KSZI [ξ] AKTÁK

MTAK TTO műhelytanulmányok

2012/4

High-End Performance or Outlier? Evaluating the Tail of Scientometric Distribution

Wolfgang Glänzel

wolfgang.glanzel@econ.kuleuven.be

➲ http://www.mtakszi.hu/kszi_aktak/

High-End Performance or Outlier? Evaluating the Tail of Scientometric Distribution

WOLFGANG GLÄNZEL

Centre for R&D Monitoring and Dept. MSI, KU Leuven, Belgium
and
Dept. Science Policy & Scientometrics, Library of the Hungarian Academy of Sciences, Budapest, Hungary

ABSTRACT

The present paper attempts to shed light on outstanding research performance using the example of citation distributions. In order to answer the question of how the analysis of outstanding performance, in general, and highly cited papers, in particular, could be integrated into standard techniques of evaluative scientometrics. Two general methods are proposed: One solution aims at quantifying the performance represented by the tail of citations distributions independently of the "mainstream", the second one, a parameter-free solution, provides performance classes for any level. Advantages and shortcoming of both methods are discussed.

1. INTRODUCTION

In an earlier study of the statistical background of scientometric indicators (Glänzel and Moed, 2012) the issue of 'outliers' was raised. While in many fields outliers can simply be discarded as being exceptions, in bibliometrics the extreme values represent the high-end of research performance and deserve therefore special attention. The authors addressed the question of in how far extreme-value statistics can serve as supplementary indicators to the standard measures but only general suggestion were given. In the present paper this issue will be deepened by combining results of extreme-value theory with statistical methods in scientometrics. As for the theory of extreme values, I mainly refer to the results by Emil Julius Gumbel and, more recently, by Jan Beirlant and his collaborators.

On the basis of the pioneering work by L. Tippet and R.A. Fisher, E.J. Gumbel published his book entitled *Statistics of Extreme* on the theory of extremes in 1958. The Gumbel distribution, which is one of the three possible extreme-value distributions and Gumbel's characteristic extreme values are closely related to this theory.

According to the Fisher-Tippet limit theorem for maxima (Fisher and Tippet, 1928), two cases are of particular interest:

i) Fréchet distribution $F(x)=\mathrm{e}^{-x^{-a}}$ for $x > 0$ for Pareto-type distribution and
ii) Gumbel distribution $F(x)=\mathrm{e}^{-\mathrm{e}^{-x}}$ for distributions of exponential type

It is not the objective of this study to analyse and discuss the properties of these particular limit distributions, but it should be mentioned that the above types illustrate that extreme values of Paretian and exponential distributions behave in a different way. This important property of distributions concerning the behaviour of their extreme values has strong effect on the evaluation of the observations in the tail of different types of empirical distributions. Most distributions in scientometrics are assumed to be of Paretian type, that is, they approximately follow a power law. In particular, scientometrics mainly deals with distributions derived from authorship or citation networks. These include publication activity, co-authorship, citation rates and number of references. In evaluative bibliometrics, publication productivity and citation impact form the most important distributions. Before the issue of outstanding performance is tackled, a short general introduction into the theory of scientometric distributions is given. All steps are illustrated by examples from real-life distributions.

2. A CONCISE DISCOURSE ON SCIENTOMETRIC DISTRIBUTIONS

One single statistic is usually not sufficient to describe the distribution of citations by papers in an adequate manner (Glänzel, 2009). Using the example of scientific journals covered by Thomson Reuters' Web of Science database, the author has shown that journals in the same field might have a similar mean citation impact although the relative frequency of their unicited papers considerably differs. Indeed, shares of uncited papers or inactive authors (f_0) and mean values (m) are the most frequently used statistics that can directly be derived from the mentioned distribution models. Although means are certainly affected by extreme values to a certain extent, most distribution models provide acceptable fits on the basis of these two statistics to the lower end and the central section of those skewed distributions that are typical of bibliometrics. The following example might illustrate this effect. Beforehand, some notations and the necessary background for this exercise is introduced.

In this context, two models with two free parameters each are chosen to describe typical skewed bibliometric distributions. The first one, the *negative binomial distribution* represents the exponential type, while the other one, the *Waring distribution* stands for the Pareto-type.

A non-negative integer-valued random variable X is said to have a negative binomial distribution, if

$$p_k = P(X=k) = \binom{N+k-1}{k}(P+1)^{-N}\left(\frac{P}{P+1}\right)^k;\ k = 0, 1, 2, \dots ,$$

where $N > 0$ and $P > 0$ are real parameters.

The probability $p_0 = P(X=0) = (P+1)^{-N}$ and the expectation $\mathrm{E}X = NP$ are of particular interest as they can be used to characterise the share of uncited papers and the mean, respectively.

A non-negative integer-valued random variable X has a Waring distribution with real parameters N and α, if

$$p_k = P(X=k) = \frac{\alpha}{N+\alpha}\cdot\frac{N}{N+\alpha+1}\cdot\dots\cdot\frac{N+k-1}{N+\alpha+k};\ k = 0, 1, 2, \dots ,$$

where $N > 0$ and $\alpha > 0$.

Analogously to the previous case one has for the Waring distribution $p_0 = \alpha/(N+\alpha)$ and $\mathrm{E}X = N/(\alpha-1)$. Note that the expected value is finite only if $\alpha > 1$.

The following tail property is quite obvious.

$$P(X=k) \approx \begin{cases} c_1 e^{c \cdot k} & \text{for the negative binomial distribution} \\ c_2 k^{-(\alpha+1)} & \text{for the Waring distribution} \end{cases},\ \text{if } k \text{ is large.}$$

c_1 and c_2 are positive real values and $c = \ln[P/(P+1)]$. Furthermore, $\lim p_{k+1}/p_k = P/(P+1) < 1$, if k tends to infinity, in the first case, and $\lim p_{k+1}/p_k = 1$in the second case.

In the following case a hypothetical sample size of $n = 1000$ is assumed, and the parameters of the two distributions are chosen so that both f_0 and m roughly coincide. The choice of $N = 0.45$ and $P = 3$ for the negative binomial, and $N = 1.9$ and $\alpha = 2.4$ for the Waring distribution results in $f_0 = 0.54$ and $m = 1.35$ vs. $f_0 = 0.56$ and $m = 1.36$, respectively. Both pairs are similar and even the lower part of the distribution does not differ significantly ($n = 1000$ and $\chi^2 \sim 10$). Figure 1 shows the similar shapes of the two distributions for their "heads" and "trunks".

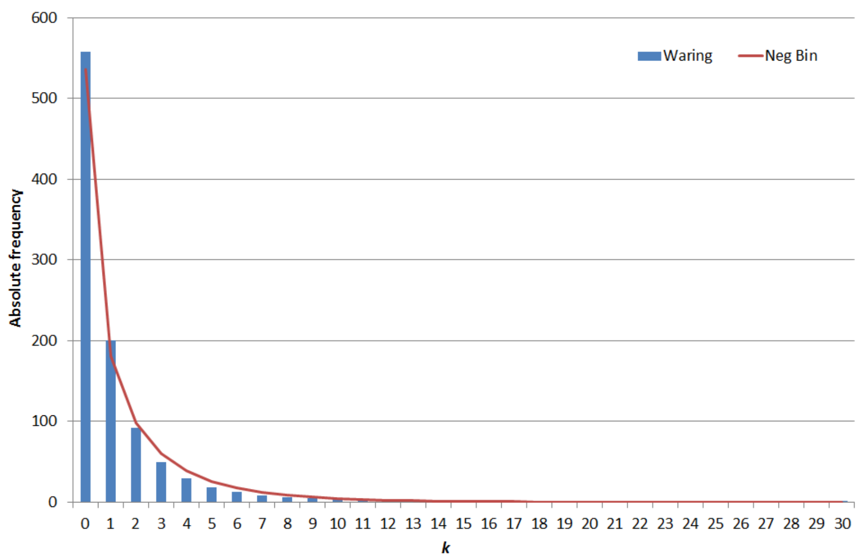

Figure 1. "Head" and "trunk" of the negative binomial and Waring distribution with similar mean and share of uncited papers

Nevertheless, the tail elements of the negative binomial distribution are of order $e^{-1.2k}$ while those of the Waring distribution are approximately $k^{-3.4}$. In order to illustrate the deviation of the tails from each other, Gumbel's characteristic extreme values have been calculated for the two distributions. For any given sample of size n, the Gumbel's characteristic extreme values (u_k) are defined as follows.

$$u_k = G^{-1}(k/n) = \sup \{x\colon G(x) > k/n\} \;;\; k = 1, 2, \ldots, n\,,$$

where $G := 1-F$ and F is the common empirical distribution function of the sample elements.

It can be shown that in the above example Gumbel's characteristic extreme value (u_1) considerably differs for the two distribution models. One obtains $u_1 = 18$ for the negative binomial and $u_1 = 42$ for the Waring distribution although the calculated frequencies hardly differ in the central section and at the lower end of the distributions.

A further real-world example is based on citation data collected for the topic "Osteoarthritis research". Data have been extracted from Thomson Reuters' *Web of Science*. The publication year was 2008 and the citation window comprises four years beginning with the publication year (i.e., 2008–2011). Since parameter estimation is based on the mean ($m = 9.437$) and the share of uncited papers ($f_0 = 11.4\%$), the shapes of both fitted distributions (negative binomial and Waring) are close to the shape of the empirical one (see Figure 2). The estimated parameters are $N = 0.88$ and $P = 10.7$ for the negative binomial model, and N=39.53 and $\alpha = 5.19$ for the Waring distribution.

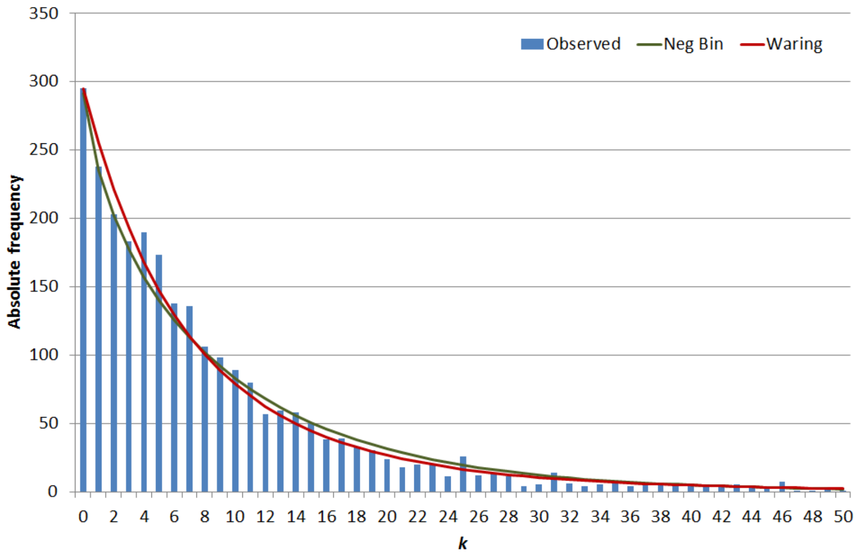

Figure 2. Fit of the negative binomial and Waring distribution to the "head" and "trunk" of the citation distribution in 'osteoarthritis research' (Publication year: 2008, citation window: 2008-2011) [Data sourced from Thomson Reuters Web of Knowledge]

For Gumbel's extreme values one obtains $u_1 = 86$ from the negative binomial model and $u_1 = 148$ from the Waring model. However, the most cited paper received 411 citations and the number of citations received by the paper ranking fourth (152 citations) roughly coincides with the "prediction". Glänzel and Schubert (1988a) have shown that the often extremely long tail can often not be explained by the underlying distribution model.

In total, only 1.3% of all papers in the set received more than 50 citations each, however, these papers jointly received 11.4% of all citations. It is straightforward to show that, in general, we have the following situation.

i) In the "exponential" case the share $p \in [0, 1]$ of the most cited papers receives a share p of all citations.

ii) In the "Paretian" case the share $p \in [0, 1]$ of the most cited papers receives a share $p^{\frac{\alpha-1}{\alpha}}$ of all citations, where α is assumed to be larger than 1.

In the above case, for instance, the exponential tail model has obviously to be rejected. In the Paretian model p=1.3% results in

a) 23.5% for α= 1.5,

b) 11.4% for $\alpha = 2$, and

c) 5.5% for $\alpha = 3$

This means, that with respect to the tail the parameter estimation for the Waring distribution on the basis of the head and trunk needs to be corrected, i.e., an α value in the neighbourhood of 2 should be assumed instead of previously estimated value of 5. However, both the exponential and the Paretian model works well for the head and trunk usually representing 95% or even more of the observations, but they are not in line with the long tail (cf. Glänzel and Schubert, 1988a). In the following sections two particular approaches to overcome this discrepancy will be discussed. The first one is based on a tail parameter, while the second one provides a parameter-free solution.

3. TAIL INDICES AS SUPPLEMENT FOR PERFORMANCE INDICATORS

3.1. The tail index of Paretian distributions

Because of the above mentioned "inconsistencies" of tail evaluation the question arises of in how far tail characteristics could supplement bibliometric indicators that are, otherwise, rather measuring the mainstream. In several areas such as insurance mathematics, where extreme values play an important role, the estimation of the tail parameter α of Pareto-type distributions has received much attention. Assume that $\{X_i\}_{i=1,\,..,\,n}$ is a sample of independent identically distributed random variables with Paretian distribution. Then the ranked sample X_i^* has the following property.

$$P(k \cdot \ln(X^*_k / X^*_{k+1}) < x) \sim 1 - e^{-\alpha \cdot x}\ ;\ k \leq k_0$$

Hence Hill's estimator (Hill, 1975) for the tail index $\gamma = \alpha^{-1}$ can be derived as the mean of the upper k elements of this series.

$$H_k = k^{-1} \sum_{i=1}^{k} \ln X_i^* - \ln X_{k+1}^*$$

It can be shown that H_k is asymptotically normally distributed (if $k << n$) with variance $1/(k\alpha^2)$. This property allows to construct confidence intervals for $\gamma = \alpha^{-1}$.

The estimation of the tail index is rather problematic since most methods, as for instance the Hill estimator too, are sensitive to the cut-off point for the tail. Mathematicians have therefore sought for alternative and more robust solutions ever since. In what follows, one simple solution will be presented that provides more robust results, but is still sensitive to the cut-off point.

3.2. QQ plots

Assume that $\{X^*_i\}_{i=1,..,n}$ is a ranked sample of the observations $\{X_i\}_{i=1,..,n}$. The exponential quantile–quantile plot (QQ plot) is then constructed as follows.

$$\left(-\ln\left(\frac{i}{n+1}\right), \ln X_i^* \right)\ ,\quad i = 1, 2, \dots n$$

Beirlant et al. (2004) have shown that in case of linearity the slope of a Pareto QQ plot approximates the Pareto tail index $\gamma = 1/\alpha$. The application of quantile plotting to scientometrics and using the Pareto tail index for the assessment of individual research performance has been proposed by Beirlant et al (2007). Since the choice of the cut-off point is always arbitrary the h-index is used for the following examples. At the macro and meso level this choice forms a good compromise but at lower levels of aggregation the h-index goes often far beyond the tail, notably in the case of citation impact of pre-eminent scientists.

Figures 3 and 4 present the QQ plots for two research topics. The tail index γ is given by the slope of the regression line. The first one is already known by the above discussion. The plot for *osteoarthritis research* shown in Figure 3 once again emphasises that the Pareto parameter α has to be assumed much less than estimated from the sample. The value of 1.6 lies even below that one suggested on the basis of the citation share received by the most cited 1.3% of publications. Also the second example is taken from the Web of Science database. The topic is *biofuels* within the subject category 'fuels and energy'. Publication year and citation window are the same as in the previous example. Here one obtains a Pareto parameter of almost exactly 2 and the fit is even better than in the previous example. Similarly to osteoarthritis research, mean and share of uncited papers, otherwise, provide good fits of the overwhelming part of the distribution for both the negative binomial and the Waring model.

In evaluative practice, the mean and the share of uncited papers could be supplemented by the tail index $\gamma = 1/\alpha$ to characterise the outstanding citation impact. Clearly, a higher γ value indicates more outstanding performance. This method is already applied by several bibliometric groups (e.g., Lietz, 2012). Of course, the question arises of what the weight of an indicator based on a minute share of publications could be, and in how far this kind of quantification could be precise enough to provide a reliable measure for the assessment of (outstanding) performance or even for ranking exercises. The almost insoluble problem of finding the optimum cut-off point for the tail of *non-Pareto* Paretian distributions and the observed incompatibility with other estimators for the same parameter (cf. Glänzel and Schubert, 1988a) makes the use of the tail index as a fine-tuned indicator for evaluative purposes questionable.

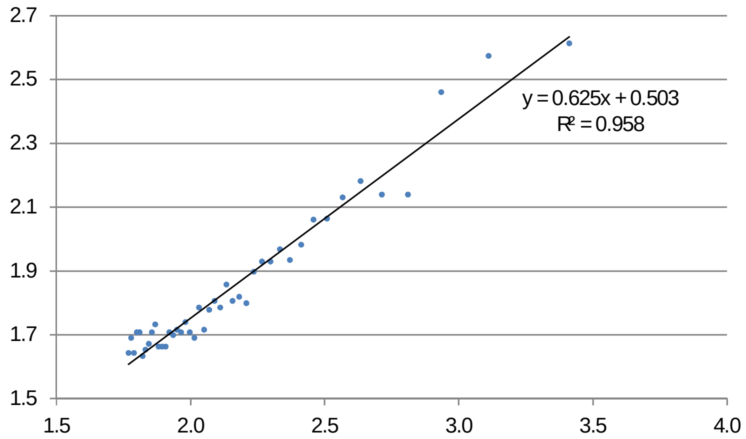

Figure 3. Pareto QQ plot based on the h-index (44) for the topic osteoarthritis research with α = 1.6 *(Publication year: 2008, citation window: 2008-2011) [Data sourced from Thomson Reuters Web of Knowledge]*

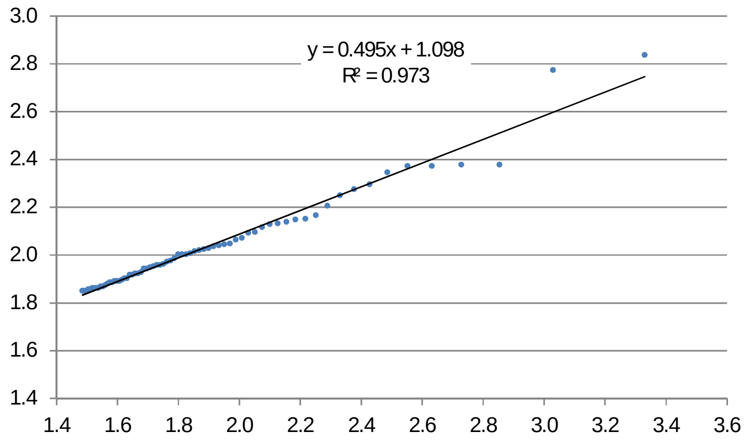

Figure 4. Pareto QQ plot based on the h-index (70) for the topic biofuel research with α = 2.0 *(Publication year: 2008, citation window: 2008-2011) [Data sourced from Thomson Reuters Web of Knowledge]*

A further issue arises from real outliers that might distort the estimates or at least affect their reliability. In such cases, only a smaller fraction of data can be used for inference in the tail. Recently statistics of extremes of *randomly censored data* (Einmahl et al., 2008) has become a new topic in probability theory to cope with suchlike issues. In the following, an alternative method, that is less sensitive to the above-mentioned contingencies than the tail index γ, will be introduced.

4. A PARAMETER-FREE SOLUTION USING CSS

A further alternative is a "reduction" of the distribution over individual items to a distribution over some essential classes representing specific sections of the original one. A solution using six classes has been suggested by Leydesdorff et al. (2011). According to their model, a pre-set set of six rank percentages is calculated on the basis of the reference distribution. Individual observations are then scored according to the percentage the publications in question belong to. Two particular problems arise from this approach, namely the arbitrariness of pre-set percentiles and the ties in both the reference distribution and the observations.

4.1. Characteristic Scores and Scales (CSS)

Another solution can be based on the method of *Characteristic Scores and Scales* (CSS) proposed by Glänzel and Schubert (1988b). Characteristic scores are obtained from iteratively truncating samples at their mean value and recalculating the mean of the truncated sample until the procedure is stopped or no new scores are obtained. The following mathematical description has been taken from a study by Glänzel (2007, p.93–94).

First put $b_0 = 0$ and $\nu_0 = n$, where n is the sample size, i.e., the number of publications. b_1 is then defined as the sample mean

$$b_1 = \sum_{i=1}^{n} \frac{X_i}{n} = \sum_{i=1}^{n} \frac{X_i^*}{\nu_o} \ .$$

The value ν_1 is defined by the following inequality

$$X^*_{\nu_1} \geq b_1 \text{ and } X^*_{\nu_1+1} < b_1 .$$

This procedure is repeated recurrently, particularly,

$$b_k = \sum_{i=1}^{\nu_{k-1}} \frac{X_i^*}{\nu_{k-1}} ,$$

and ν_k is chosen so that

$$X^*_{\nu_k} \geq b_k \text{ and } X^*_{\nu_k+1} < b_k , \; k \geq 2 .$$

The basic properties $b_0 \leq b_1 \leq \ldots$ and $\nu_0 \geq \nu_1 \geq \ldots$ are obvious from the definition. Obviously, the procedure comes to an end if $\nu_k = 1$ for some $k > 0$ is reached. The k-th class is defined by the pair of threshold values $[b_{k-1}, b_k)$ and the number of papers belonging to this class amounts to ν_{k-1}–ν_k.

The procedure is usually stopped at $k = 3$ since ν_k might otherwise become too small. The papers found in the resulting four classes are called *poorly* cited (less cited than average), *fairly* (above average but less citations than b_2), *remarkably* cited (received at least b_2 but less than b_3 citations) and *outstandingly* cited (more frequently cited than b_3).

Both the scores b_k and the scales (ν_{k-1}–ν_k) have interesting mathematical properties, which result in self-adjusting, parameter-free solutions for citation-impact assessment. The robustness of scales and classes has been analysed in the above-mentioned paper by Glänzel (2007).

According to the characterisation theorem for the Pareto distribution by Glänzel et al. (1984) the conditional expectation satisfies the condition

$$b_k = \mathbf{E}(X|X \geq b_{k-1}) \sim a \cdot b_{k-1} + b_1 ,$$

where $a = \alpha/(\alpha\text{-}1)$ and α is the free (tail) parameter of the Pareto distribution. This results by recursion in the following property

$$b_k / b_1 \approx \sum_{i=0}^{k-1} \left(\frac{\alpha}{\alpha - 1} \right)^i = \sum_{i=0}^{k-1} a^i = \frac{a^k - 1}{a - 1} .$$

These properties will serve as the groundwork for possible application to research evaluation, notably for the high-end of performance.

4.2. CSS in evaluation practice

Characteristic scores should – as all location indicators – not be used for comparison across subject areas since those depend strongly on the subject. The first score b_1 is, in fact, identical with the mean value of the empirical citation distribution. All other scores are conditional means depending also on b_1, and thus increase with growing k following a power law. Examples for this effect can be found in Glänzel (2007), again.

Within a narrow discipline, the comparison of corresponding scores is, of course, possible. This is shown using the example of papers published in the journals *Scientometrics* in the period 2002-2011. The long period was necessary to obtain sufficiently large publication sets for both the world total and the individual countries that have published in the journal. Citations have been counted from the publication year till 2012. I have selected four countries and compared their citation impact among each other as well as with the world standard.

It should be mentioned that this example primarily serves as an illustration of methodology because of some limitations due to the choice of the citation window. In particular, a paper published in 2010 or later has, of course, window less chance to reach the highest class than a paper published much earlier, but this has little effect if the distribution of publications of the countries under study over the complete period does not essentially differ. More specifically, the variable citation-window structure is to the detriment of dynamically growing countries like China since articles tend to be younger than those in the reference set.

The scores for the world total are used as the reference standard. The procedure was stopped at $k = 3$. The calculation of the equivalent scores for each country is not necessary; it is merely used an auxiliary tool for *internal benchmarking* and illustration of the possibility of intra-subject comparison. The corresponding b_i thresholds are shown in Table 1. For example, all b_i values of the US are higher than those of Spain thus representing a higher standard at all levels, and a US paper in scientometrics needs 49 citations to be considered outstandingly cited with respect to the own national standard, while in Spain 30 citations are sufficient to reach the same effect with respect to the Spanish benchmark. However, papers from both countries need to receive 55 citations each to qualify as outstandingly cited with respect to the world standard.

Table 1. CSS scores of *Scientometrics* papers for the world standard and four selected countries [Data sourced from Thomson Reuters Web of Knowledge]

Score	**World**	**USA**	**Spain**	**Belgium**	**China**
b_1	**9.2**	9.6	6.2	16.0	6.8
b_2	**26.5**	26.0	16.5	43.2	16.3
b_3	**54.5**	48.2	29.5	100.8	27.6

After these introductory considerations, the assessment of the citation impact according to performance classes will be explained. Preferably four classes should be used, where the b_i thresholds calculated from the world total are used again as reference standard. The share of a given unit's (e.g., country, region or institute) papers found in the four world classes of the reference population can be compared with the world standard as well as with other units. Note that the unit under study (and all other benchmark units as well) must be part of the reference population. The CSS scores of the world standard then serve as the benchmark (cf. Glänzel and Schubert, 1988b). If a unit's performance is a true "mirror" of the world standard, its distribution over classes is expected to coincide with that of the world.

In the present case, for instance, 13 out of 148 papers with Belgium author(s) have received more than 26 but less than 55 citations each (see Table 2). These 8.8% of all Belgian papers are considered remarkably cited. 9 papers have been cited more frequently than 54 times each. Thus 6.1% of Belgian papers in Scientometrics are outstandingly cited. The share of papers in the classes fairly, remarkably and outstandingly cited exceed their reference standards. Consequently, the remaining class of poorly cited papers contains less papers than expected on the basis of the world standard. The four countries' publication counts meeting the above criteria and the shares in the corresponding classes can be found in Table 2. The comparison among the individual countries can be interpreted analogously. The "reduced" distribution with four classes provides a quantified overview of citation impact with respect to the world standard while it keeps the peculiarities of the shape and skewness of the original citation distribution. For instance, the data for the US reveal a less polarised distribution than the reference standard.

Table 2. CSS-class shares of publications in Scientometrics for four selected counties with respect to the world standard [Data sourced from Thomson Reuters Web of Knowledge]

USA		Spain		Belgium		China		World	
Papers	Share	Papers	Share	Papers	Share	Papers	Share	Papers	Share
133	71.5%	126	81.8%	87	58.8%	109	80.1%	1047	**74.1%**
37	19.9%	21	13.6%	39	26.4%	20	14.7%	261	**18.5%**
13	7.0%	6	3.9%	13	8.8%	6	4.4%	77	**5.4%**
3	1.6%	1	0.6%	9	6.1%	1	0.7%	28	**2.0%**
186	100.0%	154	100.00%	148	100.00%	136	100.00%	1413	**100.00%**

The question arises of what indicator(s) could be built on the basis of these shares. The answer to that question is none. This is certainly a disappointment to those who wish to would like to express citation impact by one single number. However, one should resist any temptation to calculate averages, linear combinations or composite indicators *over classes*. This would result in crashing multi-dimensionality to linearity and in loosing essential information. Except for smoothening the effect of real outliers, these indicators would not provide more information than properly calculated elementary statistics.

Although shares should be compared in each discipline separately, combination *over subjects* is in principle possible. In order to avoid distortions caused by different citation behaviour in the different disciplines, classes should be determined for each individual subject, and appropriate shares should be combined on the basis of the unit's publication counts in the corresponding classes afterwards. One should, however, keep in mind that the results for large subject fields and for all fields combined calculated in this way might be affected by biases caused by deviating publication profiles of different units.

5. DISCUSSION AND CONCLUSIONS

The analysis of the high-end of scientific distributions is one of the most difficulty issues in evaluative scientometrics. And this is not merely a mathematical issue: it is fundamentally quite impossible to draw an exact borderline between "very good" and "outstanding". Furthermore, in citation analysis, outstanding performance can often not be explained by the "standard" citation behaviour (cf. Glänzel and Schubert, 1988) – and individual observations might be misleading and even distort statistics as has impressively been shown by Waltman et al. (2012). One extremely highly cited paper might even distort the ranking of universities if this is based on one single statistic. However, this is not typically a bibliometric issue. So-called censored data or data distorting extreme values of a distribution are known in several fields, e.g., in insurance mathematics (cf. Matthys et al., 2004), just to mention one example.

The analysis of the tail of publication and citation distributions, i.e., of that part, which is assumed to be linked with high performance, might help understand the mathematical rules of extreme communication behaviour and quantify outstanding performance. However, estimators of tail indices, as such, represent a very small share of the underlying population or sample and do not provide information about the characteristics of the overwhelming majority. Furthermore, these statistics are often not suited for combination with standard statistics. Percentiles or – even better – self-adjusting classes like those obtained from the CSS model allow the definition of proper performance classes and the needless integration of measures of outstanding and even extreme performance into the assessment of standard performance. Even extreme outliers like the case reported by Waltman et al. do not affect statistics because the influence of individual observation on the total publication is marginal and observation for the units under study are represented by classes instead of individual values. The only “drawback” of this method is that the calculation of a “single” indicator over classes should be avoided as this would reduce the gained added value and destroy all advantages of the method.

REFERENCES

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J. L. (2004). *Statistics of extremes: Theory and application.* Wiley.

Beirlant, J., Glänzel, W., Carbonez, A., Leemans, H., Scoring research output using statistical quantile plotting. *Journal of Informetrics*, 1 (3), 2007, 185–192.

Einmahl, J.H.J., Fils-Villetard, A. and Guillou (2007), A. Statistics of extremes under random censoring. *Bernoulli*, 14(1), 207–227.

Fisher, R.A., Tippett, L.H.C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24, 180–190.

Glänzel, W., Telcs, A., Schubert, A. (1984), Characterization by truncated moments and its application to Pearson-type distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 66, 173–183. (Correction: *Probability Theory and Related Fields*, 1987, 74, 317)

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

Glänzel, W., Schubert, A. (1988b), Characteristic Scores and Scales in assessing citation impact. *Journal of Information Science*, 14, 123–127.

Glänzel, W. (2007), Characteristic scores and scales. A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics*, 1 (1), 92–102.

Glänzel, W. (2009), The Multi-Dimensionality of Journal Impact. *Scientometrics*, 78 (2), 355–374.

Glänzel, W., Moed, H.F., (2012). *Thoughts and Facts on Bibliometric Indicators.* In: Eric Archambault, Yves Gingras, and Vincent Lariviere (Eds.), Proceedings of STI 2012 Montreal (17th International Conference on Science and Technology Indicators), Volume I, 305–318.

Gumbel, E.J. (1958), *Statistics of Extremes*. Columbia University Press, New York.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3, 1163–1174.

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.

Lietz, H. (2012), Private communication.

Matthys, G., Delafosse, E., Guillou, A., Beirlant, J. (2004), Estimating catastrophic quantile levels for heavy-tailed distributions. *Insurance Mathematics & Economics* 34 (3), 517–537.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E.C.M., Tijssen, R.J.W., van Eck, N.J., van Leeuwen, Th.N., van Raan, A.F.J., Visser, M.S., Wouters, P., (2012), *The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation*. In: Eric Archambault, Yves Gingras, and Vincent Lariviere (Eds.), Proceedings of STI 2012 Montreal (17th International Conference on Science and Technology Indicators), Volume II, 791–802.