

The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparison and institutional research assessment

Wolfgang Glänzel · Bart Thijs · Koenraad Debackere

Received: 27 September 2013 / Published online: 25 February 2014
© Akadémiai Kiadó, Budapest, Hungary 2014

Abstract The analysis of the high end of citation distributions represented by its tail provides important supplementary information on the citation profile of the unit under study. In a previous study by Glänzel (Scientometrics 97:13–23, 2013a), a parameter-free solution providing four performance classes has been proposed. Unlike in methods based on pre-set percentiles, this method is not sensitive to ties and ensures needless integration of measures of outstanding and even extreme performance into the standard tools of scientometric performance assessment. The applicability of the proposed method is demonstrated for both subject analysis and the combination of different subjects at the macro and meso level.

Keywords Characteristic scores and scales · Citation distributions · Performance classes · Crossnational · Comparison · Institutional assessment

Introduction

One of the objectives of a previous study (Glänzel 2013a) was to analyse to what extent the tail of scientometric distributions are in line with the ‘head’ and ‘trunk’ forming the major part of the distribution and to what extent ‘outliers’ might be responsible for possible deviations. Two important observations are relevant in this context. Unlike in many other fields, where outliers can simply be discarded as being exceptions, in bibliometrics extreme values represent the high end of research performance and therefore deserve special attention. One solution proposed in this study was therefore to use tail indices as a

W. Glänzel (✉) · B. Thijs · K. Debackere
Centre for R&D Monitoring (ECCOM) and Department of MSI, KU Leuven, Leuven, Belgium
e-mail: wolfgang.glanzel@kuleuven.be

W. Glänzel
Department of Science Policy & Scientometrics, Library of the Hungarian Academy of Sciences,
Budapest, Hungary

supplement to traditional citation-based performance indicators, such as the share of uncited papers to the mean citation rate. The analysis of the tail, which was based on ordered or ranked observations, can practically be uncoupled from the overwhelming rest of the empirical distribution. The second observation refers to empirical evidence concerning specific tail properties of citation distributions. Glänzel and Schubert (1988a) have shown that the often extremely long tail cannot be explained by the underlying distribution model. While extreme performance in publication activity was in line with the parameters estimated on the basis of the underlying distribution model, in the case of citation impact, the tail proved to be distinctly heavier than estimated on the basis of the head and trunk of the empirical distribution, which, in turn, usually represents 95 % (or even more) of all observations. This effect was observed even if a Paretian distribution model was assumed. This property was confirmed in the above-mentioned study by Glänzel (2013a). One solution proposed in the study was to use tail indices as a supplement to traditional citation-based performance indicators, such as the share of uncited papers and the mean citation rate. The analysis of the tail, which was based on ordered or ranked observations, can practically be uncoupled from the overwhelming rest of the empirical distribution. Most studies of the tail of scientometric distributions proceed from a Pareto model. The estimation of the tail parameter can directly be obtained from subsets of order statistics and are mostly based on the Rényi's representation (Rényi 1953). Versions of Hill's estimator (Hill 1975) and estimators based on so-called quantile–quantile plots (Kratz and Resnik 1996; Beirlant et al. 2004) are the most commonly used statistics. It has been shown that these estimators are consistent and asymptotically normally distributed. This property allows to construct confidence intervals for tail parameters. The practicability of quantile plotting in scientometrics and the use of the Pareto tail parameter for the assessment of individual research performance has been proposed, for instance, by Beirlant et al. (2007). Nevertheless, the estimation of the tail index remains rather problematic since most methods are still sensitive to the cut-off point for the tail. Since already minute changes of the tail parameter might have consequences in an evaluative context, the recommendation in the study by Glänzel (2013a) was to favour a parameter-free solution for the assessment of outstanding performance. This might also help avoid parameter conflicts resulting from estimating parameters on the basis of head and trunk of the distributions, on one hand, and from their tail, on the other hand.

Therefore, a “reduction” of the original citation distribution to performance classes on the basis of *Characteristic Scores and Scales* (CSS) introduced by Glänzel and Schubert (1988b) was proposed as an alternative parameter-free solution. Taking into account that citation standards considerably differ across the various disciplines, the method was developed for *individual subjects*. The classes obtained from this method can be applied to the comparative analysis of the citation-impact profiles of given units amongst themselves as well as with the reference standard in the given subject.

In two contributions to recent conferences (Glänzel 2013b and Glänzel et al. 2013) we showed that the application to combinations of different disciplines is indeed possible both at the national and institutional level. The present paper integrates the theoretical framework developed earlier with a demonstration of the robustness of the method for combinations of disciplines with respect to the publication year and the citation window at three different levels: a disciplinary analysis, a cross-national and a institutional comparison.

At the lower level of aggregation, particularly to the assessment of research institutions, the number of publications per unit is considerably lower than at the national level but more important is that we expect to observe more diverse research profiles. In particular,

some institutions have a specialised profile while others are truly multidisciplinary in their research activities.

We want to stress again that the calculation of a “single” indicator over these classes is not suitable as this would reduce the gained added value and thus destroy the advantages of the method. In this context, it is clear that it is not our aim to compare the application of the performance classes with other techniques that focus on single indicators like expected citation rates or top 1 or 10 % papers nor is it our intention to prove one or the other to be better.

A parameter-free solution using characteristic scores and scales (CSS)

An alternative to the tail analysis supplementing standard indicators is the “reduction” of the original citation distribution to a distribution over some essential performance classes including one or more particular classes corresponding to the high end of performance, i.e., to the tail of the original distribution. A solution using six classes has already 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.

Another solution has recently been suggested by Adams et al. (2007). The proposed classification proceeds from the mean citation rate on the basis of the world standard. The lowest class is formed by uncited papers. Other performance classes are then formed by setting thresholds at one quarter and one half of the standard for the lower performance classes and the double and quadruple of the standard for the higher classes, respectively. This procedure can be continued by extending the geometrics series based on positive and negative powers of 2. This method avoids the problem of ties but still uses preset threshold. In what follows, a self-adjusting method will be presented. The thresholds subdividing the population and samples into different performance classes are produced by the method itself and only depend on the underlying citation distribution. The sole arbitrarily chosen value is then the number of performance classes.

Characteristic scores and scales (CSS)

A self-adjusting 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 a distribution according to conditional mean values from the low end up to the high end. In particular, the scores b_k ($k > 0$) 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. Instead of the verbal description given here, an exact mathematical description can be found, e.g., in the study by Glänzel and Schubert (1988b).

First put $b_0 = 0$. b_1 is then defined as the mean of the original sample. The procedure is usually stopped at $k = 3$ since the number of papers remaining in the subsequent truncated sample might otherwise become too small. The k -th class is defined by the pair of threshold values $[b_{k-1}, b_k)$ with $k > 0$. The last and highest class is defined by the interval $[b_k, \infty)$, with usually $k = 3$. The number of papers belonging to any class is obtained from those papers, the citation rate of which falls into the corresponding half-open interval. This

definition solves the problem of ties since all papers can uniquely be assigned to one single class. In earlier studies the resulting four classes were called *poorly* cited (if less cited than average), *fairly* (if cited above average but received less citations than b_2), *remarkably* cited (if received at least b_2 but less than b_3 citations) and *outstandingly* cited (if at least as frequently cited as b_3). In the present study ‘Class k ’ ($k = 1, 2, 3, 4$) is used instead for the sake of convenience. The robustness of scales and classes has already been analysed and reported, for instance, by Glänzel in 2007. In addition, one important property should be pointed out here, particularly

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

provided the underlying distribution is of Pareto-type and α is its tail parameter. According to this property, the ratios of the k -th and the first score form a geometric series. As all location parameters, characteristic scores, too, are very sensitive to the subject field and the citation window. b_1 is, by definition, the mean value of the empirical citation distribution; all other scores are conditional means that depend on this initial value. This property is also reflected by the above approximate formula. Therefore, characteristic scores should not be used for comparison across subject areas.

Another property refers to the distribution of papers over the classes. The studies by Glänzel (2007, 2013 a, b) give empirical evidence that, in contrast to the b_k scores, this distribution over classes is strikingly stable with respect to the underlying subject field, the publication year as well as the citation window. This property makes the method useful for longitudinal and multi-disciplinary studies. Classes 1 and 2 represent “head” and “trunk” of the underlying citation distribution over individual papers. Usually, this refers to 90 % or a slightly larger share of all papers. The upper two classes, representing nearly 10 % of all papers, stand for the highly cited part of publications. Class 4, finally, covers the top 2–3 % of the corresponding population or sample and forms the most interesting category. It also contains possible outliers that have, however, no further effect on the outcomes as merely their assignment to the class but not their actual value counts. The following subsection will provide an introduction into the application of the method.

Application of characteristic scores and scales in comparative studies

After these introductory methodological remarks, the assessment of the citation impact according to performance classes will be explained in detail. This will be done in two steps. In the first step, the application to topics and disciplines is explained; thereafter the application to a combination of disciplines or even to all fields combined will be described. In the latter case a special procedure is necessary since simply forming four classes on the basis of the citation distribution in all fields combined would bias the results in favour of the life-sciences and to the detriment of mathematics and engineering sciences.

Disciplinary analysis

For the disciplinary analysis, first a brief summary of the procedure described in the already mentioned study (Glänzel 2012) is given. Again, preferably four classes should be used. First the b_k ($k = 1, 2, 3$) thresholds are calculated from the world total in the discipline or topic under study. These scores are used to define the reference standard, which is based on the four classes [b_{k-1}, b_k], $k = 1, 2, 3$ and [b_3, ∞).

Table 1 Characteristic scores of publications in 2007 and 2009 for 20 selected subfields according to the Leuven–Budapest scheme (Data sourced from Thomson Reuters Web of Knowledge)

Subfield	2007 (5-year citation window)			2009 (3-year citation window)		
	b_1	b_2	b_3	b_1	b_2	b_3
A2	6.43	13.80	21.97	2.68	6.01	10.68
B1	16.75	39.24	79.61	8.21	19.96	38.24
B2	23.05	58.33	116.72	11.34	28.96	56.28
C1	9.37	22.04	40.48	5.13	12.37	21.68
C3	11.22	24.68	42.04	5.84	12.24	20.83
C6	8.21	23.67	51.24	4.56	12.71	26.50
E1	5.04	14.75	29.83	2.37	6.64	12.60
E2	4.71	11.90	21.97	2.27	6.15	11.54
E3	6.57	17.82	34.00	4.19	11.19	21.10
G1	15.55	38.35	74.51	8.75	20.82	39.17
H1	5.21	14.36	29.83	2.41	6.66	12.88
I1	13.52	34.87	69.24	6.01	15.92	29.58
I5	16.24	41.52	84.74	7.96	19.26	39.49
M6	11.50	28.31	51.81	5.27	13.51	24.88
N1	15.28	35.38	64.73	7.18	16.92	29.77
P4	7.25	17.71	32.75	3.09	8.12	15.13
P6	7.27	20.05	43.89	4.30	12.15	26.54
R2	10.60	23.99	42.54	4.82	10.64	18.37
R4	11.42	26.19	48.62	5.49	12.65	22.50
Z3	12.80	29.48	54.96	6.36	15.25	28.88

A2 plant & soil science & technology, B1 biochemistry/biophysics/molecular biology, B2 cell biology, C1 analytical, inorganic & nuclear chemistry, C3 organic & medicinal chemistry, C6 materials science, E1 computer science/information technology, E2 electrical & electronic engineering, E3 energy & fuels, G1 astronomy & astrophysics, H1 applied mathematics, I1 cardiovascular & respiratory medicine, I5 immunology, M6 psychiatry & neurology, N1 neurosciences & psychopharmacology, P4 mathematical & theoretical physics, P6 physics of solids, R2 biomaterials & bioengineering, R4 pharmacology & toxicology, Z3 microbiology

For the demonstration, 20 out of the 60 subfields in the sciences according to the Leuven-Budapest classification scheme (see Glänzel and Schubert 2003) have been selected. Furthermore, two publication years have been chosen, 2007 with a five-year citation window (2007–2011) and 2009 with the three-year citation window 2009–2011. All journal publications indexed as article, letter, proceedings paper or review in the 2007 and 2009 volumes of Thomson Reuters’ Web of Science (WoS) have been selected and processed.

As expected, both subject and citation window have a strong effect on the actual values of the characteristic scores b_k . The lowest value has been found in A2 (plant & soil science & technology) in 2009 on the basis of a 3-year citations windows, while the highest one was observed in B2 (cell biology) in 2007 with a 5-year citation window. Increasing the citation window changed all b_k values. For the used combination of publication year and citation window, this resulted in roughly doubling the corresponding values with respect to the shorter window. The b_k values for the two WoS volumes are presented in Table 1.

Table 2 CSS-class shares of publications in 2007 and 2009 for 20 selected subfields according to the Leuven–Budapest scheme (Data sourced from Thomson Reuters Web of Knowledge)

Subfield	2007 (5-year citation window)				2009 (3-year citation window)			
	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
A2	65.2	22.6	8.1	4.2	63.3	26.0	7.1	3.6
B1	69.4	22.5	6.0	2.1	70.6	21.0	6.3	2.2
B2	72.0	20.2	5.6	2.2	71.6	20.1	5.8	2.4
C1	68.2	22.5	6.6	2.7	69.2	21.3	6.4	3.0
C3	67.4	22.2	7.5	3.0	63.6	24.9	7.7	3.9
C6	73.5	19.5	5.3	1.8	71.6	20.5	5.8	2.1
E1	73.7	18.8	5.5	2.0	71.4	19.9	6.2	2.4
E2	68.2	21.7	7.0	3.1	70.8	20.9	5.7	2.5
E3	70.7	20.2	6.3	2.9	70.9	20.6	6.1	2.4
G1	70.1	21.4	6.3	2.2	68.1	22.4	7.2	2.4
H1	72.3	20.3	5.4	1.9	71.0	20.4	6.2	2.4
I1	70.2	21.3	6.2	2.3	71.2	20.0	6.1	2.7
I5	71.9	20.4	5.4	2.2	68.7	22.8	6.1	2.3
M6	68.9	21.6	6.5	3.0	69.9	20.9	6.3	2.9
N1	69.1	21.7	6.4	2.8	69.1	21.1	6.8	3.0
P4	69.6	21.2	6.7	2.4	71.2	20.8	5.7	2.3
P6	72.4	20.7	5.3	1.7	72.8	20.4	5.2	1.6
R2	72.4	20.7	5.3	1.7	64.7	23.7	7.8	3.8
R4	68.4	22.5	6.4	2.7	67.3	22.5	7.1	3.0
Z3	68.2	22.3	6.8	2.6	69.3	22.1	6.2	2.5

A2 plant & soil science & technology, B1 biochemistry/biophysics/molecular biology, B2 cell biology, C1 analytical, inorganic & nuclear chemistry C3 organic & medicinal chemistry, C6 materials science, E1 computer science/information technology, E2 electrical & electronic engineering, E3 energy & fuels, G1 astronomy & astrophysics, H1 applied mathematics, I1 cardiovascular & respiratory medicine, I5 immunology, M6 psychiatry & neurology, N1 neurosciences & psychopharmacology, P4 mathematical & theoretical physics, P6 physics of solids, R2 biomaterials & bioengineering, R4 pharmacology & toxicology, Z3 microbiology

By contrast, the citation classes defined by the characteristic scores are by and large insensitive to both the length of the citation window and the underlying subject. Table 2 gives the corresponding values for the same subfields as above. The share of papers cited less frequently than the average (Class 1) amounts to roughly 70 %, the share of those categorised to Class 2 to about 21 % and the in the highest two classes one finds 6–7 % and 2–3 % of all publications, respectively. This coincides with the observations made by Glänzel (2007) on the basis of the 1980 volume of the *Science Citation Index* (SCI) and a 21-year citation window.

The comparison of national citation impact with the world standard can readily be done by using the above classes $[b_{k-1}, b_k]$, $k = 1, 2, 3$ and $[b_3, \infty)$ as the respective subject standard. The comparison of the distribution over classes provides a more detailed picture, notable on the high end of the performance range, than the comparison of the means and the shares of uncited papers alone. The calculation of the corresponding scores for each individual country is not necessary. The share of a given country's (or any other unit's)

Table 3 National shares of publications in the reference CSS classes in 2007 and 2009 for subfield I1 according to the Leuven–Budapest scheme (in alphabetic order) (Data sourced from Thomson Reuters Web of Knowledge)

Country	2007 (5-year citation window)				2009 (3-year citation window)			
	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
BEL	61.5	24.1	9.8	4.6	61.2	24.0	9.3	5.5
BRA	73.5	19.8	4.7	2.0	87.0	8.8	3.1	1.2
CAN	61.8	25.3	9.2	3.7	59.4	26.6	8.7	5.3
CHE	60.8	25.2	10.7	3.3	61.7	23.0	9.5	5.8
CHN	68.7	24.4	5.6	1.3	72.8	21.0	4.8	1.4
DEU	62.5	24.5	8.9	4.1	63.0	23.7	8.6	4.7
ESP	73.8	17.8	5.2	3.2	72.9	17.0	6.7	3.4
FRA	71.3	17.8	7.5	3.4	66.4	20.9	7.9	4.8
GBR	61.0	26.2	8.5	4.3	62.1	24.0	8.9	5.0
GRC	74.8	19.4	4.2	1.6	75.6	17.8	4.6	2.0
ITA	70.8	20.0	6.3	3.0	66.9	21.7	7.3	4.0
JPN	73.2	19.9	5.3	1.5	71.6	21.3	5.2	1.8
KOR	74.2	18.2	5.2	2.3	65.4	25.1	7.6	1.9
NLD	56.4	28.9	9.9	4.8	57.7	28.0	9.9	4.4
POL	71.4	20.6	4.2	3.8	82.4	10.3	3.8	3.5
SWE	59.1	27.7	10.0	3.2	60.2	24.3	9.9	5.6
TUR	92.7	6.3	0.9	0.0	93.8	4.7	1.1	0.4
TWN	78.6	17.4	2.6	1.4	76.4	16.8	5.0	1.7
USA	61.0	26.4	9.0	3.6	61.8	25.0	8.9	4.3
Total	70.2	21.3	6.2	2.3	71.2	20.0	6.1	2.7

publications found in the four performance classes of the reference population can be compared with the world standard as shown in Table 2 or with those of other countries (or other units). Note that the unit under study (and all other benchmark units as well) must be part of the reference population. If a unit under study were the true mirror of the entire population, its share in all four classes would be identical with the reference standard. Any deviation from this standard indicates a specific profile. The unit’s profile might be more or less *skewed* with higher or lower shares in the lower classes, respectively, and more or less *polarised* according as the lower/higher share of lower-class papers is compensated by a higher/lower share of upper-class papers. Such cases have been reported by Glänzel (2012) for the *Scientometrics* sample, where China had a more skewed profile than the reference standard, Belgium had a less skewed profile and the profile of the USA was somewhat less polarised than the reference standard.

In the following sections, the method will be explained on the basis of a discipline in the life sciences. In particular, the subfield ‘cardiovascular & respiratory medicine’ (I1) has been chosen. The country Belgium is used as the example unit and the publication year is 2007. 55 out of 561 papers with at least one Belgium (co-)author have received at least 35 but less than 70 citations each (cf. Table 1). These 9.8 % of all Belgian papers are considered remarkably cited (Class 3). 26 papers have been cited at least 70 times each. Thus 4.6 % of Belgian papers in the subfield cardiovascular & respiratory medicine are

outstandingly cited (Class 4). The share of papers (38.5 %) in the three Classes 2, 3 and 4 exceeds the reference standard of 29.8 %. Consequently, the remaining class of poorly cited papers (Class 1) contains fewer papers than expected on the basis of the world standard.

The indicators for the world's 20 most active countries in this subfield are presented in Table 3. 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.

The distributions over the four “performance” classes provide more detailed insight than traditional citation indicators. Clearly, Italy's distribution in this subfield reflects a more favourable situation than that of Japan in both years and Turkey has the least favourable one in the country set. The question arises of what indicators could possibly be built on the basis of these shares. Glänzel (2012) has argued that no combination or composite indicator *over classes* should be built. Except for smoothening the effect of outliers, such indicators would not provide more information than properly calculated elementary statistics. It has been stressed that, on the other hand, a combination *over subjects* is, in principle, possible, provided of course that document assignment to performance classes can be “disambiguated” in case of multiple subject assignment. In any case, classes should be determined for each individual subject first, and appropriate shares should be combined on the basis of the unit's publication counts in the corresponding classes afterwards. Also the choice of the level of aggregation of the underlying subject is crucial. If subject areas are too broad, the high end of the citation distribution is formed by papers in subjects that have, in general, a high citation standard; but theoretical or technology-oriented topics would scarcely appear in the upper classes. If, on the other hand, subjects are too narrow then the number of papers is not sufficient to form stable classes, or, in other words, the upper classes remain (nearly) empty for most units. The above 60 subfields seem to form a stable groundwork for both national and institutional assessment. In the next subsection the combination of subjects will be discussed.

CSS in all fields combined

One precondition for the application of CSS to broad science fields or to all fields combined is the unique assignment of publications to performance classes. The following example describes this problem. Assume, for instance, that a paper is assigned to two subjects, here denoted by S1 and S2. According to possibly different citation standards in the two subjects, the paper is then assigned, for instance, to Class 3 in subject S1 and to Class 4 in S2 because its citation rate does not exceed b_3 in S1 but it is greater than the corresponding threshold b_3 in S2. A direct combination can, therefore, not provide any acceptable solution. A proper subject-based fractionation must be applied such that each publication is gauged against only one individual threshold value. As argued in the study by Glänzel et al. (2009) one important consequence of multiple assignments is the necessity of fractionation by subjects and thus of calculating proper weights for the corresponding individual subject-expected citation rates. Furthermore, it was stressed that the weighting of fractional data is correct only if the sum of the individual field expectations over all publications in the system equals the citation total of the database in the combination of these fields. This will result in an ‘implicit’ classification without calculating any *common* thresholds b_k . Again, the procedure is based on an iteration, where the first step is identical with the procedure of calculating subfield-expected citation rates. A first fractionation is applied when the citation means of subfields is determined. This

is done on the basis of the respective number of subfields to which a publication is assigned. Both publications and citations are fractionated. Using the above-mentioned example, the paper contributes only as a half to the total number of publications and citations of respectively subjects S1 and S2. The second fractionation follows when individual expectations are calculated for each paper. This expectation is then the mean value of the fractionated subfield standards. As such the individual threshold for the paper in our example is a combination of the half of both means of fields S1 and S2.

In the following step of the iteration, all papers, that have received fewer citations than their individual expectation, are removed. The above procedure is repeated on the remaining set. This is done three times in total to obtain the *individual* characteristic scores b_k^* ($k = 1, 2, 3$) for each publications. All papers can now uniquely be assigned to one of the four classes. It should be mentioned in passing that, if the underlying paper set comprises only publications from one single subfield and fractionation is not required, the results will be identical with those described in the previous subsection. It is straightforward that, in this case, the individual thresholds are identical with the common characteristic scores.

One important validity aspect of this method is the appropriate subject distribution in all performance classes, notably in the highest one since this reflects outstanding performance. Thus the question arises of whether all subject fields are proportionally represented in what is considered the high end of the citation distribution. Table 4 gives the distribution of papers over major fields according to the Leuven-Budapest scheme and the field distribution of papers assigned to Class 4 in 2007 and 2009. The same citation windows as above have been used here as well. Some deviation from the complete WoS representation can be observed in both years but this deviation should not be considered a serious bias. The patterns in Table 4 are strikingly stable over time although different citation windows have been applied. All subjects can, therefore, be considered adequately represented among highly cited publications.

The distribution of papers over classes reflects the same stability as already found in the disciplinary analysis in the previous subsection (cf. Table 2). The CSS procedure in all fields combined resulted in the following distribution for the two selected WoS volumes.

- 2007 (5-year citations): 69.8 % (Class 1), 21.5 % (Class 2), 6.3 % (Class 3), 2.4 % (Class 4).
- 2009 (3-year citations): 69.7 % (Class 1), 21.4 % (Class 2), 6.4 % (Class 3), 2.5 % (Class 4).

CSS-based classes for cross-national comparison

Figure 1 gives a graphic presentation of the world standard and the national shares in the upper three classes in 2007 for the 30 most active countries in 2007 and 2009. Among these countries, Belgium, Denmark, The Netherlands and Switzerland have the highest shares in the upper three CSS classes with more than 40 % each. Norway, Sweden, UK and USA, with slightly lower values, have a similar profile. This, of course, corresponds to the lowest share of “poorly” cited papers (Class 1) since, by definition, the content of the four classes adds up to 100 %.

Besides, a similar share of Class 1 papers does not imply the same distribution over the upper classes. France and Poland in ‘cardiovascular & respiratory medicine’ (II) in 2007 might just serve as an example (see Table 3). This demonstrates again that a particular share in one class can not be used for the prediction of the distribution over the other classes.

Table 4 Distribution of publications over major fields in 2007 and 2009 according to the Leuven–Budapest scheme (Data sourced from Thomson Reuters Web of Knowledge)

Field	2007 (5-year citations)		2009 (3-year citations)	
	WoS (%)	Class 4 (%)	WoS (%)	Class 4 (%)
A	7.0	8.2	7.5	8.5
B	10.1	10.1	9.3	9.3
C	20.2	19.8	20.0	21.7
E	11.2	8.5	11.8	9.1
G	5.7	6.9	5.8	6.7
H	4.5	4.1	5.0	4.1
I	12.2	11.0	12.0	10.5
M	18.4	18.3	18.7	18.3
N	5.7	6.8	5.6	6.7
P	15.0	13.6	14.3	13.2
R	7.2	6.4	7.2	6.8
Z	10.3	9.6	10.0	9.8

A agriculture & environment, B biosciences (General, cellular & subcellular biology; genetics), C chemistry, E engineering, G geosciences & space sciences, H Mathematics, I clinical and experimental medicine I (General & internal medicine), M clinical and experimental medicine II (Non-internal medicine specialties), N neuroscience & behaviour, P physics, R biomedical research, Z biology (organismic & supraorganismic level)

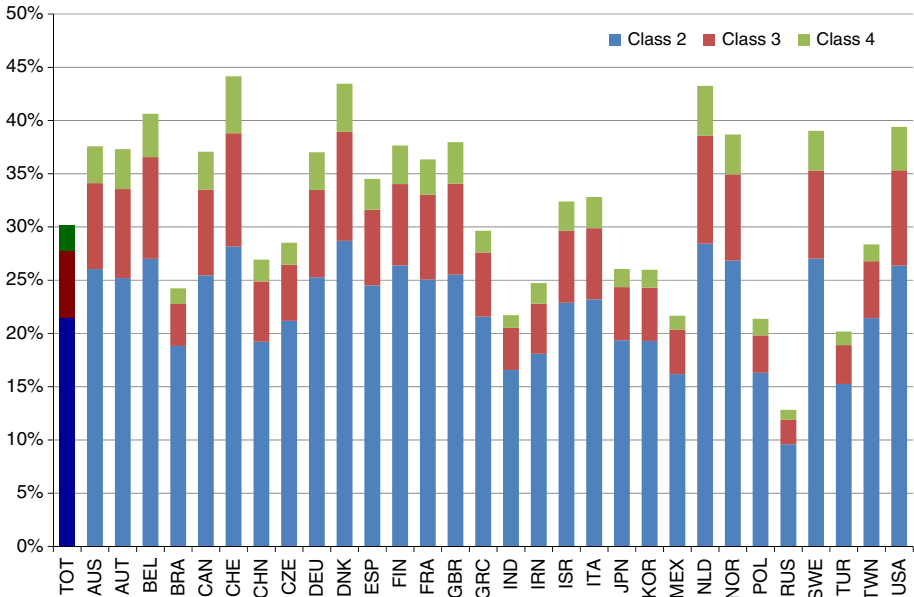


Fig. 1 The world standard (left-most column) and national shares of publications (in alphabetic order) in the upper three CSS classes in all fields combined in 2007 (5-year citation window) [Data sourced from Thomson Reuters Web of Knowledge]

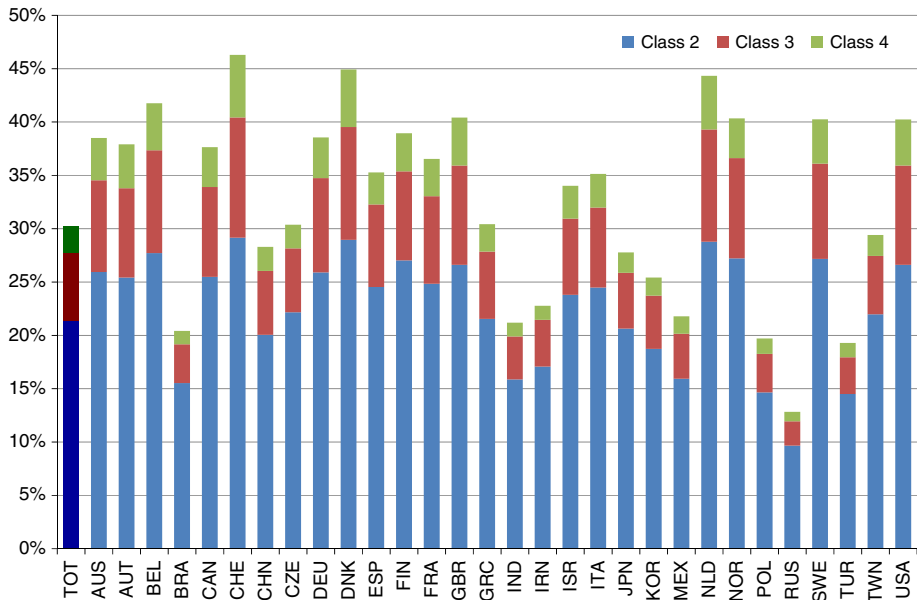


Fig. 2 The world standard (left-most column) and national shares of publications (in alphabetic order) in the upper three CSS classes in all fields combined in 2009 (3-year citation window) [Data sourced from Thomson Reuters Web of Knowledge]

Even very similar shares of Class 2 papers might go with different distributions over the two other upper classes as the comparison of the country pairs Belgium-Sweden, Finland-USA and Brazil-China in all fields combined (2007) convincingly illustrates (cf. Fig. 1).

The same presentation for the WoS volume 2009 on the basis of a three-year citation window can be found in Fig. 2. The reference standard is practically unchanged with respect to the 2007 volume with the five-year citation window. Nevertheless, a certain polarisation can be observed. UK, Italy and Switzerland (with growing shares in the upper three CSS classes), and Poland, Iran and Brazil (with decreasing shares in these classes) are the most concerned countries in this selection.

Belgium, Denmark, The Netherlands and Switzerland are the four countries with the highest standard and the lowest share of Class 1 papers in 2009 as well, and are again closely followed by the Norway, Sweden, UK and the US. The profile of Russia reflects the least favourable situation, but is along with that of Mexico and France the most stable one in the selection.

The possibility of the identification of individual highly cited papers (Class 4 publications) forms a further added value of this method.

Finally it should be mentioned, that in contrast to the “subject disambiguation” in the calculation of citation thresholds, assignment to addresses is not unique. Note that, for instance, a paper in Class 4 is counted as highly cited for both Germany and France, whenever it has co-authors from the two countries.

CSS-based classes for institutional comparison

For the institutional comparison of class profiles we have selected two universities each from 11 European countries (see Fig. 3). Although the universities’ profiles mostly mirror

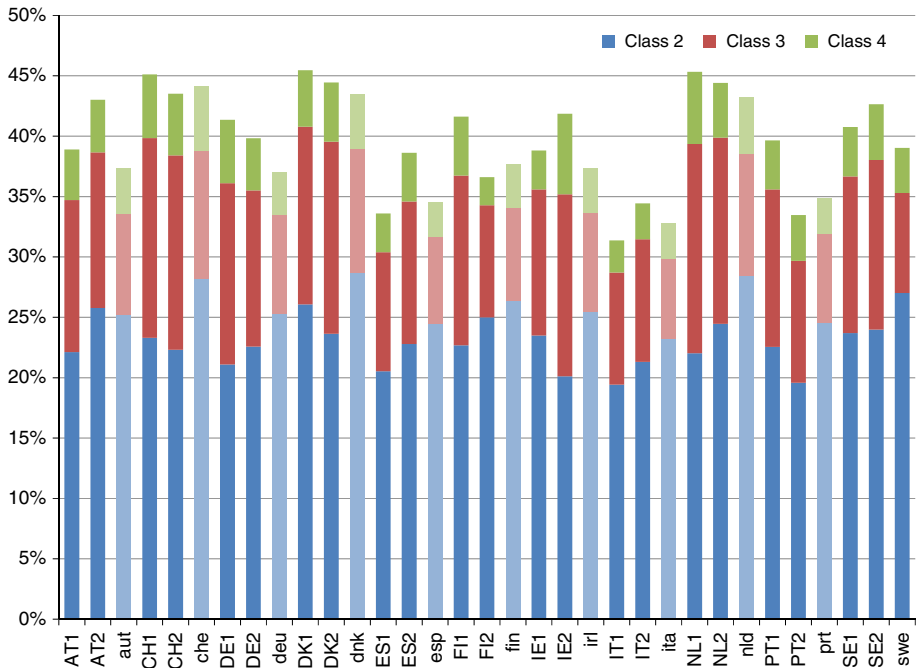


Fig. 3 Shares of publications of selected universities and countries in the *upper* three CSS classes in all fields combined in 2007 (5-year citation window) [Data sourced from Thomson Reuters Web of Knowledge]

the national patterns, we find in several cases a distinctly more favourable situation than in the national standards. This is contrasted by a less favourable situation for the 2 South-European universities IT1 and PT2 as well as to a lesser extent for ES1, FI2 and the second Swiss university in the selection (CH2). The high standards of the selected Danish and Dutch universities are worth mentioning. Finally, DK1 and PT1 are technical universities while SE1 stands for a medical university. This again substantiates the subject-independence of the method (cf. Glänzel 2013b).

Discussion and conclusions

The analysis of the high end of scientific distributions is one of the most difficult and challenging issues in evaluative scientometrics. This is, of course, not merely a mathematical issue as it is always difficult to draw a sharp borderline between “very good” and “outstanding”. Also the effect of outliers, i.e., of observations that might bias or even distort statistics, impressively shown by Waltman et al. (2012), is not typically a bibliometric issue. So-called censored data or data distorting extreme values of a distribution are known in several fields, for instance, in insurance mathematics (e.g., Matthys et al. 2004). In the proposed CSS-based method the effect of outliers is limited as the influence of individual observation on the total is marginal and observation for the units under study are represented by classes instead of individual values.

Self-adjusting classes, such as those based on CSS, allow the definition of proper performance classes without any pre-set thresholds. This is certainly one of the main

advantages of the proposed method. Another one is the needless integration of measures of outstanding performance into the assessment tools of standard performance. The method of “implicit” subject fractionation can also be used in the context of other publication and citation indicators, whenever the issue of multiple subject assignment needs to be resolved.

The studies have shown that a publication output at the meso-level suffices to provide a solid basis of interpretation and further statistical analysis. A further important property has become apparent, namely the method’s independence of the unit’s research profile. In the small meso-level sample we have found two technical universities with more favourable citation profiles than that of medical universities or than their corresponding national reference standards.

Those mathematical and empirical insights lead to some final considerations that do warrant further research and reflection. More precisely, the possibility to identify and analyse the high end of publication distributions immediately leads to the questions on its potential policy uses. It is obvious that the potential policy applications of the method developed and described in this paper require further reflection and research. No absolute value judgements are or should be expressed as to the form of a specific national or institutional distribution. The longitudinal comparison of distributions across countries or institutions does of course allow for further in-depth analyses of the underlying nature, the development stage and the growth of a country or institution’s science base. The focus on the high end of the distribution, as done in this paper, allows identifying the fraction of a country or institution’s science base that has reached extreme visibility in the global scientific arena. But once again, this does not imply a direct value judgement as to what is a “good” or a “better” profile. On the contrary, every distribution profile computed opens both the perspective and the need for a more in-depth, qualitative understanding and analysis of the dynamics of a national or institutional science base. The CSS-based method therefore has the potential to further inform and enrich the policy use of scientometric data, without aiming at or claiming absolute value judgements on the mere basis of the profiles obtained.

Acknowledgments The present study is an extended version of a article presented at the 14th International Conference on Scientometrics and Informetrics, Vienna (Austria), 15–19 July 2013 (Glänzel 2013). It also includes results presented at the 18th International Conference on Science and Technology Indicators, held in Berlin, 4–6 September 2013 (Glänzel et al. 2013).

References

- Adams, J., Gurney, K., & Marshall, S. (2007). Profiling citation impact: A new methodology. *Scientometrics*, 72(2), 325–344.
- Beirlant, J., Glänzel, W., Carbonez, A., & Leemans, H. (2007). Scoring research output using statistical quantile plotting. *Journal of Informetrics*, 1(3), 185–192.
- Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. L. (2004). *Statistics of extremes: Theory and application*. Hoboken: Wiley.
- 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. (2013a). High-end performance or outlier? Evaluating the tail of scientometric distribution. *Scientometrics*, 97(1), 13–23.
- Glänzel, W. (2013b). The application of citation-based performance classes in disciplinary and multidisciplinary assessment. In: J. Gorraiz, E. Schiebel, Ch. Gumpenberger, M. Hörlesberger, H.F. Moed (Eds.), *Proceedings of ISSI 2013—The 14th International Conference on Scientometrics and Informetrics* (pp. 109–122). Vienna, Austria.

- Glänzel, W., & Schubert, A. (1988a). Theoretical and empirical studies of the tail of scientometric distributions. In L. Egghe & R. Rousseau (Eds.), *Informetrics* (pp. 75–83). London: Elsevier Science Publisher.
- Glänzel, W., & Schubert, A. (1988b). Characteristic scores and scales in assessing citation impact. *Journal of Information Science*, *14*(2), 123–127.
- Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for bibliometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.
- Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2009). Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics*, *78*(1), 165–188.
- Glänzel, W., Thijs, B., Debackere, K. (2013). The application of citation-based performance classes in institutional research assessment. In: S. Hinze, A. Lottmann (Eds.), “Translational twists and turns: Science as a socio-economic endeavour”, Proceedings of STI 2013 (18th International Conference on Science and Technology Indicators), pp. (122–129). Berlin.
- Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, *3*, 1163–1174.
- Kratz, M., & Resnick, S. (1996). The qq-estimator of the index of regular variation. *Communications in Statistics: Stochastic Models*, *12*, 699–724.
- 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*(7), 1370–1381.
- Mathys, G., Delafosse, E., Guillou, A., & Beirlant, J. (2004). Estimating catastrophic quantile levels for heavy-tailed distributions. *Insurance Mathematics & Economics*, *34*(3), 517–537.
- Rényi, A. (1953). On the theory of order statistics. *Acta Mathematica Academiae Scientiarum Hungaricae*, *4*(3–4), 191–231.
- 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, & Vincent Larivière (Eds.), Proceedings of STI 2012 (17th International Conference on Science and Technology Indicators), (pp. 791–802). Montreal.