

# Aggregating different paper quality measures with a generalized $h$ -index

Marek Gagolewski<sup>\*,a,b</sup>, Radko Mesiar<sup>c,d</sup>

<sup>a</sup>*Systems Research Institute, Polish Academy of Sciences*

*ul. Newelska 6, 01-447 Warsaw, Poland; Email: gagolews@ibspan.waw.pl*

<sup>b</sup>*Faculty of Mathematics and Information Science, Warsaw University of Technology  
pl. Politechniki 1, 00-661 Warsaw, Poland*

<sup>c</sup>*Faculty of Civil Engineering, Department of Mathematics, Slovak University of Technology,  
81 368 Bratislava, Slovakia; Email: mesiar@math.sk*

<sup>d</sup>*Institute for Research and Applications of Fuzzy Modelling, University of Ostrava  
701 03 Ostrava, Czech Republic*

---

## Abstract

The process of assessing individual authors should rely upon a proper aggregation of reliable and valid papers' quality metrics. Citations are merely one possible way to measure appreciation of publications.

In this study we propose some new, SJR- and SNIP-based indicators, which not only take into account the broadly conceived popularity of a paper (manifested by the number of citations), but also other factors like its potential, or the quality of papers that cite a given publication. We explore the relation and correlation between different metrics and study how they affect the values of a real-valued generalized  $h$ -index calculated for 11 prominent scientometricians. We note that the  $h$ -index is a very unstable impact function, highly sensitive for applying input elements' scaling.

Our analysis is not only of theoretical significance: data scaling is often performed to normalize citations across disciplines. Uncontrolled application of this operation may lead to unfair and biased (towards some groups) decisions. This puts the validity of authors assessment and ranking using the  $h$ -index into question. Obviously, a good impact function to be used in practice should not be as much sensitive to changing input data as the analyzed one.

**Keywords:** aggregation operators, impact functions, Hirsch's  $h$ -index, quality control, scientometrics, bibliometrics, SJR, SNIP, Scopus, CITAN, R.

---

\* Corresponding author.

**This is a preprint version of the paper:**

Gagolewski M., Mesiar R., Aggregating Different Paper Quality Measures with a Generalized  $h$ -index, *Journal of Informetrics* **6**(4), 2012, pp. 566–579.

---

## 1. Introduction

The idea of applying citation analysis in scientific quality control was proposed more than 85 years ago [cf. 1]. Citations reflect the intensiveness of information use [cf. 2] and therefore may be conceived as manifestations of papers' recognition among the scientific community.

The need for assessment, ranking, or just indication of prominent individual authors appear in many contexts, e.g. in research policy, funding, and scientometrics, which aims for describing, explaining, or even predicting measurable features and characteristics of science and scientific research. Such a process classically bases on a proper aggregation of the citations number received by an author's publications. Thus, it uses some kind of combination of citations to obtain a single numeric value which is representative (in some sense) for the whole input.

Among the most popular citation indices we have the Hirsch index [3], which not only takes into concern the quality of individual papers but also their number. The so-called  $h$ -index is a symmetric, integer-valued function monotonic with respect to each aggregated variable, and also with respect to the length of the input vector [cf. 4–6, and also 7–12]. The *theory of aggregation* [cf. 13, 14], which is a rapidly-developing branch of applied mathematics, is by default concerned with algebraic properties of such operators, i.e. independent of the very nature of the input data. In real-world applications, however, we must not disregard neither the reliability nor validity of quantitative characteristics that are being summarized.

Obviously, citations are just one way to measure a paper's quality. Some other usable and important metrics may be, however, not integer-, but real-valued, or may have a different scale [cf. 15–18].

The article is organized as follows. In Sec. 2 we propose and discuss various new quality metrics of papers. They base upon SNIP and SJR indicators of journals in which the papers and/or their citations are published. In Sec. 3 we recall the notion of an impact function, i.e. an aggregation operator which may be used to summarize an author's papers quality metrics into a single, representative number. Additionally, we introduce the real-valued generalized  $h$ -index,  $H_s$ , that may be applicable to metrics with different scales. In Sec. 4 we present main empirical results; we explore the relation and correlation between different metrics, study

how they affect the rankings of authors in our sample, and check the stability of the proposed impact indices with respect to elements' scaling. Finally, Sec. 5 concludes the article.

## 2. Input Data

On January 27–February 1, 2012 we have manually gathered from Elsevier's *SciVerse Scopus* the publication output of 11 Derek de Solla Price Memorial medalists: L. Egghe, E. Garfield, W. Glänzel, P. Ingwersen, L. Leydesdorff, K.W. McCain, H.F. Moed, R. Rousseau, A.F.J. Van Raan, P. Vinkler, and M. Zitt. We have excluded from the survey scientifically inactive or deceased laureates, and those who have ambiguous author records (homonyms). The output of R.W. Rousseau (a chemist) was distinguished from R. Rousseau's. This gave 1240 documents. Then we fetched citations of all cited publications. As a result, we have obtained 9017 document records in total.

We used version SVN-1.01 of the CITAN package [19] for the R 2.15.0 statistical environment [20] to preprocess and analyze the data set.

In this paper we study the behavior of a generalized  $h$ -index used to aggregate different quality measures of papers (see Sec. 2.2) published by the 11 Price medalists. Some of the measures base on different journal quality metrics, which we review in the subsection to follow.

### 2.1. Journal Quality Metrics

Almost all commonly used nowadays journal quality metrics (abbreviated further on as JQM) are citation-based. The idea of applying citation analysis in scientific quality control dates back as far as 1927 [1, cf. also 21–23].

*Impact Factor.* An  $n$ -year Impact Factor [24], published yearly in Thomson's *Journal Citation Reports*, is the most popular measure of a journal's quality. It is defined as:

$$IF_n = \frac{\# \text{ citations to articles published in past } n \text{ years received in current year}}{\# \text{ citable papers published in past } n \text{ years}}.$$

An  $n = 2$ -year citation window is default, however some note that it is too short for disciplines in which citation impact matures slowly, which may partially be due to long publication delay, e.g. in mathematics [cf. 25, 26].

*h-index.* Just like for any author, the well-known Hirsch’s index [3] may also be calculated for journals [cf. 27]. In this approach we simply aggregate citations received by papers published in a given source during some time period. Please note that IF is an arithmetic mean, thus it is highly influenced by low- and very high-cited items. The *h-index* does not change its value if there is a high number of papers with small number of citations published in a journal; it may, on the other hand, be biased towards journals which issues appear more frequently.

The two above-mentioned tools have a very important drawback: their value depends on the citation practice in a journal’s domain. The next two journal metrics, SJR and SNIP, aim for assessing sources with citation intensity of their subject fields taken into account. They are published yearly by Elsevier and use a 3-year citation window.

*SCImago Journal Rank.* SJR [28] is inspired by the Google *PageRank* algorithm [29]. It is a prestige/influence metric based on the idea that citations have different importance; its recursive definition embodies the relative impact of sources citing a given journal. Interestingly, SJR of a journal in a highly-cited field is shared over a lot of citations, so each citation’s impact then is relatively small.

*Source-Normalized Impact per Paper.* SNIP, proposed by H.F. Moed [25], measures relative citation impact by weighting citations according to the total number of references in a subject field. It gives higher weights to citations in areas where referencing is less likely [cf. e.g. 30 for discussion]. This indicator is defined as

$$\text{SNIP} = \frac{\text{IF}_3}{\text{3-year citation potential}}.$$

SNIP is a ratio of the actual average citation rate of a journal’s papers, and the citation potential in the subject field covered by the journal. The citation potential expresses how frequently papers in the subject field cite other papers (by considering the reference lists’ lengths). The citation potential is defined as the average number of 1–3 year old number of references per source article citing a given journal normalized by dividing it by the citation potential of the median journal in the database. By applying this operation, 50% of journals has a (normalized) citation potential above one, and the other 50% below one. Compared to the IF, e.g. molecular biological and many other biomedical journals go down in the ranking and, on the other hand, mathematical, humanities, social and applied sciences journals go up.

In our study we use the 2011 SNIP and SJR, which are directly available in *Scopus* (data source: [www.info.sciverse.com/documents/files/scopus-training/resourcelibrary/xls/title\\_list.xlsx](http://www.info.sciverse.com/documents/files/scopus-training/resourcelibrary/xls/title_list.xlsx), updated October 2011). Table 1 lists titles and some basic summary statistics of journals in which the 11 Price medalists published at least 10 papers. Note that the publication of *Journal of Chemical Documentation* has been discontinued since 1974, therefore their recent SJR and SNIP are unavailable. Also please note a relatively high value of SJR for the *Journal of Informetrics* (first published in 2007), which is more mathematically-oriented than other scientometric sources.

## 2.2. Paper Quality Metrics

Assuming that publications are the most basic and important effects of scientists' activity, the assessment of the 11 Price medalists may base upon a proper aggregation of their papers' quality measures. In this paper we consider the following metrics.

*Direct number of citations ( $nC_{src}$ ).* The most common approach to measuring a paper's quality directly takes into account the number of citations it has received. Table 2 shows the results of assessing the 11 Price Medal laureates by using this method. Note that the "sLog1" indicator tends to favor authors publishing many good papers over their career rather than single "peaks". It also does not ignore any citation information, like the *h*- or *g*-index.

Interestingly, the *h*-index does not discriminate between K.W. McCain and M. Zitt, and also between P. Ingwersen and P. Vinkler. E. Garfield has the smallest proportion of cited articles (cf. the "nGe1" column), equal to 52%. This may be due to the fact that cited references in *SciVerse Scopus* go back to 1996. Garfield's astonishing research career started in the 1950s. Please also note that W. Glänzel and L. Leydesdorff most often appear in the top 3 groups (marked in bold).

Formally, citations merely reflect the intensiveness of information use [cf. 2] and therefore are measures of papers' broadly conceived **popularity** among, or even appreciation by, the scientific community. The discussion about the motives of citations (reward or persuasion) seems to have no end [cf. 32–34]. Intensive usage of this method is clearly due to its accessibility via existing bibliographic databases (of, unfortunately, limited content coverage, [cf. 35]), and objectivity — in contrast with peer-review, which may sometimes be discretionary. However, citation process needs some time; a recently published paper cannot be valued in this way.

Table 1: The most popular journals among the 11 Price laureates. “nD” denotes the number of papers, “nC” and “avgC” total and average number of citations received by their publications, respectively. Additionally, SNIP 2011 and SJR 2011 are given. 3 highest values in each column are marked in bold.

Title	nD	nC	avgC	SNIP	SJR
Scientometrics	<b>380</b>	<b>6841</b>	18.0	1.42	0.07
JASIS(T)	<b>136</b>	<b>2270</b>	16.7	2.66	0.07
Scientist	<b>83</b>	288	3.5	0.01	0.03
J Informetrics	54	434	8.0	2.09	<b>0.09</b>
Inf Processing and Management	51	657	12.9	3.48	0.05
J Inf Science	32	588	18.4	2.07	0.05
Research Policy	28	<b>1521</b>	<b>54.3</b>	<b>3.77</b>	0.06
Research Evaluation	28	177	6.3	1.13	0.05
Math and Comp Modelling	25	78	3.1	1.31	0.06
J Documentation	20	980	<b>49.0</b>	1.77	0.05
Nature	19	370	19.5	<b>12.69</b>	<b>7.77</b>
Science	10	1090	<b>109.0</b>	<b>10.56</b>	<b>5.42</b>
J Chem Documentation	10	15	1.5	NA	NA

Table 2: Basic citation-based summary statistics for the Price medalists. “maxC” is the maximal number of citations, “H” and “G” denote the  $h$ - [3] and  $g$ -index [31], respectively, “sLog1” is equal to  $\sum_i \log(1 + x_i)$  (where  $x_i$  denotes the number of citations received by the  $i$ th paper), “nGe1” and “nGe5” is the number of papers with at least 1 and 5 citations.

Author	nD	nC	avgC	maxC	H	G	sLog1	nGe1	nGe5
EGGHE L.	<b>162</b>	1566	9.7	255	18	34	248.2	128	69
GARFIELD E.	<b>211</b>	<b>3123</b>	14.8	<b>582</b>	22	<b>54</b>	227.0	110	52
GLANZEL W.	159	<b>2941</b>	18.5	143	<b>28</b>	<b>47</b>	<b>344.7</b>	<b>130</b>	<b>105</b>
INGWERSEN P.	61	1446	<b>23.7</b>	244	15	37	128.1	51	37
LEYDESDORFF L.	<b>190</b>	<b>3419</b>	18.0	<b>569</b>	<b>29</b>	<b>51</b>	<b>406.2</b>	<b>166</b>	<b>124</b>
MCCAIN K.W.	40	749	18.7	<b>284</b>	13	27	75.0	32	22
MOED H.F.	81	1871	<b>23.1</b>	156	24	40	196.8	67	58
ROUSSEAU R.	159	1916	12.1	137	23	37	<b>286.1</b>	<b>133</b>	<b>88</b>
VAN RAAN A.F.J.	91	1971	<b>21.7</b>	180	<b>27</b>	41	212.6	77	64
VINKLER P.	57	648	11.4	67	15	23	110.9	52	33
ZITT M.	29	429	14.8	55	13	20	64.6	26	20

On the other hand, it would probably be advisable to consider in some way the quality of a journal in which a paper is published. Although the Association of Science Editors [36] states that source impact factors should only be used for comparing the influence of entire journals, and not for the assessment of single papers or researchers, it is evident that peer-reviewers and editors value the **potential** of the paper. A poor paper will rather not be accepted by a good journal, however its novelty, if not recognized during a paper’s submission, may be appreciated later by the whole community. It may be thus believed that a high IF/SJR/SNIP value correlates with editors’ criticism. More authors tend to submit their papers to highly-ranked journals. As only better ones are accepted, this “recursively” raises the journal’s standards. For example, a paper to be accepted in *Science* should be conceived by editors as outstandingly innovative and decisive. Simply, a good author publishes papers in good journals.

We therefore suggest that journal quality metrics may also bring some interesting insight into paper’s assessment process, because they reflect a different dimension of its quality (“initial” potential as perceived by reviewers/editors versus “factual” impact/popularity expressed by citations). Although both factors are in overall positively correlated, it is obvious that sometimes they do not coincide.

What is more, as noted in the previous section, the main problem with citations is that they are all treated equal. Neither the prestige of the journal citing a given paper, the “quality” of the paper that cite a given paper, nor citation intensity of the field is taken into account. Thus, we propose few measures that rely upon SNIP and SJR, which are already field-normalized. All the below-discussed indicators, listed in Table 3, are based on some sensible grounds.

Table 3: Paper quality measures considered.

Paper quality measure		Assessed paper		Citations	
		Potential	Popular.	Potential	Popular.
$a$	$nC_{src}$		✓		
$b_{JQM}$	$nC_{src} \times JQM_{src}$	✓	✓		
$c_{JQM}$	$\Sigma(JQM_{cit})$		→	✓	
$d_{JQM}$	$\Sigma(nC_{cit} \times JQM_{cit})$		→	✓	✓
$e_{JQM}$	$JQM_{src} \times \Sigma(nC_{cit} \times JQM_{cit})$	✓	→	✓	✓

$nC_{src} \times SNIP_{src}$  and  $nC_{src} \times SJR_{src}$  ( $b_{SNIP}$  and  $b_{SJR}$ ). Assuming that each paper reflects at least a “typical” (properly normalized) quality of a journal in which it is

published, we may consider SNIP/SJR value in which the original paper appeared scaled by the number of citations received. In this setting, a paper published in *Science* that has one citation gains the same valuation as a paper in *Journal of Informetrics* with 5 (SNIP) or about 60 (if SJR is considered) citations.

$\Sigma(SNIP_{cit})$  and  $\Sigma(SJR_{cit})$  ( $c_{SNIP}$  and  $c_{SJR}$ ). Under the same assumption, we may measure the merit of a paper by summing SNIP/SJR values of the journals in which its citations appeared. For example, a citation from *Journal of Informetrics* is considered more important than one from *Lecture Notes on Computer Science*, consisting mainly of proceedings papers.

$\Sigma(nC_{cit} \times SNIP_{cit})$  and  $\Sigma(nC_{cit} \times SJR_{cit})$  ( $d_{SNIP}$  and  $d_{SJR}$ ). In this approach we sum the number of citations received by papers citing a given paper scaled by SNIP/SJR values in which they appeared. Therefore this method acknowledges greater importance for papers referenced by intensively cited (popular) articles that appeared in good journals.

$SNIP_{src} \times \Sigma(nC_{cit} \times SNIP_{cit})$  and  $SJR_{src} \times \Sigma(nC_{cit} \times SJR_{cit})$  ( $e_{SNIP}$  and  $e_{SJR}$ ). The last proposed quality measure is created by multiplying the previous one by the SNIP/SJR value of the journal in which the assessed paper was published. Remembering that SNIP/SJR are field normalized, this would for example increase the valuation of papers in domains such as mathematics.

If SNIP or SJR were unavailable for a given journal, we set its value to 0 (this is the case of inactive, no longer published journals).

### 3. Method

Let  $\mathbb{I} = [0, \infty]$  be a set of some paper quality measures' values and  $\mathbb{I}^{1,2,\dots} := \bigcup_{n=1}^{\infty} \mathbb{I}^n$  denote the set of all vectors (or arbitrary length) with elements in  $\mathbb{I}$ . Intuitively, each author that has published  $n \geq 1$  publications can be represented by some vector  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{I}^{1,2,\dots}$ , where  $x_i$  denotes the valuation of his/her  $i$ th paper.

From now on, let  $\mathcal{E}(\mathbb{I})$  denote the family of all **aggregation operators** in  $\mathbb{I}^{1,2,\dots}$ , i.e. all the functions from  $\mathbb{I}^{1,2,\dots}$  to  $\mathbb{I}$ . Therefore, aggregation operators merge several numerical values (e.g. quality measures of each author's publication) into a single number, representative for the whole input in some way. Theory of aggregation is a rapidly developing mathematical domain (we refer the reader to the recent state of the art monograph [13]).



### 3.1. Impact functions

Clearly, not each aggregation operator can be used in the bibliometric impact assessment process. We shall thus provide some sine qua non conditions that should be fulfilled in the domain of our interest [4, 5, cf. also 7].

**Definition 1.** An **impact function** is an aggregation operator  $F \in \mathcal{E}(\mathbb{I})$ , which:

- (I1) is nondecreasing in each variable:  $(\forall n)(\forall \mathbf{x}, \mathbf{y} \in \mathbb{I}^n) \mathbf{x} \leq \mathbf{y} \Rightarrow F(\mathbf{x}) \leq F(\mathbf{y})$ , where  $\mathbf{x} \leq \mathbf{y}$  if and only if  $(\forall i \in [n]) x_i \leq y_i$ ,
- (I2) is arity-monotonic, i.e.  $(\forall n, m)(\forall \mathbf{x} \in \mathbb{I}^n)(\forall \mathbf{y} \in \mathbb{I}^m) F(\mathbf{x}) \leq F(\mathbf{x}, \mathbf{y})$ , where  $(\mathbf{x}, \mathbf{y})$  denotes vectors concatenation, i.e.  $(x_1, \dots, x_n, y_1, \dots, y_m) \in \mathbb{I}^{n+m}$ ,
- (I3) fulfills the weak lower boundary condition:  $\inf_{\mathbf{x} \in \mathbb{I}^{1,2,\dots}} F(\mathbf{x}) = 0$ ,
- (I4) fulfills the weak upper boundary condition:  $\sup_{\mathbf{x} \in \mathbb{I}^{1,2,\dots}} F(\mathbf{x}) = \infty$ ,
- (I5) is symmetric, i.e.  $(\forall n)(\forall \mathbf{x}, \mathbf{y} \in \mathbb{I}^n) \mathbf{x} \cong \mathbf{y} \Rightarrow F(\mathbf{x}) = F(\mathbf{y})$ , where  $\mathbf{x} \cong \mathbf{y}$  if and only if there exists a permutation  $\sigma$  of the set  $\{1, \dots, n\}$  such that  $\mathbf{x} = (y_{\sigma(1)}, \dots, y_{\sigma(n)})$ .

The first two conditions correspond to the principle called “the more the better”. They state that by increasing the quality of some papers (I1), or by adding some new publications to an author’s output (I2), we never decrease the overall author’s valuation. Hence each impact function is an aggregation operator that reflects two dimensions of authors’ quality:

1. ability to write highly-rated papers,
2. overall productivity.

The boundary condition (I3) together with (I1) and (I2) implies that  $F(0) = 0$ . Additionally, (I3) and (I4) implies that an impact function can not be constant in its whole domain. According to (I5), the overall rating is not affected by the presentation order of the publications.

It is worth noting that in a classical approach to aggregation [cf. 14] only the first author’s quality component (represented by (I1)) is taken into account. In such case, non-decreasing functions fulfilling stronger boundary conditions  $((\forall n) \inf_{\mathbf{x} \in \mathbb{I}^n} J(\mathbf{x}) = 0 \text{ and } \sup_{\mathbf{x} \in \mathbb{I}^n} J(\mathbf{x}) = \infty)$  are considered.

### 3.2. The generalized $h$ -index

As the proposed paper quality measures are arbitrary real numbers, in our context we shall consider generalized versions of well-known scientometric indices which, by default, assume that aggregated information is represented by integer-valued vectors. Here we are interested in a properly modified  $h$ -index.

**Definition 2.** Let  $s > 0$  and  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{I}^{1,2,\dots}$ . The **generalized  $h$ -index** is an impact function

$$H_s(\mathbf{x}) = \bigvee_{i=1}^n x_{(n-i+1)} \wedge (si), \quad (1)$$

where  $\vee$  and  $\wedge$  denote the maximum and minimum operators, respectively,  $x_{(n-i+1)}$  is the  $(n - i + 1)$ th order statistic, i.e. the  $i$ th largest value in  $\mathbf{x}$ .

The value of this impact function may easily be calculated in R by calling `max(pmin(sort(x, decreasing=TRUE), s*(1:length(x))))`.

**Example 1.** Assume that  $n = 10$ ,  $\mathbf{x} = (0.1, 2.1, 11.2, 16.1, 1.4, 0.8, 9.7, 14.3, 9.6, 5.4)$ , and  $s = 2$ . We have:

$i$	1	2	3	4	5	6	7	8	9	10
$si$	<b>2</b>	<b>4</b>	<b>6</b>	<b>8</b>	10	12	14	16	18	20
$x_{(n-i+1)}$	16.1	14.3	11.2	9.7	<b>9.6</b>	<b>5.4</b>	<b>2.1</b>	<b>1.4</b>	<b>0.8</b>	<b>0.1</b>
$si \wedge x_{(n-i+1)}$	2	4	6	8	<b>9.6</b>	5.4	2.1	1.4	0.8	0.1

$\uparrow$   
max

Thus,  $H_2(\mathbf{x}) = 9.6$ . □

Interestingly, each such impact function is an S-statistic [5], an aggregation operator which generalizes the Ordered Weighted Maximum (OWMax) [37], well-known from the aggregation theory. Their axiomatic analysis has been performed in [4, 6, 38], and their basic stochastic properties (e.g. their asymptotic normality in an i.i.d. model) have been examined in [39].

It is easily seen that  $H_s(\mathbf{x}) \in [0, sn]$  for any  $s$  and  $\mathbf{x}$ , and for  $s \geq x_{(n)}$  we get  $H_s(\mathbf{x}) = x_{(n)} =: H_\infty(\mathbf{x})$  (maximal value). Additionally, from the last equation we have  $H_s(\mathbf{x}) = sH_1(\mathbf{x}/s)$ . Please note that the idea of elements' scaling also appeared in e.g. [18].

It may be shown (see [4] for the proof) that we get the ordinary Hirsch's  $h$ -index under the following assumptions.

**Proposition 3.** If  $x_1, \dots, x_n \in \mathbb{N}_0$  and  $s = 1$  we have

$$\begin{aligned} H_1(\mathbf{x}) &= \begin{cases} \max\{i = 1, \dots, n : x_{(n-i+1)} \geq i\} & \text{if } x_{(n)} \geq 1 \\ 0 & \text{otherwise} \end{cases} \\ &= H(\mathbf{x}), \end{aligned}$$

where  $H$  is the  $h$ -index [3].

More generally, for any  $\mathbf{x} \in \mathbb{I}^{1,2,\dots}$  it holds  $H_1(\lfloor \mathbf{x} \rfloor) = \lfloor H_1(\mathbf{x}) \rfloor = H(\mathbf{x})$ , where  $\lfloor \cdot \rfloor$  denotes the floor function.

**Example 1 (continued).** In the above example we have  $H_1(\mathbf{x}) = 5.4$  (cf. 1st and 3rd row in the table), but  $H(\mathbf{x}) = \lfloor H_1(\mathbf{x}) \rfloor = 5$ , as there are 5 observations  $\geq 5$  in the input vector.

### 3.3. Kendall's rank correlation coefficient $\tau$

In this subsection we recall the definition of Kendall's rank correlation coefficient  $\tau$ . It will be used to measure the degree of conformity between 11 authors' rankings created by applying different paper quality measures and/or impact functions. What is worth noting, the value of  $\tau$  may be interpreted quite easily, which is due to the simplicity of its definition. The well-known Pearson's  $r$  (applicable for normally distributed samples) or Spearman's  $\rho$  (also a rank-based coefficient) does not exhibit this desirable feature, but on the other hand they have some useful statistical properties when the samples are large.

**Definition 4.** Let  $\mathbf{x} = (x_1, \dots, x_n)$  and  $\mathbf{y} = (y_1, \dots, y_n)$  for some  $n$ . Kendall's rank correlation coefficient is a function

$$\tau(\mathbf{x}, \mathbf{y}) = \frac{n_{c-d}}{\sqrt{(n^* - t_x)(n^* - t_y)}}, \quad (2)$$

where  $n_{c-d} = \sum_{i=2}^n \sum_{j=1}^{i-1} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j)$  denotes the difference between the number of 'concordant' and 'discordant' pairs,  $n^* = n(n-1)/2$  is the number of all possible unique pairs of elements in each sample, and

$$\begin{aligned} t_x &= \sum_{i=2}^n \left( i - \min\{j = 1, \dots, i : x_{(j)} = x_{(i)}\} \right), \\ t_y &= \sum_{i=2}^n \left( i - \min\{j = 1, \dots, i : y_{(j)} = y_{(i)}\} \right), \end{aligned}$$

are adjustments for the so-called **tied observations**, i.e. those which have the same rank.

Please note that if there are no ties (that is, in our case, no authors receive the same valuations) in both samples, the denominator is simply equal to  $n^*$ . A tied pair is neither concordant nor discordant. What is important, each rank correlation coefficient does not change its value if we apply to each element in  $\mathbf{x}$  and  $\mathbf{y}$  some strictly increasing function (that is, it does not matter whether we correlate author's ranks or author's valuations), and if we apply the same permutation of elements in  $\mathbf{x}$  and  $\mathbf{y}$  (that is, it is independent of how we order our authors).

It may be shown that for all  $\mathbf{x}, \mathbf{y}$  it holds  $\tau \in [-1, 1]$ . Complete agreement between two rankings implies that  $\tau = 1$ . Intuitively, if we assess a group of  $n$  authors using two different methods then the higher the  $\tau$ , the more similar are the two rankings.

The following examples shall ease the interpretation of  $\tau$  values.

**Example 2.** Table 4 shows how the addition of  $d$  discordant pairs to a vector  $\mathbf{x} = (1, 2, \dots, 11)$  (cf. the above notice) affects the value of  $\tau$ . It is easily seen from Eq. (2) that for  $n = 11$  and when there are no tied observations in each sample, i.e. all of them have unique ranks, then  $\tau = 1 - 0.0(36)d$ .

Table 4: Kendall's correlation coefficients between  $\mathbf{x} = (1, 2, \dots, 11)$  and some exemplary vectors of length 11.

$\mathbf{y}$	$\tau(\mathbf{x}, \mathbf{y})$	
(1,2,3,4,5,6,7,8,9,10,11)	1.00000	(Complete agreement)
( <u>2</u> ,1,3,4,5,6,7,8,9,10,11)	0.96364	(1 discordant pair: $2 < 1$ )
(1,2, <u>4</u> , <u>3</u> ,5,6,7,8,9,10,11)	0.96364	(1 discordant pair: $4 < 3$ )
( <u>2</u> , <u>1</u> , <u>4</u> , <u>3</u> ,5,6,7,8,9,10,11)	0.92727	(2 discordant pairs: $2 < 1, 4 < 3$ )
( <u>3</u> , <u>2</u> , <u>1</u> ,4,5,6,7,8,9,10,11)	0.89091	(3 discordant pairs: $3 < 2, 3 < 1, 2 < 1$ )
( <u>4</u> , <u>3</u> , <u>2</u> , <u>1</u> ,5,6,7,8,9,10,11)	0.78182	(6 discordant pairs: $4 < 3, 4 < 2, 4 < 1,$ $3 < 2, 3 < 1, 2 < 1$ )

**Example 3.** Let  $n = 11$  and  $\mathbf{x} = (1, 2, \dots, 7, 8, 8, 10, 10)$  be a vector of ranks which has 2 pairs of tied observations (cf. below). Table 5 shows how tied observations removal from and discordant pairs addition to  $\mathbf{x}$  affect the value of Kendall's correlation coefficient. We observe that the existence or non-existence of tied observations have smaller effect on the value of  $\tau$  than the introduction of discordant pairs.  $\square$

#### 4. Empirical results

Let us first examine how the 11 Price medalists are assessed using the impact function  $H_1$  which, by Proposition 3, is equivalent to the ordinary  $h$ -index in case of integer-valued paper quality measures (e.g.  $nC_{src}$ ). The results are listed in Table 6. W. Glänzel appears in the top 3 group 7 times, L. Leydesdorff 6 times, and H.F. Moed 5 times. Please note that if we used the ordinary  $h$ -index to evaluate our authors, e.g. H.F. Moed would get the same rank as A.F.J. Van Raan for

Table 5: Kendall’s correlation coefficients between  $\mathbf{x} = (1, 2, \dots, 7, 8, 8, 10, 10)$  and some exemplary vectors of length 11.

$\mathbf{y}$	$\tau(\mathbf{x}, \mathbf{y})$	
(1,2,3,4,5,6,7, <u>8,9</u> ,10,10)	0.99070	(One pair of ties resolved)
(1,2,3,4,5,6,7,8, <u>8,10,11</u> )	0.99070	(One pair of ties resolved)
(1,2,3,4,5,6,7, <u>8,9,10,11</u> )	0.98165	(Both pairs of ties resolved)
( <u>2,1</u> ,3,4,5,6,7,8,8,10,10)	0.96226	(1 discordant pair)
( <u>2,1</u> ,3,4,5,6,7, <u>8,9</u> ,10,10)	0.95331	(+1 pair of ties resolved)
( <u>2,1</u> ,3,4,5,6,7, <u>8,9,10,11</u> )	0.94461	(+2 pairs of ties resolved)
( <u>2,1,4,3</u> ,5,6,7,8,8,10,10)	0.92453	(2 discordant pairs)
( <u>2,1,4,3</u> ,5,6,7, <u>8,9</u> ,10,10)	0.91593	(+1 pair of ties resolved)
( <u>2,1,4,3</u> ,5,6,7, <u>8,9,10,11</u> )	0.90756	(+2 pairs of ties resolved)
( <u>3,2,1</u> ,4,5,6,7,8,8,10,10)	0.88679	(3 discordant pairs)
( <u>3,2,1</u> ,4,5,6,7, <u>8,9</u> ,10,10)	0.87854	(+1 pair of ties resolved)
( <u>3,2,1</u> ,4,5,6,7, <u>8,9,10,11</u> )	0.87052	(+2 pairs of ties resolved)
( <u>4,3,2,1</u> ,5,6,7,8,8,10,10)	0.77358	(6 discordant pairs)
( <u>4,3,2,1</u> ,5,6,7, <u>8,9</u> ,10,10)	0.76639	(+1 pair of ties resolved)
( <u>4,3,2,1</u> ,5,6,7, <u>8,9,10,11</u> )	0.75939	(+2 pairs of ties resolved)

$b_{\text{SNIP}}$  (equal to 32). We see that the  $h$ -index for real-valued paper quality metrics sometimes tends to loose information that could be used to discriminate between authors, and therefore shall not be applied for this type of data.

The highest correlation with the reference ( $nC_{\text{src}}$ -based) ranking is observed for  $c_{\text{SNIP}}$ : we have  $\tau = 0.90756$ . From Tab. 5 we see that this corresponds to 2 resolved pairs of tied observations and 2 discordant pairs. Indeed, all the authors in the  $c_{\text{SJR}}$  column have unique  $H_1$ -index values (we have  $M. Zitt < K.W. McCain$  and  $P. Vinkler < P. Ingwersen$ ). Additionally, (L. Egghe, E. Garfield) and (R. Rousseau, H.F. Moed) are ordered differently.

On the other hand, we get the smallest correlation for  $e_{\text{SJR}}$ , which is not even statistically significant ( $p\text{-value}=0.12$ ). In this case we accept the hypothesis stating that the two rankings are uncorrelated.

The main problem with the proposed paper quality measures is that they use *different scales*. Gaining one citation is often much easier than obtaining e.g.  $b_{\text{SJR}}=1$ , especially when a paper was published in a “moderate quality” journal, for which SJR value is typically equal to 0.64 (see Table 7 for basic summary statistics). Let us then study the relation between different paper quality metrics more deeply.

Table 6:  $H_1$ -indices of 11 Price medalists computed under different paper quality metrics. Also given are the authors’ ranks generated by the citation-based  $h$ -index (cf. Tab. 5), and Kendall’s correlation coefficients between these ranks and ranks generated by other paper quality metrics.

Author	$a$ (rank)	SJR				SNIP			
		$b$	$c$	$d$	$e$	$b$	$c$	$d$	$e$
EGGHE L.	18 (7)	3.1	4.4	18.5	4.8	29.3	26.0	70.1	<b>80.0</b>
GARFIELD E.	22 (6)	<b>16.3</b>	<b>8.2</b>	26.8	<b>18.0</b>	25.4	25.0	56.0	48.0
GLANZEL W.	<b>28</b> (2)	6.0	<b>7.6</b>	<b>30.0</b>	7.0	<b>36.9</b>	<b>36.4</b>	<b>87.0</b>	<b>89.9</b>
INGWERSEN P.	15 (8)	4.0	4.0	16.2	5.0	18.4	21.0	34.0	31.0
LEYDESDORFF L.	<b>29</b> (1)	4.9	4.7	<b>27.6</b>	5.0	<b>42.0</b>	<b>40.0</b>	<b>92.0</b>	<b>96.0</b>
MCCAIN K.W.	13 (10)	3.0	3.0	12.1	3.0	15.0	16.0	23.0	21.0
MOED H.F.	24 (4)	<b>8.0</b>	<b>7.4</b>	<b>27.2</b>	<b>10.0</b>	<b>32.6</b>	29.7	58.0	58.0
ROUSSEAU R.	23 (5)	4.0	5.0	21.8	4.9	28.0	30.0	<b>75.0</b>	78.0
VAN RAAN A.F.J.	<b>27</b> (3)	<b>7.0</b>	6.3	26.0	<b>8.0</b>	32.0	<b>34.0</b>	64.0	67.0
VINKLER P.	15 (8)	3.0	3.6	17.6	6.0	18.0	19.0	33.0	31.0
ZITT M.	13 (10)	2.6	3.0	11.4	2.8	15.0	15.0	22.0	22.1
$\tau(a, \dots)$		0.51	0.62	0.83	0.37	0.88	0.91	0.80	0.77

Table 7: Basic summary statistics for paper quality measures (only values greater than 0 were considered); Q1 and Q3 denote the 1st and the 3rd quartile, and IQR denotes the interquartile range.

Statistic	$a$	$b_{\text{SNIP}}$	$b_{\text{SJR}}$	$c_{\text{SNIP}}$	$c_{\text{SJR}}$	$d_{\text{SNIP}}$	$d_{\text{SJR}}$	$e_{\text{SNIP}}$	$e_{\text{SJR}}$
Q1	3.00	5.32	0.21	5.50	0.22	36.32	1.58	58.70	0.10
Median	9.00	15.60	0.64	15.24	0.61	172.63	6.71	293.25	0.46
Q3	22.00	39.10	1.64	35.30	1.59	585.23	23.46	1057.54	1.69
IQR/2	9.50	16.89	0.71	14.90	0.68	274.45	10.94	499.42	0.79

#### 4.1. Relationship between different paper quality measures

Table 8 lists Spearman’s rank correlation coefficients between different pairs of paper quality measures (output of all 11 Price laureates was considered, i.e. 1240 papers). Spearman’s  $\rho$  is a non-parametric measure of association between two samples. It assesses how well the relationship between two variables can be described by a monotonic function. Values closer to 1 indicate greater degree of correlation.

In our sample all coefficients have positive sign. Unsurprisingly, we observe that the direct citations count is most weakly correlated with  $e_{\text{SJR}}$  and  $e_{\text{SNIP}}$  ( $\rho \approx 0.86$  and  $\rho \approx 0.87$ , respectively). Moreover, for  $b_{\text{SJR}}$  and  $b_{\text{SNIP}}$  we get  $\rho \approx 0.89$ . It seems that scaling a quality measure by SJR or SNIP value of a journal in which a

Table 8: Spearman’s  $\rho$  rank correlation coefficients between different paper quality measures. All coefficients are significant ( $p$ -values  $\leq 0.001$ ).

	$a$	$b_{\text{SJR}}$	$b_{\text{SNIP}}$	$c_{\text{SJR}}$	$c_{\text{SNIP}}$	$d_{\text{SJR}}$	$d_{\text{SNIP}}$	$e_{\text{SJR}}$	$e_{\text{SNIP}}$
$a$	—	0.89	0.89	<b>0.97</b>	<b>0.98</b>	0.92	0.93	0.86	0.87
$b_{\text{SJR}}$	—	—	<b>0.96</b>	0.88	0.89	0.83	0.83	0.94	0.91
$b_{\text{SNIP}}$	—	—	—	0.87	0.90	0.82	0.83	0.90	0.94
$c_{\text{SJR}}$	—	—	—	—	<b>0.97</b>	0.94	0.92	0.87	0.85
$c_{\text{SNIP}}$	—	—	—	—	—	0.93	0.94	0.87	0.89
$d_{\text{SJR}}$	—	—	—	—	—	—	<b>0.98</b>	0.92	0.90
$d_{\text{SNIP}}$	—	—	—	—	—	—	—	0.90	0.92
$e_{\text{SJR}}$	—	—	—	—	—	—	—	—	<b>0.96</b>
$e_{\text{SNIP}}$	—	—	—	—	—	—	—	—	—

paper was published (i.e. taking into account the “potential” of the assessed paper, cf. Tab. 3) “reveals” a different aspect of its quality.

On the other hand, the highest degree of correlation was detected between  $nC_{\text{src}}$  and  $c_{\text{SNIP}}$ , and between  $nC_{\text{src}}$  and  $c_{\text{SJR}}$  ( $\rho \simeq 0.98$  and  $\rho \simeq 0.97$ , respectively, which may suggest that the two pairs of metrics measure similar quality dimension; however please note that  $c_{\text{SNIP}}$  and  $c_{\text{SJR}}$  are real-valued) and between all SNIP-SJR corresponding metrics (e.g.  $b_{\text{SNIP}}$  vs  $b_{\text{SJR}}$ ).

Scatter plots for selected pairs of quality measures are depicted in Fig. 1. Please note the logarithmic scale on each axis. We see that each measure is not simply a one-to-one function of any other. This is because the metrics are manifestations of different aspects of papers’ quality. The highest variability is most often observed for small metric values. However, if we were supposed to select a class of functions that approximate the position of the points on the scatter plots, probably we would look for some linear model.

If all the proposed paper quality measures were exactly linear functions of  $nC_{\text{src}}$ , we could easily estimate equivalents (in terms of quality) of a single citation. Such numbers could be used to choose coefficients for the generalized  $h$ -index. For example, if  $nC_{\text{src}}=1$  corresponded to  $d_{\text{SNIP}}=16$  (i.e. we had  $d_{\text{SNIP}}=16a$ ), then the impact functions  $H_1$  for  $nC_{\text{src}}$  and  $H_{16}$  for  $d_{\text{SNIP}}$  would generate the same rankings.

To fit such a idealized model to our data, we have to assume that our observations were perturbed by a random “noise” term,  $\varepsilon$ . Thus, our task now is to estimate the values of coefficients  $s$  in a linear model  $y = sx + \varepsilon$ , where  $x, y$  are

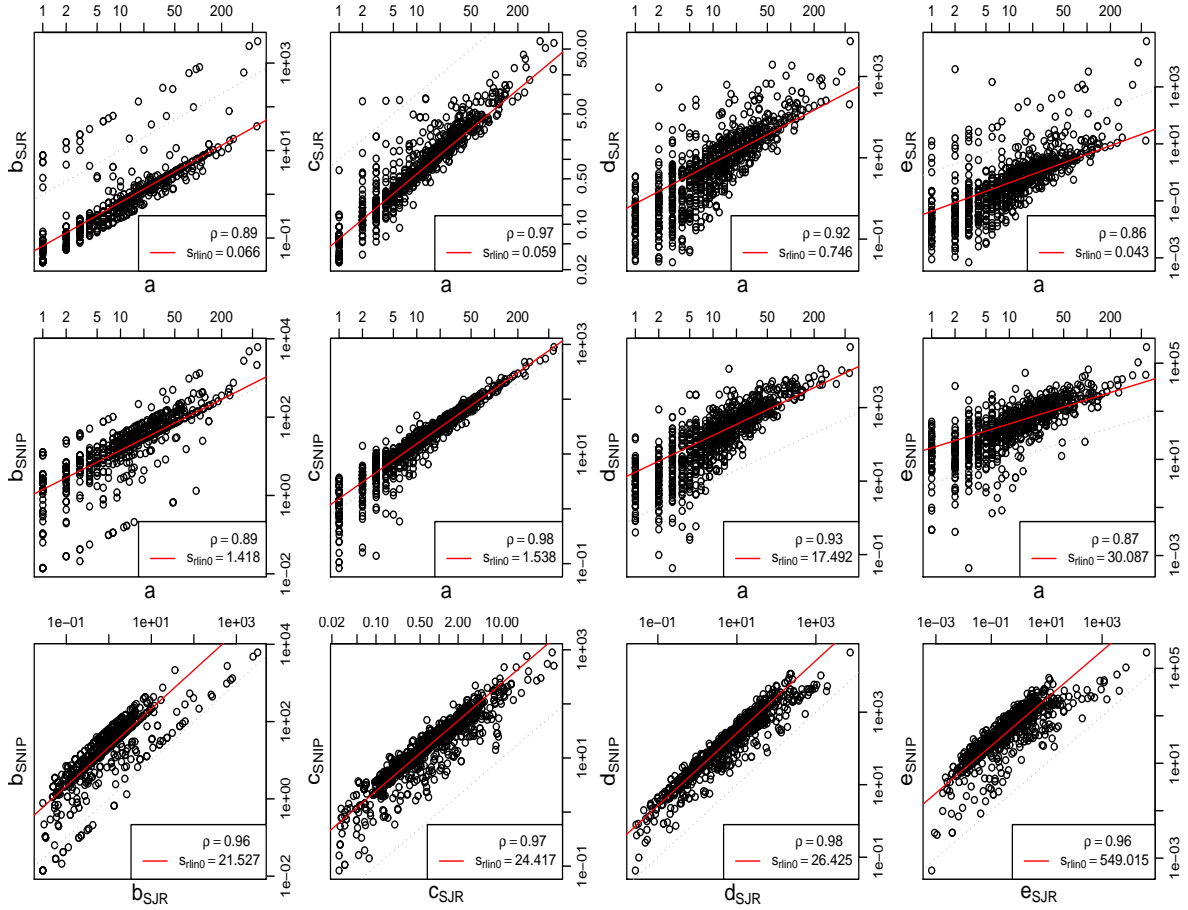


Figure 1: Scatter plots for selected pairs of paper quality measures (1240 papers).

different paper quality measures, and  $\varepsilon$  is the residual term. Please keep in mind that, however, residuals carry important, additional information, unavailable otherwise, e.g. the potential of the paper or its citations.

Table 9 lists four different estimators of coefficients  $s$  for selected pairs of paper quality measures. “ $q_{0.25}$ ” and “ $q_{0.5}$ ” denote ratios of quantiles of order 0.25 and 0.5 (median), respectively, calculated for non-zero quality measures of all the Price laureates altogether (see Tab. 7). “ $\text{rlin0}$ ” is a robust linear regression estimate for the same data. “ $\text{rlog1}$ ” is a robust nonlinear regression estimate (fitted model:  $\log y = \log x + \log s$ ) that bases on observations  $\geq 1$ .

Note that, unsurprisingly, all the SJR-based metrics indicate that there is much more “effort” required to increase their value than to obtain a single citation (cf. 1st



Table 9: 4 estimates of the coefficient  $s$  in a model  $y = sx + \varepsilon$ , where  $x, y$  denote different paper quality measures.

$x \setminus y$	$b_{\text{SJR}}$	$b_{\text{SNIP}}$	$c_{\text{SJR}}$	$c_{\text{SNIP}}$	$d_{\text{SJR}}$	$d_{\text{SNIP}}$	$e_{\text{SJR}}$	$e_{\text{SNIP}}$	method
$a$	0.070	1.773	0.073	1.832	0.528	12.107	0.033	19.568	$q_{0.25}$
	0.071	1.733	0.068	1.693	0.746	19.182	0.051	32.584	$q_{0.5}$
	<b>0.066</b>	<b>1.418</b>	<b>0.059</b>	<b>1.538</b>	<b>0.746</b>	<b>17.492</b>	<b>0.043</b>	<b>30.087</b>	rln0
	0.066	1.723	0.073	1.607	0.812	15.782	0.080	26.653	rlog1
$b_{\text{SJR}}$		25.154	$c_{\text{SJR}}$	24.982	$d_{\text{SJR}}$	22.930	$e_{\text{SJR}}$	590.91	$q_{0.25}$
		24.487		24.982		25.713		639.68	$q_{0.5}$
		<b>21.527</b>		<b>24.417</b>		<b>26.425</b>		<b>549.02</b>	rln0
		26.800		21.115		26.247		454.65	rlog1

row in Tab. 9, e.g.  $e_{\text{SJR}} \simeq 0.043a$  for the “rln0” method). On the other hand, SNIP-based metrics require less exertion (e.g.  $e_{\text{SNIP}} \simeq 549.02e_{\text{SJR}}$  for “rln0”).

In few cases these algorithms for estimating  $s$  give quite different values, e.g. for  $a$  vs  $e_{\text{SNIP}}$ . Their impact on scientists’ rankings will be examined in the next subsection.

#### 4.2. Aggregating different paper quality measures

Let us examine the rank correlation between the one-citation-equivalent generalized  $h$ -indices  $H_s$  and the ordinary citation-based Hirsch index  $H$  (see the ‘rank’ column in Table 6).

Table 10 shows Kendall’s correlation coefficients between rankings generated by ordinary citation-based  $h$ -index, and by the generalized  $h$ -index  $H_s$  for different paper quality measures (rows) and 4 different estimates of the  $s$  coefficient (columns). What is more, in the last column the greatest possible  $\tau$  (for some  $s > 0$  — which requires quite computationally intensive search over the parameter domain) for each PQM is given.

We observe that for each SNIP-based paper quality measure the correlations with respect to the proposed estimates of the  $s$  coefficient is quite high. Method “rln0” gives the greatest correlation, and “ $q_{0.25}$ ” the smallest variance of results. These two algorithms may thus be recommended for automatized selection of the scaling coefficient.

$b_{\text{SJR}}$  and  $e_{\text{SNIP}}$  give the lowest possible *best*  $\tau$ . Please note that, except for  $c_{\text{SNIP}}$  and  $d_{\text{SNIP}}$  for which we have  $\tau \simeq 0.982$ , each possible ranking generated by a generalized  $h$ -index has at least one discordant pair.

These results serve as illustrations of two important facts. Firstly, by comparing the “best” correlation with the data in Tab. 8, we see that higher association be-

Table 10: Kendall’s  $\tau$  measures of association between the Price medalists’ rankings generated by  $H_s$  for different paper quality measures (see Table 9 for the values of coefficients  $s$  obtained by the 4 discussed methods), and by ordinary citation-based  $h$ -index. ‘msd’ denotes the mean squared difference between coefficients for each estimator and ‘best’ values. All the listed coefficients are statistically significant ( $\alpha = 0.01$ ).

PQM	Estimator				
	$q_{0.25}$	$q_{0.5}$	rln0	rlog1	best
$b_{\text{SJR}}$	0.759	0.796	<b>0.804</b>	<b>0.804</b>	0.908
$b_{\text{SNIP}}$	<b>0.914</b>	0.906	0.879	0.906	0.925
$c_{\text{SJR}}$	<b>0.833</b>	<b>0.833</b>	<b>0.833</b>	<b>0.833</b>	0.935
$c_{\text{SNIP}}$	0.833	<b>0.981</b>	0.953	0.953	0.982
$d_{\text{SJR}}$	<b>0.897</b>	0.868	0.868	0.796	0.952
$d_{\text{SNIP}}$	<b>0.982</b>	0.972	<b>0.982</b>	0.972	0.982
$e_{\text{SJR}}$	<b>0.792</b>	0.648	0.717	0.648	0.943
$e_{\text{SNIP}}$	0.897	0.908	<b>0.916</b>	0.908	0.916
mean	0.864	0.864	<b>0.869</b>	0.853	—
st.dev.	<b>0.072</b>	0.108	0.085	0.105	—
$\sqrt{\text{msd}}$	<b>0.100</b>	0.121	0.101	0.129	—

tween two paper quality measures (unless it is a one-to-one function, i.e. if  $\varepsilon = 0$ ), does not necessarily imply that we obtain more concordant authors’ rankings.

Secondly, as none of the introduced paper quality measures is truly a linear function of the number of citations (we expect that the same property holds in other domains than scientometrics), in the vast majority of the cases we do not get the same ranking as the referential one (here it is a  $h$ -index-based ranking) by a simple scaling operation.

Let us study the above-mentioned phenomenon more deeply. Fig. 2 depicts the rank correlation between  $H_s$  and three different reference rankings (citation-based  $h$ - and  $g$ -indices, and  $\max(PQM)$ ) as a function of the  $s$  coefficient. Note that, as far as our sample is concerned, the generalized  $h$ -index for moderate  $s$  values correlates very weakly (if at all statistically) with  $\max(PQM)$ . The highest correlation coefficient between  $H_s$  and the citation-based  $g$ -index was obtained for  $b_{\text{SJR}}$  and  $b_{\text{SNIP}}$  ( $\approx 0.9$ ). Therefore, a perfect concordance to the  $g$ -index may not be obtained.

In the next subsection we will examine how the scaling operation affects the ranking of the Price medalists. Moreover, we will answer the question concerning

the stability of this impact function, i.e. whether small changes in  $s$  affect the authors' ordering only slightly or not. Owing to the fact that our sample is small, we will be able to observe its behavior in very detail.

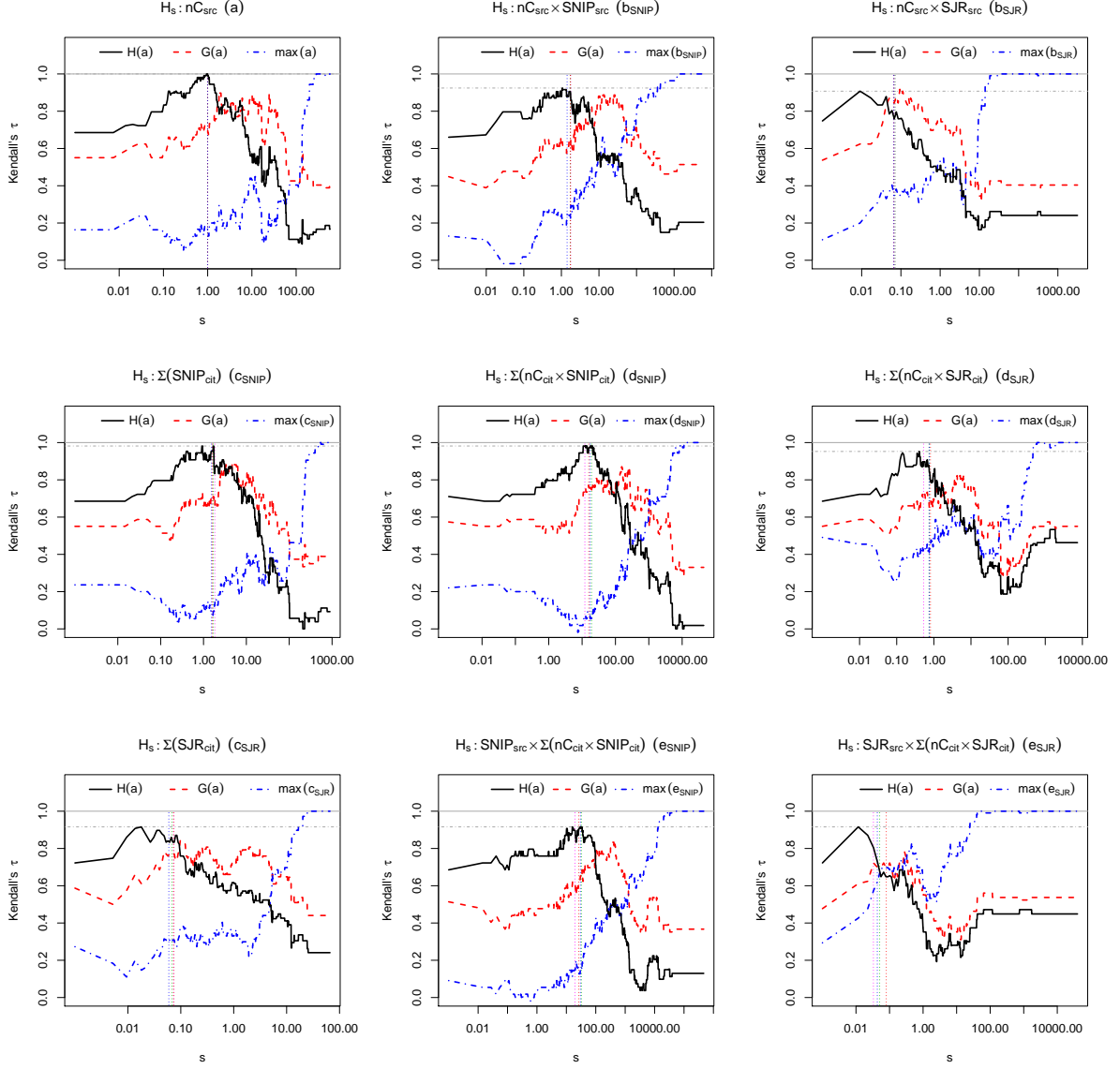


Figure 2: Kendall's  $\tau$  measures of association between the Price medalists' rankings generated by  $H_s$  (as a function of coefficient  $s$ ), and three reference rankings (assessment methods considered: the ordinary citation-based  $h$ - and  $g$ -indices, and  $\max=H_\infty$  value of given paper quality measure). Four estimated coefficients from Tab. 9 are marked with dotted vertical lines.

#### 4.3. Ranking (in)stability under different scales

In Fig. 3 we show authors' ranks generated by  $H_s$  as a function of  $s$ . For readability, we plotted the curves for 5 authors that appear most often in the top 3 group in Tab. 2 in the upper figure, and for the other authors in the bottom figure.

We observe a considerable instability of authors' assessment, which contrasts with results given in [18], where also elements' scaling effect was considered (for only 3 different coefficients). For example, L. Leydesdorff's rank ranges from 1 to 8.5; for small and large  $s$  his rank is among the highest, and for ca. 15–100 it significantly decreases.

Our analysis is not only of theoretical significance: data scalling is often performed to “normalize” citations across disciplines. Uncontrolled application of this operation may lead to unfair and biased (towards some groups) decisions. It is of course not possible to determine which coefficient is undoubtedly *best* for practical purposes. Such a selection should be performed basing on expert knowledge and their goodwill, and — importantly — without prior analysis of how it will affect our favorite authors.

Please recall that, as far as  $nC_{src}$  is concerned, we obtain the ordinary  $h$ -index by taking  $s = 1$  (see Prop. 3). A very small increase in  $s$  (which could be applied in order to normalize “scientometrical citations”) changes the ranks of right up to 8 authors (L. Leydesdorf and W. Glänzel, R. Rousseau and E. Garfield, K. W. McCain and M. Zitt, and P. Ingwersen and P. Vinkler, cf. Fig. 2a).

For different paper quality metrics similar observations can be made. Fig. 4 depicts how a relatively small increase in the  $s$  value affects the rankings. By altering its value by  $0.01 \cdot IQR/2$  we may in some cases obtain  $\tau \simeq 0.8$  (about 5 discordant pairs). Changing the way we assess the paper does not alter our conclusion: the  $h$ -index is a very unstable tool for quality control of scientific research (as far as prominent authors are concerned).

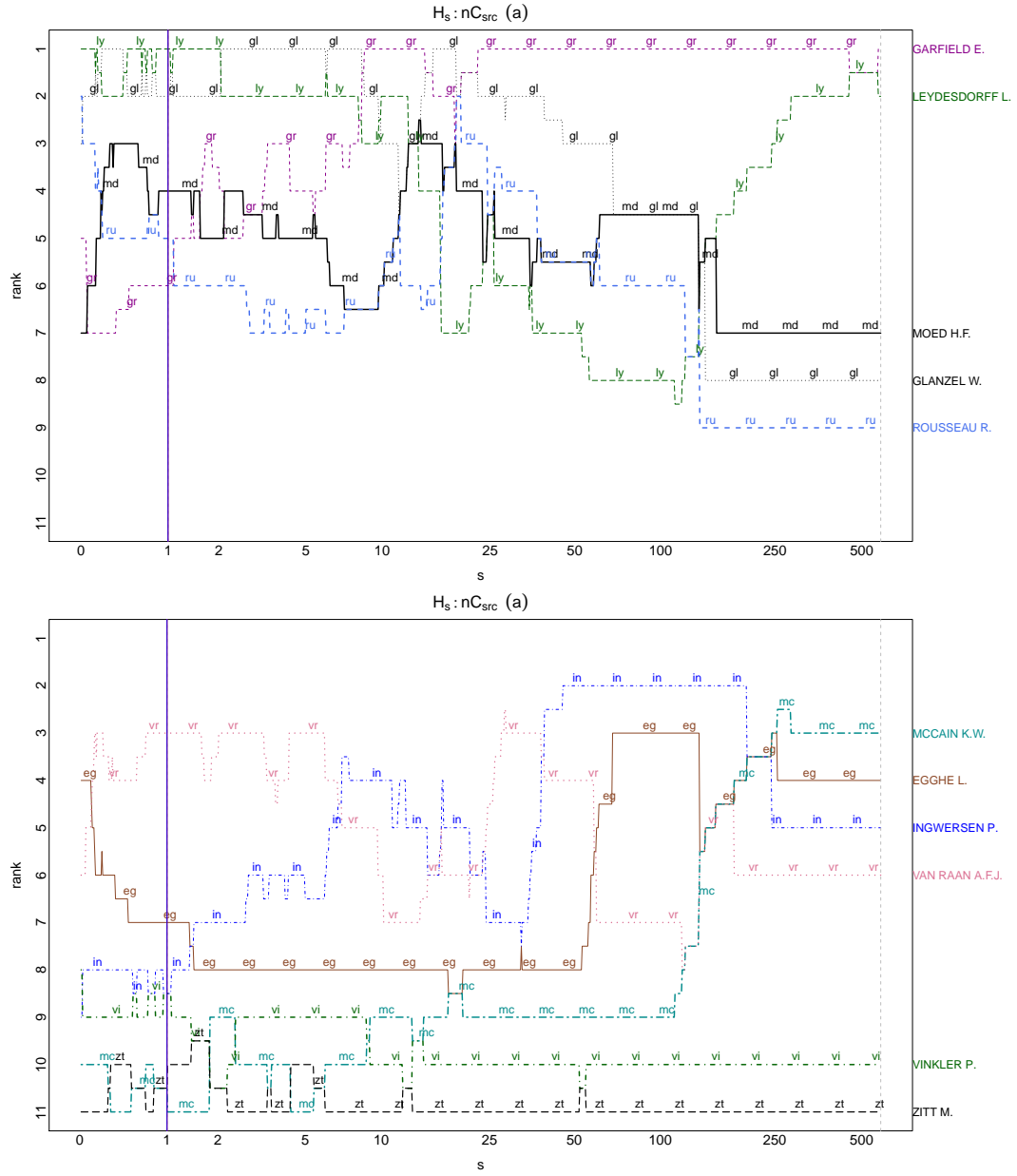


Figure 3: Ranks of 11 Price medalists generated by  $H_s$  as a function of  $s$  for  $nC_{src}$ .

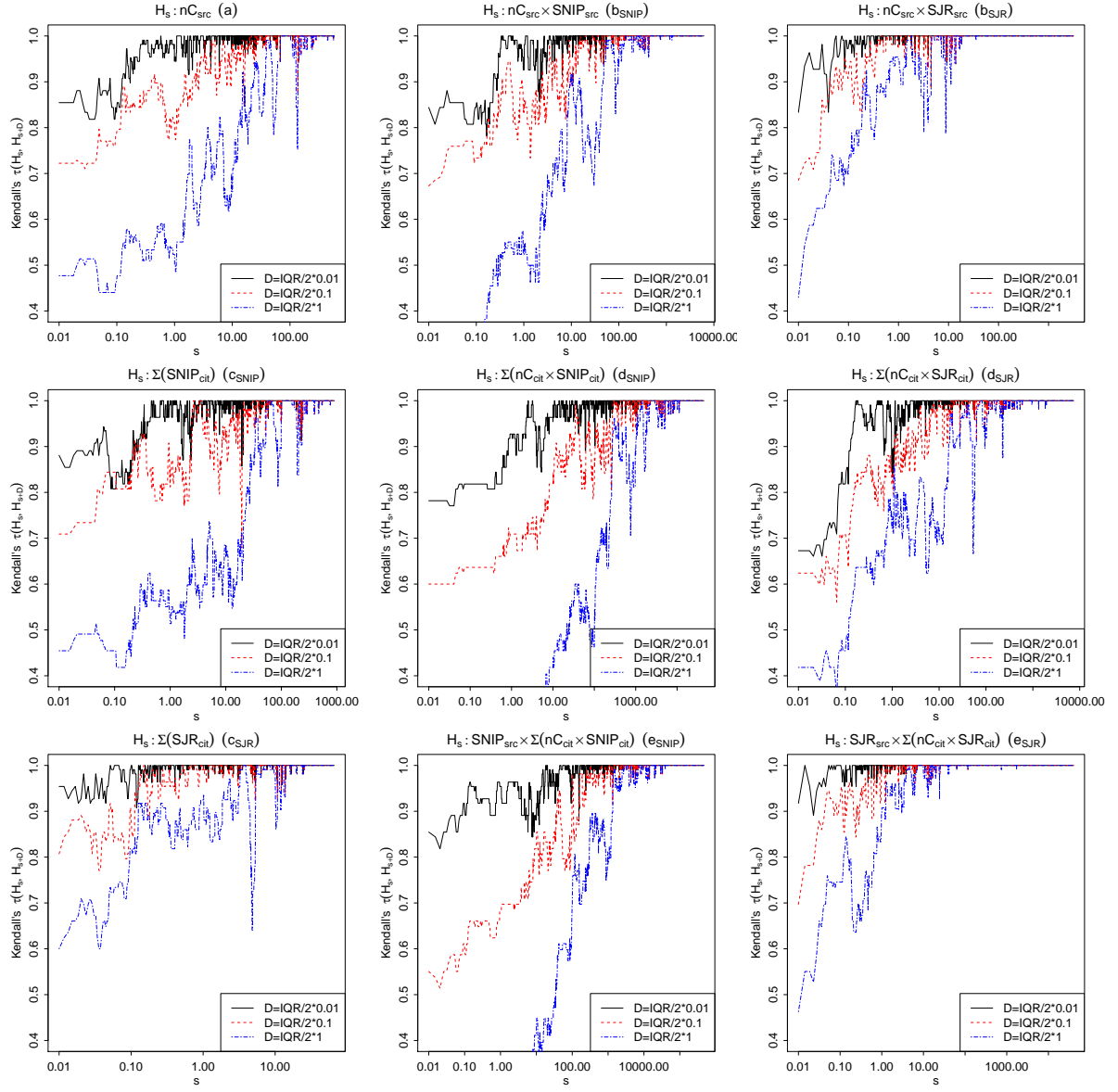


Figure 4: Kendall's correlation coefficients between the Price medalists' rankings generated by  $H_s$  and  $H_{s+D}$  as a function of  $s$ ;  $D = 0.01 \cdot IQR/2, 0.1 \cdot IQR/2, IQR/2$  (see Tab. 7).

## 5. Conclusions

In this study we proposed 8 new field-normalized SJR- and SNIP-based paper quality metrics. These measures not only take the documents' broadly conceived popularity among the scientific community into account, but also their so-called potential, and the quality of their citations. We remember that citations are merely one possible (and definitely not ideal) way to assess the quality of papers.

Moreover, we tried to answer the question of how much “effort” is required to obtain a measure equivalent to a single citation in the field of scientometrics. We should keep in our minds, however, that there is no best metric, as reality is not just a single number. The choice of a paper quality measure *matters* — by applying different metrics we were not able to replicate exactly the original *h*-index-based ranking.

Note that all the introduced measures have a cumulative nature. We could make them “actual” by applying in some way the time when the papers or citations appeared.

We noted a very high instability of author's ranking under the elements' scaling. This puts the validity of (at least prominent) authors assessment and ranking using the *h*-index into question. Obviously, a good impact function to be used in practice should not be as much sensitive to changing input data as the analyzed one.

Further research should consider the study of the behavior of a similarly transformed *g*-index (related to some impact function from the aggregation theory), definitely on a more extensive data set (however, please note that our small sample let us observe the phenomena of concern in very detail — which otherwise would not be possible) and in different scientific fields. Additionally, we indicate the need for the construction and usage of aggregation operators that concern more than one paper quality measure at a time.

*Acknowledgments.* The authors would like to express their gratitude to the anonymous referees for their helpful suggestions, and K. Fokow for proof-reading of the manuscript. R. Mesiar wishes to acknowledge support by the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070) and by the grant VEGA 1/0071/12.

---

**Please cite this paper as:**

Gagolewski M., Mesiar R., Aggregating Different Paper Quality Measures with a Generalized *h*-index, *Journal of Informetrics* **6**(4), 2012, pp. 566–579.

---



## References

- [1] P. L. K. Gross, E. M. Gross, College libraries and chemical education, *Science* 66 (1713) (1927) 385–389.
- [2] W. Glänzel, Seven myths in bibliometrics. About facts and fiction in quantitative science studies, *COLLNET Journal of Scientometrics and Information Management* 2 (1) (2008) 9–17.
- [3] J. E. Hirsch, An index to quantify individual’s scientific research output, *Proceedings of the National Academy of Sciences* 102 (46) (2005) 16569–16572.
- [4] M. Gagolewski, P. Grzegorzewski, Arity-monotonic extended aggregation operators, in: E. Hüllermeier, R. Kruse, F. Hoffmann (Eds.), *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Vol. 80, Springer-Verlag, 2010, pp. 693–702.
- [5] M. Gagolewski, P. Grzegorzewski, Possibilistic analysis of arity-monotonic aggregation operators and its relation to bibliometric impact assessment of individuals, *International Journal of Approximate Reasoning* 52 (9) (2011) 1312–1324.
- [6] M. Gagolewski, P. Grzegorzewski, Axiomatic characterizations of (quasi-) L-statistics and S-statistics and the Producer Assessment Problem, in: S. Galichet, J. Montero, G. Mauris (Eds.), *Proc. Eusflat/LFA 2011*, 2011, pp. 53–58.
- [7] G. J. Woeginger, An axiomatic characterization of the Hirsch-index, *Mathematical Social Sciences* 56 (2) (2008) 224–232.
- [8] G. J. Woeginger, An axiomatic analysis of Egghe’s  $g$ -index, *Journal of Informetrics* 2 (4) (2008) 364–368.
- [9] R. Rousseau, Woeginger’s axiomatisation of the  $h$ -index and its relation to the  $g$ -index, the  $h(2)$ -index and the  $r^2$ -index, *Journal of Informetrics* 2 (4) (2008) 335–340.
- [10] G. J. Woeginger, A symmetry axiom for scientific impact indices, *Journal of Informetrics* 2 (2008) 298–303.

- [11] A. Quesada, Monotonicity and the Hirsch index, *Journal of Informetrics* 3 (2) (2009) 158–160.
- [12] A. Quesada, More axiomatics for the Hirsch index, *Scientometrics* 82 (2010) 413–418.
- [13] M. Grabisch, E. Pap, J.-L. Marichal, R. Mesiar, *Aggregation functions*, Cambridge, 2009.
- [14] R. G. Ricci, R. Mesiar, Multi-attribute aggregation operators, *Fuzzy Sets and Systems* 181 (1) (2011) 1–13.
- [15] R. Guns, R. Rousseau, Real and rational variants of the  $h$ -index and the  $g$ -index, *Journal of Informetrics* 3 (1) (2009) 64–71.
- [16] M. Gagolewski, P. Grzegorzewski, A geometric approach to the construction of scientific impact indices, *Scientometrics* 81 (3) (2009) 617–634.
- [17] G. M. Nair, B. A. Turlach, The stochastic  $h$ -index, *Journal of Informetrics* 6 (1) (2012) 80–87.
- [18] N. J. van Eck, L. Waltman, Generalizing the  $h$ - and  $g$ -indices, *Journal of Informetrics* 2 (4) (2008) 263–271.
- [19] M. Gagolewski, Bibliometric impact assessment with R and the CITAN package, *Journal of Informetrics* 5 (4) (2011) 678–692.
- [20] R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, URL: <http://www.R-project.org> (2012).
- [21] K. Buchholz, Criteria for the analysis of scientific quality, *Scientometrics* 32 (2) (1995) 195–218.
- [22] R. N. Kostoff, The use and misuse of citation analysis in research evaluation, *Scientometrics* 43 (1) (1998) 27–43.
- [23] C. Nicolini, S. Vakula, M. Italo Balla, E. Gandini, Can the assignment of university chairs be automated?, *Scientometrics* 32 (2) (1995) 93–107.
- [24] E. Garfield, Citation indexes for science, *Science* 122 (3159) (1955) 108–111.

- [25] H. F. Moed, Measuring contextual citation impact of scientific journals, *Journal of Informetrics* 4 (3) (2010) 265–277.
- [26] I. Podlubny, Comparison of scientific impact expressed by the number of citations in different fields of science, *Scientometrics* 64 (1) (2005) 95–99.
- [27] T. Braun, W. Glänzel, A. Schubert, A Hirsch-type index for journals, *Scientometrics* 69 (1) (2006) 169–173.
- [28] B. Gonzalez-Pereira, V. P. Guerrero-Bote, F. de Moya-Anegon, A new approach to the metric of journals' scientific prestige: The SJR indicator, *Journal of Informetrics* 4 (3) (2010) 379–391.
- [29] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: Bringing order to the Web, Tech. rep., Stanford University (1998).
- [30] L. Leydesdorff, T. Opthof, Scopus' Source Normalized Impact per Paper (SNIP) versus the Journal Impact Factor based on fractional counting of citations, *Journal of the American Society for Information Science and Technology* 61 (11) (2010) 2365–2396.
- [31] L. Egghe, Theory and practise of the *g*-index, *Scientometrics* 69 (1) (2006) 131–152.
- [32] E. Garfield, Can citation indexing be automated?, in: M. E. Stevens, V. E. Giuliano, L. B. Heilprin (Eds.), *Proc. Statistical Association Methods for Mechanized Documentation*, Washington, 1964, pp. 189–192.
- [33] L. Bornmann, H.-D. Daniel, What do citation counts measure? A review of studies on citing behavior, *Journal of Documentation* 64 (1) (2008) 45–80.
- [34] P. M. Davis, Reward or persuasion? The battle to define the meaning of a citation, *Learned Publishing* 21 (2009) 5–11.
- [35] L. I. Meho, C. R. Sugimoto, Assessing the scholarly impact of information studies: A tale of two citation databases — *Scopus* and *Web of Science*, *Journal of the American Society for Information Science and Technology* 60 (12) (2009) 2499–2508.
- [36] European Association of Science Editors, EASE statement on inappropriate use of impact factors, URL: [http://www.ease.org.uk/statements/EASE\\_statement\\_on\\_impact\\_factors.shtml](http://www.ease.org.uk/statements/EASE_statement_on_impact_factors.shtml) (1998).

- [37] D. Dubois, H. Prade, C. Testemale, Weighted fuzzy pattern matching, *Fuzzy Sets and Systems* 28 (1988) 313–331.
- [38] M. Gagolewski, On the Relation Between Effort-Dominating and Symmetric Minitive Aggregation Operators, to appear in *Proc. IPMU 2012 (LNCS/LNAI/CCIS series)* (2012).
- [39] M. Gagolewski, P. Grzegorzewski, S-Statistics and Their Basic Properties, in: C. Borgelt et al (Eds.), *Combining Soft Computing and Statistical Methods in Data Analysis (AISC 77)*, Springer-Verlag, 2010, pp. 281-288.