have an absolute error AE not greater than 1, and about three-quarters (77%) of all researchers had an AE not greater than 2. As one can see from Table 1 the MAE was less than 2 (MAE = 1.92).

The precision of the approximation seems to be slightly related to the average number of citations per publication. As a general rule, we may say that the approximation works particularly well when the mean C/N (or, equivalently, the concentration) is not extremely high. Indeed, the MARE was equal to 0.172 when $C/N > 30$ (33 cases), and it was equal to 0.056 when $C/N < 30$ (98 cases). The MARE value grew to 0.233 when $C/N > 40$ (15 cases) versus a value of 0.066 for the case of $C/N < 40$ (116 cases).

It should also be noted that, as expected, high levels of C/N seemed to be related to high levels of $c_{\text{MAX}}\%$. Indeed, for the subset of scientists with $C/N < 30$, we found $c_{\text{MAX}}\% = 13.2$: that is, 13.2% of all the citations were concentrated in the single most cited paper, while for the subset of scientists with $C/N < 30$, we found that $c_{MAX}\%$ grew to 24.2%. From this, we can indirectly deduce that the geometric distribution is probably less suited to highly concentrated citation patterns. Fig. 3 illustrates the effect of different levels of the mean of the number of citations per publication (in its trimmed version, $\tilde{m}$) on the MARE. Note that for 80% of the researchers the MARE is less than 0.1. Also, we can see a lack of fit as m grows very large.

From a comparative point of view, the new formula appeared to be by far the most accurate among the different alternative formulas considered for h, in this case study. Indeed, the Pearson correlation coefficient (r) between the h-index and $T^{(m_0-1)/(2m_0-1)}$, $\sqrt{T}$, $cm_0{}^{2/3}T^{1/3}$ (that is, equivalently, the so called 'p-index') and $\tilde{h}_W$, resulted in r = 0.86, r = 0.79, r = 0.84 and r = 0.97 (see Fig. 4), respectively.

Table 1
Basic statistics for the sample of applicants: T = total number of papers; N = total number of papers cited at least once; C = total number of citations; c_{MAX} = citation count of the most cited paper; $c_{MAX}\% = \left(c_{MAX}/C\right)100\%$; m = mean of the number of citations per paper; $\tilde{m}$ = trimmed version of m; $\tilde{q}$ = trimmed estimate of q; h = h-index; $\tilde{h}_W$ = trimmed MLE of (4); AE absolute error and ARE = absolute relative error.

#	T	N	C	c_{MAX}	$c_{MAX}\%$	m	$\tilde{m}$	$\tilde{q}$	h	$\tilde{h}_W$	AE	ARE
1	80	63	1770	176	9.9	28.1	25.7	0.961	25	24	1	0.040
2	145	141	3803	184	4.8	27.0	25.9	0.961	34	35	1	0.029
3	10	9	129	37	28.7	14.3	11.5	0.913	6	5	1	0.167
4	120	107	1716	82	4.8	16.0	15.4	0.935	22	23	1	0.045
5	91	83	1535	197	12.8	18.5	16.3	0.939	20	21	1	0.050
6	24	19	550	152	27.6	28.9	22.1	0.955	10	11	1	0.100
7	80	57	1020	138	13.5	17.9	15.8	0.937	17	18	1	0.059
8	86	71	1427	131	9.2	20.1	18.5	0.946	22	22	0	0.000
9	101	74	1538	196	12.7	20.8	18.4	0.946	22	22	0	0.000
10	405	328	4309	330	7.7	13.1	12.2	0.918	31	29	2	0.065
11	138	116	2740	170	6.2	23.6	22.3	0.955	30	30	0	0.000
12	130	114	3056	213	7.0	26.8	25.2	0.960	27	32	5	0.185
13	11	9	87	23	26.4	9.7	8.0	0.875	5	5	0	0.000
14	16	12	75	14	18.7	6.3	5.5	0.820	5	5	0	0.000
15	92	82	1925	318	16.5	23.5	19.8	0.950	24	24	0	0.000
16	148	124	2753	106	3.9	22.2	21.5	0.954	28	30	2	0.071
17	183	147	7165	2706	37.8	48.7	30.5	0.967	30	40	10	0.333
18	49	38	236	31	13.1	6.2	5.5	0.820	8	8	0	0.000
19	113	98	2064	171	8.3	21.1	19.5	0.949	27	26	1	0.037
20	49	41	481	85	17.7	11.7	9.9	0.899	12	12	0	0.000
21	16	11	114	41	36.0	10.4	7.3	0.863	5	5	0	0.000
22	50	39	235	23	9.8	6.0	5.6	0.821	8	8	0	0.000
23	39	29	74	11	14.9	2.6	2.3	0.556	5	4	1	0.200
24	57	51	2816	1637	58.1	55.2	23.6	0.958	22	21	1	0.045
25	108	98	2452	296	12.1	25.0	22.2	0.955	25	28	3	0.120
26	154	117	1979	107	5.4	16.9	16.1	0.938	26	25	1	0.038
27	103	87	1851	129	7.0	21.3	20.0	0.950	27	25	2	0.074
28	31	23	298	43	14.4	13.0	11.6	0.914	9	10	1	0.111
29	96	85	1786	143	8.0	21.0	19.6	0.949	26	24	2	0.077
30	123	99	2645	191	7.2	26.7	25.0	0.960	30	30	0	0.000
31	113	101	1211	44	3.6	12.0	11.7	0.914	20	19	1	0.050
32	57	52	1913	335	17.5	36.8	30.9	0.968	22	24	2	0.091
33	64	56	1726	408	23.6	30.8	24.0	0.958	20	22	2	0.100
34	7	7	18	5	27.8	2.6	2.2	0.538	3	2	1	0.333
35	135	119	2819	302	10.7	23.7	21.3	0.953	29	29	0	0.000
36	59	51	648	99	15.3	12.7	11.0	0.909	13	14	1	0.077
37	94	77	1673	152	9.1	21.7	20.0	0.950	23	23	0	0.000
38	103	94	2468	206	8.3	26.3	24.3	0.959	24	29	5	0.208
39	75	73	5843	1820	31.1	80.0	55.9	0.982	29	37	8	0.276
40	28	17	43	8	18.6	2.5	2.2	0.543	3	3	0	0.000
41	96	76	1365	108	7.9	18.0	16.8	0.940	22	21	1	0.045
42	123	116	8147	1826	22.4	70.2	55.0	0.982	34	48	14	0.412
43	249	168	2617	179	6.8	15.6	14.6	0.932	26	26	0	0.000
44	118	101	3297	345	10.5	32.6	29.5	0.966	30	33	3	0.100
45	99	89	2091	131	6.3	23.5	22.3	0.955	26	26	0	0.000
46	62	46	741	103	13.9	16.1	14.2	0.929	16	15	1	0.063
47	66	63	2047	617	30.1	32.5	23.1	0.957	20	23	3	0.150
48	88	68	816	80	9.8	12.0	11.0	0.909	16	16	0	0.000
49	108	89	1494	155	10.4	16.8	15.2	0.934	20	21	1	0.050
50	93	64	2429	885	36.4	38.0	24.5	0.959	18	24	6	0.333
51	22	12	98	24	24.5	8.2	6.7	0.851	6	5	1	0.167
52	173	159	5537	804	14.5	34.8	30.0	0.967	34	40	6	0.176
53	67	60	1891	442	23.4	31.5	24.6	0.959	21	23	2	0.095
54	130	113	6342	1702	26.8	56.1	41.4	0.976	34	41	7	0.206
55	61	52	1403	235	16.7	27.0	22.9	0.956	19	21	2	0.105
56	156	133	2634	211	8.0	19.8	18.4	0.946	27	28	1	0.037
57	79	70	2179	240	11.0	31.1	28.1	0.964	23	27	4	0.174
58	104	85	1581	128	8.1	18.6	17.3	0.942	22	22	0	0.000
59	39	35	752	139	18.5	21.5	18.0	0.945	14	15	1	0.071
60	111	66	2342	244	10.4	35.5	32.3	0.969	25	28	3	0.120
61	52	46	1612	333	20.7	35.0	28.4	0.965	22	21	1	0.045
62	100	96	2619	168	6.4	27.3	25.8	0.961	29	30	1	0.034
63	65	53	5428	3068	56.5	102.4	45.4	0.978	27	28	1	0.037
64	174	141	3610	508	14.1	25.6	22.2	0.955	29	32	3	0.103
65	229	167	2278	224	9.8	13.6	12.4	0.919	22	24	2	0.091
66	118	100	2043	159	7.8	20.4	19.0	0.947	25	25	0	0.000
67	209	152	1251	70	5.6	8.2	7.8	0.872	18	17	1	0.056
68	162	128	2064	234	11.3	16.1	14.4	0.931	22	24	2	0.091

Table 1 (*Continued*)

#	T	N	C	c_{MAX}	$c_{MAX}\%$	m	$\tilde{m}$	$\tilde{q}$	h	$\tilde{h}_W$	AE	ARE
69	86	83	4823	1058	21.9	58.1	45.9	0.978	34	37	3	0.088
70	142	123	8126	2731	33.6	66.1	44.2	0.977	38	44	6	0.158
71	314	239	4002	154	3.8	16.7	16.2	0.938	30	32	2	0.067
72	112	95	3511	666	19.0	37.0	30.3	0.967	28	32	4	0.143
73	110	90	4319	582	13.5	48.0	42.0	0.976	28	37	9	0.321
74	79	66	2153	504	23.4	32.6	25.4	0.961	22	25	3	0.136
75	78	72	1159	116	10.0	16.1	14.7	0.932	19	19	0	0.000
76	162	134	2028	134	6.6	15.1	14.2	0.930	25	24	1	0.040
77	45	37	698	246	35.2	18.9	12.6	0.920	10	13	3	0.300
78	264	235	13916	2396	17.2	59.2	49.2	0.980	53	64	11	0.208
79	93	79	1156	87	7.5	14.6	13.7	0.927	19	19	0	0.000
80	91	82	2067	172	8.3	25.2	23.4	0.957	27	26	1	0.037
81	88	76	2771	465	16.8	36.5	30.7	0.967	27	29	2	0.074
82	91	80	1821	323	17.7	22.8	19.0	0.947	21	23	2	0.095
83	42	35	444	184	41.4	12.7	7.6	0.869	8	9	1	0.125
84	109	84	1381	87	6.3	16.4	15.6	0.936	20	21	1	0.050
85	106	98	4304	1142	26.5	43.9	32.6	0.969	27	34	7	0.259
86	152	141	3204	548	17.1	22.7	19.0	0.947	30	29	1	0.033
87	15	14	229	59	25.8	16.4	13.1	0.924	8	8	0	0.000
88	27	19	209	81	38.8	11.0	7.1	0.859	6	7	1	0.167
89	40	35	2509	882	35.2	71.7	47.9	0.979	15	22	7	0.467
90	104	77	1724	268	15.5	22.4	19.2	0.948	23	23	0	0.000
91	82	69	2355	391	16.6	34.1	28.9	0.965	22	27	5	0.227
92	261	215	3647	179	4.9	17.0	16.2	0.938	32	31	1	0.031
93	146	123	2210	155	7.0	18.0	16.8	0.941	25	26	1	0.040
94	103	73	948	91	9.6	13.0	11.9	0.916	17	17	0	0.000
95	9	5	50	40	80.0	10.0	2.5	0.600	2	2	0	0.000
96	66	63	1975	299	15.1	31.3	27.0	0.963	26	25	1	0.038
97	144	126	3157	302	9.6	25.1	22.8	0.956	30	31	1	0.033
98	111	92	2363	242	10.2	25.7	23.3	0.957	28	28	0	0.000
99	76	70	1589	129	8.1	22.7	21.2	0.953	23	23	0	0.000
100	80	70	1143	111	9.7	16.3	15.0	0.933	19	19	0	0.000
101	80	67	1264	139	11.0	18.9	17.0	0.941	20	20	0	0.000
102	67	59	1380	227	16.4	23.4	19.9	0.950	18	21	3	0.167
103	90	75	1750	194	11.1	23.3	21.0	0.952	20	24	4	0.200
104	108	79	1717	617	35.9	21.7	14.1	0.929	18	19	1	0.056
105	75	67	686	49	7.1	10.2	9.7	0.896	14	14	0	0.000
106	79	67	1362	122	9.0	20.3	18.8	0.947	20	21	1	0.050
107	16	10	399	254	63.7	39.9	16.1	0.938	6	6	0	0.000
108	79	69	1414	220	15.6	20.5	17.6	0.943	19	21	2	0.105
109	149	90	2088	177	8.5	23.2	21.5	0.953	25	26	1	0.040
110	147	135	2271	285	12.5	16.8	14.8	0.933	25	25	0	0.000
111	204	181	3431	150	4.4	19.0	18.2	0.945	31	31	0	0.000
112	108	98	1682	112	6.7	17.2	16.2	0.938	25	23	2	0.080
113	111	86	1211	67	5.5	14.1	13.5	0.926	19	19	0	0.000
114	91	61	755	59	7.8	12.4	11.6	0.914	15	15	0	0.000
115	87	82	1633	106	6.5	19.9	18.9	0.947	23	23	0	0.000
116	78	70	2801	394	14.1	40.0	34.9	0.971	23	30	7	0.304
117	42	37	1179	104	8.8	31.9	29.9	0.967	21	19	2	0.095
118	100	89	4429	683	15.4	49.8	42.6	0.977	29	37	8	0.276
119	146	107	1729	162	9.4	16.2	14.8	0.932	22	22	0	0.000
120	31	23	190	32	16.8	8.3	7.2	0.861	7	8	1	0.143
121	308	244	6302	899	14.3	25.8	22.2	0.955	38	40	2	0.053
122	59	49	1876	261	13.9	38.3	33.6	0.970	22	24	2	0.091
123	70	58	1234	79	6.4	21.3	20.3	0.951	23	21	2	0.087
124	87	80	1348	84	6.2	16.9	16.0	0.938	21	21	0	0.000
125	80	59	492	49	10.0	8.3	7.6	0.869	11	12	1	0.091
126	161	123	3323	242	7.3	27.0	25.3	0.960	32	33	1	0.031
127	79	61	1459	350	24.0	23.9	18.5	0.946	20	20	0	0.000
128	73	60	967	139	14.4	16.1	14.0	0.929	18	17	1	0.056
129	129	118	6105	775	12.7	51.7	45.6	0.978	40	44	4	0.100
130	155	133	3867	226	5.8	29.1	27.6	0.964	30	36	6	0.200
131	94	75	938	82	8.7	12.5	11.6	0.914	16	17	1	0.063
Mean	*101.9*	*84.6*	*2205.9*	*358.7*	*16.0%*	*25.0*	*20.6*	*0.93*	*21.6*	*23.1*	*1.9*	*0.09*
St dev	*62.9*	*51*	*1935*	*543*	*0.12*	*15.9*	*10.8*	*0.074*	*8.7*	*10.1*	*2.52*	*0.097*
Min	*7*	*5*	*18*	*5*	*3.6%*	*2.5*	*2.2*	*0.538*	*2*	*2*	*0*	*0.000*
Q1	*66*	*57.5*	*1157.5*	*105*	*8.0%*	*16.1*	*14.1*	*0.929*	*18*	*19*	*0*	*0.000*
Q2	*92*	*77*	*1786*	*177*	*12.5%*	*21.3*	*19.0*	*0.947*	*22*	*23*	*1*	*0.056*
Q3	*123*	*104*	*2692.5*	*326.5*	*18.5%*	*29.9*	*25.1*	*0.960*	*27*	*29*	*2*	*0.116*
Max	405	328	13916	3068	80.0%	102.4	55.9	0.982	53	64	14	0.467

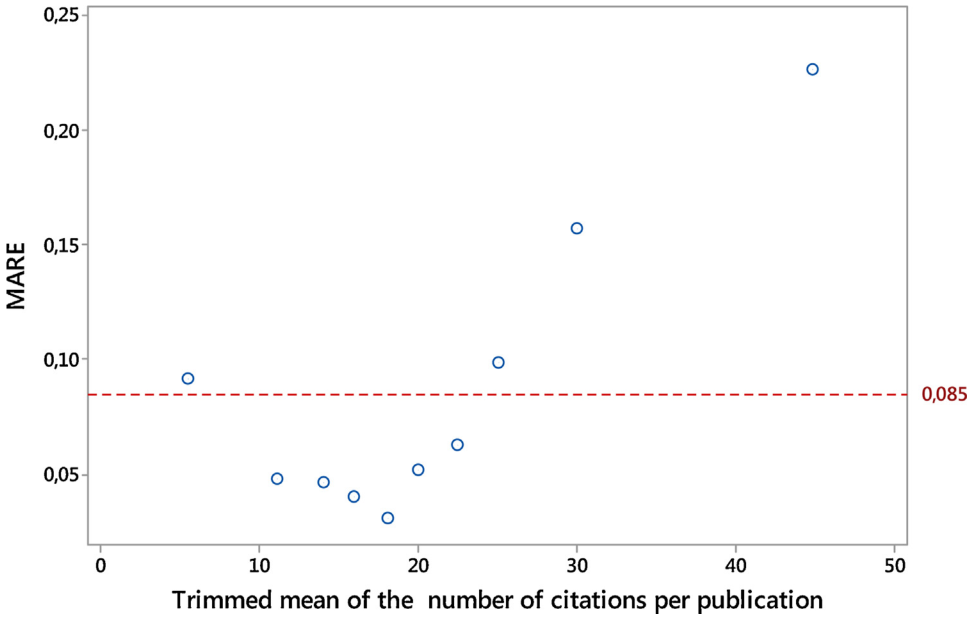

Fig. 3. Mean absolute error (MARE) of $\tilde{h}_W$ as a function of the mean of the number of citations per publication; each point refers to 13 applicants (14 cases for the first group). For 80% of the applicants the MARE is less than 0.1.

To illustrate the extent, in some cases, of the (problematic) 'king' effect, consider for example the dataset 24, with rank-citation profile: 1637, 111, 87, 85, 49, 49, 48, 42, 41, 40, 39, 36, 35, 34, 33, ... For this applicant, we find $N = 57$, $C = 2816$, $c_{MAX} = 1637$ and a very large value of $c_{MAX}\% = 58.1\%$. (The observed largest value of $c_{MAX}\%$ is 80%, and occurs for the dataset 95). Overall, note that the trimmed mean resulted, on average, in a 17.6 smaller than the original mean. Excluding the most cited publication, the h-index dropped by 8% on average.

Finally, to illustrate the dependence of h_w on the individual parameters N, C and c_{MAX}, let us consider four authors: #121(A), #25(B), #2(C) and #62(D) (see Fig. 5), who differ in their mean number of citations per publication and/or the number of publications (note that here we also take into account here the value of c_{MAX} by considering the mean number

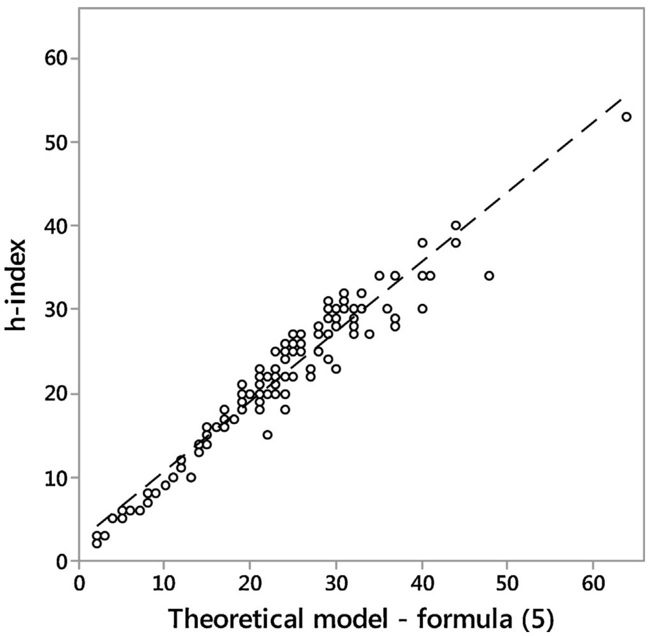

Fig. 4. Correlation of $\tilde{h}_W$ with the (true) h-index. Pearson correlation $r = 0.97$.

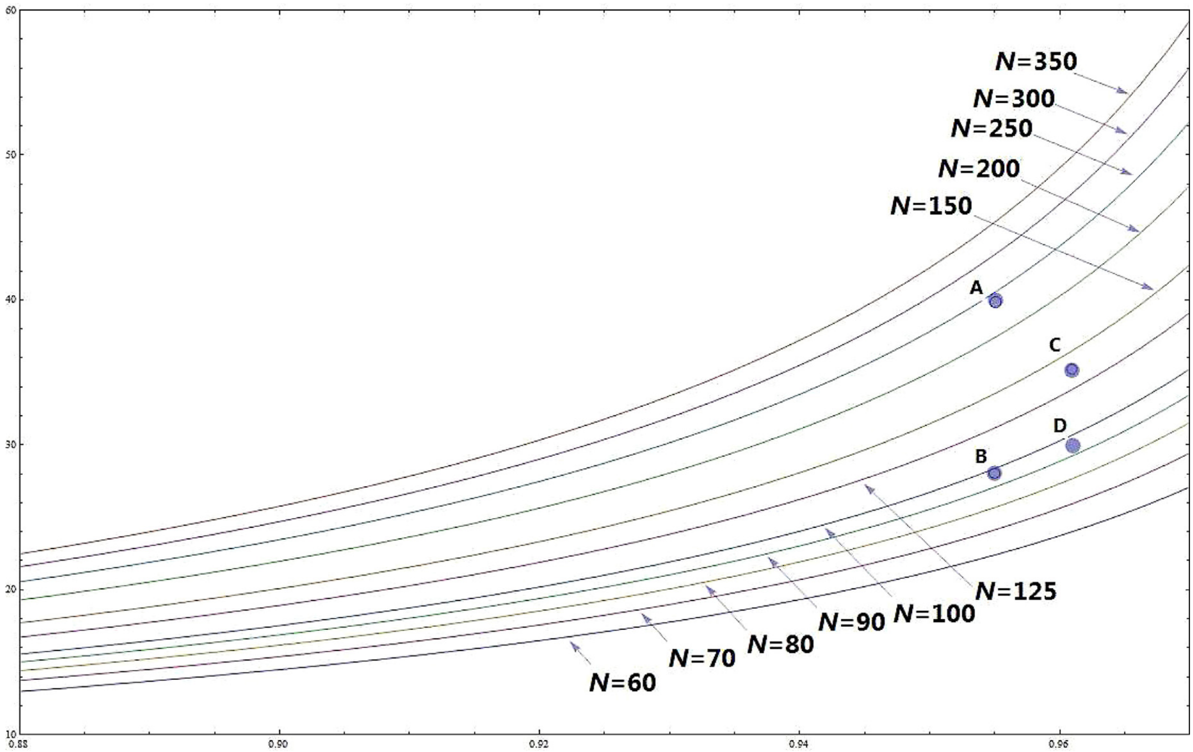

Fig. 5. Comparison between four applicants, #121(A), #25(B), #2(C) and #62(D), with similar levels of (trimmed) mean number of citations per publication and/or number of publications.

of citations per publication in its trimmed version, $\tilde{m}$). Researchers A and B have a similar value of $\tilde{m}$ (we find $\tilde{m}$ = 22.23 for both researchers, which corresponds to $\tilde{q}$ = 0.955), but a different number of publications, i.e. N = 244 for researcher A and N = 98 for researcher B. Then, formula $\tilde{h}_W$ produces a higher value for researcher A. Indeed, we find $\tilde{h}_W(\text{A}) = 40$ and $\tilde{h}_W(\text{B}) = 28$ (the observed values for the h-index are 38 and 25, respectively). Similarly, researchers C and D have a similar value of $\tilde{m}$ (one gets $\tilde{q} = 0.961$ for both researchers), but a different number of publications (N = 141 for researcher C and N = 96 for researcher D). Then formula $\tilde{h}_W$ produces a higher value for researcher C. Specifically, we find $\tilde{h}_W(\text{C}) = 35$ and $\tilde{h}_W(\text{D}) = 30$ (the observed values of the h-index are 34 and 29, respectively). Moreover, let us consider researchers B and D. They have a similar number of publication, 98 for researcher B and 96 for researcher D; but the latter has a higher level of $\tilde{m}$. Consequently, the corresponding level of $\tilde{h}_W$ is higher for researcher D. Finally, let us consider researchers C and A. The former presents a higher level of $\tilde{m}$, but a smaller number of publications. The formula $\tilde{h}_W$ states that the h-index should yield a higher value for A than for C, as indeed is actually observed. In other words, we can conclude that researchers with equal (or similar) *numbers of publications* are directly comparable – as regards the level of h – on the basis of the *mean number of citations per publication*, and *vice versa*. Moreover, increasing publications alone (or citations alone) does not have an immediate effect on the h-index, in general.

5. Conclusion

This paper has proposed a formula for the h-index which can be easily computed from three simple bibliometric indicators, namely N, C and c_{MAX}. More precisely, our formula describes the functional relationship between the h-index and the indicators: number of publications, N, and mean of an author's citations, C. The third factor, c_{MAX}, i.e. the number of citations received by the most cited paper, plays the role of a mere (but important) bias-reducing adjustment term. Indeed, the choice of a trimmed sample mean limits the bias induced by the 'big hit' problem. Our formula for the h-index involves two unknown parameters that can vary from author to author: one for normalization (N), and one that characterizes the shape of the author-specific citation distribution. This latter parameter is estimated by calculating a trimmed mean of the number of citations per publication.

To deduce our formula for the h-index, we temporarily assumed that citations follow a geometric law. To be noted is that this random variable can also be viewed as the discrete version of a special case of a Weibull (stretched exponential) distribution and also as a special case of negative binomial (Mingers & Burrell, 2006). Although the geometric distribution is perhaps too restrictive, in general, to be satisfactory as a model describing the citations over the *whole* range of the values (but, on the other hand, this was not the purpose of our study), it works well for representing the *center* of the citation distributions (while the Paretian models, instead, are well suited to the high citation end of the distribution), and this fact suffices to obtain an excellent proxy for the 'true' value of h, in the general case. To confirm this finding, in a case study we examined publication and citation data for a rather homogeneous cohort of 131 scientists. The preliminary results are encouraging: the overall fit (defined as the capacity of $\tilde{h}_W$ to reproduce the true value of h) was remarkably good, in that the predicted value $\tilde{h}_W$ was within one of the actual value h, for more than 60% of the datasets. The MARE was 0.09 for the whole sample of applicants. This value decreased to about 0.056 for those applicants with a mean number of citations per publication m not greater than 30 (as a general rule, the formula works particularly well for not very high levels of m). These findings confirm analogous positive results obtained by Burrell (2013a), on the basis of a study of the citation data sets of 15 scientists.

To conclude, owing to its dependence on a special function (the so-called Lambert W function), the presented formula $\tilde{h}_W$ is perhaps slightly less straightforward to compute, with respect to formulas such as those given by Eqs. (5), (6) and (8). Nevertheless, its computation is similarly simple, in that it needs only (the knowledge of) three standard bibliometric indicators, and its precision seems to be far better than that obtained with these alternative methods – at least in regard to the data in our analysis.

Acknowledgements

This paper has been supported in part by the Italian funds ex MURST 60% 2015 and in part through the Czech Science Foundation (GACR) under project 15-23699S, through SP2015/15, an SGS research project of VSB—TU Ostrava, and furthermore by the European Social Fund in the framework of CZ.1.07/2.3.00/20.0296.

Appendix A.

Lambert W function tabulated for values of its argument in the range from 0.5 to 10, in steps of 0.05. The reported values of the Lambert W function were computed using the command LambertW (or, equivalently, ProductLog) of the Mathematica® 9.0 software package (Wolfram Research Inc. (2012)).

y	$W(y)$	y	$W(y)$	y	$W(y)$	y	$W(y)$	y	$W(y)$	y	$W(y)$
0.50	0.352	2.10	0.875	3.70	1.160	5.30	1.360	6.90	1.516	8.50	1.643
0.55	0.377	2.15	0.886	3.75	1.167	5.35	1.366	6.95	1.520	8.55	1.647
0.60	0.402	2.20	0.897	3.80	1.174	5.40	1.371	7.00	1.524	8.60	1.651
0.65	0.425	2.25	0.908	3.85	1.181	5.45	1.376	7.05	1.529	8.65	1.654
0.70	0.448	2.30	0.918	3.90	1.188	5.50	1.382	7.10	1.533	8.70	1.658
0.75	0.469	2.35	0.929	3.95	1.195	5.55	1.387	7.15	1.537	8.75	1.661
0.80	0.490	2.40	0.939	4.00	1.202	5.60	1.392	7.20	1.541	8.80	1.665
0.85	0.510	2.45	0.949	4.05	1.209	5.65	1.397	7.25	1.546	8.85	1.669
0.90	0.530	2.50	0.959	4.10	1.216	5.70	1.402	7.30	1.550	8.90	1.672
0.95	0.549	2.55	0.968	4.15	1.222	5.75	1.407	7.35	1.554	8.95	1.676
1.00	0.567	2.60	0.978	4.20	1.229	5.80	1.413	7.40	1.558	9.00	1.679
1.05	0.585	2.65	0.987	4.25	1.236	5.85	1.418	7.45	1.562	9.05	1.683
1.10	0.602	2.70	0.997	4.30	1.242	5.90	1.423	7.50	1.566	9.10	1.686
1.15	0.619	2.75	1.006	4.35	1.248	5.95	1.428	7.55	1.570	9.15	1.689
1.20	0.636	2.80	1.015	4.40	1.255	6.00	1.432	7.60	1.574	9.20	1.693
1.25	0.652	2.85	1.024	4.45	1.261	6.05	1.437	7.65	1.578	9.25	1.696
1.30	0.667	2.90	1.033	4.50	1.267	6.10	1.442	7.70	1.582	9.30	1.700
1.35	0.682	2.95	1.041	4.55	1.273	6.15	1.447	7.75	1.586	9.35	1.703
1.40	0.697	3.00	1.050	4.60	1.280	6.20	1.452	7.80	1.590	9.40	1.706
1.45	0.712	3.05	1.058	4.65	1.286	6.25	1.457	7.85	1.594	9.45	1.710
1.50	0.726	3.10	1.067	4.70	1.292	6.30	1.461	7.90	1.598	9.50	1.713
1.55	0.740	3.15	1.075	4.75	1.298	6.35	1.466	7.95	1.602	9.55	1.716
1.60	0.753	3.20	1.083	4.80	1.304	6.40	1.471	8.00	1.606	9.60	1.720
1.65	0.767	3.25	1.091	4.85	1.309	6.45	1.475	8.05	1.610	9.65	1.723
1.70	0.780	3.30	1.099	4.90	1.315	6.50	1.480	8.10	1.614	9.70	1.726
1.75	0.792	3.35	1.107	4.95	1.321	6.55	1.484	8.15	1.617	9.75	1.730
1.80	0.805	3.40	1.115	5.00	1.327	6.60	1.489	8.20	1.621	9.80	1.733
1.85	0.817	3.45	1.123	5.05	1.332	6.65	1.494	8.25	1.625	9.85	1.736
1.90	0.829	3.50	1.130	5.10	1.338	6.70	1.498	8.30	1.629	9.90	1.739
1.95	0.841	3.55	1.138	5.15	1.344	6.75	1.502	8.35	1.632	9.95	1.742
2.00	0.853	3.60	1.145	5.20	1.349	6.80	1.507	8.40	1.636	10.00	1.746
2.05	0.864	3.65	1.153	5.25	1.355	6.85	1.511	8.45	1.640		

References

Abbas, A. M. (2012). Bounds and inequalities relating *h*-index, *g*-index, *e*-index and generalized impact factor: An improvement over existing models. *PLoS ONE*, *7*(4), e33699. http://dx.doi.org/10.1371/journal.pone.0033699

Adler, R., Ewing, J., & Taylor, P. (2009). Citation statistics. *Statistical Sciences*, *24*(1), 1–14.

Anastasiadis, A. D., deAlbuquerque, M. P., deAlbuquerque, M. P., & Mussi, D. B. (2010). Tsallis *q*-exponential describes the distribution of scientific citations—A new characterization of the impact. *Scientometrics*, *83*(1), 205–218.

Anderegg, W. R. L., Prall, J. W., Harold, J., & Schneider, S. H. (2010). Expert credibility in climate change. *Proceedings of the National Academy of Sciences*, *107*, 12107–12109.

Arnold, B. C. (1983). *Pareto distributions*. Fairland, MD: International Cooperative Publishing House.

Bertoli-Barsotti, L. (2013). Improving a decomposition of the *h*-index. *Journal of the American Society for Information Science and Technology*, *64*(7), 1522.

Bertoli-Barsotti, L., & Lando, T. (2015). A geometric model for the analysis of citation distributions. *International Journal of Mathematical Models and Methods in Applied Sciences*, *9*, 315–319.

Bletsas, A., & Sahalos, J. N. (2009). Hirsch index rankings require scaling and higher moment. *Journal of the American Society for Information Science and Technology*, *60*, 2577–2586.

Burrell, Q. L. (2007). Hirsch's *h*-index: A stochastic model. *Journal of Informetrics*, *1*(1), 16–25.

Burrell, Q. L. (2008). Extending Lotkaian informetrics. *Information Processing and Management*, *44*(5), 1794–1807.

Burrell, Q. L. (2013a). The *h*-index: A case of the tail wagging the dog? *Journal of Informetrics*, *7*, 774–783.

Burrell, Q. L. (2013b). Formulae for the *h*-index: A lack of robustness in Lotkaian informetrics? *Journal of the American Society for Information Science and Technology*, *64*(7), 1504–1514.

Burrell, Q. L. (2014). The individual author's publication-citation process: Theory and practice. *Scientometrics*, *98*(1), 725–742.

Campanario, J. M. (2010). Distribution of ranks of articles and citations in journals. *Journal of the Association for Information Science and Technology*, *61*(2), 419–423.

Csajbók, E., Berhidi, A., Vasas, L., & Schubert, A. (2007). Hirsch-index for countries based on essential science indicators data. *Scientometrics*, *73*, 91–117.

de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306.

Egghe, L. (2005a). *Power laws in the information production process: Lotkaian informetrics*. London: Academic Press.

Egghe, L. (2005b). Relations between the continuous and the discrete Lotka power function. *Journal of the American Society for Information Science and Technology*, *56*(7), 664–668.

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

Egghe, L., & Rousseau, R. (2012). The Hirsch index of a shifted Lotka function and its relation with the impact factor. *Journal of the American Society for Information Science and Technology*, *63*(5), 1048–1053.

Egghe, L., Guns, R., & Rousseau, R. (2011). Thoughts on uncitedness: Nobel laureates and fields medalists as case studies. *Journal of the American Society for Information Science and Technology*, *62*(8), 1637–1644.

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. (2007). Characteristic scores and scales. *Journal of Informetrics*, *1*, 92–102.

Glänzel, W. (2008). On some new bibliometric applications of statistics related to the *h*-index. *Scientometrics*, *77*(1), 187–196.

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.

Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences of the United States of America*, *104*(49), 19193–19198.

Iglesias, J., & Pecharroman, C. (2007). Scaling the *h*-index for different scientific ISI fields. *Scientometrics*, *73*(3), 303–320.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). (2nd ed.). *Continuous univariate distributions* (Vol. 1) New York, NY: Wiley.

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions* (3rd ed.). New York: Wiley.

Lafouge, T. (2007). The source-item coverage of the exponential function. *Journal of Informetrics*, *1*(1), 59–67.

Laherrère, J. H. (1996). Distributions de type fractal parabolique dans la Nature. *Comptes Rendus de l'Academie des Sciences*, *T.322*(Série IIa n. 7), 535–541.

Laherrère, J., & Sornette, D. (1998). Stretched exponential distributions in nature and economy: "Fat tails" with characteristic scales. *European Physical Journal B*, *2*(4), 525–539.

Lancho-Barrantes, B. S., Guerrero-Bote, V. P., & Moya-Anegon, F. (2010). The iceberg hypothesis revisited. *Scientometrics*, *85*(2), 443–461.

Lando, T., & Bertoli-Barsotti, L. (2014). A new bibliometric index based on the shape of the citation distribution. *PLoS ONE*, *9*(12), e115962. http://dx.doi.org/10.1371/journal.pone.0115962

Malacarne, L. C., Mendes, R. S., & Lenzi, E. K. (2001). *q*-Exponential distribution in urban agglomeration. *Physical Review E*, *65*, 017106.

Malesios, C. (2015). Some variations on the standard theoretical models for the *h*-index: A comparative analysis. *Journal of the Association for Information Science and Technology*, http://dx.doi.org/10.1002/asi.23410

Mansilla, R., Köppen, E., Cocho, G., & Miramontes, P. (2007). On the behavior of journal impact factor rank-order distribution. *Journal of Informetrics*, *1*, 155–160.

Martinez-Mekler, G., Martinez, R. A., del Rio, M. B., Mansilla, R., Miramontes, P., & Cocho, G. (2009). Universality of rank-ordering distributions in the arts and sciences. *PLoS ONE*, *4*, e4791.

Mingers, J., & Burrell, Q. L. (2006). Modeling citation behavior in Management Science journals. *Information Processing and Management*, *42*, 1451–1464.

Nicholls, P. T. (1987). Estimation of Zipf parameters. *Journal of the American Society for Information Science*, *38*, 443–445.

Perc, M. (2010). Zipf's law and log-normal distributions in measures of scientific output across fields and institutions: 40 years of Slovenia's research as an example. *Journal of Informetrics*, *4*, 358–364.

Petersen, A. M., Stanley, H. E., & Succi, S. (2011). Statistical regularities in the rank-citation profile of scientists. *Scientific Reports*, *1*, 181. http://dx.doi.org/10.1038/srep00181

Prathap, G. (2010a). Is there a place for a mock *h*-index? *Scientometrics*, *84*, 153–165.

Prathap, G. (2010b). The 100 most prolific economists using the *p*-index. *Scientometrics*, *84*, 167–172.

Prathap, G. (2014). The Zynergy-index and the formula for the *h*-index. *Journal of the Association for Information Science and Technology*, *65*(2), 426–427.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B*, *4*, 131–134.

Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today*, *58*(1), 49–54.

Rousseau, B., & Rousseau, R. (2000). LOTKA: A program to fit a power law distribution to observed frequency data. *Cybermetrics*, *4*(1.).

Schubert, A., & Glänzel, W. (2007). A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics*, *1*, 179–184.

Schubert, A., Korn, A., & Telcs, A. (2009). Hirsch-type indices for characterizing networks. *Scientometrics*, *78*(2), 375–382.

Shalizi, R. C. (2007). *Maximum likelihood estimation for q-exponential (Tsallis) distributions*. (ar***X***iv:math/0701854v2).

Stringer, M. J., Sales-Pardo, M., & Amaral, L. A. N. (2008). Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE*, *3*(2), e1683.

Tsallis, C. (1988). Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics*, *52*, 479.

Tsallis, C., & de Albuquerque, M. P. (2000). Are citations of scientific papers a case of nonextensivity? *European Physical Journal B*, *13*(4), 777–780.

van Raan, A. F. J. (2001). Two-step competition process leads to quasi power law income distribution. Application to scientific publication and citation distribution. *Physica A*, *298*, 530–536.

van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, *67*(3), 491–502.

Vinkler, P. (2009). The π-index: A new indicator for assessing scientific impact. *Journal of Information Science*, *35*, 602–612.

Vinkler, P. (2010). The π-index: A new indicator to characterize the impact of journals. *Scientometrics*, *82*(3), 461–475.

Wallace, M. L., Larivière, V., & Gingras, Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics*, *3*(4), 296–303.

Weibull, W. (1951). A statistical distribution of wide applicability. *Journal of Applied Mechanics*, *18*, 293–297.

Wolfram Research Inc. (2013). *Wolfram Mathematica*. Documentation Center. ⟨http://reference.wolfram.com/language/ref/ProductLog.html⟩ (ProductLog).

Wolfram Research Inc. (2012). *Mathematica, Version 9.0*. Champaign, IL: Wolfram Research, Inc.

Ye, F. Y. (2009). An investigation on mathematical models of the *h*-index. *Scientometrics*, *81*, 493–498.

Ye, F. Y. (2011). A unification of three models for the *h*-index. *Journal of the American Society for Information Science and Technology*, *62*(1), 205–207.