Scientometrics (2013) 96:381–394
DOI 10.1007/s11192-012-0898-z

Opinion paper: thoughts and facts on bibliometric indicators

Wolfgang Glänzel · Henk F. Moed

Received: 18 October 2012 / Published online: 21 November 2012
© Akadémiai Kiadó, Budapest, Hungary 2012

Abstract This paper aims at contributing to the on-going discussion about building and applying bibliometric indicators. It sheds light on their properties and requirements concerning six different aspects: deterministic versus probabilistic approach, application-related properties, the time dependence, normalization issues, size dependence and network indicators.

Keywords Indicators · Mathematical statistics · Normalization

Introduction

The use of bibliometric indicators emerged from the application of scientometric methods to the evaluation of research. Indicators aim at characterizing and assessing units of analysis by quantitative methods based on generic measures or on the quantification of expert opinions. Basic requirements are robustness, validity of measurement and application as well as reproducibility. Thus indicators should be insensitive to marginal changes in the aspects they aim to measure, should be meaningful measures of what they are applied to and, of course, under the same conditions and using the same data and methods, the same indicator values should be obtained. Furthermore another issue emerges, namely that correctness of application has two sides. An indicator might be meaningful but formally not well-defined, or, conversely, it might be formally-mathematically correct but not a meaningful measure. In both cases one should refrain from the application of such an indicator.

W. Glänzel (✉)
Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven, Belgium
e-mail: Wolfgang.Glanzel@econ.kuleuven.be

W. Glänzel
Department Science Policy and Scientometrics, LHAS, Budapest, Hungary

H. F. Moed
SciVal Dept., Elsevier, Amsterdam, Netherlands
e-mail: h.moed@elsevier.com

In bibliometrics, finally, a further specific issue arises, namely that of outliers. While in many fields of application outliers are simply discarded as being exceptions, in bibliometrics ‘outliers’ are often part of the high-end of research performance and deserve certainly special attention. The question arises in how far extreme value statistics can serve as supplementary indicators to the standard measures.

In the following sections, we shed light on important properties of bibliometric indicators from the viewpoint of six aspects concerning their definition and use. To each of them a separate section will be devoted.

Deterministic versus probabilistic approach

Beyond doubt, the *deterministic approach* is the easiest way to process simple counts of raw data and measurements to indicators. Mostly elementary mathematical operations (e.g., shares, averages, ratios) are applied; more sophisticated techniques such as transformations are the preferred tools for data presentation and visualization. Beyond the application of rather simple methods, the interpretation of more complex measures and constructs, for instance, composite indicators, becomes increasingly problematic. Yet, observations and measurement are in practice subject to a variety of influences, most of which are not apparent or at least not quantifiable and thus not directly measurable. Even social processes, although more dependent on the actual and individual constellation and therefore less well reproducible if the same individuals consciously take action under the same conditions, seem to be influenced by random effects. It is the complexity of social interactions itself that yields the effect that one usually interprets as randomness. In bibliometrics, too, the events we become aware of, seem to be random as they are conditioned by a plethora of superposing actions, processes and effects; communication, mobility, collaboration, publications and citations all are subject to these effects. Even the unlikely might happen with positive probability, if the number of “trials” is large enough, as Stanisław Lem has impressively described in his novel “The Chain of Chance” (Lem 1978).

Beyond the determinism in social processes, a probabilistic component causes all events to happen with a certain likelihood, that the *impossible event* might happen anyway, that the *almost sure event* might fail to happen and that this becomes measurable as well. Thus the *probabilistic approach* assumes random effects and provides stochastic methods. Events might imply other ones with a *certain probability*. This approach broadens the spectrum of applications and interpretation. While validity of measurement can still be guaranteed in the deterministic approach, validity of the actual application remains problematic even if appropriate deterministic models are used. For instance, the question as to what level of aggregation data could be broken down to, or what the error rate tolerance of a measurement might be, can only be answered by the probabilistic approach. In particular, this helps to construct confidence intervals (which is important for ranking) and limitations concerning the level of aggregation (for the evaluation of individuals or teams).

The stochastic nature of bibliometric indicators

The calculation of shares for a unit of assessment—for instance, the world share of articles published from a particular country—plays an important role in the deterministic approach. One important advantage of the probabilistic approach is that shares of units of analysis that take certain values, or values in given ranges, can be considered part of empirical distributions. As will be explained below, such *relative frequencies* have important *asymptotic* properties providing *unbiased* and *consistent estimators* of the assumed corresponding theoretical distribution function and their moments. Similarly to relative frequencies, means of empirical distributions are unbiased and consistent estimators for the corresponding expected value.

Another important feature of the stochastic approach is the introduction of a time-dependent parameter resulting in *stochastic processes* that are able to reflect the changes of probabilities, moments and empirical values in time. As early as in 1976, Dieks and Chang introduced a mathematical model describing citing as a stochastic process. The model enabled one to ascertain what differences in citation rates are to be considered significant, i.e., not caused by mere chance, with a certain probability. The total number of citations was assumed to follow a Poisson distribution, often used to describe a situation in which many events occur with a low probability, and independently of one another. The parameter of the Poisson distribution itself could be assumed to be a random variable (according to the subject, age, social status, etc. of the author) at any time so that a *compound process* (e.g., a negative binomial or generalized Waring process) is obtained (cf. Burrell 2005).

An often raised question concerns the use of means if the underlying distribution is not only non-Gaussian, but also skewed and integer-valued. The answer consists of two parts.

1. Is the mean to be used to represent the individual observation?
2. Or is the mean to be used for estimation or for comparison with other means from similar distributions?

In the first case, the mean should certainly *not* be used if the underlying distribution is very skewed, and has a long tail. The same applies, under certain conditions, to symmetrical distributions as well, such as the normal distributions with large standard deviation, (e.g., if the *coefficient of variation* is greater than 50 %). In the second case—and this is the essential one for building and using indicators—this question is closely related to the number of free parameters of the underlying distribution. Distributions with one free parameter, e.g., the Poisson, the geometric, exponential or Pareto distribution are uniquely determined by the *expectation*, provided this is finite. Two Poisson distributions with the same expected value are necessarily identical. This implies that the comparison of two means makes sense only if the two underlying distributions are of the same type and have only one free parameter. Even if there are skewed, these distributions are best described by their mean. At the same time the mean often provides an efficient, unbiased estimator for the parameter. However, if the number of free parameters is greater than one, additional information is needed. This holds for the normal distribution as well. In order to illustrate this, we use the following example from zoology. For a housecat we "compare" two statistics. (1) The average number of kittens per litter amounts to about four. (2) The average number of legs per kitten amounts to about four. While the probability of three or five kitten in a litter is quite large, the probability of a kitten born with three or five legs is, fortunately, minute. Without applying any particular test, one could state that a litter size of five does not significantly deviate from the expectation, but a new-born kitten with five legs does. Although the underlying distributions are almost symmetrical and have the same expected value, their *standard deviations* completely differ. In the case of the normal distribution the standard deviation forms the second free parameter.

Distribution of means and shares

The use of means in comparative analysis gave often rise to a further misunderstanding: comparing the means of two different sets is not the same as comparing individual observations in the sets, for which the means might not be representative, because of their skewness and large *variation*. One of the base properties of sample means is that those have an approximately *normal distribution* and converge to the expectation of the underlying common distribution, provided this distribution belongs to the attraction domain of the normal distribution. This property, which is a consequence of the *Central Limit Theorem*, even holds for skewed and integer-valued distributions (cf. Glänzel 2010). The convergence speed, of course, depends on several properties of the distributions such as the shape and continuity. In the literature thresholds between 30 and 50 observations are indicated for acceptable approximation in practice and the application of tests requiring normality. In particular, Vincze (1974) mentions 40 for the context of the Welch-test. Furthermore, the standard deviation of the mean equals the standard deviation of the common distribution divided by the square root of the sample size. The same applies to the share of uncited papers with respect to the probability that a paper is not cited (Glänzel and Moed 2002). In order to illustrate this, we have selected 20 random samples from the Belgian publication output in 2004 as indexed in the *Science Citation Index Expanded* of Thomson Reuters' *Web of Science* (WoS). Only so-called citable documents, that is, articles, letters, reviews and proceedings papers published in journals have been taken into account. Citations have been counted for the 3-year citation window 2004–2006. In total 20 samples have been drawn representing about 1 % each of the Belgian publication activity as reflected by the WoS database. Since the samples do not overlap, their union represents about 20 % of the Belgian total. The sample means and the shares of uncited papers are presented in Table 1.

Both statistics seem to be randomly distributed around the common expectations. According to a characterization theorem for Gaussian distributions, a random variable X has a normal distribution *iff* the equality $D_X(x) = \mathrm{m}\cdot d_X(x) + \sigma^2$ holds for some real value m and positive real value σ, where $d_X(x) = \mathrm{E}(X \mid X \geq x) - x$, $D_X(x) = \mathrm{E}(X^2 \mid X \geq x) - x\cdot\mathrm{E}(X \mid X \geq x)$ for all real x, $m = \mathrm{E}(X)$ and $\sigma = [\mathrm{E}(X^2) - \mathrm{E}(X)^2]^{½}$ (Glänzel 1990). Substituting the functions $D_x(x)$ and $d_X(x)$ by the corresponding sample statistics, a normality test for random variables with unknown parameters is obtained. Figure 1 shows the results for both the mean values and the relative frequencies of uncitedness. The correlation is in both cases very strong. Consequently the statistics are approximately normally distributed. A simple test also substantiates that the totals in Table 1 ($\bar{x} = 5.788$ and $f_0 = 21.8$ %) representing 19.5 % of the Belgian publications in 2004 do not deviate

Table 1 Sample means and shares of uncited papers (1 % of Belgian publications in 2004 with 3-year citation window) [Data sourced from Thomson Reuters Web of Knowledge]

k	n	$\bar{x}$	f_0 (%)	k	n	$\bar{x}$	f_0 (%)	k	n	$\bar{x}$	f_0 (%)
1	126	5.016	20.6	8	113	5.265	21.2	15	108	5.028	16.7
2	94	7.160	23.4	9	115	5.652	18.3	16	108	4.102	25.0
3	133	5.639	20.3	10	117	5.538	20.5	17	130	6.362	24.6
4	122	5.951	19.7	11	122	5.385	19.7	18	137	4.569	28.5
5	126	6.262	22.2	12	149	6.913	23.5	19	114	4.456	19.3
6	112	5.768	17.0	13	103	4.641	29.1	20	110	6.473	21.8
7	128	6.992	22.7	14	145	7.807	21.4	Total	2,412	5.788	21.8

k sample number, n sample size, $\bar{x}$ sample mean, f_0 share of uncited papers

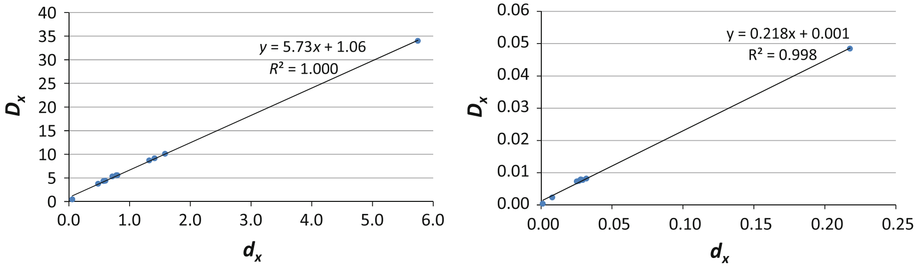

Fig. 1 Plot of truncated moments for 20 sample means and shares of uncited papers (1 % of Belgian publication in 2004 with 3-year citation window) [data sourced from Thomson Reuters Web of Knowledge]

significantly from the mean citation rate and share of uncited papers of the total Belgian publication output which amounted to 5.72 and 21.5 %, respectively. The absolute values of the corresponding *w*-statistics (cf. Glänzel and Moed 2002) amount to 0.471 and 0.377, respectively. Both values are far below the critical value of 1.96 corresponding to the confidence level of 0.95.

The above tests refer to random samples. In real life, however, statistics are built for units of analysis that are far from being considered random samples. In such cases, the possible significance of deviation can be interpreted as an indication that the units in question are biased in one or the other direction, and cannot simply be considered a random sample of the same population and representing its standard. Furthermore, the assumption of one single free parameter does not reflect reality (see e.g., Glänzel 2009) of bibliometric practice. Based on the experience, at least two parameters are needed to model publication activity or citation impact. Taking into account that both phenomena change over time, and thus form stochastic processes, the time windows should always be compatible in comparative analysis. The issue of the multidimensionality of the parameter space can be solved by the use of several independent statistics. As a workaround mean and share of (un)cited papers might be used as has been done in our example, although these two statistics cannot be considered completely independent (cf. Glänzel 2009).

More sophisticated derivatives of empirical distributions provide better insight into the statistical properties of bibliometric processes than a mean or relative frequency could do. These derivatives, zones usually derived from a reference distribution, such as the citation distribution of a scientific journal or a complete discipline. These zones might be self-adjusting as in the case of the *Characteristic Scores and Scales* (CSS, see Glänzel and Schubert 1988) or predefined by given percentiles (Leydesdorff et al. 2011). The first approach proceeds from iteratively truncated moments and usually results, when applied to citation impact, in four self-adjusting zones (classes) ranging from poorly till outstandingly cited. This method, which has interesting mathematical properties in the case of *Paretian* distributions, as has been shown in the above-mentioned paper, can be applied to grade individual publications as well as to compare citation-impact statistics and distributions of given units of analysis with the corresponding reference standard. The second method proceeds from a pre-set set of six rank percentages calculated for the reference distribution. Individual observations are then scored according to the rank percentage the publications in question belong to. The proposed R(6) indicators is then defined as the average score over the papers by the unit of analysis. Since the authors found objections to the use of averages, in a later paper they revised their approach by summing up the rank scores obtained from the rank percentages of the reference distribution (Leydesdorff and

Bornmann 2011). Their new indicator is called *Integrated Impact Indicators* (I3). In both cases the impact of a given unit is characterized by one single indicator again.

Properties proposed for correct applications

In this section we point at some basic requirements on indicators from the viewpoint of statistical functions and estimators for moments, probabilities or parameters. In this respect, one has to distinguish between two essential issues, a methodological one and the implications for application.

Consistency of indicators

Indicators should have at least one of the following important properties. The first one is *consistency*. An estimator of a parameter is said to be (weakly) consistent if it converges in probability to the true value of the parameter it estimates. In verbal terms, estimates should improve as the number of observations increases, that is, the more observations we have, the smaller the error. In practice, this means that the lesser the tolerance, the more reliable conclusions one can draw. A somewhat weaker requirement is *asymptotic unbiasedness*. An estimator of a parameter is said to be *asymptotically unbiased* if its expectation converges to the true value of the parameter. For instance, means are unbiased and consistent estimators of the expected values and relative frequencies of the corresponding probabilities, as has already mentioned above.

Important implications for the application to bibliometric indicators are, that there are limitations concerning the size of the underlying set the indicators represent. There are no clearly defined "lower bounds" for application, but as a rule of thumb a value of 50 is suggested as minimum value for approximate properties such as "normality" of the distribution of means and relative frequencies (cf. 'Distribution of means and shares' above). In the above example a sample size of the order of magnitude of 100 was used and provided acceptable results. Another implication is that seemingly large deviations between different values of the same indicator need not necessarily significant. This is often a consequence of large standard deviations that are in bibliometrics usually caused by the superposition of small sizes and skewed distributions. In terms of ranking according to indicators, ties might occur where indicators have actually taken different values (see Glänzel 2010).

Recently a new interpretation of 'consistency' was proposed by Waltman et al. (2011). According to their definition, which should not be confused with consistency in a mathematical-statistical sense (see paragraphs above), an indicator is consistent if the order relation between the indicator values of two identically sized sets will not change if the same value is added to both sets. Even stricter conditions for indicators have been introduced by Rousseau and Leydesdorff (2011). According to their requirement of 'ranking invariance' (with respect to non-cited items) the rank of two sets of not necessarily the same size should not change if the same amount of items with zero value is added. Intuitively both requirements sound justified. However, most statistics fail to meet these criteria. For instance, it is not difficult to show that the median and percentiles are not 'consistent' in the sense of the definition by Waltman et al. and that neither means nor medians and percentiles are 'ranking invariant'. So what is wrong with classical statistics? According to Rousseau and Leydesdorff their requirement tells against the use of means in the case of highly skewed distributions. Yet, 'ranking invariance' has the same effect with respect to rarely or highly cited items as well. It is a truism that small sets are sensitive to changes, whereas large sets can more easily absorb new items that deviate from their profiles. But here statistics meets life. In particular, the real problem is that these two requirements are *not* formulated as *asymptotic* properties and thus they pave the way for questionable application of indicators at the small scale, where (mathematical) statistics should not be used at all.

One should also be aware that rigid application of such formal criteria may lower the level of sophistication of composite indicators. For instance, as regards the journal metric Source Normalized Impact per Paper (SNIP, see 'Normalization issues' below) it is true that the score of a journal may decrease when a journal obtains an additional citation. But this can also be conceived as a sign of sophistication of relative, composite indicators: this one "additional citation" provides at the same time additional information on a journal's subject field, and can for instance reveal that the citation potential in the field is higher (or lower) than was previously estimated without the citation.

High-end versus the common run?

One of the *key issues* in present-day bibliometrics is that researchers in our field, even though they recognise the necessity of rather complex approaches, tend to condense their statistics into one single indicator at the end, but demand solutions for inference at the individual level at the same time. This of course forms an insoluble conflict: on the one hand, measures based on larger sets are projected on the linear scale (e.g., to allow for ranking) and, on the other hand, individual cases are sought to be assessed using the same or similar measures. In order to solve this paradox, the question as to how to treat the high-end of a skewed distribution (e.g., the highly cited papers representing only a minute share of the total) should be separated from those statistical questions that are based on large numbers. Quantiles (percentiles) or CSS-type zones can be used for assessing the high-end. Furthermore, the tail behaviour of bibliometric distributions can be specified using tail indices based on even the same distribution model as used for other statistics. Unfortunately, estimators of tail indices and extreme-value statistics are often not unbiased although there are methods to build asymptotically unbiased or at least bias-reduced estimators in some cases (cf. Peng 1998). This is the price for dealing with extreme values.

A key theoretical issue is why citation distributions are skewed, Moed (2005) claims that the development of a theoretical-conceptual framework reaches beyond the view of research output as a collection of individual papers, and proposes to conceive research articles as elements from coherent publication ensembles of research groups carrying out a research programme. Citing authors acknowledging a research group's works do not distribute their citations evenly among all papers emerging from its programme, but rather cite particular papers that have become symbols or 'flags' of such a programme. Citations to these flag papers can be conceived as citations to the entire oeuvre and to the programme embodied in it (cf. Moed 2005, pp 216–218). A stochastic model of citation should take these processes into account (see 'Conclusions and perspectives').

Time perspective and bibliometric indicators

In this section we will discuss further important opportunities but also caveats that result from the time perspective. Even reference standards for bibliometric measures are by no means physical constants. They are, among others, subject to changes in time. This time perspective is twofold: on the one hand, identically built and structured indicators provide different results if they are calculated in different instants of time and, on the other hand, indicators calculated in the same instant of time will change as time elapse. The first perspective is called *time series.* The perhaps most popular time series in bibliometrics is the *Impact Factor Trend* provided by Thomson Reuters' Journal Citation Reports (JCR) for decades.

Time-dependence of bibliometric measures

The second perspective refers to *processes.* Prominent examples are the changing publication behaviour in a scientist's career or the changing citation impact of a paper in different citation windows. While quite a range of standard tools is available for the analysis of time series, the second involves risks and pitfalls that might result in severe distortion of indicators built for measurement and assessment.

The first pitfall refers to the choice of citation windows and to the so-called "age normalization" as, for instance, proposed in the context of the h-index. Neither is the growth of citation impact linear (the annual increments are not constant), nor can the time window simply shifted along the time axis. The latter property is a result of the inhomogeneity of the underlying processes. Both effects can be easily observed for the publication activity of authors or citation rates of paper sets using the "analysis tools" of Elsevier's SCOPUS and Thomson Reuters' Web of Science. Also the diagram in Fig. 2 visualises both effects using the example of four selected journals (Annals of Mathematics [AMATH], Physics Letters B [PYSLB], Cell and JACS) representing four different fields of the sciences. The diagram shows the annual change of citations received by papers published in 1980 during 21 years beginning with the publication year. One of the most important and obvious consequences of the non-linearity of bibliometrics growth processes is the invalidity of normalization by age or the length of the citation window.

Diachronous versus synchronous indicators

A second consequence of the above-mentioned two properties is the incompatibility or complementarity of diachronous (prospective) indicators with synchronous (retrospective) indicators. Since citation impact of recent publications based on sufficiently large citation

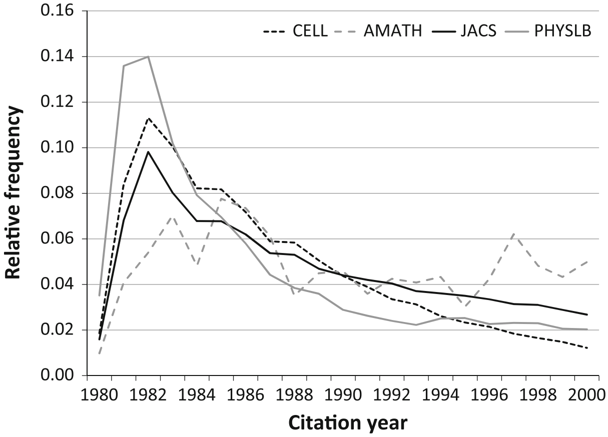

Fig. 2 Annual change of citations for four selected journals (Publication year 1980) (data sourced from the CD edition of Thomson Reuters Science Citation Index)

windows point to the future, and is therefore unknown, the superposition of shifted publication years with fixed citation year and thus variable citation windows is often used as substitute. This means instead of the diachronous 5-year citation impact with fixed publication year, say, 2011 and citation window 2011–2015, for instance, the citation year (2011) is fixed and the publication year moves over several years (e.g., 2006–2010 as in the case of the JCR 5-year impact factor, or 2007–2010 in SCImago's 4-year based 'cites per doc' journal indicator). This approach has become popular as it provides kind of long-term citation impact. Yet, these indicators are synchronous measures and they point to the past. In principle, this approach is certainly useful but involves the risk of wrong interpretation and application. But valid synchronous indicators of citation impact are feasible if one properly takes into account the age distribution of cited articles of the unit under assessment. The use of the 2-year impact factor issued for citation year *x* as kind of reference standard for the publication year *x* is one of the most common improper practices. But even if properly used, synchronous indicators tend to cause distortions notably in the case of small units. This might be illustrated by the following example. The annual publication output of dynamically growing or a declining small research unit might distinctly deviate from the general trend of the reference standard. The above-mentioned nonlinearity and non-homogeneity of the superposing citation processes might then result in a bias. Diachronous indicators should therefore be favoured in the evaluation of research performance. On the other hand, it must be noted that diachronous approaches can be easily distorted by changes in coverage of the database in which the citation analysis is carried out. Properties of diachronous and synchronous indicators have been analysed and discussed further, e.g., by Ingwersen et al. (2001), Glänzel (2004), Frandsen and Rousseau (2005).

Normalization issues

A further commonly known issue is the necessity of normalization, notably in a multidisciplinary environment. Communication behaviour differs considerably among and even within the various subject fields. This applies to both publication activity and citation impact. The diagram on the right-hand side of Fig. 2 shows the large deviation of impact among top journals of four different science areas. Subject normalization aims at improving the validity of indicators and one of its main field of application is at present (journal) citation indicators. Recently two paradigms emerged, namely normalization *before* and *after* citations are counted. The second approach, which is already in use since the early 1980s, does not require any recalculation of the original citation rates, and has therefore gained popularity. This popularity is contrasted by severe flaws. We mention subject normalization just as an example in this context. In bibliographic databases most journals or individual documents are assigned to more than one discipline or subject. This implies that subject normalization of documents or journals has to adjusted to the thematic environment in which the indicator is to be used. In extreme cases, interdisciplinarity might thus make this type of normalization meaningless since it is done at the cited side. This type of normalization is consequently called *cited-side* (Zitt and Small 2008), target (Moed 2010) or a posteriori normalization. An interesting property of an a posteriori solution is nevertheless worth mentioning: it transforms the distribution of journal impact measures within each discipline to a lognormal distribution (Beirlant et al. 2007) and thus facilitates the use of standard percentiles.

Zitt and Small (2008), Zitt (2010), Moed (2010, 2011), Leydesdorff and Opthof (2010), and Glänzel et al. (2011) have run another path towards citation normalization. Citations are immediately normalized from the *citing side* before indicators are built and this type is also referred to as "citing side"(Zitt and Small 2008), "source" (Moed 2010) or "a priori" normalization. Unlike a posteriori (cited-side) normalization, the latter type is expected to result in citation-based indicators that are largely insensitive to subject-specific peculiarities.

In fact, a strong feature of citing side or source normalized citation impact indicators is that they do account for disparities in citation potential among subject fields, but their calculation does not depend upon an a priori subject field classification, such as the journal subject field classifications currently available in Web of Science or Scopus. A subject field is defined as the collection of articles (Moed 2010) or journals (Zitt and Small 2008) citing a particular journal. Zitt and Small (2008) and Moed (2010) apply their methods to the analysis of journals, although these methods can be easily extended to other units of analysis. Zhou and Leydesdorff (2011) apply their approach to citing side normalization to the assessment of university departments.

SNIP constitutes a bridge between the 'classical' journal metric—of which the Thomson-Reuters journal impact factor is the exemplar- and the citing side normalization methodologies. SNIP is a ratio of a journal's classical metric in the numerator, and in the denominator a normalized measure of the citation potential in the subject field covered by that journal. Citation potential reflects the frequency at which papers in a subject field cite one another, but also takes into account the degree to which a subject field is covered by the citation database in which it is calculated. The normalization of the citation potential ensures that the subject field covered by the median journal in the database (in terms of citation potential) has a normalized citation potential of one. Consequently, compared to the classical metric, 50 percent of journals has a SNIP value that is lower than that of the classical metric (e.g., molecular-biological journals), while for another 50 percent it is higher (for many journals in mathematics, humanities and parts of social sciences and engineering).

An important issue is the extent to which such a delimitation can lead to bias, and how such a bias can be accounted for (e.g., Zitt 2010; Moed 2011; Leydesdorff and Opthof 2010; Waltman 2011). One possible source of bias is the fact that articles in the subject field covered by a unit of analysis but *not* citing that unit, are *not* taken into account. Moreover, as pointed out by Zitt (2011), citing side field normalization does not account for the growth of the literature in a subject field.

Size dependence

When the first citation measures were introduced, one argument in favour of using arithmetic means was the possibility of comparing journals regardless of their size (Garfield and Sher 1963). But mean values have their limitations too. Calculated at the level of research groups they may be affected by publication strategies: groups selectively publishing only their best work tend to obtain a higher mean than groups distributing their findings among a large number of articles. That mean values are in fact not always size independent may also have to do with closeness of small communities and self-citations, but does not affect the validity of the original idea. The idea of using "size dependent" indices to measure research performance already arose since bibliometrics found that a combination of research output with impact would better reflect performance. Alvarez and Pulgarin (1996) proposed an improvement of the journal impact factor using the psychometric *Rasch model*. Their model has not found much response in the bibliometric community. More success was accorded to the *h-index* suggested by the physicist J.E. Hirsch (2005). This indicator provides a simple combination of a scientist's publication activity and citation impact. Thanks to its simplicity it has immediately found interest in the public, and received positive reception both in the physics community and the scientometrics literature. Despite its popularity, the h-index challenged criticism too. The fact that a scientist with more citations (in total and on an average) than a colleague might have a lower h-index although both have published the same amount of papers, the "defiance" of the small-is-beautiful principle (i.e., that the small elite might score worse than the large mediocre, cf. Glänzel 2006) and other apparently counterintuitive properties of this measure have soon fostered scepticism. Nevertheless, several scientometricians with mathematical background have studies the otherwise interesting statistical properties of this indicator. Beirlant and Einmahl (2007) have proven the statistical *consistency* and the *asymptotic normality* of the empirical h-index and Barcza and Telcs (2009) have found the formula of the probability distribution of the h-index. Despite its apparent simplicity, the h-index proved to be a complex indicator with non-trivial properties that implicate a careful use of this measure.

A similar combination of publication output and citation impact has recently been presented by Leydesdorff and Bornmann (2011). Their *Integrated Impact Indicator* (I3) is defined as the *sum* over the rank scores of the unit under study obtained from the rank percentages of the reference distribution (cf. 'Deterministic versus probabilistic approach' above). I3 has been suggested as alternative measure of journal impact. This indicator has similar properties concerning small sets and citations as described above in the context of the h-index.

Network indicators

In the classical, linear model, publications are considered as separate entities, and citations as separate events. But both citing and cited articles or authors may have all kinds of relationships that need to be taken into account when assessing citation impact of a unit of assessment. At the cited side, a research group's publications can be viewed as elements of a coherent research program, and individual publications may be symbols of the program as a whole. At the citing side, one particular article or author may cite a group's oeuvre more than once; such citations are not independent.

The following example may clarify the relevance of taking into account the relationships between cited and between citing papers. If one target paper is cited by one source article, its cites-per-paper ratio obviously amounts to one. But if the target's content is distributed among two papers both published by the same author(s), and if the same is true for the citing article, while each of these cites both target papers, the absolute number of received citations of the target oeuvre amounts to 4, and the cites-per-paper ratio to 2. In this way, authors debating a specific subject and formally publishing each single step in a sequence of arguments in a separate journal paper, may easily obtain not only higher absolute numbers of citations but also higher 'raw' cites-per-paper values than authors who carry out their debate mainly via informal channels and who publish in the end one single paper summarizing the debate and presenting its conclusions.

Some bibliometric indicators can be derived from network representation. Typical examples are co-authorship and citation links providing the measures of co-publication activity and citation impact. The degree distributions of the vertices in co-authorship and citation graphs directly lead to those indicators that have been discussed in the previous section. Algebraic manipulations of matrices representing these networks can then be used to analyse structural aspects of the networks and to build “higher-level” indicators. So-called ‘scientometric transaction matrices’ have been introduced and analysed by de Solla Price (1981) based on the citation flow among units. From the historic viewpoint we would also like to mention the ‘citation influence’ methodology suggested by Pinski and Narin (1976) and the Markov-chain approach by Geller (1978). These models have been applied to citation transactions among scientific journals. This model assumes that journals have different weights on the basis of the number of their citations or references. A citation has a larger weight if it is received from a journal with higher impact. Since this impact also depends on citations received by other journals, the solution can only be found by iteration. Proceeding from similar considerations, more recently Brin and Page (1998) developed the he ‘Google PageRank’, an algorithm for scoring websites and page ranking, that is based on the analysis of web links. The Sciago Journal Rank indicator (SJR) (cf. Gonzalez-Pereira et al. 2010) and Thomson Reuters’ Eigenfactor/Article Influence Score (see Bergstrom et al. 2008) are based on similar algorithms.

Network based indicators, such as indicators derived from ‘citation influence’ and PageRank-type algorithms have one important property that might be essential for their application to bibliometrics. Units (e.g., journals) entering the network might influence indicator values of those units with which they are *not directly* linked. Validity requirement concerning network indicators should therefore be considered from a completely different perspective as is usually adopted in the case of traditional statistical functions. Network indicators still need further research to understand their power and limitations.

Conclusions and perspectives

Following the model by Dieks and Chang (1976) introduced in ‘The stochastic nature of bibliometric indicators’ one approach could be to pick up the notions embodied in their paper, and further develop these into a model that would be able to deal with aggregations of papers. A first major challenge would be to take into account at the cited side the relationships that exist between the articles published by a particular unit under assessment, in terms of whether they are from the oeuvre of the same research, since Dieks and Chang focused on the level of individual articles. A second task would be to operationalize the concept of independent citations. Chang and Dieks found that the number of first authors citing a paper in a particular year more neatly follows a Poisson distribution than the number of “crude” citations. During the past decades, co-authorship has increased so strongly, and scientific collaboration networks have become so dense, that it is questionable whether counting the number of citing first authors in a year is still the most appropriate method today.

Furthermore, instead of searching for new indicators, a possible approach in the assessment of a set of n publications made by a particular unit of assessment is to draw random samples each of size n from the population in which the analysis is made, characterise the distribution of sample means, localise the unit’s average citation impact in the distribution of sample means, and determine the probability what the unit’s impact deviates from a random sample. In this way one is able to assess whether the citation-per-paper ratio of a particular unit of assessment deviates from that of a random sample drawn from the sub-population of documents being similar in terms of age, subject and type published by the unit.

In the past ISI's Science Citation Index (SCI) was the standard in citation analysis. Its content coverage was selective, and contained only the most import journals with a citation impact exceeding a certain minimum threshold. SCOPUS has a broader coverage and indexes also journals with a lower citation impact and a more local or national relevance. Thomson Reuters' Web of Science also expands its content coverage towards national or regional journals. At the same time, both indexes increased their coverage of conference proceedings and books. Bibliometricians must analyse the consequences for their citation based methods and indicators of these significant changes in content coverage. Indicators should be developed reflecting the effect of changes in the degree of content selectiveness or changes in content coverage over time. One approach is the extension of the notions of Pinski and Narin (1976) of weighting citations by prestige of the citing articles (see 'Network indicators'), and applying these to the assessment of research groups or departments rather than scientific journals.

Acknowledgments This is a version of a paper presented by the first author at the 17th International Conference on Science and Technology Indicators (Montreal, 5–8 September 2012). The full proceedings are published as: E. Archambault, Y. Gingras, V. Lariviere (Eds.), Proceedings of STI 2012 Montreal, Science-Metrix and OST, Montréal, Quebec, Canada. The authors would like to thank Dr. Bart Thijs (ECOOM/KU Leuven) for the statistical simulation based on the Belgian samples.

References

Alvarez, P., & Pulgarin, A. (1996). Application of the Rasch model to measuring the impact of scientific journals. *Publishing Research Quarterly, 12*, 57–64.

Barcza, K., & Telcs, A. (2009). Paretian publication patterns imply Paretian Hirsch index. *Scientometrics, 81*, 513–519.

Beirlant, J., Einmahl, J. H. J. (2007). Asymptotics for the Hirsch Index. CentER Discussion Paper #2007-86. Accessible at http://arno.uvt.nl/show.cgi?fid=65780.

Beirlant, J., Glänzel, W., Carbonez, A., & Leemans, H. (2007). Scoring research output using statistical quantile plotting. *Journal of Informetrics, 1*, 185–192.

Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The eigenfactor (TM) metrics. *Journal of Neuroscience, 28*, 11433–11434.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*, 107–117.

Burrell, Q. L. (2005). The use of the generalized Waring process in modelling informetric data. *Scientometrics, 64*, 247–270.

Dieks, D., & Chang, H. (1976). Differences in impact of scientific publications: some indices from a citation analysis. *Social Studies of Science, 6*, 247–267.

Frandsen, T. F., & Rousseau, R. (2005). Article impact calculated over arbitrary periods. *JASIST, 56*, 58–62.

Garfield, E., & Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation, 14*(3), 195–201.

Geller, N. L. (1978). On the citation influence methodology of Pinski and Narin. *Information Processing and Management, 14*, 93–95.

Glänzel, W. (1990). Some consequences of a characterization theorem based on truncated moments. *Statistics, 21*, 613–618.

Glänzel, W. (2004). Towards a model for diachronous and synchronous citation analyses. *Scientometrics, 60*, 511–522.

Glänzel, W. (2006). On the opportunities and limitations of the H-index. *Science Focus, 1*, 10–11.

Glänzel, W. (2009). The multi-dimensionality of journal impact. *Scientometrics, 78*, 355–374.

Glänzel, W. (2010). On reliability and robustness of scientometrics indicators based on stochastic models. An evidence-based opinion paper. *Journal of Informetrics, 4*, 313–319.

Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. *Scientometrics, 53*, 171–193.

Glänzel, W., & Schubert, A. (1988). Theoretical and empirical studies of the tail of scientometric distributions. In L. Egghe, & R. Rousseau (Eds.), Informetrics 87/88 (pp. 75–83). Elsevier Science Publisher B.V.

Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2011). A priori vs. a posteriori normalisation of citation indicators. The case of journal ranking. *Scientometrics, 87*, 415–424.

Gonzalez-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegon, F. (2010). A new approach to the metric of journals' scientific prestige: the SJR indicator. *Journal of Informetrics, 4*, 379–391.

Hirsch, J. E. (2005). An index to quantify an individual's scientific output. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 16569–16572.

Ingwersen, P., Larsen, B., Rousseau, R., & Russell, J. (2001). The publication-citation matrix and its derived quantities. *Chinese Science Bulletin, 46*, 524–528.

Lem, S. (1978). *The chain of chance*. New York: Harcourt Brace Jovanovich.

Leydesdorff, L., & Bornmann, L. (2011). Integrated Impact Indicators compared with impact factors: an alternative research design with policy implications. *JASIST, 62*, 2133–2146.

Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables on citation analysis one more time: principles for comparing sets of documents. *JASIST, 62*, 1370–1381.

Leydesdorff, L., & Opthof, T. (2010). Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *JASIST, 61*, 2365–2369.

Moed, H. F. (2005). Citation analysis in research evaluation. Springer, Dordrecht, 346 pp, ISBN 1-4020-3713–9.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics, 4*, 265–277.

Moed, H. F. (2011). The source normalized impact per paper is a valid and sophisticated indicator of journal citation impact. *JASIST, 62*, 211–213.

Peng, L. (1998). Asymptotically unbiased estimators for the extreme-value index. *Statistics & Probability Letters, 38*, 107–115.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications. *Information Processing and Management, 12*, 297–312.

Price, D. D. (1981). The analysis of square matrices of scientometric transaction. *Scientometrics, 3*, 55–63.

Rousseau, R., & Leydesdorff, L. (2011). Simple arithmetic versus intuitive understanding. *ISSI Newsletter, 7*, 10–14.

Vincze, I. (1974). *Mathematical Statistics*. Eötvös University Budapest, (4th ed.) (in Hungarian).

Waltman, L. (2011). *Personal communication*.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: some theoretical considerations. *Journal of Informetrics, 5*, 37–47.

Zhou, P., & Leydesdorff, L. (2011). Fractional counting of citations in research evaluation: a cross- and interdisciplinary assessment of the Tsinghua University in Beijing. *Journal of Informetrics, 5*, 360–368.

Zitt, M. (2010). Citing-side normalization of journal impact: a robust variant of the audience factor. *Journal of Informetrics, 4*, 392–406.

Zitt, M. (2011). Behind citing-side normalization of citations: some properties of the journal impact factor. *Scientometrics* 89, 329–344.

Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: the audience factor. *JASIST, 59*, 1856–1860.