This is a preprint of: A. Cena, M. Gagolewski, G. Siudem, B. Żogała-Siudem,
Validating citation models by proxy indices, Journal of Informetrics 16(2), 101267, 2022,
doi:10.1016/j.joi.2022.101267

Validating citation models by proxy indices

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

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

ARTICLE INFO

Keywords:
science of science
bibliometric indices
scientometrics
citation models
power law

ABSTRACT

There are many approaches to the modelling of citation vectors of individual authors. Models may serve different purposes, but usually they are evaluated with regards to how well they align to citation distributions in large networks of papers. Here we compare a few leading models in terms of their ability to correctly reproduce the values of selected bibliometric indices of individual authors. Our recently-proposed three-dimensional model of scientific impact serves this purpose equally well as the discrete generalised beta distribution and the log-normal models, but has fewer parameters which additionally are all easy to interpret. We also indicate which indices can be predicted with high accuracy and which are more difficult to model.

1. Introduction

In the era of big data, the title of the seminal book *Little Science, Big Science* by Price (1963) takes on a new meaning. *Big* can refer to science in its entirety, as described by the huge amounts of bibliometric data available nowadays. *Small* science, on the other hand, can be seen as a well-defined fragment of that rich and complex reality at a precisely chosen level of detail. Bibliometric research is interested in any data granularity, from macro- to microscopic: from studying the very general topological structure of citation networks (Chen and Redner, 2010), through modelling scientific fields (Herrera et al., 2010; Guevara et al., 2016; Battiston et al., 2019), institutions (Wuchty et al., 2007; Shen and Barabási, 2014; Sziklai, 2021), journals (Mingers and Yang, 2017), or individuals (Siudem et al., 2020; Ionescu and Chopard, 2013; Żogała-Siudem et al., 2016; Burrell, 2007a; Egghe and Rousseau, 2006), to considering a single-paper perspective, e.g., identifying the distribution of the number of citations (Thelwall and Wilson, 2014; Thelwall, 2016b,a) and its dynamics (Eom and Fortunato, 2011; Néda et al., 2017). Some fruitful attempts to introduce links between different abstraction levels include (Golosovsky and Solomon, 2012; Golosovsky, 2019).

Here we shall focus on the dependencies between entire citation records at an author level and their synthesised or compressed versions, see Fig. 1 for an illustration. In our recent work (Siudem et al., 2020) we have proposed a new agent-based model to describe an author's citation record – instead of storing data related to dozens of papers, we can determine three easily interpretable parameters that allow for recreating the original list. We should therefore now address the question of how well does it compare to other representations. As a baseline for this evaluation, we have selected three other popular models: the power-law, the log-normal, and the discrete generalised beta distribution (DGBD) ones. The latter two are very flexible as they involve many parameters that are optimised for when fitting them to data.

As there is no consensus on how to quantify the models' goodness of fit to true citation sequences, we have decided to use selected bibliometric indices as a proxy for this very purpose. The indices are assumed to be valid summaries of different characteristics of citation vectors and their being accurately predicted by a given model is taken as a working proof of the model's ability to reproduce the original vector faithfully. This is in line with the typical use of indices, see Sec. 2.1 for a detailed discussion. We shall check if more complex models can predict the indices better than the 3DSI model and also indicate which indices are easier or more difficult to estimate.

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

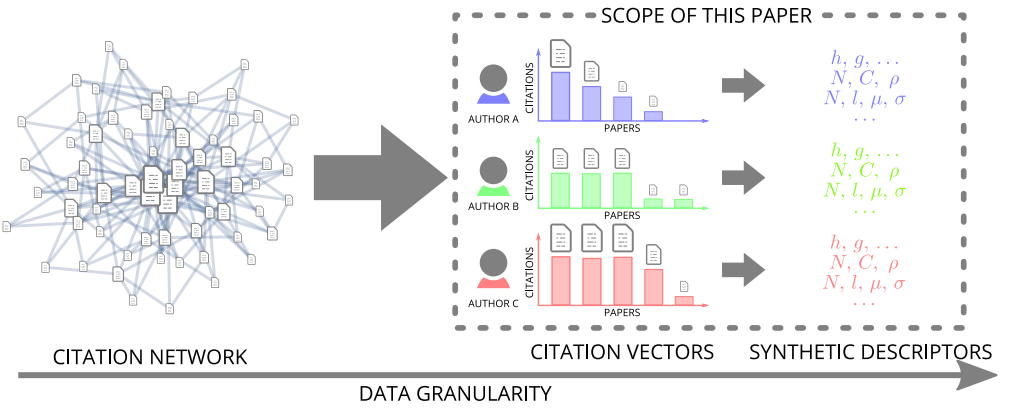

Figure 1: Different levels of granularity of bibliometric data: from the whole network, through citation vectors, to author-level descriptors.

What remains is structured as follows. First we review the impact indices and citation vector models that we use in this study. In particular, we derive some new approximations to the recently-proposed three-dimensional model of scientific impact (Siudem et al., 2020). Then, we describe the model fitting procedure, i.e., how to identify the model parameters that fit a citation vector best with respect to one of the assumed criteria. Next, we examine how well do the introduced models align with real-world data from the DBLP database. Finally, we verify to what extent a good overall fit corresponds to an accurate prediction of various data aggregates, including the h- and the g-index.

2. Methods

2.1. Bibliometric indices

Bibliometric indices are most often used to numerically summarise, rank, or assess the performance of institutions (Sziklai, 2021; Siwinski et al., 2021), journals (Mingers and Yang, 2017; Waltman, 2016), and individuals (Hirsch, 2005; Bornmann et al., 2008; Dorogovtsev and Mendes, 2015). Despite its sometimes being difficult and troublesome (Risi et al., 2019), they can also be employed for estimating the chances of success of individual papers or authors (Havemann and Larsen, 2015).

Here, however, we shall consider a quite different use case: how well or badly can we predict the values of bibliometric indices based on different citation models.

In our work we focus on the author-level indices, i.e., we consider the track record of a single scientist. This usually is represented as a citation vector, i.e., a list of the numbers of citations that each of their N papers have received (ordered from the highest to the lowest), compare Fig. 1. More formally, a citation vector is a sequence $\mathbf{x} = (x_1, \ldots, x_N)$, where x_k denotes the citation count of the k-th most cited paper,

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

Let us recall the definitions of some popular bibliometric indices whose aim is to provide a simple numerical summary of the above.

First, the h-index (Hirsch, 2005), given by the equation

$$h(\mathbf{x}) = \max\left\{H = 1, \ldots, N \,:\, x_H \geq H\right\}$$

is a measure that aims to take into account not only the overall quality of the papers but also their number. Simply put, an author has the h-index equal to H, if H of their N papers have at least H citations each, and the other $N - H$ papers have no more than H citations each.

Over the years, many extensions of the h-index have been proposed. Some of them have been inspired by a similar idea, like for example the g-index (Egghe, 2006), given by

$$g(\mathbf{x}) = \max\left\{G = 1, \ldots, N \,:\, \sum_{i=1}^{G} x_i \geq G^2\right\}.$$

Note that the g-index is the greatest number G such that the top G articles received at least G^2 citations altogether.

Further, the w-index (Woeginger, 2008), defined as

$$w(\mathbf{x}) = \max\left\{W = 1, \dots, N : \ x_i \geq W - i + 1 \text{ for all } i \leq W\right\}$$

is the length of the side of the largest isosceles right-angle triangle that can be fit under the citation curve (note that the h-index is the side of the largest square, compare (Gagolewski and Grzegorzewski, 2009)).

Some authors proposed further measures that either employed the h-index directly or were computed solely based on the so-called h-core (Burrell, 2007b; Alonso et al., 2009) – the set of the $h(\mathbf{x})$ most cited papers, $x_1, x_2, \dots, x_{h(\mathbf{x})}$. Let us review some examples of such an approach.

Alonso et al. (2010) proposed the use of the geometric mean of the h- and the g-index, to introduce some balance between these two, called the hg-index,

$$hg(\mathbf{x}) = \sqrt{h(\mathbf{x})\, g(\mathbf{x})}.$$

The o-index (Dorogovtsev and Mendes, 2015) is the average between the h-index and the top-cited paper's value, i.e.,

$$o(\mathbf{x}) = \sqrt{h(\mathbf{x})\, x_1}.$$

Here, we note the high dependency of this measure upon just a single observation from the citation vector, which results in the indicator's high variance.

Next, the r-index (Jin et al., 2007) takes the square root of the sum of the h-core,

$$r(\mathbf{x}) = \sqrt{\sum_{i=1}^{h(\mathbf{x})} x_i}.$$

We also analyse the $i10$-index (which is used in, amongst others, the Google Scholar database), being the number of papers with at least 10 citations:

$$i10(\mathbf{x}) = \max\{i = 1, \dots, N \ : \ x_i \geq 10\}.$$

Finally, the s-index, is simply the sum of logarithms of all citations

$$s(\mathbf{x}) = \sum_{i=1}^{N} \log(x_i + 1).$$

This index serves as the maximum likelihood estimator in the Pareto distribution family, see, e.g., Arnold (2015).

Of course, uncountably many impact indices or combinations thereof can be introduced, see, e.g., (Gagolewski and Mesiar, 2014; Egghe and Rousseau, 2020, 2021). Nevertheless, the above are a quite representative sample of the most notable ones, and we believe they form a solid basis for an interesting investigation that we shall perform below.

2.2. Citation models

Whilst the focus of impact indices is often on providing an interpretable numerical summary, citation models aim to recreate the original vectors in their entirety, using a few underlying parameters. Below we shall review some noteworthy models whose common feature is that they are *solvable*, i.e., we are able provide compact formulas for the predicted number of citations to the k-th most cited paper, denoted by $\hat{x}_k$, $k = 1, \dots, N$, for a given N.

Historically, the power law was amongst the first models of this kind (Sec. 2.2.1). It can be generalised in a few different ways, e.g., in the form of the discrete generalised beta distribution (DGBD, Sec. 2.2.3). Another possible generator is based on the survival function of the log-normal distribution (Sec. 2.2.2). We will treat them as a baseline for the evaluation of our recently proposed 3DSI model and its approximations (Sec. 2.2.4).

Note that while many other models exist in the literature, e.g., (Burrell, 2007a, 2014; Egghe and Rousseau, 2006; Ionescu and Chopard, 2013; Żogała-Siudem et al., 2016; Malesios, 2015; Schubert and Schubert, 2019), many of them are non-solvable in terms of providing an analytic formula for the rank-size distribution, depend on way too many parameters, or are too similar to the ones discussed herein.

2.2.1. Power-law model

The observation that order statistics for real citation data approximately scale according to a power function (a straight line on the log-log-scale) is often referred to as the Zipf, Lotka, or power law (Newman, 2005; Egghe and Rousseau, 2006; Egghe, 2009; Egghe et al., 2009; Egghe and Rousseau, 2012; Schubert and Schubert, 2019). This yields the following formula for the number of citations to the k-th most cited paper

$$\hat{x}_k^{\text{PowerLaw}}(N, \alpha, \gamma) = \frac{\gamma}{k^\alpha}, \quad \text{where } \gamma > 0,\ \alpha > 0. \tag{1}$$

In practice, the scale parameter $\gamma > 0$ and the exponent $\alpha > 0$ need to be estimated from data, see below for discussion.

2.2.2. Log-normal model

Despite the fact that the power law provides a nice first approach to the modelling of citation distributions, various papers pointed out its limitations (Eom and Fortunato, 2011; Thelwall and Wilson, 2014; Thelwall, 2016a,b; Néda et al., 2017; Brito and Navarro, 2021). Let us note, however, that most of these works (with the exception of (Brito and Navarro, 2021)) focus on the citation distribution of the whole network. We, on the other hand, are dealing with the subset restricted to a considered author (see Fig. 1). Keeping that in mind, let us derive a rank distribution of x_k, assuming that citations follow a log-normal distribution, see (Thelwall, 2016a,b).

Let the survival function of the shifted log-normal distribution be denoted with

$$S_{l,\mu,\sigma}(x) = 1 - \Phi\left(\frac{\log(x-l) - \mu}{\sigma}\right),$$

where Φ is the cumulative distribution function of the standard normal distribution with parameters μ and σ, i.e., $\Phi(x) = \int_{-\infty}^{x} e^{-t^2/2}/\sqrt{2\pi}\, dt$. Additionally, for the sake of the increased model's flexibility, l is the location shift. In other words, if $\log X$ is a random variable following $N(0, \sigma)$ and $Y = (\log X - l)/e^{\mu}$ is its shifted and scaled version, then $S_{l,\mu,\sigma}(x) = \Pr(Y > x)$. Then, the estimates of the number of citations of the k-th most cited paper can be determined based on the inverse of survival function $S_{l,\mu,\sigma}$

$$\hat{x}_k^{\text{LogNorm}}(N, l, \mu, \sigma) = S^{-1}_{l,\mu,\sigma}\left(\frac{k}{N+1}\right), \quad \text{where } l, \mu \in \mathbb{R}, \sigma > 0. \tag{2}$$

Please note that even though such an approach relies upon some well known objects from probability theory, one cannot presume that in the real world the citations are independent and randomly distributed. However, the above approximation can sometimes work surprisingly well, see (Brito and Navarro, 2021).

2.2.3. DGBD model

Another approach, proposed in (Petersen et al., 2011), involves the use of the discrete generalised beta distribution (DGBD) which has been observed to fit a wide range of data types well, see, e.g., (Naumis and Cocho, 2008; Martínez-Mekler et al., 2009; Margellou and Pomonis, 2021; Ghosh et al., 2021). In case of citation vectors, we can express the number of citations of the k-th most-cited paper as:

$$\hat{x}_k^{\text{DGBD}}(N, A, a, b) = A\frac{(N+1-k)^b}{k^a}, \quad \text{where } A > 0,\ a > 0,\ b \geq 0. \tag{3}$$

Note that the above reduces to Eq. 1 when $b = 0$, and therefore can be seen as a generalisation of the power law. The increased flexibility (note the additional parameter), allows for a better fit to highly cited papers as well as to the tail of the empirical distribution.

2.2.4. 3D model for scientific impact (3DSI)

Another approach to citation vectors' modelling is presented by the 3DSI model which we have recently proposed in (Siudem et al., 2020). The idea standing behind this model relies upon an assumption that citations of a given author are distributed amongst their papers through two mechanisms: due to the preferential attachment rule and purely at random. The model depends on three interpretable parameters: the number of published papers N, the overall number of citations C and the preferential-to-accidental ratio $\rho \in (0, 1)$. The agent-based model is built in N steps: starting from an empty citation vector, in each step a new paper is added and $\frac{C}{N}$ citations are being distributed. Each time

$(1-\rho)\frac{C}{N}$ citations are allocated at random and the remaining $\rho\frac{C}{N}$ ones due to the rich-get-richer rule, see (Siudem et al., 2020) for more details.

The estimated number of citations of the k-th most cited paper is then given by

$$\hat{x}_k^{\text{3DSI}}(N,C,\rho) = \frac{1-\rho}{\rho}\frac{C}{N}\left(\prod_{i=k}^{N}\frac{i}{i-\rho}-1\right) = \frac{1-\rho}{\rho}\frac{C}{N}\left[\frac{\Gamma(k-\rho)}{\Gamma(k)}\frac{\Gamma(N+1)}{\Gamma(N+1-\rho)}-1\right], \tag{4}$$

where $\rho \in (0,1)$ and Γ denotes the gamma function (i.e., the Euler integral of the second kind), $\Gamma(x) = \int_0^\infty t^{x-1}\,e^{-t}\,dt$, $x > 0$. The advantage of this model is that, as we have stated above, all its underlying parameters are easily interpretable, with C denoting the total number of citations, $C = \sum_{k=1}^{N}\hat{x}_k$, and ρ controlling the degree of the skewness of the citation distribution.

Let us now derive two different approximations to this model. Note that the gamma functions appearing in Eq. 4 can be expanded asymptotically via the well-known asymptotic relation described in (Gautschi, 1959)

$$\frac{\Gamma(n-p)}{\Gamma(n)} \approx n^{-p}. \tag{5}$$

Firstly, let us replace $\frac{\Gamma(N+1)}{\Gamma(N+1-\rho)}$ (after an appropriate transformation of Eq. 5) by $N^\rho \approx (N+1)^\rho$. Note that this formula is quite accurate even for relatively small N (e.g., for $N = 5$ the relative absolute error between N^ρ and $\frac{\Gamma(N+1)}{\Gamma(N+1-\rho)}$ does not exceed 10^{-4} for any $\rho \in (0,1)$). This results in the first approximation, i.e.,

$$\hat{x}_k^{\text{3DSI-app1}}(N,C,\rho) = \frac{1-\rho}{\rho}\frac{C}{N}\left[\frac{\Gamma(k-\rho)}{\Gamma(k)}N^\rho-1\right]. \tag{6}$$

Next, we can use the same approach to simplify the second gamma function, i.e., $\Gamma(k)$. Even though this transformation will be problematic for small values of k, it gives a nice, power-law-like equation

$$\hat{x}_k^{\text{3DSI-app2}}(N,C,\rho) = \frac{1-\rho}{\rho}\frac{C}{N}\left[\left(\frac{N}{k}\right)^\rho-1\right]. \tag{7}$$

We will validate the quality of these approximations below.

2.2.5. Harmonic model

Another approximation of the 3DSI model stems from the study of the limiting behaviour of Eq. 4, i.e., when $\rho \to 0$. It corresponds to the situation where no preferential factor is involved in the underlying random process. Interestingly, when it comes to real data, this phenomenon occurs quite often (empirically, see Sec. 3.3). Let

$$\hat{x}_k^{\text{Harmonic}}(N,C) = \frac{C}{N}\sum_{i=k}^{N}\frac{1}{i}. \tag{8}$$

From now on, we shall refer to the above as the harmonic model. Note that even when $\rho \approx 0$, the citation vector is still moderately skewed as older papers are being cited more often anyway. The preferential part is however needed to increase the distribution's skewness even further; it will turn out (see Sec. 3) that this harmonic model cannot capture the nature of many of the citations vectors well.

2.3. Parameter estimation

In order to be able to identify the parameters of a model that reproduces a given citation vector $x_1,\ldots,x_N$ best, an appropriate penalty function should be chosen. Due to the skewed (compare the rich-get-richer rule) nature of the citation vectors, the fitting will be applied with regards to the logarithms of the inputs so as to avoid over-fitting to the highly cited works.

In order for this to be feasible (recall that $\log 0 = -\infty$), we need to omit all the uncited papers from the analysis. Therefore, from now on we shall assume that $x_N > 0$. Note that such a removal does not influence the values of the bibliometric indices studied herein, with the exception of the g-index. Of course, another option would be to rely on fitting to $\log(x_1+1),\ldots,\log(x_N+1)$, similarly as in (Thelwall, 2016b). Overall, our preliminary analyses have

indicated that the models are oftentimes more accurate when actually N denotes the number of papers that have been referenced at least once, and not just the total size of an author's output.

Let $\hat{x}_1^M(\boldsymbol{\theta}), \ldots, \hat{x}_N^M(\boldsymbol{\theta})$ be a citation vector generated by means of a model M with p parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. We will assume that N (the number of papers) and C (the total number of citations) are always determined directly from the sample and are never optimised for. Therefore, for instance, in the 3DSI model we have $p = 3$, $\boldsymbol{\theta} = (N, C, \rho)$, but the search space has only 1 degree of freedom, $\Theta = \{N\} \times \{C\} \times (0, 1)$. Also, in the harmonic model, $p = 2$ but $\Theta = \{N\} \times \{C\}$ has no free parameters at all.

The cost function to minimise over $\boldsymbol{\theta} \in \Theta$ has the form:

$$\frac{1}{N} \sum_{k=1}^{N} \ell \left(\left(\log \hat{x}_k^M(\boldsymbol{\theta}) - \log x_k \right)^2 \right), \tag{9}$$

where $\ell : [0, \infty) \to [0, \infty)$ is some loss function.

In what follows, the minimisation is performed numerically by means of `optimize.least_squares` from the `scipy` package (Virtanen et al., 2020) for Python, using the Trust Region Reflective algorithm (based on Branch et al., 1999). We considered the following loss functions:

- standard least squares (L_2 loss):
  - *linear*: $\ell(z) = z$,
- smooth approximations to L_1 losses:
  - *soft_l1*: $\ell(z) = 2\sqrt{1 + z} - 2$,
  - *Huber*: $\ell(z) = z$ for $z \geq 1$ and $2\sqrt{z} - 1$ otherwise,
- to weaken the influence of the outliers:
  - *Cauchy*: $\ell(z) = \log(1 + z)$,
  - *arctan*: $\ell(z) = \arctan z$.

Note that Thelwall (2016b) used ordinary least squares regression (corresponding to $\ell(z) = z$, *linear* above) applied over $\log(x_k + 1)$. We shall restrict ourselves to the *Cauchy* loss ($\ell(z) = \log(1 + z)$), as this was our choice also in (Siudem et al., 2020), and it is generally more robust. Overall, however, we have observed that the choice of the loss function does not significantly affect the average relative errors of the reproduced bibliometric indices – which in fact is a metric of our focus in the next section.

Furthermore, to lessen the risk of getting stuck in a suboptimal solution, we shall employ 5 restarts from random initial guesses and choose the solution that yields the smallest objective.

Figures 2 and 3 depict two example citation vectors along with some models minimising the cost function based on the Cauchy loss where, respectively, we observe a quite good and a pretty bad fit.

3. Results

3.1. Data

Empirical analysis presented in this paper was conducted upon the DBLP v12 bibliography database (Tang et al., 2008), see `https://aminer.org/citation`. The DBLP database consists of 4,398,138 citation vectors of computer science authors. We have used the DBLP author IDs, however, it is possible that a single researcher has been assigned multiple IDs – no author disambiguation algorithm is completely accurate.

Most authors have a small number of papers with few citations. Therefore, to enable a proper model construction and validation, we have restricted our analysis to the subset of researchers who published at least 5 cited papers and their $i10$-index was greater than 0. This resulted in $\eta = 348{,}941$ citation records representing $\approx 8\%$ of the original database. Also, as stated above, we have omitted all papers with 0 citations, as they are problematic when performing computations on the log-scale.

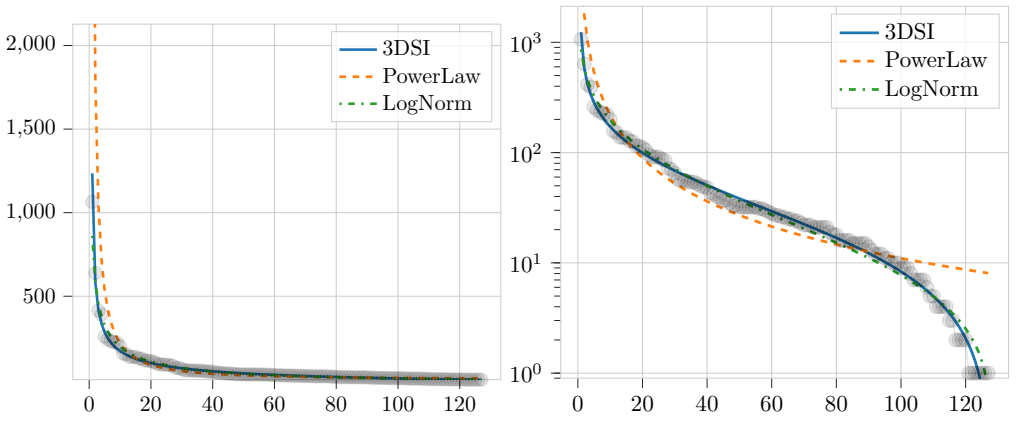

Figure 2: An example citation vector (citations to the k-th most cited paper x_k as a function of paper index k) to which the 3DSI and log-normal models yield a good fit (left: citation counts on a linear scale; right: on a log scale).

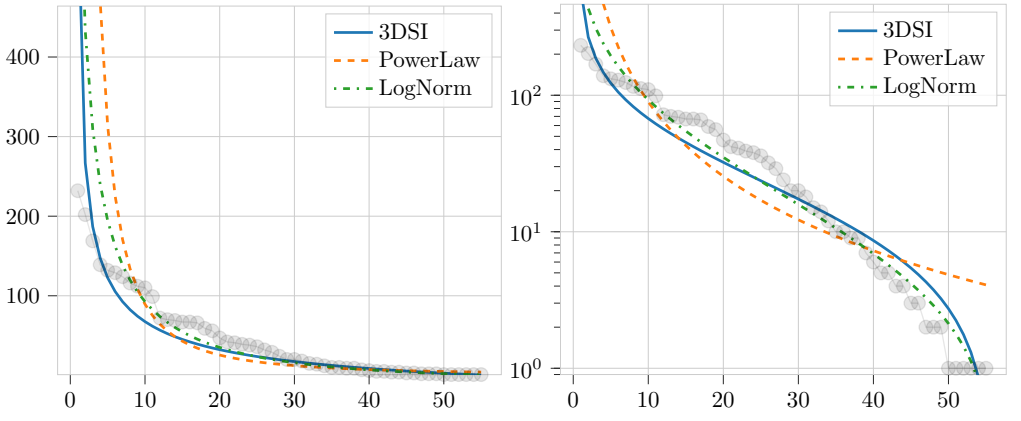

Figure 3: An example citation vector where the fitted models poorly reflect the underlying data.

3.2. Fitting models to data

As we have mentioned in Sec. 2.2, the models studied herein differ considerably not only in the number of parameters that describe them, but also in their interpretability. Both the DGBD and log-normal models require fitting of 3 additional parameters (recall that N is taken directly from the input sample), which makes them the most complex ones and increases the likelihood that they will fit the citation vectors better – as a general rule, we can expect that the more degrees of freedom a model has, the greater its flexibility.

The harmonic model, on the other hand, relies only on N and C that are directly taken from the sample – no further optimisation is made. We may expect that this model will fit real-world data poorly, unless they are indeed generated by a process resembling the theoretical one.

The 3DSI model and its approximations result from a particular citation generation process (in each iteration, a paper is added to a simulated author's output and C/N citations are distributed amongst the existing papers according to a weighted combination of the (ρ) rich-get-richer rule and ($1-\rho$) sheer chance). Thanks to this, all their parameters are well-interpretable. Recall that these models require optimising with regard to just one parameter, ρ.

For each citation vector $\mathbf{x}_i$, $i = 1, \ldots, \eta$, and every citation model M, we have applied the optimisation procedure described above in order to identify the model parameters $\hat{\boldsymbol{\theta}}_i^M$ that minimise the cost function given by Eq. 9. This way, we have determined the best fitting models.

The box and whisker plots in Fig. 4 depict the empirical distribution of the values of the cost function (Eq. 9 with ℓ being the Cauchy loss) at the identified minima for each model. As expected, on average, the DGBD and log-normal models give the best fits. This comes as no surprise as these models have three degrees of freedom, therefore they

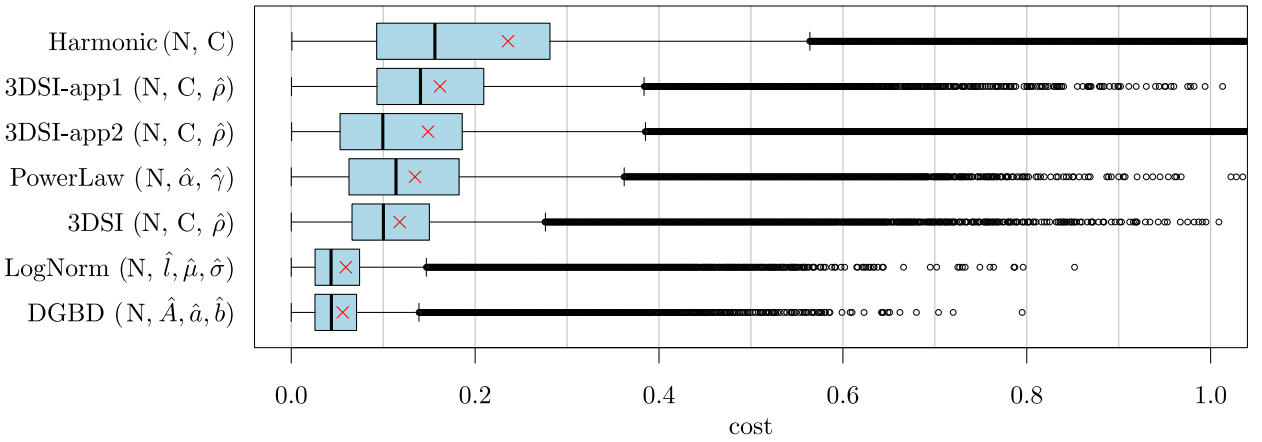

Figure 4: Box plots of the cost function values (Eq. 9 with the Cauchy loss) across all citation vectors for each model. Data are sorted with respect to the average costs (marked with ×). As expected, the two models with the greatest number of free parameters are the most flexible, and hence they yield the best fits. However, as we argue below, a good *overall* fit to the whole vector does not necessarily allow for an accurate prediction of bibliometric indices and vice versa.

can be expected to be able to adapt to the citation curves quite well. Not so far behind we can find the 3DSI model followed by its two approximations and the power-law model. The worst alignments were obtained for the harmonic model which, however, was not fitted using any numerical optimisation of the cost function.

3.3. Predicting indices

Fig. 4 focuses on the *overall* discrepancy between the citation vectors and models thereof. As we mentioned earlier, in practice it is difficult to define how a "good fit" should be identified, though. Should a small prediction error, averaged across all individual papers be preferred over a close fitting to the tail of the distribution (which often consists of the newest papers that are yet to be cited) or to a few most cited papers (which naturally are subject to the most variability)?

Looking only at the value of the cost function, it is not possible to determine to which part of the vector the curves fit best. However, above we have reviewed some metrics that aggregate the whole citation curve into a single number that reflects one of its (possibly many) characteristics. Let us then study how well the fitted models enable us to predict these numerical summaries.

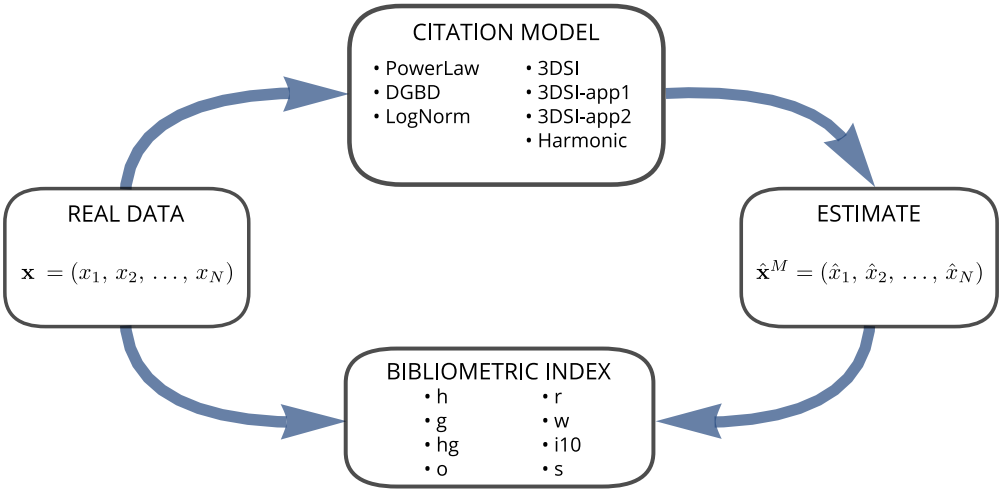

Figure 5: Real vs. estimated citation vectors and the corresponding bibliometric indices.

The main idea behind our analysis is presented in Fig. 5. We take a citation vector $\mathbf{x}_i$ and compute the true values of each citation index $I \in \{h, g, hg, o, r, w, s, i10\}$. Then, for each model M we identify the parameter vector $\hat{\boldsymbol{\theta}}_i^M$ that minimises the loss function. Based on these estimates, we build the predicted citation vector $\hat{\mathbf{x}}^M(\hat{\boldsymbol{\theta}}_i^M)$ and compute the predicted bibliometric indices.

In order to compare the results obtained, we shall use the relative estimation error of an index I for the i-th citation vector and the best fitted model M,

$$E_i(I, M) = \frac{\left| I\left(\mathbf{x}_i\right) - I\left(\hat{\mathbf{x}}^M\left(\hat{\boldsymbol{\theta}}_i^M\right)\right)\right|}{\left|I\left(\mathbf{x}_i\right)\right|}, \tag{10}$$

which is the absolute difference between the true value of the bibliometric index (taken directly from the citation vector) and the predicted index (computed based on the best fitted model), scaled by the index size. For example, a relative error of 0.05 means that the predicted index differs by 5% from the actual one. Of course, an ideally fitted model yields an estimation error of 0.

Fig. 6 gives the arithmetic means and medians of the relative estimation errors of all the considered indices aggregated over all η vectors, for every bibliometric index I and citation model M. The models have been ordered by the value of the relative error averaged over all the indices. Below we discuss the results in very detail.

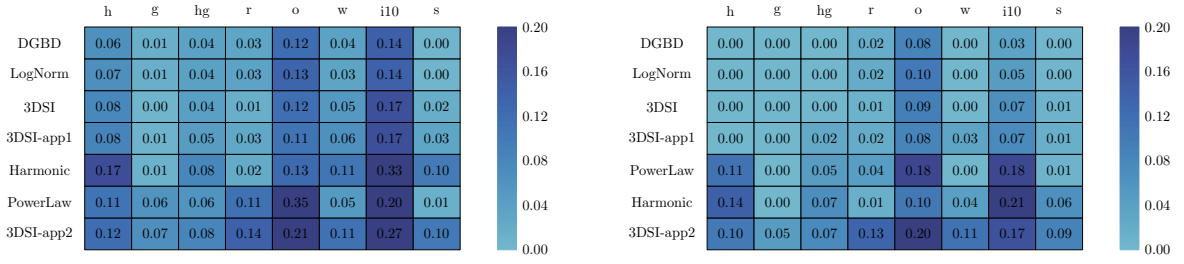

Figure 6: Average (left) and median (right) relative estimation errors of different citation indices for different models. DGBD, log-normal, 3DSI, and the first approximation thereof have a similar predictive performance.

3.3.1. 3DSI, DGBD, and log-normal models

The top three models: 3DSI, DGBD, and log-normal, give a very similar error profile (with the 4th model closely following them) and seem to be indistinguishable, yielding better results for a few metrics and slightly worse for others.

Let us start with a comment on the h-index as it is the most popular metric used to assess the quality of scientific impact. The average relative errors for the 3 top models were similar, ca. 7–8%. The left side of Fig. 7 presents a direct comparison of these models, showing the bands ranging from the 5th to the 95th percentile of the predicted $h(\hat{\mathbf{x}})$ as a function of the observed (true) $h(\mathbf{x})$. The plot includes only the vectors with $h(\mathbf{x}) \leq 90$ as outside this interval the relevant data points become too scarce.

The 3DSI model slightly underestimates the h-index in cases where it is high: the blue area tends to lie below the identity line. For the DGBD model, 60% of vectors, mostly the shorter ones, reproduce the h-index perfectly, i.e., they yield $h(\mathbf{x}) - h(\hat{\mathbf{x}}) = 0$. The accuracy of the log-normal model reaches 58% and for the 3DSI model it is equal to 56%. Overall, we can say that all the three models tend to capture the characteristic of the citation vector that the h-index measures quite well.

Our attention is particularly drawn to the surprisingly good prediction of the g-index and s-index values. In fact, the 3DSI model gives the relative error for the g-index prediction of just 0.008 (less then 1%), with 94% of vectors reproducing this metric exactly. The two other models give the relative error of ca. 0.015 with ca. 85% of the values fitted perfectly, see the right side of Fig. 7 for a comparison between $g(\mathbf{x})$ and $g(\hat{\mathbf{x}})$. Note that the blue 5th-95th percentile band, corresponding to the 3DSI model, is in fact practically invisible due to the aforementioned high predictive capabilities. On the other hand, the DGBD model tends to slightly overestimate the g-index, whereas the log-normal

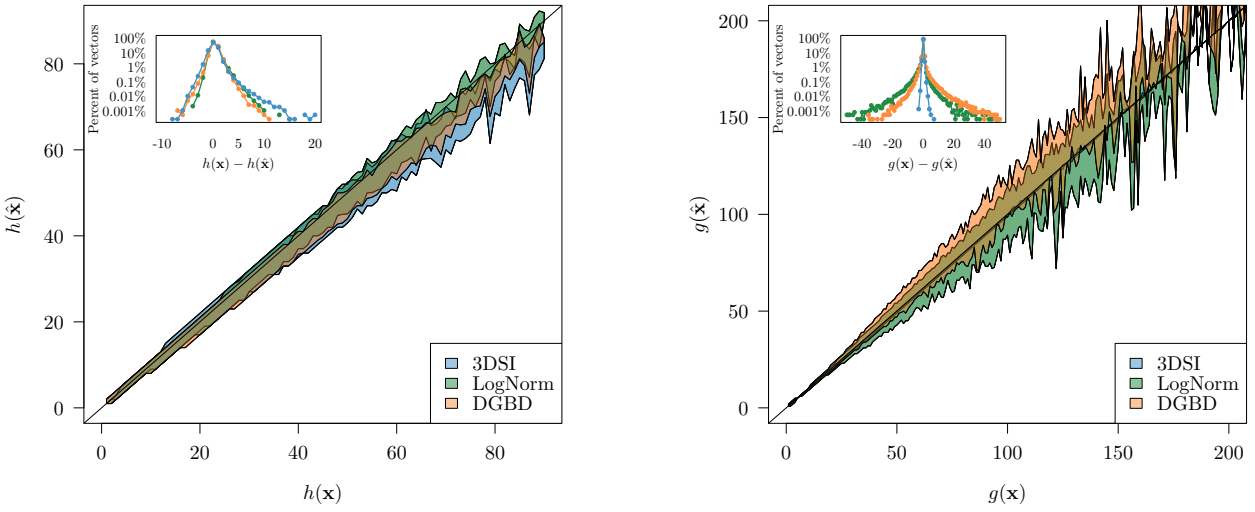

Figure 7: Predicted vs. actual indexes. For each fixed h and g, the interval where 90% of the predicted $\hat{h}$ and $\hat{g}$ fall is depicted. The insets show the empirical distributions of errors (note the logarithmic scale). The 3DSI model slightly underestimates the larger h-indices, but it has a near-perfect accuracy for the g-index.

one underestimates it. As for the s-index, the fit is very good for almost all models, with DGBD, Lognorm, and 3DSI yielding 0.003, 0.004, and 0.015, respectively. This might partially be explained by the fact that the optimisation of model parameters is done on the logarithms of citation vectors, thus the index itself is being optimised here as well.

Almost equally high precision is observed for the r-index, where the average relative errors are equal to 1.9%, 3.3%, 3.5% for the 3DSI, log-normal, and DGBD models, respectively. Contrary to the g-index, however, the values of this measure are not discrete (i.e., natural numbers), thus one should not expect perfect fits. Very low relative errors are also obtained for the hg-index (which is not surprising as it depends on the g-index) and the w-index.

The worst fits were obtained for the i10- and o-indices, and this observation applies to all the models. We should note that one of the reasons for this is the skewness of the real vectors as that those measures focus only on a few high-cited papers whose exact values are not predicted well. What is more, the o-index highly depends on one particular value in the citations vector which as such is prone to high variability. Even though the o-index combines the most cited paper with another metric, the h-index, the high variance of x_1 still has a huge impact on its properties. Recall that the fitting procedure is performed on the log scale and it is based on the Cauchy loss which damps down the influence of any potential outliers, so that x_1 does not have a huge influence during the optimisation process.

3.3.2. Power-law and approximations to 3DSI

In this paper we are also interested in examining how good the approximations of the 3DSI model are. The relative error values for 3DSI-app1 in Fig. 6 are slightly worse than, but still comparable to, the ones obtained for the original 3DSI. This is in agreement with the above observation that this approximation is indeed very accurate even for relatively small values of N. We also compared the fitted values of the ρ parameter in the case of 3DSI and 3DSI-app1 and they were also similar.

The situation worsens in the case of 3DSI-app2, which suggests that despite the pleasing formula obtained for the citation vector, it does not reflect the real data well. The errors here are visibly larger. What is more, in the majority (82%) of fitting tasks, the cost function was minimised at $\rho \approx 0$, whereas for the 3DSI model this proportion was equal to 35%.

The power-law model yields results which are quite comparable to 3DSI-app2 in terms of the indices' prediction accuracy, which puts it amongst the relatively inaccurate data modelling techniques. Note also that it has two parameters optimised for instead of one for all the 3DSI variants.

3.3.3. Harmonic model

As mentioned above, in many cases the optimal value for the ρ parameter was 0. Recall that $\rho \approx 0$ corresponds to a situation where the influence of the preferential component is neglected and the focus is on the accidental part. The

proposed Harmonic model represents exactly this case. Here, both N and C are directly taken from the sample, with no numerical optimisation in place.

Unfortunately, this model gave the worst fits to the citation vectors (Fig. 4). Surprisingly, though, it predicts the g- and the r-index quite well, but the h-index particularly badly. Seeking the reasons for such a behaviour is out of scope for this paper and we leave it as an interesting question for future research.

4. Conclusion

Overall, a good model is the one that not only fits data well *on average*, but also agrees on the essential data characteristics, aggregates, or numerical summaries. The 3DSI, DGBD, and log-normal models have a similar accuracy, but it is the recently-proposed 3DSI (Siudem et al., 2020) that is the simplest:

- it has the smallest number of adjustable parameters (ρ),
- each parameter has an intuitive interpretation (number of papers N, number of citations C, ratio of preferentially-to-accidentally attached citations ρ which controls the skewness of the resulting distribution).

Note that two of the considered models have an unambiguous microscopic interpretation:

- the power-law model can be seen as stemming from the assumption that all citations are subject to preferential attachment,
- in the harmonic model every new citation is distributed in a purely accidental manner.

There are indices (g, r, o) which are better reproduced based only on the accidental-harmonic mechanism, while others (h, w) better fit under the preferential-power-law one.

Finally, note that each bibliometric index focuses on a slightly different aspect of a citation vector, therefore there is some inherent variability in how well it can be reproduced by the citation models. We have observed the following regularities:

- usually the best fitted models give better fits for the bibliometric indices, however there are some exceptions: see, e.g., the power-law model and the w-index,
- the indices g, s, and r (and to a lesser degree w) are reproduced surprisingly well, even by the models with far from perfect overall fits (harmonic, 3DSI-app2, power-law), the reasons behind this should be further investigated in the future,
- some models behave particularly poorly on certain indices (e.g., h or w),
- it is generally difficult to pinpoint the top-cited papers, which suffer from high variance, hence the performance of the i10- and o- indices is low.

Acknowledgement

The project was partially funded (AC and GS) by the POB Research Centre Cybersecurity and Data Science of Warsaw University of Technology within the Excellence Initiative Program – Research University (ID-UB). This research was also supported (MG) by the Australian Research Council Discovery Project ARC DP 210100227. We thank the reviewers for their useful remarks that helped improve the manuscript.

CRediT authorship contribution statement

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

Battiston, F., Musciotto, F., Wang, D., Barabási, A.L., Szell, M., Sinatra, R., 2019. Taking census of physics. Nature Reviews Physics 1, 89–97.

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.

Branch, M., Coleman, T., Li, Y., 1999. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. SIAM Journal on Scientific Computing 21, 1–23.

Brito, R., Navarro, A.R., 2021. The inconsistency of h-index: A mathematical analysis. Journal of Informetrics 15, 101106. doi:10.1016/j.joi.2020.101106.

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

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

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

Chen, P., Redner, S., 2010. Community structure of the physical review citation network. Journal of Informetrics 4, 278–290. doi:10.1016/j.joi.2010.01.001.

Dorogovtsev, S., Mendes, J., 2015. 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., Liang, L., Rousseau, R., 2009. A relation between h-index and impact factor in the power-law model. Journal of the American Society for Information Science and Technology 60, 2362–2365. doi:10.1002/asi.21144.

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

Egghe, L., Rousseau, R., 2012. Theory and practice of the shifted Lotka function. Scientometrics 91, 295–301.

Egghe, L., Rousseau, R., 2020. Polar coordinates and generalized h-type indices. Journal of Informetrics 14, 101024. doi:10.1016/j.joi.2020.101024.

Egghe, L., Rousseau, R., 2021. The h-index formalism. Scientometrics doi:10.1007/s11192-020-03699-9.

Eom, Y.H., Fortunato, S., 2011. Characterizing and modeling citation dynamics. PLOS ONE 6, 1–7. doi:10.1371/journal.pone.0024926.

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

Gagolewski, M., Mesiar, R., 2014. Monotone measures and universal integrals in a uniform framework for the scientific impact assessment problem. Information Sciences 263, 166–174. doi:10.1016/j.ins.2013.12.004.

Gautschi, W., 1959. Some elementary inequalities relating to the gamma and incomplete gamma function. Journal of Mathematics and Physics 38, 77–81. doi:10.1002/sapm195938177.

Ghosh, A., Shreya, P., Basu, B., 2021. Maximum entropy framework for a universal rank order distribution with socio-economic applications. Physica A: Statistical Mechanics and its Applications 563, 125433. URL: https://www.sciencedirect.com/science/article/pii/S0378437120307603, doi:https://doi.org/10.1016/j.physa.2020.125433.

Golosovsky, M., 2019. Citation Analysis and Dynamics of Citation Networks. Springer.

Golosovsky, M., Solomon, S., 2012. Stochastic dynamical model of a growing citation network based on a self-exciting point process. Phys. Rev. Lett. 109, 098701.

Guevara, M.R., Hartmann, D., Aristarán, M., Mendoza, M., Hidalgo, C.A., 2016. The research space: Using career paths to predict the evolution of the research output of individuals, institutions, and nations. Scientometrics 109, 1695–1709.

Havemann, F., Larsen, B., 2015. Bibliometric indicators of young authors in astrophysics: Can later stars be predicted? Scientometrics 102, 1413–1434.

Herrera, M., Roberts, D.C., Gulbahce, N., 2010. Mapping the evolution of scientific fields. PloS one 5, e10355.

Hirsch, J.E., 2005. An index to quantify individual's scientific research output. Proceedings of the National Academy of Sciences 102, 16569–16572.

Ionescu, G., Chopard, B., 2013. An agent-based model for the bibliometric h-index. Eur. Phys. J. B 86, 426.

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.

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 66, 2384–2388. doi:10.1002/asi.23410.

Margellou, A.G., Pomonis, P.J., 2021. Beyond Zipf's law: Pore ranking in solids by beta distributions. Microporous and Mesoporous Materials 317, 110987. URL: https://www.sciencedirect.com/science/article/pii/S138718112100113X, doi:https://doi.org/10.1016/j.micromeso.2021.110987.

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.

Mingers, J., Yang, L., 2017. Evaluating journal quality: A review of journal citation indicators and ranking in business and management. European Journal of Operational Research 257, 323–337. doi:10.1016/j.ejor.2016.07.058.

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:10.1016/j.physa.2007.08.002.

Néda, Z., Varga, L., Biró, T.S., 2017. Science and Facebook: The same popularity law! PLOS ONE 12, 1–11. doi:10.1371/journal.pone.0179656.

Newman, M.E., 2005. Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46, 323–351.

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.

Price, D.J., 1963. Little science, big science. Columbia Univ. Press, New York.

Risi, J., Sharma, A., Shah, R., Connelly, M., Watts, D.J., 2019. Predicting history. Nature Human Behaviour 3, 906–912.

Schubert, A., Schubert, G., 2019. All along the h-index-related literature: A guided tour, in: Glänzel, W., Moed, H.F., Schmoch, U., Thelwall, M. (Eds.), Springer Handbook of Science and Technology Indicators. Springer International Publishing, Cham, pp. 301–334. doi:10.1007/978-3-030-02511-3_12.

Shen, H.W., Barabási, A.L., 2014. Collective credit allocation in science. Proceedings of the National Academy of Sciences 111, 12325–12330.

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.

Siwinski, W., Holmes, R., Kopanska, J., 2021. IREG Inventory of International University Rankings. IREG Observatory on Academic Ranking and Excellence. Warsaw-Brussels. URL: www.ireg-observatory.org/en/inventory-international-rankings.

Sziklai, B.R., 2021. Ranking institutions within a discipline: The steep mountain of academic excellence. Journal of Informetrics 15, 101133. doi:10.1016/j.joi.2021.101133.

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., 2016a. Are the discretised lognormal and hooked power law distributions plausible for citation data? Journal of Informetrics 10, 454–470. doi:10.1016/j.joi.2016.03.001.

Thelwall, M., 2016b. 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:10.1016/j.joi.2015.12.007.

Thelwall, M., Wilson, P., 2014. Distributions for cited articles from individual subjects and years. Journal of Informetrics 8, 824–839. doi:10.1016/j.joi.2014.08.001.

Virtanen, P., et al., 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17, 261–272. doi:10.1038/s41592-019-0686-2.

Waltman, L., 2016. A review of the literature on citation impact indicators. Journal of Informetrics 10, 365–391. doi:10.1016/j.joi.2016.02.007.

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

Wuchty, S., Jones, B.F., Uzzi, B., 2007. The increasing dominance of teams in production of knowledge. Science 316, 1036–1039.

Żogała-Siudem, B., Siudem, G., Cena, A., Gagolewski, M., 2016. Agent-based model for the h-index – Exact solution. European Physical Journal B , 21.

Please cite this paper as:
A. Cena, M. Gagolewski, G. Siudem, B. Żogała-Siudem,
Validating citation models by proxy indices, Journal of Informetrics 16(2), 101267, 2022,
doi:10.1016/j.joi.2022.101267