Please cite this paper as: Żogała-Siudem B., Cena A., Siudem G., Gagolewski M., Interpretable reparameterisations of citation models, Journal of Informetrics 17(1), 101355, 2023, DOI:10.1016/j.joi.2022.101355

Interpretable reparameterisations of citation models

Barbara Żogała-Siudem[a,*], Anna Cena[b], Grzegorz Siudem[c] and Marek Gagolewski[d]

[a]*Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warsaw, Poland*
[b]*Warsaw University of Technology, Faculty of Mathematics and Information Science, ul. Koszykowa 75, 00-662 Warsaw, Poland*
[c]*Warsaw University of Technology, Faculty of Physics, ul. Koszykowa 75, 00-662 Warsaw, Poland*
[d]*Deakin University, Data to Intelligence Research Centre, School of IT, Geelong, VIC 3220, Australia*

ARTICLE INFO

Keywords:
science of science
bibliometric indices
informetrics
citation models
interpretability

ABSTRACT

This paper aims to find the reasons why some citation models can predict a set of specific bibliometric indices extremely well. We show why fitting a model that preserves the total sum of a vector can be beneficial in the case of heavy-tailed data that are frequently observed in informetrics and similar disciplines. Based on this observation, we introduce the reparameterised versions of the discrete generalised beta distribution (DGBD) and power law models that preserve the total sum of elements in a citation vector and, as a byproduct, they enjoy much better predictive power when predicting many bibliometric indices as well as partial cumulative sums. This also results in the underlying model parameters' being easier to fit numerically. Moreover, they are also more interpretable. Namely, just like in our recently-introduced 3DSI (three dimensions of scientific impact) model, we have a clear distinction between the coefficients determining the total productivity (size), total impact (sum), and those that affect the shape of the resulting theoretical curve.

1. Introduction

Evaluation of scientific accomplishments is at the heart of bibliometrics, where citation indices are used as a proxy for the quality of one's research output. Even though the number of existing bibliometric indices is already very large (Wildgaard et al., 2014), with the modifications of the h-index (Hirsch, 2005) alone constituting a significant part thereof, many new measures are still being proposed every year (Poirrier et al., 2021; Li et al., 2021). This makes analysing the existing indices in various theoretical settings a relevant topic, as it may shed light on their advantages as well as limitations.

Citation models aim to express the whole citation record (e.g., of an author) through very few parameters. In (Cena et al., 2022) we have studied and compared a number of different approaches, including via the 3DSI (three dimensions of scientific impact; Siudem et al., 2020), DGBD (discrete generalised beta distribution; Petersen et al., 2011), power-law, and harmonic models. As a measure of the quality of fit, we have assumed that a model is practically useful, if it is able to reproduce a chosen set of key citation indices reasonably well.

We have observed that a model's flexibility in terms of how well it reproduces the individual elements in a citation vector *on average*, does not always imply an accurate reproduction of bibliometric measures. The reverse might not always be true, either. Models with more parameters can overfit to the part of the citation

*Corresponding author
Email addresses: zogala@ibspan.waw.pl (B. Żogała-Siudem); anna.cena@pw.edu.pl (A. Cena); grzegorz.siudem@pw.edu.pl (G. Siudem); m.gagolewski@deakin.edu.au (M. Gagolewski)
URL: http://cena.rexamine.com (A. Cena); http://if.pw.edu.pl/~siudem (G. Siudem); https://www.gagolewski.com (M. Gagolewski)
ORCID(s): 0000-0002-2869-7300 (B. Żogała-Siudem); 0000-0001-8697-5383 (A. Cena); 0000-0002-9391-6477 (G. Siudem); 0000-0003-0637-6028 (M. Gagolewski)

curve where there are papers with a small number of citations. As the tail of the rank-size distribution is not taken into consideration by most of the popular citation measures, this puts these models at a disadvantage.

This paper aims to find the reasons why some citation models can predict a set of specific bibliometric indices extremely well. One such measure is the g-index (Egghe, 2006), which the 3DSI model reproduced exactly for 94% of the DBLP citation vectors analysed in (Cena et al., 2022). Further, we will study this phenomenon in relation to some other popular indices, such as the r-index (Jin et al., 2007), a-index (Jin et al., 2007), o-index (Dorogovtsev and Mendes, 2015a), and hg-index (Alonso et al., 2010).

Following the line of research that notes the high degree of correlation between these bibliometric indices and some functions of the total number of citations (e.g., Schreiber, 2008; De Visscher, 2011), we argue that citation models should rather focus on reproducing the total sum of citations exactly. It turns out that due to the highly skewed nature of the citation distribution, our approach will allow for a quite accurate estimation of the number of citations received by the first few highly cited papers of a given author.

Thus, we introduce the versions of the DGBD and power law models that preserve the sums of the target vectors, which have more interpretable parameters and reproduce many citation indices better. Furthermore, they are easier to apply in practice, as the number of parameters that need to be fit numerically is smaller.

Here is the outline of the paper. Section 2 introduces the citation models, the bibliometric indices, and the dataset studied. Section 3 discusses how different partial sums of the top-cited items in a citation vector correlate with various indices. In Section 4, we introduce the reparametrisations of the DGBD and power law models that have more interpretable parameters and preserve the total citation count. Furthermore, we show that they usually lead to a significant improvement in the bibliometric indices' reproduction quality. Section 5 analyses the benefits of fitting models to cumulative sums instead of original rank-size distributions in more detail. The paper is concluded in Section 6.

2. Methods

Assume that a citation vector of someone who authored N papers is represented with a sequence $\mathbf{x} = \big(x_1, \dots, x_N\big)$, where x_k denotes the citation count of the k-th most cited item, i.e.,

$$x_1 \geq x_2 \geq \cdots \geq x_N \geq 0.$$

In this section, we are going to outline some popular approaches to the modelling of such vectors that we have studied in (Cena et al., 2022). We will also recall a set of bibliometric indices which are frequently used to summarise citation sequences into just one number.

2.1. Citation models

It is frequently observed that the number of citations received by a paper follows the power law (which is also known as the law of Lotka or the law of Zipf; see Egghe and Rousseau, 2006; Egghe, 2009). In this setting, the formula for the number of citations to the k-th most cited paper takes the form

$$\hat{x}_k^{\text{PowerLaw}}(N, \alpha, \gamma) = \frac{\gamma}{k^\alpha},$$

for some $\alpha > 0$ and $\gamma > 0$.

Despite the popularity of the power law model, in practice, it is not very accurate. One way to improve its predictive abilities is to add a parameter that makes the citation curve more adaptable to empirical data. And thus, the DGBD (discrete generalised beta distribution) model (Naumis and Cocho, 2008; Martínez-Mekler et al., 2009) is described by the formula

$$\hat{x}_k^{\text{DGBD}}(N, A, a, b) = A\frac{(N+1-k)^b}{k^a},$$

where $A > 0$, $a > 0$ and $b > 0$. It can easily be seen that the above reduces to the power law equation when $b = 0$.

Unfortunately, the parameters of the DGBD model are not easily interpretable. It is also unclear which mechanisms could lead to the emergence of citations distributed in such a way. The recently proposed 3DSI (three dimensions of scientific impact) model (Siudem et al., 2020) does not have such shortcomings. It is given by

$$\hat{x}_k^{\mathrm{3DSI}}(N, C, \rho) = \frac{1-\rho}{\rho}\frac{C}{N}\left(\prod_{j=k}^{N}\frac{j}{j-\rho} - 1\right), \tag{1}$$

where C is the total number of citations and $\rho < 1$ is the ratio of preferentially to accidentally attached citations; see (Gagolewski et al., 2022) for the interpretation in the case of $\rho < 0$. Also, the left subfigure of Fig. 6 depicts how this parameter affects the shape of the resulting curve.

Furthermore, the harmonic model (Cena et al., 2022) is a special case of 3DSI where only the accidental part of the citation granting process is considered, i.e., the limiting case as $\rho \to 0$. It is defined as

$$\hat{x}_k^{\mathrm{Harmonic}}(N, C) = \frac{C}{N}\sum_{l=k}^{N}\frac{1}{l}.$$

Contrary to the other models, the above does not require any sophisticated fitting procedures. This is because the number of papers N and the total number of citations C can be taken directly from the input sample.

In other cases, in (Cena et al., 2022) we suggested minimising the differences in the logarithms of the citation vectors via the nonlinear least squares fitting with respect to the Cauchy loss ($\mathcal{L}(\epsilon) = \log(1+\epsilon^2)$) so as to minimise the impact of the outliers. The objective function then takes the form

$$F_1(\mathbf{x}, \hat{\mathbf{x}}^{\mathrm{M}}) = \frac{1}{N}\sum_{k=1}^{N}\log\left(1 + \left(\log(x_k) - \log(\hat{x}_k^M)\right)^2\right) \tag{2}$$

and is minimised with respect to the parameters of the underlying model M (either PowerLaw, DGBD, or 3DSI). The value of F_1 at the minimum can be used as a measure of the overall goodness of fit to the citation curve. Of course, other loss functions than the Cauchy one can be employed. However, in our previous study (Cena et al., 2022), we observed that they did not yield significantly different results.

2.2. Bibliometric indices

Many bibliometric indices have been proposed to summarise a citation vector into just one number. Here, we study a similar set of measures as in (Cena et al., 2022) and (Gagolewski et al., 2022), whose definitions are presented in Table 1.

The h-index (Hirsch, 2005), takes into consideration both the quantity and quality of papers and is defined as the largest number H for which H papers by a given author have at least H citations each. This measure gained popularity rather quickly, and now it is universally used in many decision-making contexts.

One of its most noteworthy modifications is the g-index (Egghe, 2006), defined as the greatest number G for which the G most highly cited papers have at least G^2 citations in total. Unfortunately, such a definition often yields g equal to the number of papers, N (in the case of the dataset we use here, this was true for about 50% of the vectors). Thus, we will compute this metric assuming that the citation vector is padded with 0s; e.g., $g(100) = g(100, 0, 0, \dots) = 10$. This way, adding papers with 0 citations does not change the value of the index (e.g., the h-index has the same property).

The h-index can be described geometrically as the length of the side of the largest square, which fits under the citation curve. The w-index (Woeginger, 2008) is based on a similar concept: we want to find the largest right triangle with legs of equal lengths. The length of the longest possible leg is then the value of w.

Table 1
Bibliometric indices studied in this paper

Name	Definition	Source
h-index	$h(\mathbf{x}) = \max\left\{h = 1, \dots, N \,:\, x_h \geq h\right\}$	(Hirsch, 2005)
g-index	$g(\mathbf{x}) = \max\left\{g \in \mathbb{N} \,:\, \sum_{i=1}^{g} x_i \geq g^2\right\}$, $x_{N+1} = x_{N+2} = \dots = 0$	(Egghe, 2006)
r-index	$r(\mathbf{x}) = \sqrt{h(\mathbf{x})\, a(\mathbf{x})} = \sqrt{\sum_{i=1}^{h(\mathbf{x})} x_i}$	(Jin et al., 2007)
a-index	$a(\mathbf{x}) = \frac{1}{h(\mathbf{x})} \sum_{i=1}^{h(\mathbf{x})} x_i$	(Burrell, 2007; Alonso et al., 2009)
$p20$	$p20(\mathbf{x}) = \sum_{i=1}^{0.2N} x_i$	
rms	$rms(\mathbf{x}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} x_i^2}$	
o-index	$o(\mathbf{x}) = \sqrt{h(\mathbf{x})\, x_1}$	(Dorogovtsev and Mendes, 2015b)
hg-index	$hg(\mathbf{x}) = \sqrt{h(\mathbf{x})\, g(\mathbf{x})}$	(Alonso et al., 2010)
s-index	$s(\mathbf{x}) = \sum_{i=1}^{N} (1 + \log(x_i))$	
m-index	$m(\mathbf{x}) = \text{median}\left\{x_1, x_2, \dots, x_{h(\mathbf{x})}\right\}$	(Bornmann et al., 2008)
w-index	$w(\mathbf{x}) = \max\left\{w = 1, \dots, N \,:\, x_i \geq w - i + 1 \text{ for all } i \leq w\right\}$	(Woeginger, 2008)
$i10$	$i10(\mathbf{x}) = \max\{i = 1, \dots, N \,:\, x_i \geq 10\}$	

Many other indices are based directly on the value of $h(\mathbf{x})$, e.g., the hg-index (Alonso et al., 2010), which is simply the geometric mean of the h- and g-indices or the o-index (Dorogovtsev and Mendes, 2015a), being the geometric mean of h and x_1. Other measures are calculated on the subset of a citation vector, called the h-core, consisting of the $H = h(\mathbf{x})$ most cited papers. Amongst the examples of such metrics, we find the a-index (Jin et al., 2007), which is the arithmetic mean of h-core, the m-index (Bornmann et al., 2008), where instead of the mean, the median is taken, as well as the r-index (Jin et al., 2007), where the square root of the sum is calculated. Even though all these metrics seem to be very similar, they differ considerably when it comes to analysing their behaviour on real data sets (Cena et al., 2022; Gagolewski et al., 2022).

We also study some measures which consider the full input vector, such as the s-index being the sum of its logarithms whose function serves as the maximum likelihood estimator in the Pareto distribution family (see (Arnold, 2015; Siudem et al., 2022)), and the root mean squares of the whole sample, which can be utilised in the method of moments estimator therein.

Additionally, the sum of citations in the first 20% of the papers is analysed together with the $i10$ measure, which counts the number of papers that have at least 10 citations.

Note that, with the exception of a and m, all indices are nondecreasing with respect to the dominance relation (Woeginger, 2008) $(X_1, \dots, X_N) \preceq (X'_1, \dots, X'_{N'})$ which holds whenever $N \leq N'$ and $X_i \leq X'_i$ for every $i \leq N$; see (Wu and Zhang, 2017; Gagolewski, 2013) for discussion. In particular, $h(10, 4, 2) = 2 < h(11, 4, 3) = 3$, but $a(10, 4, 2) = 7 > a(11, 4, 3) = 6$ and $m(10, 4, 2) = 7 > m(11, 4, 3) = 4$. Except for the $i10$-, s-, and $p20$-indices, all measures yield the value of H for a vector featuring H repeated H times, e.g., $h(3, 3, 3) = 3$. Therefore, they can be thought of as being on the same scale as the elements of the input vector.

2.3. Database

We will analyse the DBLP v12 bibliography database (Tang et al., 2008) (see `https://aminer.org/citation`), which consists of 4,398,138 citation vectors of individual scientists. The vast majority of authors have very few published papers which have a small number of citations. We have thus decided to limit our analysis to those with at least 5 cited papers and the $i10$-index greater than 0. It resulted in 348,956 citation sequences (which is still a very large number). We omitted papers with no citations, because the case of

$x_k = 0$ would be problematic during optimisation/computations on the log-scale.

Note that another way to deal with the problem with zeroes could be by replacing x_k with $x_k + 1$ as in (Thelwall, 2016). Nevertheless, our analyses suggested that models are often more accurate when we use the former approach, i.e., when N is the number of papers that have been cited at least once (see also Sec. 2.3 in (Cena et al., 2022)). Moreover, in this case, the overall number of citations C is not perturbed. Also, note that most citation indices do not take uncited papers into account anyway.

3. Bibliometric indices vs top-cited papers

To measure the accuracy of the prediction of a bibliometric index, we take into consideration the relative error averaged over all vectors in the whole database $\mathbf{X}$, i.e.,

$$\mathrm{Err}(I, M) = \frac{1}{|\mathbf{X}|} \sum_{\mathbf{x}\in\mathbf{X}} \frac{|I(\mathbf{x}) - I(\hat{\mathbf{x}}^{\mathrm{M}})|}{I(\mathbf{x})}, \tag{3}$$

where I is a selected citation index and $\hat{\mathbf{x}}^{\mathrm{M}}$ denotes the vector predicted by a citation model M (e.g., 3DSI or DGBD) fitted to a given vector $\mathbf{x}$ by minimising F_1 given by Eq. (2).

For example, when we take a citation vector $\mathbf{x}_{\mathrm{example}} = (69, 17, 16, 14, 11, 10, 9, 8, 8, 6, 6, 4, 4, 3, 1)$ (see also Fig. 5) and the model $M = $ 3DSI, we can minimise F_1 in order to find the best value of the parameter ρ. As a result, in this case, we obtain $\rho = -0.12$ and then we can predict the citation sequence $\hat{\mathbf{x}}^{\mathrm{3DSI}} \approx (37.13, 27.74, 22.47, 18.74, 15.83, 13.44, 11.39, 9.60, 8.01, 6.57, 5.26, 4.06, 2.94, 1.90, 0.92)$ using Eq. (1). Now we can calculate any bibliometric index on this estimated citation vector. For instance, for $I = h$, we get $h(\mathbf{x}) = h(\hat{\mathbf{x}}^{\mathrm{3DSI}}) = 8$.

This procedure was performed for all the aforementioned models, indices, and vectors. Then, the results were averaged over citation sequences for all authors, yielding $\mathrm{Err}(I, M)$ for each model M and index I.

Figure 1 compares what the relative estimation error of different bibliometric indices looks like across different models (similar data to those presented in (Cena et al., 2022), but this time with more indices studied). The positions of the models on the x-axis correspond to the overall goodness of fit as measured by the Cauchy loss (F_1) averaged over all citation vectors.

Overall, we can distinguish two groups of indices: the ones for which the estimation quality slightly worsens as the cost value grows (blue line segments: m, $i10$, h, w, and s-indices, plus hg being somewhat a borderline case), and the ones for which such behaviour is somewhat perturbed (red lines: rms, $p20$, a, o, g, and r).

In terms of the average loss, we observe that the DGBD model fits the citation vectors best. However, compared to 3DSI, they both have a similar performance as far as the bibliometric index reproduction quality is concerned. On the other hand, the Harmonic model is the worst one with respect to the average Cauchy loss, but actually outperforms PowerLaw when it comes to index prediction.

Some measures can be approximated through the 3DSI model with very good accuracy. This is especially visible for the g-index (0.004 mean relative error) and the r-index (0.014). What is even more interesting, also the Harmonic model enjoys good predictive abilities in these two cases, where it outperforms DGBD.

Our main aim in this research study is to find the reasons for the above and to reparameterise the DGBD and PowerLaw models so that they can reproduce some of the indices much better.

Most bibliometric measures strongly focus on the most cited papers and marginalise or even completely discard the less-cited ones. However, the way they take the top-cited papers into account differs from index to index.

Firstly, let us note that the h-index does not take the actual value of the elements in a citation sequence into account as long as they exceed a certain threshold. Knowing that, for example, $h(\mathbf{x})$ is equal to 6, we

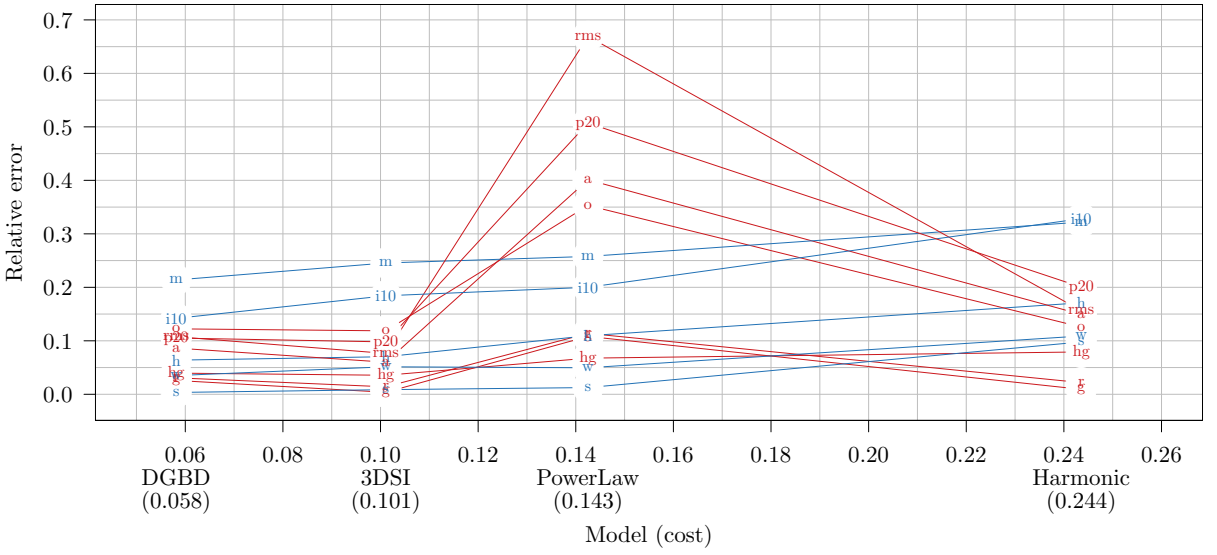

Figure 1: Relative approximation errors for the bibliometric indices (Eq. (3)) across different models as a function of the averaged Cauchy loss (Eq. (2)). The line segments are added to guide the eye. It turns out that the cost function, measuring the *overall* goodness of fit of a model, is a weak predictor of the index reproduction quality.

can imply that the total number of citations C is at least $6^2 = 36$, but in practice, this is merely a loose lower bound. Therefore, we do not expect h to correlate with C too highly. By definition, the $i10$ and w-indices have similar behaviour.

On the other hand, some indices highly rely on the actual value of the sum of the first few top-cited items. By definition, this is definitely the case for g, r, a, $p20$, and o.

Figure 2 presents the cumulative sums of 3 quite representative citation vectors, $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$. Overall, the cumulative counts curve tends to flatten quite quickly, which means that only a few first papers carry the most citation load. This is nicely captured by the g-index, which is usually reached at the point where $x_1 + \cdots + x_g$ is already close to C (98%C for $\mathbf{x}_1$, and 100%C for both $\mathbf{x}_2$ and $\mathbf{x}_3$). As it was noted in (De Visscher, 2011), the g-index is often very highly correlated with the square of the overall number of citations, $\sqrt{C}$. We observe that this is often true even for small sample sizes.

Figure 3 gives the Spearman rank correlation coefficients between the total citation count as well as different configurations of the sums of the top-cited papers and all of the considered indices. In fact, the g-index is not the only measure that is highly correlated with some monotone function of C: this is also the case for r, $p20$, o (which is a function of x_1), and hg (as it is a function of g).

Furthermore, the total sum can actually be quite accurately expressed as a monotone function of the top 5 cited papers, $C5$, and vice versa. In particular, from Figure 4 we read that for 50% of the vectors, $C5$ is already equal to at least 86%C.

We also observe that *rms* (note that the square emphasises the largest observations), a, and o highly correlate with x_1 (the maximum, i.e., the sum of top 1 papers). Figure 4 reveals that the citation distributions are so highly skewed that for 50% of the vectors the maximum alone is responsible for at least 33% of the total citation load.

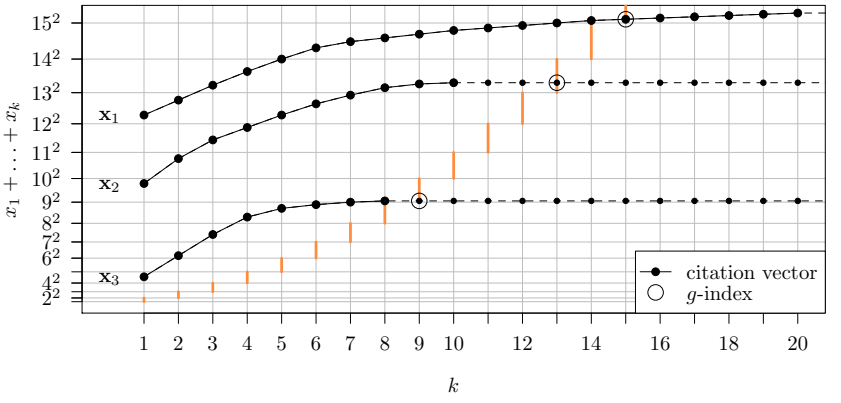

Figure 2: Cumulative sums (i.e., $(x_1, x_1 + x_2, x_1 + x_2 + x_3, \ldots)$) of three example citation vectors (padded with trailing 0s). Each circle denotes the value of the g-index, with the orange vertical lines depicting the segments $(k^2, (k+1)^2)$ for each k to guide the eye. As the empirical data distribution is usually highly skewed, the cumulative sums tend to saturate quickly, i.e., it is expected that the sum of a few top-cited papers should be close to the total sum.

	w	m	h	s	i10	x_1	rms	a	o	hg	p20	r	g	C5	C10	p70	p80	p90	C	
x_1	0.53	0.80	0.63	0.66	0.72	1.00	0.96	0.97	0.96	0.83	0.93	0.93	0.93	0.96	0.93	0.91	0.91	0.91	0.91	x_1
C5	0.66	0.87	0.77	0.79	0.85	0.96	0.92	0.97	0.98	0.93	0.96	0.99	0.99	1.00	0.99	0.98	0.98	0.98	0.98	C5
C10	0.73	0.86	0.83	0.85	0.89	0.93	0.87	0.94	0.98	0.96	0.98	1.00	1.00	0.99	1.00	0.99	0.99	0.99	0.99	C10
p70	0.77	0.84	0.85	0.89	0.90	0.91	0.83	0.92	0.98	0.97	0.98	0.99	0.99	0.98	0.99	1.00	1.00	1.00	1.00	p70
p80	0.77	0.84	0.86	0.89	0.90	0.91	0.83	0.91	0.98	0.97	0.98	0.99	0.99	0.98	0.99	1.00	1.00	1.00	1.00	p80
p90	0.77	0.84	0.86	0.89	0.91	0.91	0.83	0.91	0.98	0.97	0.98	0.99	0.99	0.98	0.99	1.00	1.00	1.00	1.00	p90
C	0.77	0.84	0.86	0.89	0.91	0.91	0.83	0.91	0.98	0.97	0.98	0.99	0.99	0.98	0.99	1.00	1.00	1.00	1.00	C
	w	m	h	s	i10	x_1	rms	a	o	hg	p20	r	g	C5	C10	p70	p80	p90	C	

Figure 3: Spearman's rank correlation coefficients between the total number of citations C, the maximal number of citations x_1, the sum of the top T observations CT (or N if $T > N$), the sum of the top $P\%$ of the greatest items pP, and all the bibliometric indices. Some indices, more than others, require the top-cited papers to be approximated well.

4. Reparameterising the DGBD and PowerLaw models

The above data suggests that a good theoretical model does not necessarily have to fit to all the items in a citation vector well on average (i.e., elementwisely). Instead, it should rather focus on reproducing the total or at least top citation counts.

Figure 5 illustrates this phenomenon. The black line segments represent the citation sequence, the blue dashed lines give the fitted 3DSI model, and the red dotted ones depict the approximation via DGBD. From the right subplot (showing the logarithm of the values in the citation sequence) we read that DGBD fits the

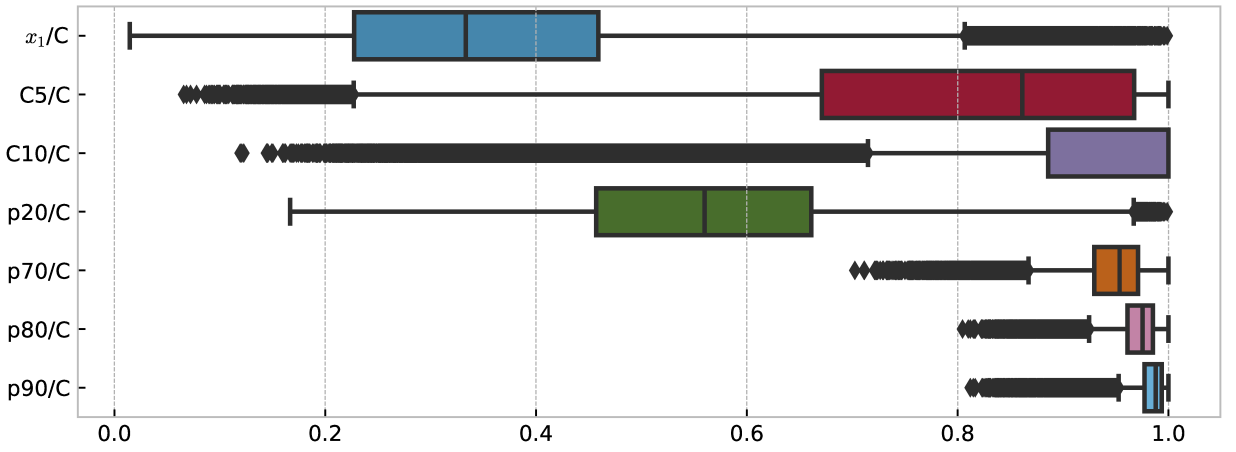

Figure 4: Box-and-whisker plots depicting the distributions of the ratios of various cumulative sums of the top-cited papers to the total sum across all vectors in the database. Due to the citation distributions' being highly skewed, only a few top-cited papers are needed to predict the total sum well.

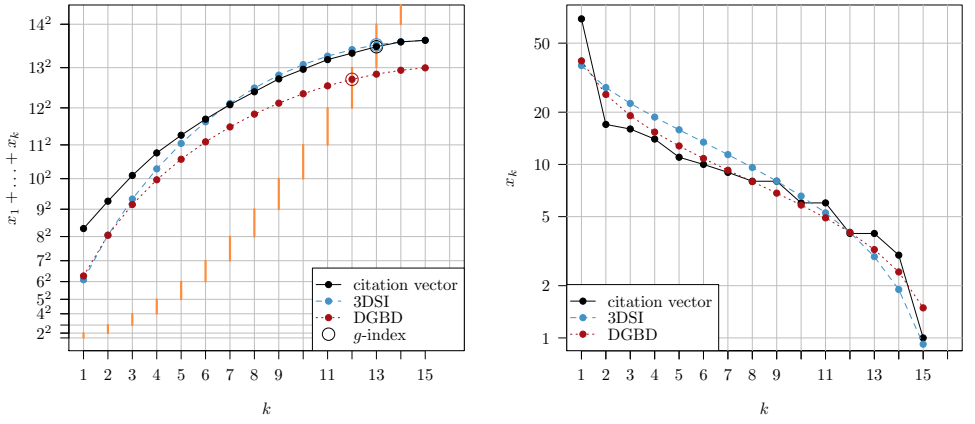

Figure 5: Example vector of one author: its cumulative sums (lefthand side) and the original version on the log-scale (right) together with the best-fitted 3DSI and DGBD models. The former approximates the sums of the top-cited papers better.

original data better. However, it is the 3DSI model that better reflects the cumulative sums (lefthand side).

What distinguishes the 3DSI and Harmonic models from the DGBD and PowerLaw ones is that the former two reproduce the sum of the elements in the original sequence exactly, i.e.,

$$C = \sum_{k=1}^{N} x_k = \sum_{k=1}^{N} \hat{x}_k^{\text{3DSI}} = \sum_{k=1}^{N} \hat{x}_k^{\text{Harmonic}}.$$

This is beneficial in terms of reproducing some of the bibliometric indices. It makes the fitting procedure much easier: the C value can be plugged-in to the formula directly, and then it does not need to be optimised for when minimising the cost function F_1. What is more, both the 3DSI and the Harmonic model have the parameter C, which is easily interpretable. The shape of the 3DSI curve depends furthermore on one

additional free parameter ρ (the smaller its value is, the flatter the citation sequence becomes; see Fig. 6 and its description).

It turns out that we can modify the DGBD and the PowerLaw models quite easily so as to retain the sum C of the original vector.

For the PowerLaw model, it is sufficient to replace the parameter γ with

$$\gamma = \frac{C}{\sum_{j=1}^{N} \frac{1}{j^\alpha}},$$

which results in what we shall refer to as the PowerLaw2 model

$$\hat{x}_k^{\text{PowerLaw2}}(N, C, \alpha) = \frac{C}{k^\alpha \sum_{j=1}^{N} \frac{1}{j^\alpha}}.$$

It holds $\sum_{k=1}^{N} x_k^{\text{PowerLaw2}} = C$. This model has only one free additional parameter, $\alpha > 0$, which controls the shape of generated citation curve (see the middle subfigure of Fig. 6).

It is interesting to note that, unlike the 3DSI model, the PowerLaw2-generated curves have only one inflection point (as compared to two in the former case), yielding flatter citation curves.

Similarly, for the DGBD model, it is enough to use the following as a substitute for the A parameter

$$A = \frac{C}{\sum_{j=1}^{N} \frac{(N+1-j)^b}{j^a}},$$

which leads to the DGBD2 model defined by the equation

$$\hat{x}_k^{\text{DGBD2}}(N, C, a, b) = \frac{C}{\sum_{j=1}^{N} \frac{(N+1-j)^b}{j^a}} \frac{(N+1-k)^b}{k^a}.$$

The reparameterised model has two free parameters, $a > 0$ and $b > 0$ (instead of three). The additional value C is taken directly from the sample. Both a and b influence the shape of the citation curve (see the right subfigure of Fig. 6). The former affects the first part of the citation curve (the most highly cited papers), whereas the latter controls the inflection in the tail of the rank-size distribution.

To emphasise the differences between these models, Fig. 7 compares the citation curves scaled in such a way that x_1 was either equal to 50, 200, or 500. For the 3DSI and PowerLaw2 models, once the values of x_1, N and C are fixed, there is only one value of ρ and α, which gives a vector enjoying such constraints. For DGBD2, however, the solution is nonunique, and thus we depict a representative sample thereof (grey curves).

The greater flexibility of DGBD2 is clearly visible here: as a generalisation of PowerLaw2, it can generate not only quite flat rank-size distributions, but also ones that transfer more citation load onto the top-cited papers. Interestingly, the 3DSI model generates quite different curves. Which model is a winner obviously depends on whether it is able to cover most of the actually observable citation shapes and not overfit to the

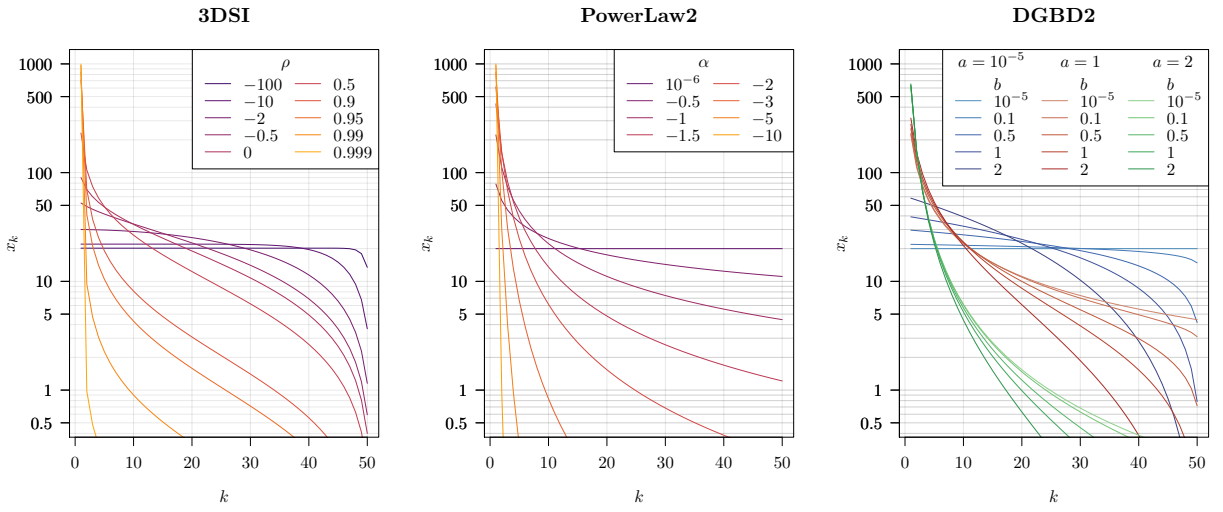

Figure 6: A comparison of how different parameters affect the shape of the theoretical citation curves generated by the 3DSI, PowerLaw2, and DGBD2 models (for fixed N and C). Note the log-scale on the y-axis. The DGBD2 model is the most flexible, but it has two adjustable parameters as compared to one in the two other cases.

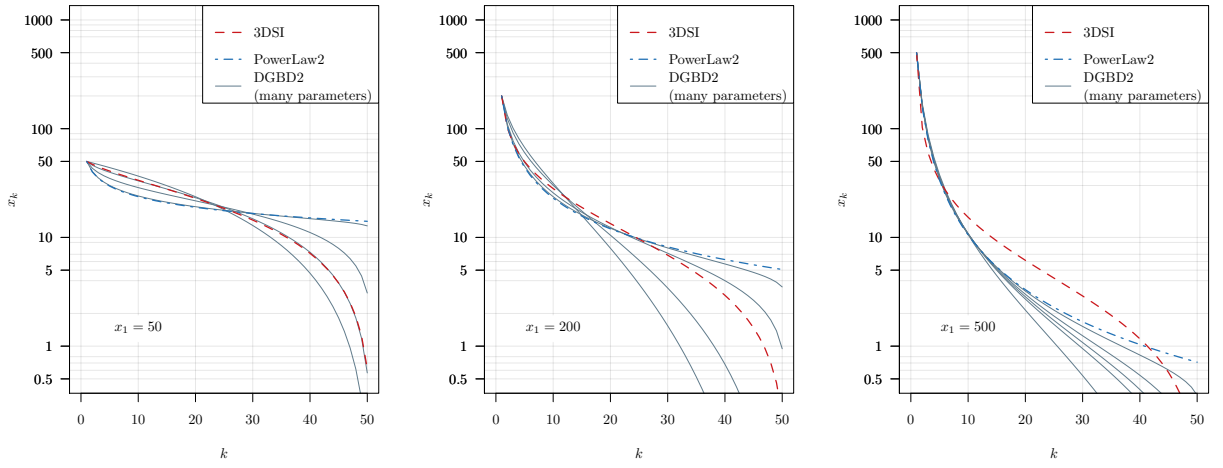

Figure 7: A comparison of shapes of different theoretical citation curves yielding a specific x_1 (for fixed N and C).

parts of the citation curve that are of smaller importance (whose risk is increased if the number of parameters is higher).

Thanks to the above reparameterisations, we have fewer parameters that need to be optimised for during the fitting of the model, and a guarantee that the total citation count C is preserved. Figure 8 shows the relative estimation error for other pivotal partial sums of top elements. We note that models with a fixed C are superior to the remaining ones, also in the most extreme case of estimating the top cited papers (x_1).

Figure 9 shows how the fixing of C affects the bibliometric index reproduction quality. In one group of indices (right subplot), namely g, a, o, r, rms, and $p20$ (and to some extent hg), the gains are significant

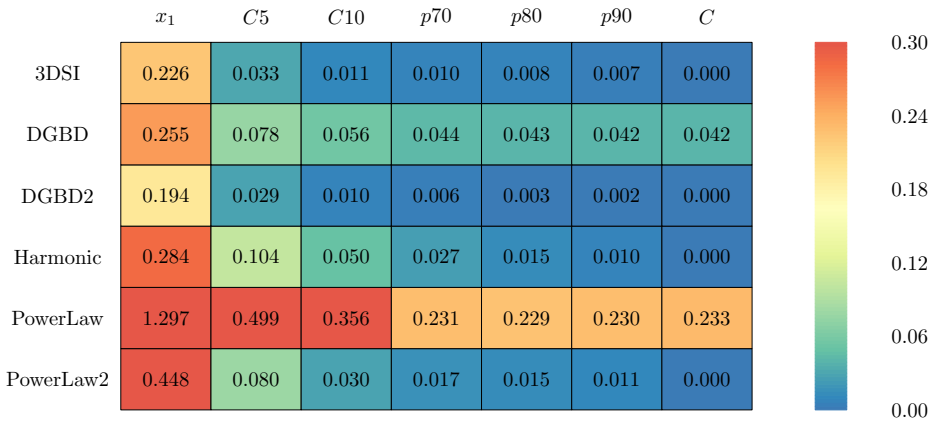

Figure 8: Relative prediction error for some cumulative sums of the citation sequences, where CT denotes the sum of the top T observations (or N if $T > N$) and pP is the sum of $P\%$ of the greatest items.

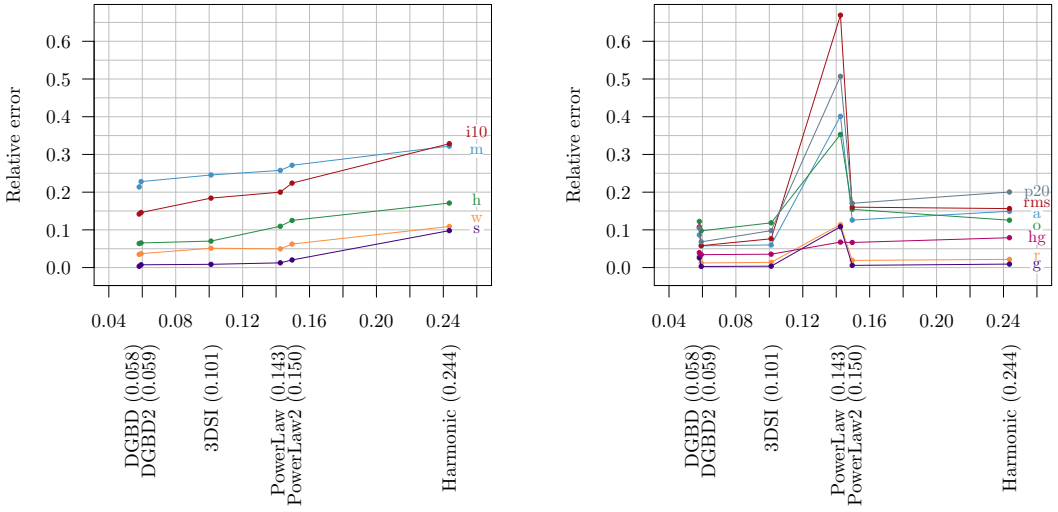

Figure 9: Relative prediction errors for different indices across all the models, whose placement on the x-axis corresponds to the value of the cost function F_1. The lefthand side plot depicts the indices for which fixing C worsens (but only slightly) their reproduction quality. For the measures on the righthand side, we see a significant improvement.

for both DGBD2 and PowerLaw2 models. At the same time, for m, h, w, s, and $i10$ (lefthand side), the worsening is minimal.

In terms of the mean cost F_1, fixing C only increases this error measure slightly. For DGBD, we raise the cost from 0.0581 to 0.0593. The difference between PowerLaw and PowerLaw2 is also very small (0.1425 vs 0.1496).

We can thus conclude that the models with fixed C are not only easier to fit, but also are more useful in terms of predicting the chosen bibliometric indices. At the same time, their loss of precision with respect to other goodness of fit measures is small.

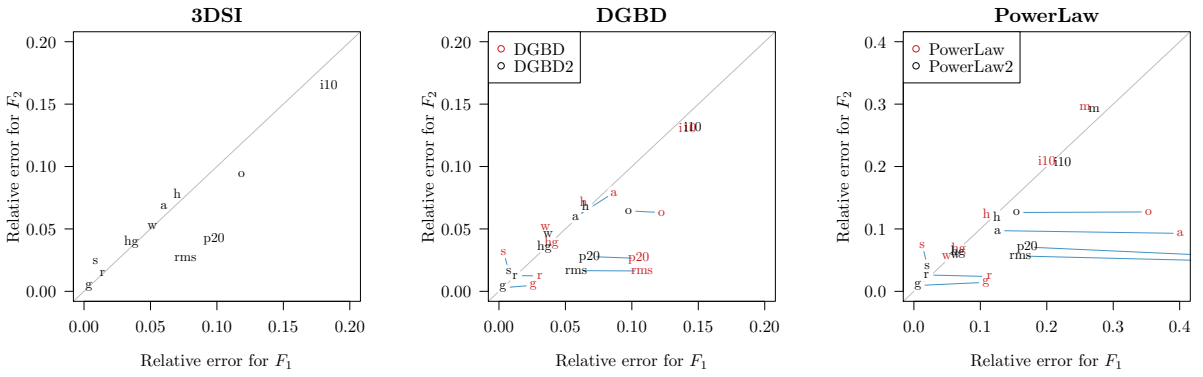

Figure 10: Relative prediction errors for bibliometric indices across different models and whether the F_1 (Eq. (2)) or F_2 (see Eq. (4)) cost function was minimised. Indices lying below the $y = x$ line benefit from a better fit to the cumulative sums, which is particularly the case for the DGBD and PowerLaw models not preserving C (line segments between the indices were added for readability).

5. Minimising costs based on cumulative sums

We have noted that for some indices, it helps if they are reproduced by the models that fit to the cumulative sums better. Inspired by this observation, let us verify whether maximising the goodness of fit defined by means of such partial sums could be beneficial.

Let us then define another cost function

$$F_2(\mathbf{x}, \hat{\mathbf{x}}^M) = \frac{1}{N} \sum_{k=1}^{N} \left(\sum_{j=1}^{k} x_j - \sum_{j=1}^{k} \hat{x}_j^M \right)^2, \tag{4}$$

which is the mean square difference between the consecutive cumulative sums.

As in the case of F_1, for each model M, we seek the underlying parameters that result in the F_2 cost function's being as small as possible. Figure 10 shows where the relative prediction errors for all the indices are smaller, when we choose to minimise with respect to the cumulative sums (F_2) instead of the original rank-size distribution (F_1). The case where a point lies above the $y = x$ line means that the original fitting procedure was better for a particular index. When it lies below it, then the cumulative fit was more beneficial.

For instance, in the left subplot that includes the results for the 3DSI model, the relative error for the o-index and the original cost being minimised is $F_1 = 0.119$, which is larger than when the fitting to the cumulative sums is conveyed, namely, $F_2 = 0.094$.

Additionally, we have two variants of the DGBD and the PowerLaw models, where the data points for the C-preserving and original model versions are – for readability – connected by line segments to highlight how the reparameterisation affects the quality of the indices' reproduction.

The error values for m-index were omitted from 3DSI as they were significantly larger than the other ones (3DSI: $F_1 = 0.25$ and $F_2 = 0.23$; DGBD: $F_1 = 0.21$ and $F_2 = 0.41$; DGBD2: $F_1 = 0.23$ and $F_2 = 0.24$), which would harm the readability. In the PowerLaw case, note that the scale on the axes is different as the errors are much more substantial (still, some indices lie outside the bounding box in the PowerLaw case: $p20$ with $F_1 = 0.51$ and $F_2 = 0.05$ as well as rms with $F_1 = 0.67$ and $F_2 = 0.04$).

Overall, in the case of the C-preserving models (3DSI, DGBD2, PowerLaw2), the difference between the two fitting methods is only noticeable for $p20$ (which in fact is a partial sum) as well as rms and o. The last two strongly correlate with x_1 (however, another measure with this property, the a-index, does not gain much from the new fitting method).

In the case of the original versions of PowerLaw and DGBD (with a free, fittable parameter of scale), reproducing cumulative sums instead of the rank-size distribution is very beneficial for g, r, rms, $p20$, o, and a, which confirms our observations regarding the advantages of preserving C (or other partial sums).

Finally, it is worth noting that occasionally the method we used (`scipy.optimize.least_squares` in Python, which implements the Trust Region Reflective algorithm) for minimising F_2 had some convergence problems, which needed to be remedied by restarting it from many more random initial points than in the case of F_1 (in many cases, a dozen or so restarts were needed to generate one solution). Thus, overall, the problem of fitting with respect to F_2 is more difficult and the development of some robust algorithms for doing so should be researched in the future.

6. Conclusions

In (Siudem et al., 2020) we have proposed a new technique for modelling citation vectors. It can not only predict the particular values in a sequence quite well, but also, as argued in (Cena et al., 2022), reproduce various bibliometric indices with a small error. Most notably, some indices were found to be much easier to estimate than others, regardless of the overall goodness of fit to the actual vector.

In this paper, we investigated the reasons for this behaviour. Namely, we argued that by reparameterising a model in such a way that it preserves the total sum, we obtain good predictive power for the indices as a byproduct. This is because the new model can reflect other partial sums of the top items much better.

We note that the new versions of the DGBD and Power Law models have fewer adjustable (numerically) parameters, making the fitting procedure less tedious and more robust. Also, they are now easier to interpret: N controls the number of items, C is the total citation count (parameter of scale), and the remaining parameters are responsible for the shape of the distribution.

As a topic for further research, a similar reparametrisation of other citation models could also be conveyed. Also, other than bibliometric datasets should be studied.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the anonymous referees for useful remarks that helped improve the paper. This research was supported by the Australian Research Council Discovery Project ARC DP210100227 (MG) and by the POB Research Centre Cybersecurity and Data Science of Warsaw University of Technology within the Excellence Initiative Program – Research University (ID-UB) (AC and GS).

CRediT authorship contribution statement

Barbara Żogała-Siudem: Conceptualisation of this study, Methodology, Data Curation, Software, Investigation, Visualisation, Writing – Original draft preparation, Writing – Revision. **Anna Cena:** Conceptualisation of this study, Methodology, Data Curation. **Grzegorz Siudem:** Conceptualisation of this study, Methodology, Writing – Original draft preparation. **Marek Gagolewski:** Conceptualisation of this study, Methodology, Data Curation, Software, Investigation, Visualisation, Writing – Original draft preparation.

References

Alonso, S., Cabrerizo, F., 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, 273 – 289. doi:10.1016/j.joi.2009.04.001.

Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E., Herrera, F., 2010. hg-index: a new index to characterize the scientific output of researchers based on the h- and g-indices. Scientometrics 82, 391–400. doi:10.1007/s11192-009-0047-5.

Arnold, B.C., 2015. Pareto Distributions. Chapman and Hall/CRC, New York, NY, USA. doi:10.1201/b18141.

Bornmann, L., Mutz, R., Daniel, H.D., 2008. Are there better indices for evaluation purposes than the *h*-index? A comparison of nine different variants of the *h*-index using data from biomedicine. Journal of the American Society for Information Science and Technology 59, 830–837. doi:10.1002/asi.20806.

Burrell, Q.L., 2007. On the *h*-index, the size of the Hirsch core and Jin's *a*-index. Journal of Informetrics 1, 170 – 177. doi:https://doi.org/10.1016/j.joi.2007.01.003.

Cena, A., Gagolewski, M., Siudem, G., Żogała-Siudem, B., 2022. Validating citation models by proxy indices. Journal of Informetrics 16, 101267.

De Visscher, A., 2011. What does the *g*-index really measure? Journal of the American Society for Information Science and Technology 62, 2290–2293.

Dorogovtsev, S., Mendes, J., 2015a. Ranking scientists. Nature Physics 11, 882–883. doi:10.1038/nphys3533.

Dorogovtsev, S., Mendes, J., 2015b. Ranking scientists. Nature Physics , 882.

Egghe, L., 2006. Theory and practise of the *g*-index. Scientometrics 69, 131–152.

Egghe, L., 2009. Lotkaian informetrics and applications to social networks. Bull. Belg. Math. Soc. Simon Stevin 16, 689–703.

Egghe, L., Rousseau, R., 2006. An informetric model for the hirsch-index. Scientometrics 69, 121–129. doi:10.1007/s11192-006-0143-8.

Gagolewski, M., 2013. Scientific impact assessment cannot be fair. Journal of Informetrics 7, 792–802. doi:10.1016/j.joi.2013.07.001.

Gagolewski, M., Żogała-Siudem, B., Siudem, G., Cena, A., 2022. Ockham's index of citation impact. Scientometrics 127, 2829–2845. doi:10.1007/s11192-022-04345-2.

Hirsch, J.E., 2005. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences 102, 16569–16572. doi:10.1073/pnas.0507655102.

Jin, B., Liang, L., Rousseau, R., Egghe, L., 2007. The R- and AR-indices: Complementing the *h*-index. Chinese Science Bulletin 52, 855–863. doi:10.1007/s11434-007-0145-9.

Li, S., Shen, H., Bao, P., Cheng, X., 2021. h_u-index: a unified index to quantify individuals across disciplines. Scientometrics 126, 3209–3226.

Martínez-Mekler, G., Martínez, R.A., del Río, M.B., Mansilla, R., Miramontes, P., Cocho, G., 2009. Universality of rank-ordering distributions in the arts and sciences. PLOS ONE 4, 1–7. doi:10.1371/journal.pone.0004791.

Naumis, G., Cocho, G., 2008. Tail universalities in rank distributions as an algebraic problem: The beta-like function. Physica A: Statistical Mechanics and its Applications 387, 84 – 96. doi:https://doi.org/10.1016/j.physa.2007.08.002.

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

Poirrier, M., Moreno, S., Huerta-Cánepa, G., 2021. Robust *h*-index. Scientometrics 126, 1969–1981.

Schreiber, M., 2008. An empirical investigation of the *g*-index for 26 physicists in comparison with the *h*-index, the *a*-index, and the *r*-index. Journal of the American Society for Information Science and Technology 59, 1513–1522.

Siudem, G., Nowak, P., Gagolewski, M., 2022. Power laws, the Price model, and the Pareto type-2 distribution. Physica A: Statistical Mechanics and its Applications 606, 128059. doi:10.1016/j.physa.2022.128059.

Siudem, G., Żogała-Siudem, B., Cena, A., Gagolewski, M., 2020. Three dimensions of scientific impact. Proceedings of the National Academy of Sciences 117, 13896–13900. doi:10.1073/pnas.2001064117.

Tang, J., et al., 2008. ArnetMiner: Extraction and mining of academic social networks, in: Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008), pp. 990–998.

Thelwall, M., 2016. The discretised lognormal and hooked power law distributions for complete citation data: Best options for modelling and regression. Journal of Informetrics 10, 336 – 346. doi:https://doi.org/10.1016/j.joi.2015.12.007.

Wildgaard, L., Schneider, J.W., Larsen, B., 2014. A review of the characteristics of 108 author-level bibliometric indicators. Scientometrics 101, 125–158.

Woeginger, G.J., 2008. An axiomatic characterization of the Hirsch-index. Mathematical Social Sciences 56, 224–232.

Wu, Q., Zhang, P., 2017. Some indices violating the basic domination relation. Scientometrics 113, 495–500. doi:10.1007/s11192-017-2475-y.