Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 92–102
Journal of INFORMETRICS An International Journal
www.elsevier.com/locate/joi

Characteristic scores and scales
A bibliometric analysis of subject characteristics based on long-term citation observation☆

Wolfgang Glänzel [a,b,*]

[a] *Steunpunt O&O Statistieken, K.U. Leuven, Dekenstraat 2, B-3000 Leuven, Belgium*
[b] *IRPS, Hungarian Academy of Sciences, Budapest, Hungary*

Received 3 July 2006; received in revised form 5 October 2006; accepted 19 October 2006

Abstract

In an earlier paper by Glänzel and Schubert [Glänzel, W., & Schubert, A. (1988a). Characteristic scores and scales in assessing citation impact. *Journal of Information Science*, *14*(2), 123–127; Glänzel, W., & Schubert, A. (1988b). Theoretical and empirical studies of the tail of scientometric distributions. In L. Egghe, & R. Rousseau (Eds.), *Informetrics*: *Vols. 87/88*, (pp. 75–83). Elsevier Science Publisher B.V.], a method for classifying ranked observations into self-adjusting categories was developed. This parameter-free method, which was called method of characteristic scores and scales, is independent of any particular bibliometric law. The objective of the present study is twofold. In the theoretical part, the analysis of its properties for the general form of the Pareto distribution will be extended and deepened; in the empirical part the citation history of individual scientific disciplines will be studied. The chosen citation window of 21 years makes it possible to analyse dynamic aspects of the method, and proves sufficiently large to also obtain stable patterns for each of the disciplines. The theoretical findings are supplemented by regularities derived from the long-term observations.

Keywords: Citation analysis; Long-term citation impact; Disciplinary citation impact; Pareto distribution; Extreme values; Truncated moments

1. Introduction

The classical studies of long-term citation impact by Abt (1981) and Garfield (1996, 1998a, 1998b) have shown that long observation periods are indispensable for obtaining reliable and stable results on citation processes. From the methodological viewpoint, such analyses are important in studying disciplinary citation impact, ageing-related issues, first-citation distributions (see, for instance, Egghe & Rao, 2001; Glänzel & Schoepflin, 1995; Rousseau, 1994) and citation-succession processes (see Glänzel, 1992; Glänzel & Schoepflin, 1995), the phenomenon of delayed recognition (e.g., Garfield, 1980; Glänzel & Garfield, 2004; Glänzel, Schlemmer, & Thijs, 2003; van Raan, 2004) and in identifying highly cited papers. Informetric long-term studies reveal regularities and provide tools that are nevertheless important for scientometric applications, too. The predictive power of models for citation processes (Glänzel & Schubert, 1995),

☆ Present study is partially based on a talk delivered at the "10th Nordic Workshop on Bibliometrics, Informetrics and Research Policy" (Glänzel, 2005).

* Present address: Steunpunt O&O Statistieken, K.U. Leuven, Dekenstraat 2, B-3000 Leuven, Belgium.
E-mail address: Wolfgang.Glanzel@econ.kuleuven.be.

1751-1577/$ – see front matter
doi:10.1016/j.joi.2006.10.001

the methodological foundation of finding optimum citation windows for evaluative purposes and the determination of subject-specific scores are only some of those applications.

In the present paper, we will use a method for classifying ranked observations into self-adjusting categories developed by Glänzel and Schubert (1988a, 1988b). The method, which is based on the analysis of rank-specific subcategories of ranked observations, was called characteristic scaling. Within the framework of this model an optionally refinable scale with variable scores is used for the characterising eminence of citation impact. It can be shown that if the grouping derived by our procedure is complete for the ranked observations it has several interesting properties, the most important of which is the grouping's independence of the common scale. Such scale- and parameter-independent classification methods are very important in scientometric practice since citation distributions are highly sensitive to subject particularities (see, for instance, Glänzel & Moed, 2002). Moreover, instead of the usual comparison of statistical functions like means, medians or percentages of a sub-sample with the corresponding values of the complete sample or of the population, this method directly allows the benchmarking of the sub-sample through the comparison of its class properties with that of the population. This, of course, provides a more complex approach than the usual comparison of "averages".

In the present paper we will first extend and generalise the investigation of the theoretical properties of the method under the condition that the underlying citation distribution is a *Pareto distribution of the second kind* at any time beginning with the publication year. In order to obtain stable results and to be able to analyse the dynamics of the *characteristic scores and scales* the study is based on a 21-year citation window. We will also demonstrate that the parameter α (and its transformation $q = (1 - 1/\alpha)^{\alpha}$) derived from our model describes important subject-specific characteristics of citation distributions. Although the method of *characteristic scores and scales* itself represents a parameter-free approach to citation-impact classification, it provides the characteristic parameter of the underlying distribution. Furthermore, the analysis of the citation process gives empirical evidence that this parameter is time-invariant.

2. The theory of 'characteristic scores and scales'

In what follows, we briefly summarise the definition of the *characteristic scores and scales* according to Glänzel & Schubert (1988a, 1988b). Consider a set of n papers published in a given subject field. The observed citations $\{X_i\}_{i=1}^{n}$ received by each paper in a given time period are then ranked in descending order $X_1^* \geq X_2^* \geq \cdots \geq X_n^*$, where X_1^* denotes the citation rate of most frequently cited paper in the set and X_n^* consequently the number of citations the least cited paper has received.

2.1. Definition

In order to develop a method for subdividing the sample into classes we define appropriate thresholds based on the following recursion. First put $\beta_0 = 0$ and $\nu_0 = n$. β_1 is then defined as the sample mean:

$$\beta_1 = \sum_{i=1}^{n} \frac{X_i}{n} = \sum_{i=1}^{n} \frac{X_i^*}{\nu_o} \tag{1}$$

The value ν_1 is defined by the following inequality:

$$X_{\nu_1}^* \geq \beta_1 \quad \text{and} \quad X_{\nu_1+1}^* < \beta_1 \tag{2}$$

This procedure is repeated recurrently, particularly:

$$\beta_k = \sum_{i=1}^{\nu_{k-1}} \frac{X_i^*}{\nu_{k-1}} \tag{3}$$

and ν_k is chosen so that:

$$X_{\nu_k}^* \geq \beta_k \quad \text{and} \quad X_{\nu_k+1}^* < \beta_k, \quad k \geq 2 \tag{4}$$

The properties $\beta_0 \leq \beta_1 \leq \ldots$ 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 kth class is defined by the pair of threshold values $[\beta_{k-1}, \beta_k)$ and the number of papers belonging to this class amounts to $\nu_{k-1} - \nu_k$. From the definition defining procedure it follows that each given publication set determines its own groups.

2.2. Properties

Though no underlying rule is necessary for arranging the sample, important properties of the threshold values as well as of the size of the classes determined through these thresholds can be derived for special citation distributions. Since citation distributions are integer valued, the unique determination of the theoretical values of the characteristic thresholds introduced above would always result in an integer approximation. Therefore, we have used an approximation based on a continuous distribution model in an earlier study (Glänzel, Telcs, & Schubert, 1984). In particular, we have found an important basic property in the case of Paretian distributions. The citation rate X of a paper has a Paretian distribution if

$$G(x):=1 - F(x) = P(X \geq x) \approx c(N + x)^{-\alpha} \tag{5}$$

for large $x>0$ and some positive value c, where N and α are positive real parameters and F denotes the distribution function of the random variable X. For $\alpha>1$ the citation rate X has a finite expectation so that one can define the following conditional expectation:

$$b_k = \begin{cases} 0 & k = 0 \\ \dfrac{\sum_{i \geq b_{k-1}} i P(X = i)}{\sum_{i \geq b_{k-1}} P(X = i)} & k > 0 \end{cases} \tag{6}$$

According to the characterisation theorem for Paretian distributions by Glänzel et al. (1984) the conditional expectation satisfies the condition $b_k = E(X|X \geq b_{k-1}) \sim \{\alpha/(\alpha - 1)\} b_{k-1} + b_1$, which, in turn, results by recursion in the following property:

$$b_k \approx b_1 \sum_{i=1}^{k-1} \left\{ \frac{\alpha}{\alpha - 1} \right\}^i \tag{7}$$

with b_1 being the expected value and α the characteristic parameter of the Pareto-approximation. Since the normalised scores $b_k^* = b_k/b_1$ depend only on the characteristic parameter we have called the method *characteristic scaling* and the values $b_k^{(*)}$ as well as the corresponding statistics $\beta_k(\beta_k/\beta_1)$ themselves *characteristic scores*. If the expectation of the citation distribution is finite, that is, if $\alpha>1$ (or, in the finite case, if $\alpha<0$) the values β_k defined above are estimators of the theoretical values b_k.

In the present study we will go a step farther. First we modify the model as follows. Instead of the Pareto-approximation for large x (see Eq. (1)) we will use the Pareto distribution as the underlying model for received citation rates. This does, because of the assumed continuity of the random variable, is not quite in keeping with the discreteness of citation rates, but can readily be used, for instance, as a continuous approximation of its discrete analogue, namely the Waring distribution (cf. Glänzel et al., 1984).

The general form of Pareto distribution, also referred to as *Pareto distribution of the second kind* or *Lomax* distribution, can be obtained from the infinite beta distribution if one of the parameters is chosen 1 (see, e.g., Johnson, Kotz, & Balakrishnan, 1994). In particular, we say that the non-negative random variable X has a Pareto distribution (of the second kind) if:

$$G(x) = P(X \geq x) = \frac{N^\alpha}{(N + x)^\alpha}, \quad \text{for all } x \geq 0 \tag{8}$$

where N and α are positive real parameters. Alternatively, the parameter transformation $a = \alpha/(\alpha - 1)$ is used as well.

In a second step, instead of the ranked sample elements $X_{\nu_k}^*$ the corresponding theoretical values, namely Gumbel's so-called characteristic kth extreme values (Gumbel, 1958) will be used. The authors have shown that for large samples ($n \gg 1$) and relatively small $k \ll n$ a small correction of these extreme values results in modified Gumbel's extreme values according to Glänzel and Schubert (1988a, 1988b) which actually form the median of the corresponding order statistic. However, for reason of simplicity of calculation we will not apply this correction. If the distribution function F of the random variable X is absolutely continuous and strictly monotonous (as, for instance, the Pareto distribution) we can define the characteristic kth extreme value as $u_k := G^{-1}(k/n)$. Assuming these properties of the distribution function we can then define the theoretical group size m_k through the characteristic extreme values and the characteristic thresholds b_k as $u_n := b_k$. Thus, we can consider the statistics β_k and ν_k estimators of the corresponding theoretical values b_k and m_k. The following property is obvious:

$$G(b_k) = G(u_{m_k}) = \frac{m_k}{n} \tag{9}$$

In other words, the theoretical class sizes $m_{k-1} - m_k$ $(k > 0)$ can be derived from the distribution function as follows:

$$m_k - m_{k-1} = n\{G(b_k) - G(b_{k-1})\} \tag{10}$$

If $\alpha > 1$ the expected value of the Pareto distribution is finite and can be expressed by $b_1 = N/(\alpha - 1)$, or alternatively by $b_1 = N(a - 1)$ using $a = \alpha/(\alpha - 1) > 1$. Hence, and from the characterisation theorem for Pearson-type distributions (Glänzel et al., 1984) we obtain the characteristic thresholds b_k by the following recursion:

$$b_k = E(X|X \geq b_{k-1}) = ab_{k-1} + b_1 = N(a-1)\left(\sum_{i=0}^{k-1} a^i\right) = N(a^k - 1) \tag{11}$$

Hence, we have

$$\frac{b_k}{b_{k-1}} = \frac{a^k - 1}{a^{k-1} - 1} = a + \frac{1}{\sum_{i=0}^{k-2} a^i} \tag{12}$$

The following three properties are obvious:

(i) $b_2^* = \frac{b_2}{b_1} = a + 1$

(ii) $\frac{b_k}{b_{k-1}} \in \left(a,\ \frac{a+1}{k-1}\right)$ if $k > 1$

(iii) $\frac{b_k}{b_{k-1}} \sim a$ if $k \gg 1$

For the class size $(m_{k-1} - m_k)$ we obtain the following important properties through the characterisation theorem and the definition of Gumbel's extreme values:

$$G(u_{m_k}) = G(b_k) = N^{\alpha}(N + N(a^k - 1))^{-\alpha} = a^{-k\alpha} = \left[\frac{\alpha - 1}{\alpha}\right]^{k\alpha} = \left(\frac{1-1}{\alpha}\right)^{k\alpha} = q^k \tag{13}$$

where $q := (1 - 1/\alpha)^{\alpha}$. Consequently, we have

$$\frac{m_k}{m_{k-1}} = \frac{m^{k-1} - m_k}{m^{k-2} - m^{k-1}} = q = \left(\frac{1-1}{\alpha}\right)^{\alpha} \quad \text{for all } k > 1 \tag{14}$$

The $\alpha - q$ relationship is presented in Table 1. The most relevant α range in bibliometrics is shaded. $\alpha = 1$ corresponds to a Lotka-type distribution.

In addition to the Pareto distribution we also have a look at one of the important limiting cases, namely the *exponential distribution*. In particular, if N, $\alpha \to \infty$ and $N/\alpha \to \lambda$ for some finite real $\lambda > 0$ we obtain $G(x) = N^{\alpha}/(N + x)^{\alpha} \to e^{-x/\lambda}$. Furthermore, $\alpha \to \infty$ implies $q = (1 - 1/\alpha)^{\alpha} \to e^{-1}$ (cf. Table 1). Analogously to Eqs. (11)–(14) we obtain the properties of the characteristic scores and classes in the case of the exponential distribution as follows:

$$b_k = E(X|X \geq b_{k-1}) = kb_1 = k\lambda \quad \text{and} \quad b_k^* = \frac{b_k}{b_1} = k \tag{15}$$

$$\frac{b_k}{b_{k-1}} = \frac{k}{k-1} \sim 1 \quad \text{if } k \gg 1 \tag{16}$$

$$G(u_{m_k}) = G(b_k) = e^{(-\lambda k/\lambda)} = e^{-k} = q^k, \quad \text{where } q = e^{-1} \tag{17}$$

$$\frac{m_k}{m_{k-1}} = \frac{m_{k-1} - m_k}{m_{k-2} - m_{k-1}} = q = e^{-1} \quad \text{for all } k > 1 \tag{18}$$

Table 1
q as a function of α. The values most frequently observed in bibliometrics are shaded

α	q	α	q	α	q	α	q	α	q
1.00	0.0000	3.00	0.2963	5.00	0.3277	10	0.3487	30	0.3617
1.10	0.0715	3.10	0.2990	5.25	0.3298	11	0.3505	35	0.3626
1.20	0.1165	3.20	0.3015	5.50	0.3316	12	0.3520	40	0.3632
1.30	0.1486	3.30	0.3038	5.75	0.3333	13	0.3533	45	0.3638
1.40	0.1731	3.40	0.3060	6.00	0.3349	14	0.3543	50	0.3642
1.50	0.1925	3.50	0.3080	6.25	0.3363	15	0.3553	55	0.3645
1.60	0.2082	3.60	0.3099	6.50	0.3376	16	0.3561	60	0.3648
1.70	0.2213	3.70	0.3117	6.75	0.3388	17	0.3568	65	0.3650
1.80	0.2323	3.80	0.3133	7.00	0.3399	18	0.3574	70	0.3652
1.90	0.2418	3.90	0.3149	7.25	0.3409	19	0.3580	75	0.3654
2.00	0.2500	4.00	0.3164	7.50	0.3419	20	0.3585	80	0.3656
2.10	0.2572	4.10	0.3178	7.75	0.3428	21	0.3589	85	0.3657
2.20	0.2636	4.20	0.3191	8.00	0.3436	22	0.3594	90	0.3658
2.30	0.2692	4.30	0.3204	8.25	0.3444	23	0.3597	95	0.3659
2.40	0.2743	4.40	0.3216	8.50	0.3451	24	0.3601	100	0.3660
2.50	0.2789	4.50	0.3227	8.75	0.3458	25	0.3604	150	0.3666
2.60	0.2830	4.60	0.3238	9.00	0.3464	26	0.3607	200	0.3670
2.70	0.2868	4.70	0.3249	9.25	0.3470	27	0.3610	250	0.3671
2.80	0.2902	4.80	0.3258	9.50	0.3476	28	0.3612	300	0.3673
2.90	0.2934	4.90	0.3268	9.75	0.3482	29	0.3614	∞	0.3679

The first important property of the Pareto distribution (and the exponential distribution as a limiting case too) is the independence of the group size of the parameter N (or λ in the case of the exponential distribution) while in both cases the characteristic scores are proportional to these parameters. Furthermore, the ratio of the size of subsequent classes is constant according to Eqs. (14) and (18). The ratios of class sizes $(m_{k-1} - m_k)$ with respect to the lowest class $(m_0 - m_1)$ consequently form a geometric series q:q^2:,. . .,:q^k, which might be considered an *inverse Bradford law*. Unlike in the original Bradford law, the number of produced items in the zones or classes, that is the number of citations, is not constant. This law will form the theoretical base for the following empirical study.

3. Characteristic scores and scales in practice

The above-mentioned properties of characteristic scores and scales and of the classes defined by them hold for continuous Pareto distributions. Informetric distributions such as publication activity and citation impact are however discrete integer-valued distributions resulting from processes with continuous or discrete time parameter. In order to obtain robust and interpretable models that can also describe the changing shape of the distribution as time elapses, for instance, simple birth processes (Glänzel & Schoepflin, 1994) or mixtures of a Poisson process with appropriate continuous distributions such as the Gamma distribution (Burrell, 1990; Burrell & Cane, 1982) have been assumed. Both models result in negative binomial processes that can describe important features such as changing citation impact in time or the ageing of information but fail to model the tail properties of informetric distributions (Glänzel & Schubert, 1988a, 1988b). The assumption of generalised Yule processes (Schubert & Glänzel, 1983) or further mixture, e.g., with a Pareto distribution result in processes having a (generalised) Waring or Irwin distribution (e.g., Karlis & Xekalaki, 2005; Xekalaki, 1983) at any time. These models describe the tail properties of informetric distribution in an appropriate manner, and their tail can be approximated by the Pareto distribution. Therefore, one could expect that the above model fits sufficiently well empirical data. In what follows, we will prepare the scores and scale to define an appropriate set of classes (zones) to characterise citation distributions and to set thresholds to distinguish poor, fair, remarkable and outstanding citation impact. It is a consequence from their definition, characteristic scores and scales are particularly suited to assessing excellent and outstanding citation rates. As has shown in the previous section, the number of scores is finites since the procedure stops if the last group is empty. The "natural" number of scores and groups therefore varies from distribution to distribution. However, experience shows that in the application to most citation distributions the use of three or four classes has satisfactory results. If more "fine-tuning" is required, the number of classes can, of course, be arbitrarily extended by adding further classes and the scale can thus be optionally refined. In order to obtain the four classes we proceed as follows.

In this empirical section we first determine the first three scores on basis of the ranked sample ($\beta_0 = 0$ by definition). Class 1 is formed by interval $[\beta_0, \beta_1)$. Its elements are less frequently cited than the average, those of categories 2 through 4 (intervals $[\beta_1, \beta_2)$, $[\beta_2, \beta_3)$ and $[\beta_3, \infty)$) are more cited than the average. Papers of the latter ones were called 'fairly cited', 'remarkably cited' and 'outstandingly cited', respectively, while the elements of the first class were called 'poorly cited' (cf., Glänzel & Schubert, 1988a, 1988b). In this manner, the original (theoretically infinite but practically finite) distribution is reduced to a finite distribution keeping essential properties of the original one and taking as many values as classes are used. Then we are ready to analyse the statistical properties of the empirical classes obtained from procedure described above.

In order to apply this method to empirical data all "citable" papers indexed in the 1980 volume of the *Science Citation Index* of Thomson-ISI (Philadelphia, PA, USA) have been collected. The set of "citable" papers includes the document type *article*, *letter*, *note* and *review*. The data set includes about 450,000 papers that have been assigned to subfields according to the Leuven/Budapest classification scheme consisting of 12 major fields and 60 subfields in the sciences (cf. Glänzel & Schubert, 2003). Citation counts have been determined on the basis of an item-by-item procedure using special identification-keys made up of bibliographic data elements such as part of the first author's name, the publication year, the volume number and first page of publication. Citations to each paper have been cumulated from the publication year till all individual years in the period 1980–2000. Finally citations have been aggregated at the level of the sixty subfields and for all fields combined.

Fig. 1 shows the evolution of relative group sizes on basis of cumulative citation rates received by the papers published in 1980 in all fields combined. The shares of the classes in the total is stabilising very soon after publication. The strikingly 'irregular' behaviour in the publication year 1980 could be explained with the fact that this year is still incomplete for most publications; a paper published in November or December has obviously less chance to be cited in the same year than a paper published in the first quarter of the year. This might also have effect on citation patterns in the subsequent year. However, already in 1983, that is, in the third year after publication, class shares do not change essentially any more. The class sizes form an extremely skewed distribution; 20 years after publication about one quarter (74.7%) of all papers received less-than-average cited, a bit less than one fifth (18.5%) are fairly cited,

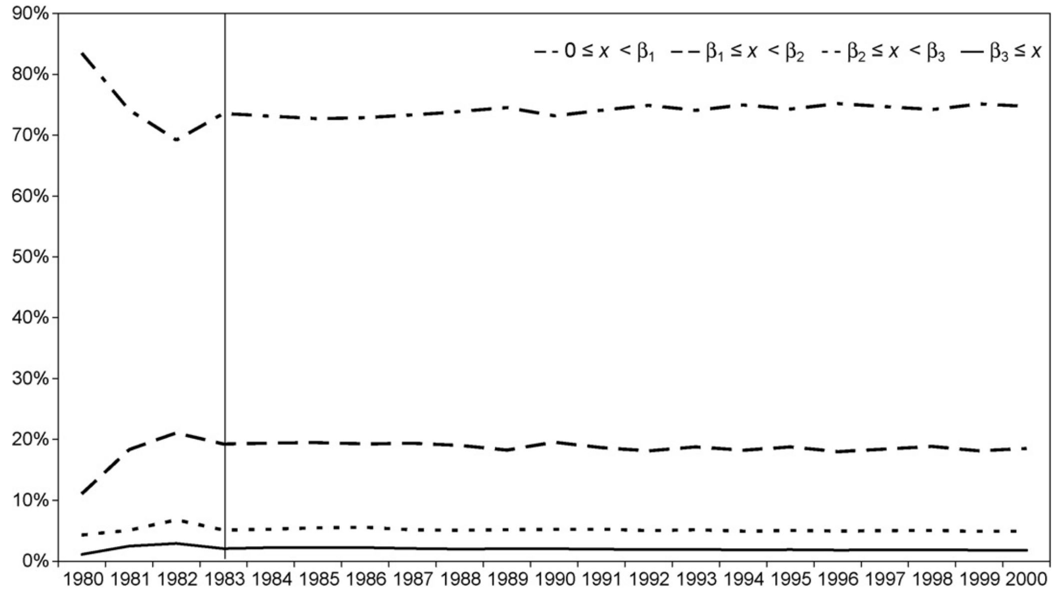

Fig. 1. Evolution of relative group sizes on basis of cumulative citation rates received by papers published in 1980 in the period 1980–2000 (all fields combined).

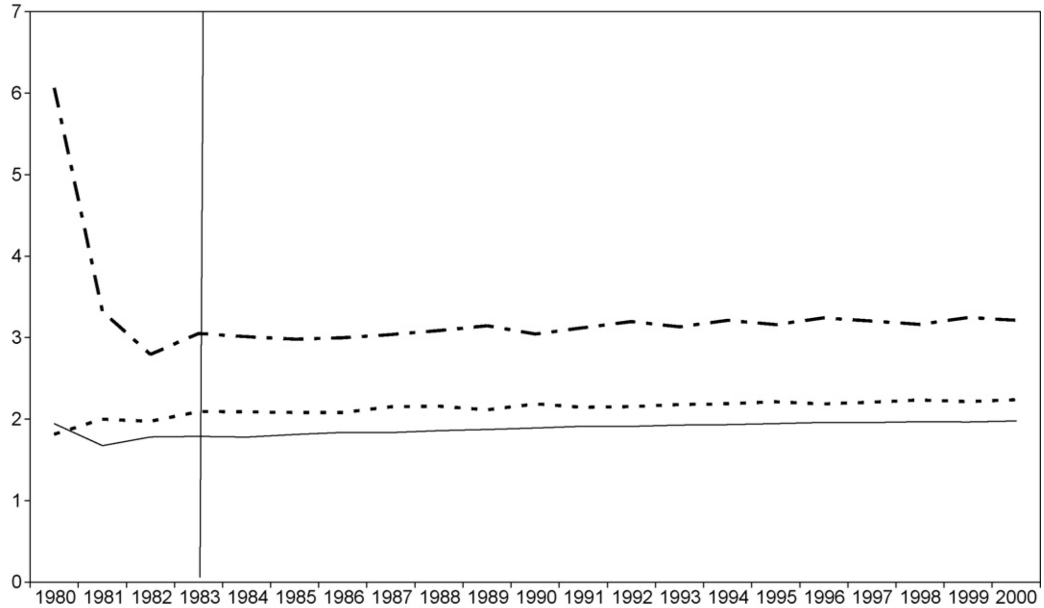

Fig. 2. Evolution of β_k/β_{k-1} ratios on basis of cumulative citation rates received by papers published in 1980 in the period 1980–2000 and in all fields combined ($k=2$ dashed line, $k=3$ dotted line, $k=4$ solid line).

practically one twentieth (4.9%) was remarkably cited and roughly 2% (1.8%) attracted outstanding citation impact. The class-size evolution in the 60 subfields parallels these patters. Stabilisation can be observed from the fourth year on and the sizes themselves are of similar order in most disciplines, too. The approximation 75% (size of class 1), 18% for class 2, 5% for class 3 and 2% for class 4 can be used as a rule of thumb for most disciplines, but should – as all informetric laws – be applied to research evaluation with the utmost caution. A more detailed discussion of the group properties of 12 out of the 60 subfields will form a part of the following section.

Fig. 2, which presents the evolution of β_k/β_{k-1} ratios for the papers published in 1980 in all fields combined, shows similar patterns of early stabilisation. For this exercise we have determined also the score β_4. The evolution of ratios shows a similar picture of stability beginning with the third year after publication. Although the ratios of subsequent score are similar in the individual subfields, the scores themselves vary considerably among the disciplines. Both the class sizes and the β_k/β_{k-1} ratios do not depend on parameter N (see previous section). Since citation processes converge quite slowly to their limiting distribution (cf. Glänzel & Schoepflin, 1995), the time stability of the two statistics shown in Figs. 1 and 2 indicates that N is the time-dependent whereas α does not depend on time.

4. Further results

On basis of the 1980 volume of the Science Index characteristic scores and scales were calculated for each individual citation window 1980 (1 year) through 1980–2000 (21 years). Twelve subfields out of the total of 60 subfields have been selected. Every subfield represents one major field each. The selection is presented in Table 2.

The sizes of these subfields range from $n=3727$ in *aquatic sciences* to $n=19{,}603$ in *pharmacology and toxicology*. All relevant statistics based on all possible citation windows have been determined; in particular the characteristic scores β_k (for $k=1,\ldots,4$) and the shares of the classes in the total $(\nu_{k-1}-\nu_k)/n$ for $k=1,\ldots,3$ and ν_3/n for $k=4$, where these shares are denoted by 'class$_k$'. Here we just mention in passing that Σ_k class$_k=\nu_0/n=1$. In addition, the parameter $q=(1-1/\alpha)^\alpha$ has been estimated according to the property (13), namely as $\tilde{q}=(\nu_k)^{1/k}$ for $k\geq 2$. The mean of the three $\tilde{q}$ values were used as the final estimator of the parameter q for the individual subfields. This estimator is denoted by $\bar{q}$. The corresponding parameter α of the assumed Pareto distribution can be found in Table 1. All above-mentioned statistics for the full citation period 1980–2000 are presented in Table 3.

The statistics presented in Table 2 substantiate that the Pareto approximation works sufficiently well for most subfields. While this approximation provides quite stable results, e.g., for the $\tilde{q}$ estimates of the subfields A4, M4 and

Table 2
List of selected subfields, subfield codes and majors fields to which the subfields are assigned

Code	Subfield	Major field
A4	Food and animal science and technology	Agriculture and environment
Z2	Aquatic sciences	Biology (organismic and supraorganismic level)
B1	Biochemistry/biophysics/molecular biology	Biosciences (general, cellular and subcellular biology; genetics)
R4	Pharmacology and toxicology	Biomedical research
I5	Immunology	Clinical and experimental medicine I (general and internal medicine)
M4	Ophthalmology/otolaryngology	Clinical and experimental medicine II (non-internal medicine specialties)
N1	Neurosciences and psychopharmacology	Neuroscience and behaviour
C1	Analytical, inorganic and nuclear chemistry	Chemistry
P6	Physics of solids, fluids and plasmas	Physics
G1	Astronomy and astrophysics	Geosciences and space sciences
E2	Electrical and electronic engineering	Engineering
H1	Applied mathematics	Mathematics

G1, the $\tilde{q}$ values of the other subfields seem to tendentiously increase or decrease with growing index k. These trends clearly show that the Pareto model can be considered only a rough approximation. The corresponding α parameter ranges between $\alpha = 2.0$ for B1 (biochemistry/biophysics/molecular biology) and 3.5 for M4 (ophthalmology/otolaryngology).

According to the theoretical considerations of the previous section, group sizes do not depend on the parameter N while the characteristic scores are a linear function of this parameter. The corresponding scores of samples with similar $\alpha(q)$ values might therefore considerably differ as the examples H1 (applied mathematics) and B1 (biochemistry/biophysics/molecular biology) show. For instance, a paper, which has received 50 citations in the period 1980–2000, would classify as 'outstandingly cited' in H1 (applied mathematics), but only 'remarkably cited' in C1 (analytical, inorganic and nuclear chemistry) and even just 'fairly cited' in I5 (immunology).

In principle, we can distinguish four basic types according to their citation standard (lower or higher threshold values β_k) and to their class-size distributions. A classification of these properties is presented in Table 4. While the first characteristics depend on both parameters, the latter distribution depends on $\alpha(q)$ alone. Eight of the selected subfield show specific profiles; the pairs M4 and A4, N1 and G1, H1and E2 as well as B1 and I5 have similar

Table 3
Characteristic scores and scales for 12 subfields based on the 21-year citation windows 1980–2000 (n = sample size, class_k = kth class size/n, $\tilde{q} = (\nu_k)^{1/k}$, $\bar{q} = \text{mean}(\tilde{q})$)

k	A4($\bar{q} = 0.291, n = 8531$)			Z2($\bar{q} = 0.298, n = 3727$)			B1($\bar{q} = 0.257, n = 28,161$)			R4($\bar{q} = 0.284, n = 19,603$)		
	β_k	Class_k	$\tilde{q}$	β_k	Class_k	$\tilde{q}$	β_k	Class_k	$\tilde{q}$	β_k	Class_k	$\tilde{q}$
1	9.62	0.710	n/a	18.42	0.700	n/a	29.79	0.738	n/a	16.64	0.719	n/a
2	26.57	0.205	0.290	46.64	0.207	0.300	85.85	0.196	0.262	46.63	0.199	0.281
3	52.67	0.060	0.292	87.21	0.068	0.304	196.55	0.051	0.257	93.71	0.058	0.285
4	92.63	0.025	0.291	159.86	0.024	0.289	437.63	0.016	0.250	166.15	0.023	0.286

k	I5($\bar{q} = 0.262, n = 8811$)			M4($\bar{q} = 0.310, n = 5342$)			N1($\bar{q} = 0.294, n = 9691$)			C1($\bar{q} = 0.288, n = 16,508$)		
	β_k	Class_k	$\tilde{q}$	β_k	Class_k	$\tilde{q}$	β_k	Class_k	$\tilde{q}$	β_k	Class_k	$\tilde{q}$
1	27.00	0.741	n/a	11.66	0.692	n/a	27.46	0.713	n/a	14.24	0.711	n/a
2	79.17	0.191	0.259	30.05	0.211	0.308	74.15	0.202	0.287	37.30	0.206	0.289
3	174.06	0.050	0.262	54.52	0.067	0.311	142.38	0.058	0.292	72.60	0.060	0.288
4	335.88	0.019	0.266	85.76	0.030	0.311	233.90	0.027	0.301	129.39	0.023	0.285

k	P6($\bar{q} = 0.262, n = 14,330$)			G1($\bar{q} = 0.286, n = 5472$)			E2($\bar{q} = 0.254, n = 11,498$)			H1($\bar{q} = 0.254, n = 6675$)		
	β_k	Class_k	$\tilde{q}$	β_k	Class_k	$\tilde{q}$	β_k	Class_k	$\tilde{q}$	β_k	Class_k	$\tilde{q}$
1	14.10	0.743	n/a	23.24	0.712	n/a	7.36	0.759	n/a	6.51	0.755	n/a
2	41.36	0.187	0.257	63.23	0.207	0.288	25.10	0.176	0.241	21.65	0.181	0.245
3	88.34	0.051	0.264	127.96	0.058	0.285	55.85	0.047	0.255	49.66	0.047	0.255
4	167.20	0.019	0.266	232.95	0.023	0.285	104.87	0.019	0.265	98.24	0.018	0.263

Table 4
Four basis types of characteristic scores and scales according to thresholds and class sizes

Properties	Lower citation standard	Higher citation standard
Less skewed class-size distribution (α, q large)	M4, A4	N1, G1
More skewed class-size distribution (α, q small)	H1, E2	B1, I5

Table 5
Time dependence of parameter N as reflected by the change of characteristic scores in time

	Field							
	M4		A4		H1		E2	
	1983	2000	1983	2000	1983	2000	1983	2000
β_1	2.55	11.66	2.24	9.62	1.35	6.51	2.23	7.36
β_2	6.41	30.05	5.96	26.57	4.00	21.65	7.20	25.10
β_3	11.08	54.52	9.46	52.67	6.52	49.66	14.57	55.85
β_4	16.60	85.76	14.30	92.63	10.22	98.24	23.30	104.87
	Field							
	N1		G1		B1		I5	
	1983	2000	1983	2000	1983	2000	1983	2000
β_1	6.74	27.46	6.72	23.24	17.09	29.79	8.95	27.00
β_2	16.11	74.15	16.12	63.23	46.14	85.85	22.32	79.17
β_3	29.04	142.38	28.78	127.96	96.57	196.55	44.18	174.06
β_4	45.52	233.90	43.83	232.95	182.98	437.63	78.19	335.88

profiles each where the distribution characteristics of the first and fourth and the second and third, respectively, can be considered opposite to each other. Most subfields are, however, somewhere in between these 'extreme' profiles. The class sizes of the individual subfields somewhat differ from the above-mentioned rule of thumb (75–18–5–2%), although the deviations from this 'standard' are not dramatic. However, the variation among the characteristic scores of the individual subfields is – according to the expectation on basis of subject specific citation standards – quite considerable.

Finally, we have a look at the time dependence of parameter N. The characteristic scores depend on both parameters N and α (see Eq. (11)). In Table 5, the characteristic scores of the eight subfields shown in Table 4 are presented for the two periods 1980–1983 (4-year window) and 1980–2000 (21-year window). The first four disciplines represent a lower citation standard (see upper part of Table 5) whereas the second group in the lower part of Table 5 stands for higher standards. The different 'growth rates' of the β_k values can be explained with the different ageing of the subfields (cf. Glänzel & Schoepflin, 1995). The deviation of the growth in the mathematics discipline from that in the engineering subfield is most striking (see Table 5).

5. Conclusions

The characteristic scores and scales defined by Glänzel & Schubert (1988a, 1988b) proved an interesting self-adjusting informetric tool that can be used for evaluative scientometric purposes as well. It can, for instance, be applied to journal and subject analysis. The identification of publication subsets according to their citation impact and gauging the mean citation rates of given subsets against a larger characteristic scale are likewise possible. Both scores and class sizes have interesting mathematical properties if a Pareto distribution for the underlying citation distribution is assumed. The size of the classes defined by the characteristic scores as well as the ratios of subsequent scores proved stable beyond an initial citation period of about 3 years. This observation provides on one hand empirical evidence for the time independence of the characteristic parameter (α) of the Pareto distribution and, on the other hand, it shows that the particular choice of the citation windows is – except for a short initial period – not important for class sizes. This does, of course, not imply that the elements of these classes, that is, the individual papers also form stable clusters.

The statement “once highly (poorly) cited always highly (poorly) cited” does not hold. While the size of the classes does not essentially change as time elapses, the composition of the classes might do.

The advantage of being able to determine these characteristic classes on basis of a shorter period and than to apply them to a larger citation window is maybe the most interesting property. The second important property is the 75–18–5–2 property, namely, that about 75% of the papers published in 1980 were poorly cited, 5% and 2% of the papers were ‘remarkably cited’ and ‘outstandingly cited’, respectively. However, this reflects the situation in 1980. Scientific communication has considerable changed during the last two decades; research collaboration and consequently also co-authorship has intensified, individual publication activity and citation impact have generally increased (see Persson, Glänzel, & Danell, 2004). The fact that the journal and disciplinary citation impact is growing but not all subjects are concerned to the same extent, points also to structural changes. The following three subfields may serve just as an example. The mean citation rate of subfield P6 (physics of solids, fluids and plasmas) based on a 3-year citation windows has increased from the publication year 1980 to the publication year 2002 by 168.4%. The increase of the 3-year citation impact in R4 (pharmacology and toxicology) was by far more moderate; it amounts to 38.2% in 2002. By contrast, the change of 6.1% in E2 (electrical and electronic engineering) is almost negligible. One can therefore expect that the above-mentioned 75–18–5–2 rule could somewhat change if the exercise is repeated, say, for papers published in 2000.

Besides its informetric value the method of characteristic scores and scales has implications also for research evaluation. The most important one is that the method provides a rule where thresholds could be set for the identification of highly cited papers within a scientific discipline and that this can be done whether on basis of characteristic scores which increase with growing citation window or on basis of the class sizes which do not change essentially beyond a certain initial citation period. Unknown scores can even be estimated from the lowest citation rate in the set of 2% and 7% (=2 + 5%) most cited papers, respectively. This method can easily be used as reference standard as well. Thus, the high-end of the national or institutional citation distribution can be gauged against the field standard. The share of subject relevant papers of a given country, region or institution found in one of the ‘upper classes’ can thus enlighten the user on the possible correspondence of the high-end of their citation distribution with the reference standard.

The Pareto approach used in the theoretic section provides powerful tools for evaluative bibliometrics; the α parameter (and its derivative q) characterises both the class-size distributions and the slope of the underlying citation distribution. Jointly with an ‘impact measure’ it can be used to distinguish four basic types of citation standard. Furthermore, the characteristic parameter proved to be time-independent, that is, it can be applied to any larger citation window once estimated for a given initial period.

It should be stressed again that the results obtained from this methods are of statistical value. The method of characteristic scores and scales should be applied to publication sets of reasonable size, that is, if the grouping procedure is complete for the ranked observations. It is certainly not designed for the assessment of individual papers, either, since those do not necessarily remain in their classes as time elapses, and might, therefore, be replaced by others.

Acknowledgements

The author wishes to thank Professor Jacqueline Leta for her help in applying the Leuven-Budapest subject classification scheme to the 1980 volume of the Science Citation Index. He also wishes to thank the anonymous referees for their corrections, critical comments and valuable suggestions.

References

Abt, H. A. (1981). Long-term citation histories of astronomical papers. *Publications of the Astronomical Society of the Pacific*, *93*(552), 207–210.

Burrell, Q. L. (1990). Using the gamma-Poisson model to predict library circulations. *Journal of the American Society for Information Science*, *41*(3), 164–170.

Burrell, Q. L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics*, *52*(1), 3–12.

Burrell, Q. L., & Cane, V. R. (1982). The analysis of library data. *Journal of the Royal Statistical Society Series A*, *145*, 439–471.

Egghe, L., & Rao, I. K. R. (2001). Theory of first-citation distributions and applications. *Mathematical and Computer Modelling*, *34*(1–2), 81–90.

Garfield, E. (1980). Premature discovery or delayed recognition. Why? *Current Contents*, *21*, 5–10.

Garfield, E. (1996). How can impact factors be improved? *British Medical Journal*, *313*(7054), 411–413.

Garfield, E. (1998a). Long-term vs. short-term journal impact: Does it matter? *Scientist*, *12*(3), 11–12.

Garfield, E. (1998b). Long-term vs. short-term impact. Part II. *Scientist*, *12*(14), 12–13.

Glänzel, W. (1992). On some stopping times on citation processes, from theory to indicators. *Information Processing and Management*, *28*(1), 53–60.

Glänzel, W. (2005). Methodological questions of subject characteristics in long-term citation analysis. An informetric study with implication for research evaluation. In *Paper presented at the 10th Nordic workshop on bibliometrics, informetrics and research policy held in Stockholm*.

Glänzel, W., & Garfield, E. (2004). The myth of delayed recognition. *The Scientist*, *18*(11), 8–9.

Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. *Scientometrics*, *53*(2), 171–193.

Glänzel, W., & Schoepflin, U. (1994). A stochastic model for the ageing of scientific literature. *Scientometrics*, *30*(1), 49–64.

Glänzel, W., & Schoepflin, U. (1995). A bibliometric study on aging and reception processes of scientific literature. *Journal of Information Science*, *21*(1), 37–53.

Glänzel, W., & Schubert, A. (1988a). Characteristic scores and scales in assessing citation impact. *Journal of Information Science*, *14*(2), 123–127.

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

Glänzel, W., & Schubert, A. (1995). Predictive aspects of a stochastic model for citation processes. *Information Processing and Management*, *31*(1), 69–80.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.

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.

Glänzel, W., Schlemmer, B., & Thijs, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics*, *58*(3), 571–586.

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

Irwin, J. O. (1975a). Generalized Waring distribution. Part 1. *Journal of the Royal Statistical Society Series A*, *138*, 18–31.

Irwin, J. O. (1975b). Generalized Waring distribution. Part 2. *Journal of the Royal Statistical Society Series A*, *138*, 204–227.

Irwin, J. O. (1975c). Generalized Waring distribution. Part 3. *Journal of the Royal Statistical Society Series A*, *138*, 374–384.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). (2nd ed.). *Continuous univariate distributions Vol. 1*, New York: John Wiley & Sons, Inc.

Karlis, D., & Xekalaki, E. (2005). Mixed Poisson distributions. *International Statistical Review*, *73*(1), 35–58.

Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, *60*(3), 421–432.

Rousseau, R. (1994). Double exponential models for 1st-citation processes. *Scientometrics*, *30*(1), 213–227.

Schubert, A., & Glänzel, W. (1983). Statistical reliability of comparisons based on the citation impact of scientific publications. *Scientometrics*, *5*(1), 59–74.

van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics*, *59*(3), 467–472.

Xekalaki, E. (1983). The univariate generalized Waring distribution in relation to accident theory—proneness, spells or contagion. *Biometrics*, *39*(4), 887–895.