Journal of Informetrics 6 (2012) 80–87

Contents lists available at SciVerse ScienceDirect

Journal of Informetrics

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

The stochastic *h*-index

Gopalan M. Nair[a], Berwin A. Turlach[a,b,*]

[a] School of Mathematics and Statistics (M019), The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia
[b] Centre for Applied Statistics (M019), The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia

ARTICLE INFO

Article history:
Received 7 June 2011
Received in revised form
26 September 2011
Accepted 27 September 2011

Keywords:
h-Index
Stochastic *h*-index
Rational *h*-index
Real *h*-index
Bibliometric indicators
Stochastic model

ABSTRACT

A variant of the *h*-index, named the stochastic *h*-index, is proposed. This new index is obtained by adding to the *h*-index the probability, under a specific stochastic model, that the *h*-index will increase by one or more within a given time interval. The stochastic *h*-index thus extends the *h*-index to the real line and has a direct interpretation as the distance to the next higher index value. We show how the stochastic *h*-index can be evaluated and compare it with other variants of the *h*-index which purportedly indicate the distance to a higher *h*-index.

1. Introduction

Hirsch (2005) introduced the so-called *h*-index to measure the impact of a researcher's publications. A researcher has an *h*-index of h_0 if h_0 of his or her papers are each cited at least h_0 times and the remaining papers are each cited h_0 or less times. Since its proposal, the *h*-index has been intensively studied and various modifications have been suggested; for recent reviews see, among others, Rousseau (2008), Alonso, Cabrerizo, Herrera-Viedma, and Herrera (2009) and Egghe (2010).

By definition, the *h*-index is an integer and hence has limited discriminative power to distinguish between researchers whose career are at a similar stage. It also conveys no information about the ease with which a researcher could move from an *h*-index of h_0 to $h_0 + 1$ (or larger) within a given time interval.

To overcome these issues, we propose in this paper the stochastic *h*-index h_s. For a researcher with *h*-index h_0, the stochastic *h*-index h_s is defined as $h_s = h_0 + \mathfrak{p}$ where $\mathfrak{p}$ is the probability that, based on current publications, the *h*-index of the researcher increases to $h_0 + 1$ or more within one time unit (typically a year). Thus, the stochastic index h_s does not give an indication what the researcher's *h*-index would be at the end of the time unit. Rather, it adds to the *h*-index the chance, $\mathfrak{p}$, that the researcher has an *h*-index of $h_0 + 1$ or more at the end of the year, which can be used to differentiate researchers by their (short term) potential as researchers in addition to the *h*-index's ability to differentiate the long term potential of researchers (Hirsch, 2007). Details of our proposal are given in Section 2, where we also show how h_s can be determined under the stochastic model used by Burrell (2007b).

Our proposal is by no means the first attempt to extend the *h*-index beyond integer values by adding a fractional component that can be interpreted as measuring the distance to the next higher index value. Ruane and Tol (2008) proposed the rational *h*-index h_{rat} to give the *h*-index a finer structure. For a researcher with *h*-index h_0 the rational *h*-index is

* Corresponding author at: School of Mathematics and Statistics (M019), The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia.
E-mail addresses: Gopalan.Nair@uwa.edu.au (G.M. Nair), Berwin.Turlach@gmail.com (B.A. Turlach).

1751-1577/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2011.09.006

$h_{\text{rat}} = h_0 + 1 - m/(2h_0 + 1)$, where m is the minimum number of citations that the researcher needs to achieve an h-index of $h_0 + 1$, i.e. if future citations happen according to the best-possible scenario for this researcher. The proposal is based on the fact that the worst-possible citation pattern for a researcher with h-index h_0 is to have h_0 papers, each cited h_0 times and all remaining papers cited not at all. In this case, in the best possible scenario, $2h_0 + 1$ additional citations would lead to an h-index of $h_0 + 1$. Thus, $2h_0 + 1$ is an upper bound for the minimum number of additional citations that a researcher needs to increase her or his h-index by 1 and this motivates the denominator in the fraction used to define h_{rat}.

An alternative approach is given by the real h-index h_{r} proposed by Rousseau (2006). To describe h_{r} and to fix notation, assume a researcher has N publications in total and publication i has n_i citations with $n_1 \geq n_2 \geq \cdots \geq n_N$. Rousseau (2006) proposed to connect the points (i, n_i) by linear interpolation and to define h_{r} as the point on the abscissa where this linear interpolant intersects the 45° line. More details on h_{r}, and a discussion on how it relates to h_{rat}, are given in Guns and Rousseau (2009) who show that for a researcher with h-index h_0 the real h-index is

$$h_{\text{r}} = \frac{(h_0 + 1)n_{h_0} - h_0 n_{h_0+1}}{1 - n_{h_0+1} + n_{h_0}}$$

.

Finally, it should be noted that the r_p and l_p indices, $1 \leq p \leq \infty$, proposed by Gągolewski and Grzegorzewski (2009) are indices that may have non-integer values. But, it is unclear how these indices relate to the h-index, except for the r_∞-index which is equal to the h-index. And some variations of the h-index (see, among others, Alonso et al., 2009) also lead to indices with non-integer values. As most, if not all, of these indices are not constructed to encode in their fractional part the distance to the next higher h-index value we will not discuss them further.

In Section 2 we define the stochastic h-index h_{s}. From the definition it will be clear that it can be interpreted as a distance to the next higher h-index. Section 3 illustrates how the stochastic h-index h_{s} can easily be evaluated on publications and citations data that is readily available, and compares this index with the rational h-index h_{rat} and the real h-index h_{r} on real and fictitious citation data. While the latter two indices also extend the h-index beyond integer values, we show that some issues arise with their interpretation if one attempts to interpret them as distance to the next higher h-index. Some concluding comments are offered in Section 4.

2. The stochastic *h*-index

In order to define the stochastic h-index, we first have to choose a stochastic model for the publication/citation process. In this paper, our choice is to use the stochastic model considered by Burrell (2007b) which is a well established informetric way of modelling the publication/citation process. In fact, we do not need the full model considered by Burrell (2007b) but shall only assume the following.

Assumptions (Stochastic model): Any particular publication i acquires citations according to a Poisson process of rate λ_i and the Poisson processes are independent of each other. Here, λ_i denotes the mean citations per unit time following publication, called the *citation rate*.

This is essentially assumption 2 of the stochastic model used by Burrell (2007b) except for the independence assumption. However, although not explicitly stated it is clear from the calculations in Burrell (2007b) that he also assumes that the citation processes are independent of each other. It should be acknowledged that the assumption of a constant citation rate for each paper is rather strong. Many of the points that are raised in Section 4 (concluding remarks) of Burrell (2007b) naturally apply to our model too. Thus, while our stochastic model might be a bit simplistic, the assumption of constant citation rates allows us to define and to evaluate the stochastic index h_{s} (see Sections 2.1 and 3.1) based on cumulative citation counts only. We are now ready to define the stochastic h-index h_{s}.

Definition of h_{s}: Consider a researcher with h-index h_0. The stochastic h-index of this researcher is defined as $h_0 + \mathfrak{p}$, where $\mathfrak{p}$ is the probability that, based on current publications, the h-index of the researcher increases to $h_0 + 1$ or more within one time unit (typically a year).

In practice, it will be easier in most cases to calculate the stochastic h-index h_{s} as $h_0 + 1 - \mathfrak{q}$ where $\mathfrak{q}$ is the probability that, based on current publications, the h-index of the researcher is still h_0 after one time unit. To show how this probability can be calculated, we need to introduce some further notation.

For a researcher with h-index h_0, let L_0 denote the number of papers with $h_0 + 1$ or more citations and let $\mathcal{L}_1 = \{L_0 + 1, \ldots, N\}$ denote the set of indices for the papers with at most h_0 citations. Finally, set $K = h_0 - L_0$. Observe that, based on the current N publications, the h-index of the researcher will increase to $h_0 + 1$ or more if and only if the number of publications, with index in $\mathcal{L}_1$, that acquire a citation count of $h_0 + 1$ or more is at least $K + 1$. Thus, the stochastic h-index h_{s} can be calculated as

$$h_{\text{s}} = h_0 + \mathfrak{p} = h_0 + 1 - \mathfrak{q} = h_0 + 1 - \sum_{i=0}^{K} \mathfrak{q}_i$$

where $\mathfrak{q}_i$ is the probability that exactly i publications, with index in $\mathcal{L}_1$, acquire a citation count of at least $h_0 + 1$ within the next time unit.

To determine the $\mathfrak{q}_i$ note that under our stochastic model the distribution of the number of citations M_i of publication i during the next time unit follows a Poisson distribution with parameter λ_i, and that these citation counts are independent of each other. Thus, for $i \in \mathcal{L}_1$, let $m_i = h_0 - n_i$ denote the number of additional citations that publication i needs to reach a citation count of h_0, and let $p_i = P(M_i > m_i)$ denote the probability that this citation count is exceeded within the next time unit. Finally, set $q_i = 1 - p_i = P(M_i \leq m_i)$. Then

$$\begin{aligned}
\mathfrak{q}_0 &= \prod_{i=L_0+1}^{N} q_i = \prod_{i \in \mathcal{L}_1} q_i \\
\mathfrak{q}_1 &= \sum_{i=L_0+1}^{N} \left(p_i \prod_{\substack{j \in \mathcal{L}_1 \\ j \neq i}} q_j \right) \\
\mathfrak{q}_2 &= \sum_{i=L_0+1}^{N-1} \sum_{j=i+1}^{N} \left(p_i p_j \prod_{\substack{k \in \mathcal{L}_1 \\ k \neq i, k \neq j}} q_k \right) \\
\mathfrak{q}_3 &= \sum_{i=L_0+1}^{N-2} \sum_{j=i+1}^{N-1} \sum_{k=j+1}^{N} \left(p_i p_j p_k \prod_{\substack{l \in \mathcal{L}_1 \\ l \neq i, l \neq j, l \neq k}} q_l \right)
\end{aligned}$$

and so forth.

2.1. Evaluating the stochastic h-index

To evaluate the stochastic h-index h_s, we only need an estimate for the citation rate of each paper in the set $\mathcal{L}_1$. From this we can obtain estimated probabilities $\hat{p}_i$ and $\hat{q}_i$, then estimates for the $\mathfrak{q}_i$'s and, finally, for $\mathfrak{q}$. If t_{0i} denotes the year in which publication i was published and t_c the year for which the stochastic h-index should be calculated (2010 in the following), then we propose to estimate λ_i by the observed citation rate $\hat{\lambda}_i = n_i/(t_c - t_{0i} + 1)$.

While, under the assumption of a constant citation rate, it is intuitively appealing to use the observed citation rate of each paper as an estimate for its citation rate, an obvious problem is that papers are published throughout the year. Hence, $t_c - t_{0i} + 1$ is presumably in most cases an overestimate for the number of time units in which publication i could acquire citations; hence $\hat{\lambda}_i$ will typically underestimate the citation rate λ_i, especially for relatively recent publications. We will return to this point during one of the examples discussed in the next section. On the other hand, we agree with a point made by one of the reviewers that for older papers that might no longer be cited, or with a lower frequency than in the early years after their publication, the observed citation rate might be an overestimate of the actual citation rate.

Another issue that needs to be addressed when using the observed citation rates is how publications with a citation count of $n_i = 0$ should be handled. Clearly an estimate of $\hat{\lambda}_i = 0$ is undesirable as it would imply that publication i will have with probability one a citation count of zero during the next time unit. One way of defining $\hat{\lambda}_i$ when $n_i = 0$ would be to use the results of Burrell (2002) to construct an appropriate estimator.

Alternatively, one could use a standard result from probability theory (see, among others, Casella & Berger, 2002; Mukhopadhyay, 2006) on how to update an estimate for the success probability in a sequence of Bernoulli trials, i.e. experiments in which the only possible outcomes are "Success" and "Failure", as the results of these trials become known one by one. Initially, if we assume that any value between zero and one is equally likely to be the success probability, we would estimate before the first trial the success probability to be 1/2. If the first trial results in a success, then we would update our estimate for the success probability from 1/2 to 2/3 and to 1/3, otherwise. Assume that the first trial resulted in a success, then if the second trial results in a success we would update our estimate for the success probability from 2/3 to 3/4 and to 2/4 = 1/2, otherwise. Likewise, if the first trial resulted in a failure, then if the second trial results in a success, we would update our estimate for the success probability from 1/3 to 2/4 = 1/2 and to 1/4, otherwise. In general, after k trials we would estimate our success probability to be (*number of observed successes* + 1)/(k + 2). In particular, if all the trials resulted in a success, then our estimate for the success probability would be $(k+1)/(k+2)$. We propose to use this result to calibrate the citation rate for a paper that has not been cited yet. That is, if we regard whether or not a paper is cited in any give time unit as a Bernoulli trial with, somewhat awkwardly, the outcome of acquiring no citations being viewed as "Success" in terms of the previous discussion, then if $n_i = 0$ after $k = t_c - t_{0i} + 1$ time units, we use as estimated citation rate $\hat{\lambda}_i = -\log((t_c - t_{0i} + 2)/(t_c - t_{0i} + 3))$ so that the probability of this paper acquiring no citations in the next time interval is $P(M_i = 0) = (t_c - t_{0i} + 2)/(t_c - t_{0i} + 3)$.

As a reviewer pointed out, it is natural to require that a publication with $n_i = 1$ always has a higher estimated citation rate than a publication with $n_i = 0$ (when both publications appeared in the same year). With our choice for the estimated

citation rates of papers with no citations, this desirable property is indeed guaranteed. Note that the (natural) logarithm is a concave function and, hence, $x \geq \log(1+x)$ for all $x > -1$. In fact, equality holds only for $x = 0$ and for all other $x > -1$ strict inequality holds. Thus it follows that

$$-\log\left(\frac{t_c - t_{0i} + 2}{t_c - t_{0i} + 3}\right) = \log\left(\frac{t_c - t_{0i} + 3}{t_c - t_{0i} + 2}\right) = \log\left(1 + \frac{1}{t_c - t_{0i} + 2}\right) < \frac{1}{t_c - t_{0i} + 2} < \frac{1}{t_c - t_{0i} + 1},$$

where the quantity on the left would be the estimated citation rate of the paper if $n_i = 0$, and the quantity on the right would be the estimated citation rate if $n_i = 1$. In the examples used in the following section, see Tables A.1–A.4, this property can also be observed.

3. Case studies and discussion

In this section we shall first illustrate how the variants of the h-index considered here, h_{rat}, h_{r} and h_{s}, can be evaluated on readily available publications and citations data and then discuss some properties of these indices. The data used in this section were collected from the Scopus data base (http://www.scopus.com) and are the publications and citations data of some early- to mid-career statisticians who are working, or have worked, at one of the Australian Go8 universities.[1] In each case we restricted ourself to data to the end of 2010.

3.1. Evaluating the indices

The data for our first example, researcher A, are given in Table A.1. In this example the various indices considered here are easy to calculate. The h-index at the end of 2010 of this researcher is 10. In the best case scenario, one additional citation, namely a further citation of publication 11, will result in an h-index of 11. Hence, the rational h-index h_{rat} of this researcher is $h_{\text{rat}} = 11 - 1/21 \approx 10.952$. The real h-index h_{r} is $h_{\text{r}} = 10.5$. Finally, in this case $L_0 = 10$, $K = 0$ and thus the stochastic h-index h_{s} is $h_{\text{s}} = 11 - \mathfrak{q}_0$. Given the estimated citation rates for the various papers this evaluates to $h_{\text{s}} \approx 10.828$. At the time of writing, this researcher still has an h-index of 10.

For researcher B, whose data are given in Table A.2, the calculations are slightly more involved. In this case the h-index of the researcher at the end of 2010 is 6, $L_0 = 5$ and, hence, $K = 1$. Thus, the stochastic h-index h_{s} of this researcher is $h_{\text{s}} = 7 - \mathfrak{q}_0 - \mathfrak{q}_1$. In this case, $\mathfrak{q}_0$ and $\mathfrak{q}_1$ are estimated to be 0.005576 and 0.07085, respectively. This leads to a stochastic h-index of $h_{\text{s}} \approx 6.924$. Note that this researcher would need, in the best case scenario, two additional citations to reach an h-index of 7. Thus the rational h-index is $h_{\text{rat}} = 7 - 2/13 \approx 6.846$ and the real h-index is $h_{\text{r}} = 6$. At the time of writing, this researcher has an h-index of 8.

In most cases K will be either zero or one, but occasionally larger values are encountered. An example is researcher C, whose data are given in Table A.3. For this researcher the h-index is 11 at the end of 2010, $L_0 = 8$ and, hence, $K = 3$. Thus, the stochastic h-index h_{s} of this researcher is $h_{\text{s}} = 12 - \mathfrak{q}_0 - \mathfrak{q}_1 - \mathfrak{q}_2 - \mathfrak{q}_3$. For this researcher, $\mathfrak{q}_0$, $\mathfrak{q}_1$, $\mathfrak{q}_2$ and $\mathfrak{q}_3$ are estimated to be, respectively, 0.00245, 0.0463, 0.2725 and 0.4921, leading to a stochastic h-index of $h_{\text{s}} \approx 11.187$. Note that this researcher would need, in the best case scenario, eight additional citations to reach an h-index of 12. Thus the rational h-index is $h_{\text{rat}} = 12 - 8/23 \approx 11.652$ and the real h-index is $h_{\text{r}} = 11$. At the time of writing, this researcher has an h-index of 11.

Our final example illustrates the problem of getting good estimates for the citation rates λ_i of papers that have only recently been published. For researcher D, whose data are given in Table A.4, the h-index at the end of 2010 is 11. Here $L_0 = 11$ and $K = 0$, thus the stochastic h-index h_{s} of this researcher is $h_{\text{s}} = 12 - \mathfrak{q}_0 \approx 11.796$. Note that this researcher needs, in the best case scenario, only one more citation to achieve an h-index of 12. Hence the rational h-index is $h_{\text{rat}} = 12 - 1/23 \approx 11.957$. Finally, the real h-index for this researcher is $h_{\text{r}} = 11.8$. At the time of writing, this researcher has an h-index of 13. Looking at the current citation counts in more detail, one notices that publication 12 from 2003 has now 13 citations, which might not be surprising. The other citation that jumped to 13 citations, and pushed the h-index to 13, is publication 17 from 2009; given its estimated citation rate of 2.5 publications per year it is somewhat surprising that this publication acquired 8 citations in half a year. However, looking more closely at the citation pattern it becomes apparent that this publication was published sometime in 2009 and received no citation in that year and all 5 citations were acquired in 2010. Thus, it is clear that $\hat{\lambda}_{17} = 2.5$ seriously underestimates the citation rate of this paper which in turn depresses the stochastic h-index h_{s}. This example demonstrates that there is some argument to change the denominator in the definition of $\hat{\lambda}_i$ to $t_c - t_{0i} + 0.5$ or $\max(t_c - t_{0i}, 1)$, but ultimately one will always have problems to obtain reliable and good estimates for the citation rate λ_i of a recently published paper due to the limited amount of data available.

3.2. Observations on the indices

Given the examples in Section 3.1, some comments are warranted on whether the various variants of the h-index considered in this paper do have an interpretation as measuring the distance to the next index value.

[1] The Group of Eight (Go8) is a coalition of Australian universities. More details are available at http://www.go8.edu.au/.

First, note that by its definition, the rational h-index h_{rat} is still essentially discrete. For researchers with an h-index of h_0, it can be viewed as adding $2h_0$ sub-categories to allow further discrimination between these researchers, but its discriminatory powers are still limited, especially for small values of h_0. This lack of discriminatory power can be illustrated by considering some fictitious examples. Assume researcher E has h_0 publications with a citation count of h_0+1 or more, one publication with citation count of h_0 and all other publications have a publication count of zero. Contrast this with researcher F, who has h_0 publications with a citation count of h_0+1 or more, and all other publications have a citation count of h_0. Both these researchers will have an h-index of h_0 and a rational h-index of $h_{\text{rat}} = h_0 + 1 - 1/(2h_0+1)$ as they need only one more citation to reach an h-index of h_0+1. However, researcher E needs that *one particular* publication (the one with publication count h_0) acquires this additional citation, while in researcher F case the additional citation necessary to reach an h-index of h_0+1 could be acquired by *any* of his publications with a citation count of h_0. The rational h-index h_{rat} is not able to distinguish between these two researchers while the stochastic h-index h_s would be larger for researcher F (if for both researchers the publications with h_0 citations or less were published in the same time span).

Secondly, the rational h-index h_{rat} is based on future citations being acquired according to a best-case scenario for reaching the next higher index value, and in that sense it does measure the distance to the next index value. However, in practice a citation pattern according to this best-case scenario will most likely not eventuate and, thus, the rational h-index h_{rat} gives an over optimistic indication of the researcher's chance to reach the next index value. This is illustrated by the example of researcher C in Section 3.1 whose rational h-index $h_r \approx 11.652$ seems to paint an over optimistic picture.

Thirdly, we would argue that the real h-index h_r cannot be interpreted as measuring the distance to the next higher index value. As noted by Guns and Rousseau (2009), for a researcher with h-index h_0 and $n_{h_0} = h_0$ the real h-index is $h_r = h_0$. This is illustrated by the example of researcher B in Section 3.1 where the real h-index fails to indicate the ease with which this researcher was able to acquire a higher h-index, while h_{rat} and h_s give more realistic judgements of the researcher's proximity to higher h-index values.

What information does the real h-index h_r encode? Assume a researcher has an h-index of h_0, and write $n_{h_0} = h_0 + m_e$ ($0 \le m_e$) and $n_{h_0+1} = h_0 - m_d + 1$ ($1 \le m_d \le h_0 + 1$). Then the real h-index of this researcher is

$$h_r = \frac{(h_0+1)n_{h_0} - h_0 n_{h_0+1}}{1 - n_{h_0+1} + n_{h_0}} = h_0 + \frac{m_e}{m_e + m_d} \tag{1}$$

Thus, it is clear that if $m_e = 0$ then $h_r = h_0$, no matter what the value of m_d is. Furthermore, depending on m_e being larger, smaller or equal to m_d the real h-index h_r is, respectively, larger, smaller or equal to $h_0 + 0.5$. In particular, $h_r = h_0 + 0.5$ as long as $m_e = m_d$. However, it is arguable that as $m_e = m_d$ increases, the researcher would be further away from the next higher h-index value. Both other indices considered here do recognise this situation. If $m_e = m_d \ge 1$, then the rational h-index is $h_{\text{rat}} = h_0 + 1 - m_d/(2h_0+1)$ which decreases as m_d increases. Note that in this situation we have, necessarily, $L_0 = h_0$, $K = 0$ and the stochastic index would be $h_s = h_0 + 1 - q_0$. Now, holding all other data constant, as m_d increases $m_{h_0+1} = h_0 - n_{h_0+1} = m_d - 1$ increases too and, hence p_{h_0+1} would decrease, and q_{h_0+1} and q_0 would increase, leading to a decrease in h_s.

As another extreme example, take a researcher who has $h_0 - 1$ publications with M or more citations, one publication with M citations, and all his other publications have zero citations. If $M \ge h_0$, then the h-index of this researcher is h_0 and the real h-index is $h_r = h_0 + (M - h_0)/(M+1)$, since $m_e = M - h_0$ and $m_d = h_0 + 1$. Note that, if the citation pattern of this researchers continues, i.e. his top-ranked h_0 papers acquire further citations and all his other publications do not acquire any citations, then M will increase and so will his real h-index h_r. In fact, with increasing M the real h-index h_r can get arbitrarily close to $h_0 + 1$, even though the researcher is "far away" from reaching an h-index of $h_0 + 1$. The rational h-index h_{rat} in this situation would be stuck on $h_{\text{rat}} = h_0 + 1 - (h_0+1)/(2h_0+1) = h_0 + h_0/(2h_0+1)$, which does not depend on M and is slightly less than $h_0 + 0.5$ as soon as h_0 is moderately large. Thus, the rational h-index h_{rat} gives a realistic, albeit perhaps optimistic, judgement on the distance of the researcher's h index to $h_0 + 1$. Given the numeric examples in Tables A.1–A.4, it is clear that for publications with a citation count of zero, and from quite modest value of h_0 onwards, the probabilities p_i and q_i that flow into the calculation of the stochastic h-index h_s are, for all practical purposes, zero and one, respectively. Hence, in this example the stochastic h-index h_s would be h_0; at most with a very small probability added. The stochastic h-index h_s correctly recognises that if the citation pattern of this researcher continues, then the researcher has very little chance to reach an h-index of $h_0 + 1$, or, in other words, that the researcher is far away from an h-index of $h_0 + 1$.

It is also clear from Eq. (1) that as long as citation counts are integers or rationals (e.g. fractional counts that take the number of co-authors into account), the real h-index h_r is a rational number and does not extend the h-index to the real line. However, the real h-index h_r adds many more sub-categories, and hence a finer structure, to the h-index than the rational h-index h_{rat}. However, for a researcher with h-index h_0, this finer structure is not used to measure her or his distance to an index value of $h_0 + 1$. Rather, the real h-index encodes the relationship between the number of excess citations m_e of the publication with rank h_0 and the number of citations m_d missing from the publication with rank $h_0 + 1$ to get this publication to a citation count of $h_0 + 1$.

Finally, from the examples and discussion above, it is clear that the rational h-index h_{rat} summarises in its fractional part some information about the $h_0 + 1$ most cited papers of the researchers, namely how many citations some of these publications have to acquire additionally to lift the researcher's h-index from h_0 to $h_0 + 1$. As just shown, the real h-index h_r encodes in its fractional part some information about the number of citations that the publications ranked h_0 and $h_0 + 1$ have acquired. Thus, the additional information about the publications/citations distribution that is made available by these

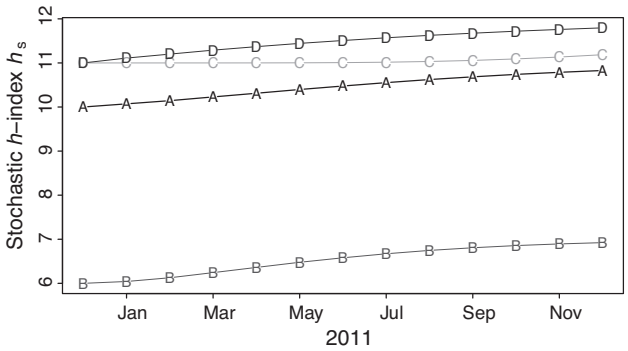

Fig. 1. Development of the stochastic h-index h_s over 2010 for the four researchers considered in Section 3.1. The graph shows the stochastic h-index of each researcher at the end of each month.

indices is somewhat limited. By way of contrast, the stochastic h-index h_s encodes in its fractional part information about the tail behaviour of the publications/citations distribution of the researcher. Specifically, it encodes information on the $N - L_0$ publications that all have a citation count of h_0 or less by calculating, under a certain stochastic model, the probability that more than $K = h_0 - L_0$ of these publications acquire a citation count of $h_0 + 1$ within the next time unit.

3.3. Evaluating for fractional time units

Our assumption that any particular publication i acquires citations according to a Poisson process of rate λ_i implies that the number of new citations within a future time unit is Poisson distributed with parameter λ_i. It also implies that the number of new citations in c future time units, where c might be a fraction, is Poisson distributed with parameter $c\lambda_i$. Thus, it is possible to calculate how the stochastic h-index h_s develops over time. The calculations are exactly the same as those described in Section 3.1 and laid out in Tables A.1–A.4 except that $c\hat{\lambda}_i$ is used to calculate $\hat{p}_i$ and $\hat{q}_i$. Fig. 1 shows the result if the stochastic h-index h_s is calculated for the end of each month in 2010 for the four researchers considered in Section 3.1. Alternative approaches for looking at short term changes in the h-index are discussed in Rousseau and Ye (2008) and Burrell (2009); the long term development of the h index is discussed in Hirsch (2005) and Burrell (2007a, 2007b).

4. Concluding remarks

This paper proposes the stochastic h-index h_s, which truly extends the h-index to the real line, and we demonstrated how this index can easily be calculated in practice. Variations of this index are possible and subject of further research.

For example, as defined in this paper, the stochastic h-index h_s only takes current publications into account. One could use the model for the publications process of Burrell (2007b) to forecast the number of publications during the next time unit and include these in the calculation of the index. Though, given the results in Tables A.1–A.4 it is unlikely that such recent publications would materially influence the stochastic h-index h_s.

Other possible extensions of this index are to relax the assumptions of the stochastic model under which it is defined. That is to either remove the assumption that the citation counts are independent of each other or the assumption that the citation rate of each paper is constant, or both. Relaxation of these assumptions would be possible, and are under investigation, but then the calculations of the stochastic h-index h_s would become more complicated, requiring more detailed observations on the citation counts for each paper. For example, to relax the assumption of constant citation rates, arguably the strongest assumption of the stochastic model used here, one could use the citation counts of a paper for each year after its publication and some forecasting technique, e.g. exponential smoothing (see, among others, Hyndman, Koehler, Ord, & Snyder, 2008), to predict its citation count/rate for the following year.

By definition, the stochastic h-index h_s has an interpretation of how far a researcher's h-index is away from the next higher value under a specific stochastic model. It would be of interest to see if the idea behind the stochastic h-index can be extended to other indices such as the g-index Egghe (2006a, 2006b) or the w-index Woeginger (2008).

Acknowledgements

The authors would like to thank the editor and two reviewers for their valuable comments and constructive suggestions.

Appendix A.

This appendix shows the data used in Section 2 to illustrate the stochastic h-index h_s on readily available publications and citations data. The data were collected from the Scopus data base (http://www.scopus.com) and are the publications

Table A.1
Publications and citations pattern for researcher *A*: i denotes the rank of each publication, t_{0i} its year of publication, n_i the number of citations until the end of 2010, $\hat{\lambda}_i$ the (estimated) rate of citation per year, m_i the maximum number of additional citations allowed to stay at a citation count of h_0 or less, $\hat{p}_i$ is the (estimated) probability that the citation count for the next time unit exceeds m_i, and $\hat{q}_i = 1 - \hat{p}_i$. The emphasised line defines the h-index h_0 for this researcher.

i	t_{0i}	n_i	$\hat{\lambda}_i$	m_i	$\hat{p}_i$	$\hat{q}_i$	i	t_{0i}	n_i	$\hat{\lambda}_i$	m_i	$\hat{p}_i$	$\hat{q}_i$
1	2004	770	110.000				17	1999	5	0.417	5	0.000	1.000
2	2000	124	11.273				18	1996	4	0.267	6	0.000	1.000
3	2000	110	10.000				19	2006	3	0.600	7	0.000	1.000
4	2001	55	5.500				20	2005	3	0.500	7	0.000	1.000
5	2005	39	6.500				21	2002	3	0.333	7	0.000	1.000
6	1997	36	2.571				22	1999	3	0.250	7	0.000	1.000
7	2004	34	4.857				23	1997	3	0.214	7	0.000	1.000
8	1997	17	1.214				24	2008	2	0.667	8	0.000	1.000
9	1997	13	0.929				25	2004	2	0.286	8	0.000	1.000
10	*1996*	*11*	*0.733*				26	2009	1	0.500	9	0.000	1.000
11	1998	10	0.769	0	0.537	0.463	27	2007	1	0.250	9	0.000	1.000
12	2005	9	1.500	1	0.442	0.558	28	2004	1	0.143	9	0.000	1.000
13	1997	9	0.643	1	0.136	0.864	29	2010	0	0.405	10	0.000	1.000
14	1999	8	0.667	2	0.030	0.970	30	2007	0	0.182	10	0.000	1.000
15	2008	7	2.333	3	0.207	0.793	31	2000	0	0.080	10	0.000	1.000
16	1999	6	0.500	4	0.000	1.000							

Table A.2
Publications and citations pattern for researcher *B*, see Table A.1 for an explanation of the columns.

i	t_{0i}	n_i	$\hat{\lambda}_i$	m_i	$\hat{p}_i$	$\hat{q}_i$	i	t_{0i}	n_i	$\hat{\lambda}_i$	m_i	$\hat{p}_i$	$\hat{q}_i$
1	2004	41	5.857				13	2006	4	0.800	2	0.047	0.953
2	2002	25	2.778				14	2008	3	1.000	3	0.019	0.981
3	2004	16	2.286				15	2006	3	0.600	3	0.003	0.997
4	2008	12	4.000				16	2009	2	1.000	4	0.004	0.996
5	2006	10	2.000				17	2007	2	0.500	4	0.000	1.000
6	*2008*	*6*	*2.000*	*0*	*0.865*	*0.135*	18	2009	1	0.500	5	0.000	1.000
7	2007	6	1.500	0	0.777	0.223	19	2007	1	0.250	5	0.000	1.000
8	2006	6	1.200	0	0.699	0.301	20	2010	0	0.405	6	0.000	1.000
9	2008	4	1.333	2	0.151	0.849	21	2010	0	0.405	6	0.000	1.000
10	2007	4	1.000	2	0.080	0.920	22	2010	0	0.405	6	0.000	1.000
11	2007	4	1.000	2	0.080	0.920	23	2009	0	0.288	6	0.000	1.000
12	2007	4	1.000	2	0.080	0.920							

Table A.3
Publications and citations pattern for researcher *C*, see Table A.1 for an explanation of the columns.

i	t_{0i}	n_i	$\hat{\lambda}_i$	m_i	$\hat{p}_i$	$\hat{q}_i$	i	t_{0i}	n_i	$\hat{\lambda}_i$	m_i	$\hat{p}_i$	$\hat{q}_i$
1	2004	42	6.000				22	2006	3	0.600	8	0.000	1.000
2	1999	20	1.667				23	2005	3	0.500	8	0.000	1.000
3	2000	19	1.727				24	2010	2	2.000	9	0.000	1.000
4	2007	15	3.750				25	2008	2	0.667	9	0.000	1.000
5	2005	14	2.333				26	2007	2	0.500	9	0.000	1.000
6	2005	13	2.167				27	2006	2	0.400	9	0.000	1.000
7	2006	12	2.400				28	2009	1	0.500	10	0.000	1.000
8	2004	12	1.714				29	2008	1	0.333	10	0.000	1.000
9	2006	11	2.200	0	0.889	0.111	30	2007	1	0.250	10	0.000	1.000
10	2006	11	2.200	0	0.889	0.111	31	2005	1	0.167	10	0.000	1.000
11	*2002*	*11*	*1.222*	*0*	*0.705*	*0.295*	32	2005	1	0.167	10	0.000	1.000
12	2009	7	3.500	4	0.275	0.725	33	2010	0	0.405	11	0.000	1.000
13	2007	7	1.750	4	0.033	0.967	34	2010	0	0.405	11	0.000	1.000
14	2006	7	1.400	4	0.014	0.986	35	2009	0	0.288	11	0.000	1.000
15	2004	7	1.000	4	0.004	0.996	36	2008	0	0.223	11	0.000	1.000
16	2006	6	1.200	5	0.002	0.998	37	2007	0	0.182	11	0.000	1.000
17	2001	6	0.600	5	0.000	1.000	38	2007	0	0.182	11	0.000	1.000
18	2000	6	0.545	5	0.000	1.000	39	2006	0	0.154	11	0.000	1.000
19	1997	6	0.429	5	0.000	1.000	40	2006	0	0.154	11	0.000	1.000
20	2009	5	2.500	6	0.014	0.986	41	2005	0	0.134	11	0.000	1.000
21	2003	4	0.500	7	0.000	1.000	42	2004	0	0.118	11	0.000	1.000

Table A.4
Publications and citations pattern for researcher *D*, see Table A.1 for an explanation of the columns.

i	t_{0i}	n_i	$\hat{\lambda}_i$	m_i	$\hat{p}_i$	$\hat{q}_i$	i	t_{0i}	n_i	$\hat{\lambda}_i$	m_i	$\hat{p}_i$	$\hat{q}_i$
1	2006	248	49.600				13	2003	9	1.125	2	0.105	0.895
2	2005	122	20.333				14	2006	8	1.600	3	0.079	0.921
3	2002	97	10.778				15	1997	8	0.571	3	0.003	0.997
4	2005	88	14.667				16	2004	7	1.000	4	0.004	0.996
5	2003	52	6.500				17	2009	5	2.500	6	0.014	0.986
6	2000	37	3.364				18	2008	4	1.333	7	0.000	1.000
7	2005	34	5.667				19	2007	3	0.750	8	0.000	1.000
8	2004	29	4.143				20	2008	2	0.667	9	0.000	1.000
9	2005	22	3.667				21	2007	1	0.250	10	0.000	1.000
10	2007	21	5.250				22	2010	0	0.405	11	0.000	1.000
11	*2004*	*15*	*2.143*				23	2008	0	0.223	11	0.000	1.000
12	2003	11	1.375	0	0.747	0.253	24	2007	0	0.182	11	0.000	1.000

and citations data of some early- to mid-career statisticians who are working, or have worked, at one of the Australian Go8 universities. In each case we restricted ourself to data to the end of 2010 (Tables A.1–A.4).

References

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E. & Herrera, F. (2009). *h*-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, *3*(4), 273–289. doi:10.1016/j.joi.2009.04.001

Burrell, Q. L. (2002). Will this paper ever be cited. *Journal of the American Society for Information Science an Technology*, *53*(3), 232–235. doi:10.1002/asi.10031

Burrell, Q. L. (2007a). Hirsch index or Hirsch rate? Some thoughts arising from Liang's data. *Scientometrics*, *3*(1), 19–28. doi:10.1007/s11192-006-1774-5

Burrell, Q. L. (2007b). Hirsch's *h*-index: A stochastic model. *Journal of Informetrics*, *1*(1), 16–25. doi:10.1016/j.joi.2006.07.001

Burrell, Q. L. (2009). Some comments on "A proposal for a dynamic h-type index" by Rousseau and Ye. *Journal of the American Society for Information Science and Technology*, *60*(2), 418–419. doi:10.1002/asi.20969

Casella, G. & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Pacific Grove, USA: Duxbury.

Egghe, L. (2006a). An improvement of the *h*-index: The g-index. *ISSI Newsletter*, *2*(1), 8–9.

Egghe, L. (2006b). Theory and practice of the *g*-index. *Scientometrics*, *9*(1), 131–152. doi:10.1007/s11192-006-0144-7

Egghe, L. (2010). The Hirsch-index and related impact measures. *Annual Review of Information Science and Technology*, *44*, 65–114.

Gągolewski, M. & Grzegorzewski, P. (2009). A geometric approach to the construction of scientific impact indices. *Scientometrics*, *81*(3), 617–634. doi:10.1007/s11192-008-2253-y

Guns, R. & Rousseau, R. (2009). Real and rational variants of the *h*-index and the *g*-index. *Journal of Informetrics*, *3*(1), 64–71. doi:10.1016/j.joi.2008.11.004

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. doi:10.1073/pnas.0507655102

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. doi:10.1073/pnas.0707962104

Hyndman, R. J., Koehler, A. B., Ord, J. K. & Snyder, R. D. (2008). *Forecasting with exponential smoothing: The state space approach. Springer series in statistics*. Berlin: Springer-Verlag.

Mukhopadhyay, N. (2006). *Introductory statistical inference. Vol. 187 of STATISTICS: Textbooks and monographs*. Boca Raton, USA: Chapman & Hall/CRC.

Rousseau, R. (2006). Simple models and the corresponding *h*- and *g*-index, e-LIS: ID 6153. http://eprints.rclis.org/archive/00006153/.

Rousseau, R. (2008). Reflections on recent developments of the *h*-index and h-type indices. In H. Kretschmer, & F. Havemann (Eds.), *Proceedings of WIS 2008, Berlin: Fourth international conference on webometrics, informetrics and scientometrics & ninth COLLNET meeting.* Humboldt-Universität zu Berlin, Institute for Library Information Science (IBI), http://www.collnet.de/Berlin-2008/RousseauWIS2008rrd.pdf, (pp. 1–9).

Rousseau, R. & Ye, F. Y. (2008). A proposal for a dynamic h-type index. *Journal of the American Society for Information Science and Technology*, *59*(11), 1853–1855. doi:10.1002/asi.20890

Ruane, F. & Tol, R. S. J. (2008). Rational (successive) *h*-indices: An application to economics in the Republic of Ireland. *Scientometrics*, *75*(2), 395–405. doi:10.1007/s11192-007-1869-7

Woeginger, G. J. (2008). An axiomatic characterization of the Hirsh-index. *Mathematical Social Sciences*, *56*(2), 224–232. doi:10.1016/j.mathsocsci.2008.03.001