Journal of Informetrics 12 (2018) 401–415

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

The lognormal distribution explains the remarkable pattern documented by characteristic scores and scales in scientometrics

Gabriel-Alexandru Vîiu

Research Institute of the University of Bucharest – Social Sciences Division, University of Bucharest, Panduri 90, Bucharest 050663, Romania

ARTICLE INFO

Article history:
Received 3 November 2017
Received in revised form 7 February 2018
Accepted 7 February 2018

Keywords:
Citation analysis
Characteristic scores and scales (CSS)
Lognormal distribution
Universality claim

ABSTRACT

Characteristic scores and scales (CSS) – a well-established scientometric tool for the study of citation counts – have been used to document a striking phenomenon that characterizes citation distributions at high levels of aggregation: irrespective of scientific field and citation window empirical studies find a persistent pattern whereby about 70% of scientific papers belong to the class of poorly cited papers, about 21% belong to the class of fairly cited papers, 6% to that of remarkably cited papers and 3% to the class of outstandingly cited papers. This article aims to advance the understanding of this remarkable result by examining it in the context of the lognormal distribution, a popular model used to describe citation counts across scientific fields. The article shows that the application of the CSS method to lognormal distributions provides a very good fit to the 70–21–6–3% empirical pattern provided these distributions are characterized by a standard deviation parameter in the range of about 0.8–1.3. The CSS pattern is essentially explainable as an epiphenomenon of the lognormal functional form and, more generally, as a consequence of the skewness of science which is manifest in heavy-tailed citation distributions.

1. Introduction

Citation analysis is an essential component of evaluative scientometrics and it is becoming the default mode chosen by policy and decision makers for the exploration and assessment of scientific research. Faced with the pressures of practical requirements incumbent in international university rankings, national evaluation processes as well as institutional appraisal oriented towards funding and promotion decisions, the field of scientometrics has become increasingly saturated over the past decades with a myriad of aggregated indicators which purport to capture scientific performance by combining in often arbitrary and idiosyncratic ways the basic building blocks of citation analysis – published papers and citation counts. However, despite the popular success of some metrics, professional scientometricians have consistently warned against the proliferation of single-number citation-based indicators such as the Hirsch index or impact factor and have sought instead to promote more complex evaluation tools that maintain a multidimensional, pluralistic view of performance (see for instance Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015; Moed & Halevi, 2015).

Against the pitfalls of aggregated indicators one tool increasingly advocated for evaluation and policy use is the method of characteristic scores and scales (Glänzel & Schubert, 1988; Schubert, Glänzel, & Braun, 1987) which offers a straightforward

E-mail addresses: gabriel.viiu@icub.unibuc.ro, gabriel.viiu@yahoo.com

https://doi.org/10.1016/j.joi.2018.02.002
1751-1577/

way of benchmarking the citation performance of individual units of assessment relative to their peers as well as to the overall population of reference based on a common framework of algorithmically constructed performance classes. Since their development in the late 1980s characteristic scores and scales – henceforth CSS – have attracted increased attention from the scientometric community and have become an informative mechanism for conducting evaluations and comparisons at various levels of analysis (journals, institutions, countries). The most important result that has emerged from the continued application of this method over the past years is a remarkable empirical regularity detected in the context of aggregated citation counts: irrespective of scientific field and citation window CSS tend to uncover an extraordinarily stable distribution of papers across predefined classes of citedness. Virtually all empirical studies using CSS (see Section 2.1 below) show that within most fields of science about 69–70% of papers seem to be *poorly* cited (i.e. are included in a citedness class I), 21% of papers seem to be *fairly* cited (i.e. belong to class II), only about 6–7% seem to be *remarkably* cited (class III), and only about 2–3% seem to be *outstandingly* cited (class IV). Viewed against the general background of known discipline heterogeneity – a heterogeneity attributable among others to variations in size, age, publication frequency, citation culture and other field specific characteristics – the discipline and time window invariance of CSS represent an intriguing topic of investigation: why is it that in spite of the many well documented specificities of each scientific field virtually all fields of science are shown by CSS to be fundamentally similar in that they share an approximate 70–21–6–3% distribution of their papers across the four CSS citation performance classes?

This article aims to advance the understanding of the remarkable pattern documented by characteristic scores and scales in scientometrics by examining it in the context of the lognormal distribution, a popular model used to describe citation counts across scientific fields. In technical terms, the article aims to show that the application of the CSS method to observations drawn from lognormal distributions consistently yields a remarkable fit to the typical CSS empirical pattern as long as these distributions are characterized by a standard deviation parameter close to a value of 1. The results of the article cast a new light on the empirical regularity documented by CSS and support the conclusion that the CSS pattern may be circumscribed to the broader phenomenon of the skewness of citation counts across the sciences. Fortunately, this latter phenomenon has already been addressed in the literature and is more readily explainable than the CSS pattern.

The article is structured as follows: Section 2 provides the theoretical and empirical background of the paper, reviewing the operation of CSS and the previous empirical literature that substantiates the remarkable CSS pattern which constitutes the focus of the present study; this section also discusses the use of the lognormal distribution in scientometrics, reviews the contentious universality claims associated with this distribution and articulates the research questions investigated in the paper; Section 3 describes the methodological approach of the article which encompasses some preliminary mathematical considerations and a description of an analytical process involving application of the CSS method to synthetic data derived from hypothetical lognormal distributions within the R language and environment for statistical computing; Section 4 presents and discusses the results while a final Section 5 offers some concluding remarks.

2. Theoretical and empirical background

2.1. Characteristic scores and scales and the 70–21–9% pattern

The operation of the CSS technique revolves around a simple recursion of arithmetic means applied at the level of some n papers published in a particular field of science.[1] The corresponding citations to these papers – $\{X_i\}_{i=1}^{n}$ – are first sorted in descending order to obtain a list of the form $X_1 \geq X_2 \geq \ldots \geq X_n$. Parameters $\beta_0 = 0$ and $\nu_0 = n$ are defined to derive the characteristic scores and scales of the citation distribution and β_1 is given by the initial sample mean of the full distribution of citations:

$$\beta_1 = \sum_{i=1}^{n} \frac{X_i}{n} = \sum_{i=1}^{n} \frac{X_i}{\nu_0} \tag{1}$$

with ν_1 jointly defined by

$$X_{\nu_1} \geq \beta_1 \text{ and } X_{\nu_{1+1}} < \beta_1 \tag{2}$$

The procedure can be iterated in the form

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

to define subsequent sub-sample means with the understanding that ν_k is chosen so that

$$X_{\nu_k} \geq \beta_k \text{ and } X_{\nu_{k+1}} < \beta_k, k \geq 2. \tag{4}$$

[1] The subsequent presentation follows the account given in Glänzel (2010, pp. 704–705).

Although in theory the iterated truncation based on sub-sample means can continue indefinitely, practical application usually limits the procedure to at most five classes of citedness (Schubert et al., 1987): class 0 made up by *uncited* papers, the class of *poorly cited* papers defined on the interval (β_0, β_1), *fairly cited* papers defined on $[\beta_1, \beta_2)$, *remarkably cited* papers defined on the interval $[\beta_2, \beta_3)$ and, finally, the class of *outstandingly cited* papers defined as belonging to the interval $[\beta_3, \infty)$. These five classes can further be collapsed into only four by merging uncited papers with poorly cited ones and even to three classes by further merging outstanding papers with remarkable ones.

The application of CSS is potentially very diverse but it usually considers high levels of aggregation, most often entire scientific fields which in practice tend to be identified with Web of Science subject categories. The method was originally proposed and subsequently employed as a tool for the comparative evaluation of scholarly journals within different fields of science, based on the number of citations to their papers (Glänzel, 2011; Glänzel & Schubert, 1988; Schubert et al., 1987) but more recently it has also been employed to assess the research productivity of academic departments and individual scholars (Abramo, D'Angelo, & Soldatenkova, 2017; Perianes-Rodriguez & Ruiz-Castillo, 2014; Ruiz-Castillo & Costas, 2014). The analysis of citation distributions within and across different scientific fields has nonetheless remained the prevalent use of the method and it is in this specific area that the CSS approach has yielded the most interesting and at the same time unexpected general result, namely that citation counts across the sciences are not only highly skewed (a feature already recognized at least since Seglen's 1992 work) but also very similar in overall shape to one another. This insight can be found in several large scale studies which use citation records from the Web of Science.

In the original paper outlining the CSS method (Schubert et al., 1987) the then 114 scientific fields covered by the Institute for Scientific Information's *Journal Citation Reports* were studied within a citation window of five years; although nowhere explicitly mentioned by the authors of the work, a careful retrospective analysis of the annex to that paper shows that, on average across the 114 fields, about 72% of papers were allocated to the poorly cited class (including uncited papers), 19% to the fairly cited, 6% to the remarkably cited and 3% to the outstandingly cited class. A later study (Glänzel, 2007) focusing on 60 subfields across a citation window extending up to 21 years found a similar trend whereby about 75% of papers were poorly cited, 18% fairly cited, 5% remarkably cited and 2% outstandingly cited. A subsequent comprehensive study (Albarrán & Ruiz-Castillo, 2011) using only three instead of four CSS citation classes and focusing on a sample of 3.9 million articles (five year citation window, 22 broad fields) reported very similar results: 70% of papers in the poorly cited class (including uncited papers), 21% in the fairly cited class and 9% within the conglomerate class pooling together remarkable and outstanding papers. This approximate 70–21–9% pattern – which has become a hallmark result of the CSS approach – is also reported in a related study (Albarrán, Crespo, Ortuño, & Ruiz-Castillo, 2011) which uses two alternative classification schemes to group articles across granular disciplines and broader fields of science.

A further large scale study that confirms the empirical validity of the 70–21–9% rule is Li, Radicchi, Castellano, and Ruiz-Castillo (2013) where about 2.9 million publications indexed in the Web of Science between 1980 and 2004 and grouped within 172 subject categories are analyzed; the overall data covered in this study actually comprise six yearly datasets and it is apparent that the 70–21–9% rule is closely observed by the articles from the more recent years (1995, 1999, 2004) than by the older ones which have had a longer time to accumulate citations (1980, 1985, 1990); for these older articles an approximate 73–19–8% configuration is reported. A somewhat more limited study focusing on 20 subfields and two distinct publication years (2007 with a five year citation window and 2009 with a three year window) again confirmed the 70–21–9% rule, not only at the level of each individual field but also when combining the papers from all fields (Glänzel, Thijs, & Debackere, 2014).

More recently a 69–22–9% pattern was reported for citation counts as well as for Mendeley readership counts based on about 1.1 million articles published in 2012 and classified into 30 disciplines (Costas, Perianes-Rodriguez, & Ruiz-Castillo, 2016). Glänzel (2011) had also reported some nuanced results based on a restricted sample of papers published in 2006 in only three selected fields (with a three year citation window): papers in biophysics/molecular biology closely followed the 70–21–9% pattern but those in applied mathematics showed a 75–18–7% distribution and those in electrical and electronic engineering deviated substantially, having a 63–25–12% configuration. A more recent study also considering a restricted sample of papers published over the 2009–2013 period (Vîiu, 2017) found further support for the more typical 70–21–9% pattern in four Web of Science subject categories.

While variations across smaller scale studies are to be expected, a final large scale study that confirms the 70–21–9% pattern deserves separate mention due to its markedly different methodological approach: whereas most of the studies previously mentioned worked within the framework of the predefined Web of Science categories Ruiz-Castillo and Waltman (2015) take a more innovative approach that involves determining scientific fields of variable granularity via algorithmic clustering: based on 3.6 million articles from 2005 to 2008 (a subset arrived at from a more comprehensive pool of about 9.4 million publications from 2003 to 2012) up to 12 distinct classification systems are constructed with between 231 and 11,987 significant clusters (i.e. clusters having at least 100 publications); remarkably, for most of these 12 granularity levels the 70–21–9% pattern is obeyed quite closely, significant departures occurring only in the more fine-grained classifications (granularity levels 9–12) which have a high prevalence of small clusters and where an approximate 67–22–11% pattern seems to prevail.

The preceding paragraphs make it clear that by now there is a substantial body of work attesting to the recurrence of the CSS pattern embodied by the approximate 70–21–9% distribution when focusing on citation counts aggregated at the level of scientific fields. While virtually all studies confirm the fact that citation distributions across narrow fields as well as broader disciplines are skewed and fundamentally similar, an explanation for why the specific 70–21–9% result keeps emerging has

yet to be provided. The recurrence of this pattern seems to have acquired the status of an ultimate result, one so eagerly integrated into current scientometric lore that it defies further inquiry. However, in this article the competing premise is pursued, namely that instead of viewing the CSS pattern as a final result which does not require further exploration, the remarkable result incumbent in the application of the CSS algorithm can and should be understood further, not merely by further empirical confirmation of the pattern across scientific fields, but by considering whether it could naturally emerge from an underlying statistical distribution that could plausibly be used to describe the citation counts within each of these fields.

Since the application of the CSS method is essentially interlinked with the more general problem of statistical analysis of citation counts it seems warranted to explore the CSS pattern from the standpoint of the existing statistical models proposed and successfully used in the previous literature to describe citation distributions. In other words, the key to advancing our understanding of the CSS pattern should be found in the body of work devoted to the mathematical modelling of citation counts. Since a variety of distributions have been proposed for this task the fundamental question to be answered is whether or not there exists a specific statistical distribution such that the application of the CSS algorithm to this distribution yields the precise 70–21–9% pattern (or an equivalent 70–21–6–3% form if considering four classes instead of only three). If such a distribution can be identified then the pattern documented by CSS in citation analysis could be re-conceptualized as a necessary epiphenomenon of the application of CSS to the specific distribution in question, rather than as a stand-alone result.

2.2. Modelling citation counts and the lognormal distribution

There are several types of statistical distributions which have been proposed to model citation counts in an accurate manner. These include the power law distribution (Price, 1976) and the so-called "hooked" or shifted power law distributions (Thelwall & Wilson, 2014), the negative binomial and the Waring distribution (Glänzel, 2009), a modified Bessel function (Van Raan, 2001), the double exponential-Poisson (Vieira & Gomes, 2010), a stretched-exponential and a form of the Tsallis q-exponential function (Wallace, Larivière, & Gingras, 2009), stopped sum distributions (Low, Wilson, & Thelwall, 2016), the generalized inverse Gaussian-Poisson distribution (Sichel, 1992) and, last and most influentially, the lognormal distribution. Recent studies (Thelwall, 2016b, 2016c; Thelwall & Wilson, 2014) indicate that of these and other distributions, the one that is best able to capture the *full spectrum* of citation counts is the (discretized) lognormal. This is an important point because the application of the CSS method always takes into account the full set of citation counts and the resulting 70–21–9% pattern emerges from this premise, not from partial sets which disregard zero citations or citation counts below a certain value.[2] For this reason this article will concentrate only on the lognormal distribution but the general approach employed here could conceivably be applied to other distributions as well.

While there are many scientific areas where lognormal processes seem to prevail – aerobiology, ecology, environment, geology and mining, medicine, linguistics, social sciences and economics among others (Limpert, Stahel, & Abbt, 2001) – in the case of scientometrics the lognormal distribution has been used for some time to model a wide range of phenomena. In general the distribution has been used under its standard two parameter continuous form or under discretized versions to successfully model citation count data of individual journals in a single year (Stringer, Sales-Pardo, & Nunes Amaral, 2010), citations of individual researchers and academic departments across multiple years (Moreira, Zeng, & Amaral, 2015) and citation counts across broad scientific subject categories (Thelwall, 2016a). The lognormal distribution was also found to adequately capture the distribution of citation-based indicators, including the h and g-index (Perc, 2010), a generalized h index (Wu, 2013), as well as the total research impact (Tori) indicator (Kurtz & Henneken, 2017). This distribution has also been found to offer a good fit for citation age data (Burrell, 2002; Egghe & Rao, 1992; Matricciani, 1991), for the number of references made in scientific papers (Egghe & Rao, 2002; Morris, 2005), as well as for Mendeley readership counts (Thelwall & Wilson, 2016).

A feature that strongly individualizes the lognormal distribution within the scientometric landscape is the fact that it has been a recurring vehicle for several controversial universality claims. There are in fact several such claims which target different scientometric topics: the distribution of *normalized* citation counts across different scientific fields, the distribution of *raw* citation counts across different scientific fields, the distribution of normalized citation counts across *institutions*, as well as the distribution of scientific *productivity* have all been claimed to follow a lognormal pattern.

Radicchi, Fortunato, and Castellano (2008) initiated the lognormal universality debate for citation counts by arguing that if one divides the citations of individual papers by the average of the field (thereby obtaining a relative indicator labelled c_f), then inter-field variability essentially vanishes and a universal lognormal curve emerges irrespective of the scientific discipline; this curve is characterized by a σ^2 parameter of 1.3 (equivalent to a σ of about 1.14). The universality claim of Radicchi and his coauthors was further elaborated and extended to the level of individual journals (Castellano & Radicchi, 2009) and it was further scrutinized at the micro level of chemistry subfields (Bornmann & Daniel, 2009) where the advantages of the c_f indicator were discussed compared to z-scores.

[2] Power law models are typically fitted with success only to the higher tail of citation distributions which means that they fail to capture citation counts in the lower tail, including uncited articles which are relevant for CSS.

There is also substantial empirical evidence against this universality claim, most notably Albarrán et al. (2011) who refute the claim based on a large sample of 3.7 million articles from 219 Web of Science categories (the original universality claim was based on a very limited sample of only 14 fields), Waltman, van Eck, and van Raan (2012) who show that important deviations from the lognormal hypothesis are especially common for fields with a low average number of citations (for instance social sciences) as well as Thelwall and Wilson (2014) who argue based on citation data from 20 Scopus categories that the lognormal distribution does not offer a universal fit. Despite the persuasive rebuttal of the initial claim, in a follow-up study (Radicchi & Castellano, 2012) the authors of the universality thesis reinforce their argument and extend their claims to include raw citation counts by appealing to a novel methodological approach: they resort to aggregating the citations to papers across all scientific fields to then derive a transformed citation count for each individual field, based on the cumulative distribution of citations within the all-sciences aggregated set which is taken as a reference; by studying the properties of the transformed citation counts for the individual fields Radicchi and Castellano conclude that raw citation distributions seem to be universal in the sense of being part of the same family of univariate distributions (i.e. the log-location-scale family which includes the lognormal and Weibull distributions).

Whereas the original lognormal universality claim was derived based on citation data from Web of Science subject categories, Evans, Hopkins, and Kaube (2012) confirm the claim at the level of a specific research institute, at the sub-level of departments, but also for data from the arXiv e-print archive. Perianes-Rodriguez and Ruiz-Castillo (2016) also verify the universality claim at the institutional level, focusing on 500 universities from the 2013 edition of the Leiden Ranking; these authors point out that in the case of universities the universality claim is untenable. Chatterjee, Ghosh, and Chakrabarti (2016) also verify the universality claim at the level of institutions (only 42 in their case) and ultimately find that although it seems to be present "universality is not very strong, and holds only in an approximate sense" (p. 9).

Finally, evidence in favor of another type of lognormal universality – contingent on a σ parameter of 0.94 ± 0.23 – has been provided recently for the scientific productivity of scholars working in either soft or hard scientific disciplines (Bonaccorsi et al., 2017).

2.3. Research questions

Setting aside the issues of productivity and institutional performance, the universality claims relevant to the present article are the ones concerning the distribution of citation counts across scientific fields. Based on the paragraphs from the previous section we may distinguish between a *strong* universality claim and a *weak* universality claim. The former is characterized by the fact that it not only specifies a functional form which citation counts in all fields of science allegedly follow (i.e. lognormal), but it further specifies concrete parameters of the functional form (i.e. σ of about 1.14). The weak universality claim on the other hand is more diffuse: it only asserts that citation counts follow the general lognormal functional form.

Although significant evidence has been offered against the universality claims (especially against the strong version, as explained above) it is important to recognize that these claims, developed under the strong version in the context of normalized citation counts and then extended – under the weaker version – to raw citation counts, have immediate consequences for the CSS empirical pattern, a fact which seems to have been overlooked in the previous literature: if the lognormal universality claims are accurate – i.e. if the lognormal distribution indeed offers a good depiction of raw empirical citation counts across scientific fields, as many recent studies in fact suggest – then it must also explain the CSS pattern since this pattern itself ultimately also professes the (near)universal similarity in shape of raw citation count distributions across the very same scientific fields.[3] In other words, if the lognormal universality claims are true then the CSS pattern should be a specific manifestation of universality, and the pattern can essentially be understood as a consequence of lognormality rather than as a stand-alone result.

To advance our understanding of the CSS pattern it is necessary to explore within a formal mathematical and statistical framework the following inter-related research questions:

(1) Does the 70–21–9% CSS pattern – or a pattern that is reasonably close to this one – actually arise under the specific scenario of the lognormal universality claim?
(2) Assuming that in general the lognormal distribution is a plausible model for raw citation counts, under what specific circumstances (i.e. parametrizations) does the CSS pattern emerge?

The first of these questions is directed towards the strong version of the lognormal universality claim (it is focused on the lognormal distribution with the specific σ parameter of 1.14) while the second is concerned with the weaker version (i.e. it starts from the softer premise that citations could be modelled by a general lognormal functional form, but this form could have many different parametrizations, not only a rigid σ of 1.14). Note that while the universality claims play an important role in articulating the research questions (and therefore serve as a convenient expositional device for the results) they are

[3] Note that although CSS were devised to be applied to *raw* citation counts, their application to *normalized* values yields the exact same results in terms of the distribution of papers across the citation classes because the CSS method is scale-independent.

not an essential premise of the paper. Aside from the existence of the CSS pattern the essential premise of the paper is the idea that the lognormal functional form is a very plausible model for empirical citation counts.

3. Methodological notes

3.1. Preliminary calculations

An important point to consider towards the end of exploring the CSS pattern in the context of the lognormal distribution is the idea that given its fixed, algorithmic nature, the CSS method will naturally produce specific configurations across its performance classes when applied to particular distributions having a known functional form and a predefined parametrization. Consider as an intuitive illustration the case of the most widely known distribution in statistics, i.e. the normal distribution. We know that for this distribution perfect symmetry relative to the mean and identity of the three indicators of central tendency (arithmetic mean, median and mode) are defining characteristics. Starting from this information alone it becomes evident that application of the CSS algorithm to *any* normal distribution would *always* allocate 50% of the observations to the lowest ranking class; clearly this is far below the 70% value which we would expect based on the CSS empirical pattern and therefore, irrespective of the composition of the other classes, we can conclude that the normal distribution can only offer a very inadequate fit for empirical citation counts. We can therefore also reject the possibility that the typical CSS pattern could originate in normally distributed citation counts.

The same reasoning can be used to investigate the explanatory power of other types of distributions, bearing in mind their specific properties and considering reasonable parametrizations. For the specific case of the lognormal distribution we can start from the known functional form of the probability density function $p(x)$ which, for the standard two-parameter lognormal distribution with mean μ_λ and standard deviation σ_λ, is given by[4]

$$p(x) = \frac{1}{x\sigma_\lambda\sqrt{2\pi}} e^{-\frac{\left(\mathrm{Log}[x]-\mu_\lambda\right)^2}{2\sigma_\lambda^2}} \tag{5}$$

Note now that integrating this function over the interval $[0,\mu]$ essentially yields the share of observations that fall within the CSS class of poor performers; if we call this quantity C_p and consider the case when the parameters of the lognormal distribution are taken to be $\mu_\lambda = 0$ and $\sigma_\lambda = 1$ we have

$$C_p = \int_0^{\mu} \frac{e^{-\frac{1}{2}\mathrm{Log}[x]^2}}{x\sqrt{2\pi}} dx, \tag{6}$$

and, since for the lognormal distribution the arithmetic mean $\mu = e^{\mu_\lambda + \frac{\sigma_\lambda^2}{2}}$, C_p becomes

$$C_p = \int_0^{\sqrt{e}} \frac{e^{-\frac{1}{2}\mathrm{Log}[x]^2}}{x\sqrt{2\pi}} dx \tag{7}$$

The integral in Eq. (7) evaluates[5] to approximately 0.69146, meaning 69.15% of the observations derived from such a distribution fall in the CSS class of poor performers. This is remarkably consistent with the CSS pattern discussed in Section 2.2 and is what we should expect to obtain if the continuous lognormal distribution (with $\mu_\lambda = 0$ and $\sigma_\lambda = 1$) were to offer a good overall fit for citation data. However, as also mentioned in the previous section, according to the strong universality claim citation counts in various scientific fields are modelled by lognormal distributions whose σ_λ parameter corresponds to about 1.14. For this specific case, holding $\mu_\lambda = 0$,[6] the probability density function $p(x)$ becomes

$$p(x) = \frac{0.34995 e^{-0.38473\mathrm{Log}[x]^2}}{x} \tag{8}$$

[4] The defining characteristic of the lognormal distribution is the fact that the natural logarithm of the raw values follows a normal distribution. Note in this context that the use of the λ subscript is meant to help avoid confusion between raw values and logarithmic ones: in the following equations μ denotes the arithmetic mean of the raw values (equivalent to the threshold value that separates the CSS class of poorly cited papers from that of fairly cited papers) while μ_λ denotes the mean of the same values in logarithmic space.

[5] Calculations performed using the Wolfram|Alpha computational engine (https://www.wolframalpha.com/).

[6] Note that the μ_λ parameter is in fact inconsequential from the perspective of CSS class composition; for instance, given a fixed $\sigma_\lambda = 1$, there is no difference in the C_p value between a lognormal distribution with $\mu_\lambda = 0$ and a distribution with $\mu_\lambda = 1$ or 15: they all evaluate to about 0.69146 because in each case $C_p = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{1}{2\sqrt{2}}\right)\right)$, where erf(x) is the error function. This property follows from the fact that the μ_λ parameter "affects only the location of the distribution. It does not affect the variance or the shape (or any property depending only on differences between values of the variable and its expected value)" (Johnson, Kotz, & Balakrishnan, 1994, p. 208).

and

$$C_p = \int_0^{1.91516} \frac{0.34995 e^{-0.38473 \mathrm{Log}[x]^2}}{x} dx \quad (9)$$

The integral in Eq. (9) evaluates to approximately 0.71566, meaning 71.57% of the observations derived from such a distribution fall in the CSS class of poorly cited papers. This value is also essentially consistent with the CSS pattern.

3.2. Exploring the CSS pattern with synthetic data

It is obvious that by incrementally shifting the σ_λ parameter potentially infinitely many parametrizations of the lognormal distribution can be considered and for each specific case a corresponding configuration of the CSS classes can be determined. A critical question which then arises is to what extent departures of the σ_λ parameter from the baseline value specified by the strong universality claim (1.14) still yield CSS class configurations that are reasonably consistent with the CSS pattern. Numerical evaluation of integrals following the structure outlined above could answer the question regarding the percent share of observations that fall in the poorly cited class under various parametrizations. However, in addition to the C_p value we must also determine the other quantities of interest, namely the percentages of papers in the fairly cited class C_f, in the remarkably cited class C_r, and in the outstandingly cited class C_o.

An effective way of determining the CSS class configurations produced by various parametrizations of the lognormal distribution is to iteratively generate large samples of synthetic data conforming to hypothetical distributions constructed following the specific parametrizations, apply the CSS algorithm to each sample and then identify the values towards which convergence occurs. To answer the research questions of the present article this process was undertaken in the R language and environment for statistical computing (R Core Team, 2016) using the *rlnorm* function and custom code designed to implement the CSS algorithm with four classes. The *rlnorm* function generates a desired number of random observations that conform to a continuous lognormal distribution whose μ_λ and σ_λ parameters are specified in advance.

For the present article the μ_λ parameter was fixed to 1 and values of the σ_λ parameter between 0.15 and 2.00 were tested, in 0.05 increments; more extreme σ_λ values of 3, 4 and 5 were also considered leading to a total of 41 distinct parametrizations. To ensure reliability of the results 10000 distinct synthetic samples consisting of 1000, 10000, 30000 and 50000 observations[7] each were generated for every individual lognormal parametrization; then, the CSS method was applied to each of the 10000 samples and the resulting C_p, C_f, C_r and C_o quantities were recorded. The basic R code used for the application of the CSS algorithm to the synthetic lognormally distributed data is available as **S**upplementary material 1. As a final step, the mean C_p, C_f, C_r and C_o values across the 10000 synthetic samples considered for each parametrization were computed together with the coefficient of variation (CV) for each of the four quantities across the 10000 samples. Minimum, maximum, 1st and 3rd quartile as well as median values could also be considered for the four quantities of interest but the CV is a more concise measure of the variability of these quantities and it was therefore the sole statistical indicator to be retained. The full results for each specific scenario are provided as Supplementary material 2; see also Table 1 for specific examples.

One additional aspect to be accounted for when investigating the CSS pattern in the context of the lognormal distribution is the fact that citation counts are discrete quantities, not continuous ones. To address this issue, for each of the data simulations described in the previous paragraph a counterpart was also considered based on a discrete version of the lognormal distribution. To obtain this discrete version lognormally distributed real values obtained in R with the *rlnorm* function were rounded to the nearest integers[8] and the resulting values were offset with 1 to address the fact that lognormal distributions cannot accommodate null values since the natural logarithm is available only for positive numbers.

Overall, given the 10000 iterations for each of the 41 parametrizations and given the 4 sample size variations and the two circumstances of continuous and discrete variable types a total of 328 distinct scenarios were explored leading to 74.62 billion synthetic observations. The following section reports the results obtained by applying the CSS method to these synthetic data.

[7] These sample size values were selected because they roughly approximate the number of publications that constitute actual scientific fields throughout a reasonable citation window ranging from three to five years, as usually employed in studies using the CSS method. For example, for the current 252 Web of Science subject categories considered throughout the five year period 2009–2013 the median number of publications (articles, reviews and letters) is about 30000, while the mean is about 47000. While most categories – about two thirds (162 of the 252) – are made up of fewer than 50000 records, there is significant variation across the 252 categories: about one seventh have fewer than 10000 records (Slavic literature has the minimum number: 675) throughout this five year window but another one seventh have in excess of 100000 records (multidisciplinary materials science has the maximum number of items: 403246).

[8] The idea of approximating a discrete lognormal distribution by rounding continuous lognormal real values is also used by Brzezinski (2015). Note that alternative strategies are available to obtain discrete lognormally distributed observations (Thelwall, 2016d) but these are more computationally demanding and cumbersome to implement in R.

Table 1
Percentage of observations in each CSS class and corresponding coefficient of variation across selected parametrizations of continuous lognormal distribution.

		C_p	C_f	C_r	C_o
$\sigma_\lambda = 0.50$	%	59.87	25.45	9.41	5.27
	CV	0.00	0.01	0.01	0.02
$\sigma_\lambda = 0.55$	%	60.84	25.06	9.10	5.00
	CV	0.00	0.01	0.01	0.02
$\sigma_\lambda = 0.60$	%	61.80	24.66	8.80	4.74
	CV	0.00	0.01	0.01	0.02
$\sigma_\lambda = 0.65$	%	62.74	24.26	8.50	4.49
	CV	0.00	0.01	0.01	0.02
$\sigma_\lambda = 0.70$	%	63.68	23.86	8.21	4.26
	CV	0.00	0.01	0.01	0.02
$\sigma_\lambda = 0.75$	%	64.62	23.44	7.91	4.03
	CV	0.00	0.01	0.01	0.02
$\sigma_\lambda = 0.80$	%	65.54	23.02	7.63	3.81
	CV	0.00	0.01	0.01	0.02
$\sigma_\lambda = 0.85$	%	66.46	22.60	7.35	3.59
	CV	0.00	0.01	0.01	0.03
$\sigma_\lambda = 0.90$	%	67.37	22.16	7.08	3.39
	CV	0.00	0.01	0.02	0.03
$\sigma_\lambda = 0.95$	%	68.26	21.73	6.81	3.20
	CV	0.00	0.01	0.02	0.03
$\sigma_\lambda = 1$	%	69.15	21.29	6.54	3.02
	CV	0.00	0.01	0.02	0.03
$\sigma_\lambda = 1.05$	%	70.02	20.85	6.29	2.85
	CV	0.00	0.01	0.02	0.03
$\sigma_\lambda = 1.10$	%	70.88	20.41	6.04	2.68
	CV	0.00	0.01	0.02	0.04
$\sigma_\lambda = 1.14$	%	71.56	20.05	5.84	2.55
	CV	0.00	0.01	0.02	0.04
$\sigma_\lambda = 1.20$	%	72.57	19.51	5.55	2.37
	CV	0.00	0.01	0.02	0.04
$\sigma_\lambda = 1.25$	%	73.40	19.06	5.32	2.22
	CV	0.00	0.01	0.02	0.04
$\sigma_\lambda = 1.30$	%	74.21	18.61	5.09	2.08
	CV	0.00	0.01	0.02	0.05
$\sigma_\lambda = 1.35$	%	75.01	18.17	4.87	1.96
	CV	0.00	0.01	0.02	0.05
$\sigma_\lambda = 1.40$	%	75.80	17.72	4.65	1.83
	CV	0.00	0.01	0.02	0.06
$\sigma_\lambda = 1.45$	%	76.57	17.27	4.44	1.72
	CV	0.00	0.01	0.03	0.06
$\sigma_\lambda = 1.50$	%	77.34	16.82	4.24	1.61
	CV	0.00	0.01	0.03	0.06

4. Results

4.1. The CSS pattern across lognormal distributions

Table 1 presents the mean values of the C_p, C_f, C_r and C_o quantities obtained across the 10000 applications of the CSS method to each of the corresponding 10000 synthetic samples derived from continuous lognormal distributions under various σ_λ parameter values (note that for conciseness results for only 21 of the 41 parametrizations considered are shown). The table is based on the scenario involving sample sizes of 30000 observations but it must be noted that the results obtained under the four sample sizes of 1000, 10000, 30000 and 50000 observations are virtually identical; therefore, to avoid redundancy, the discussion within this section and the figures referenced below also rely only on the scenarios involving sample sizes of 30000 observations.

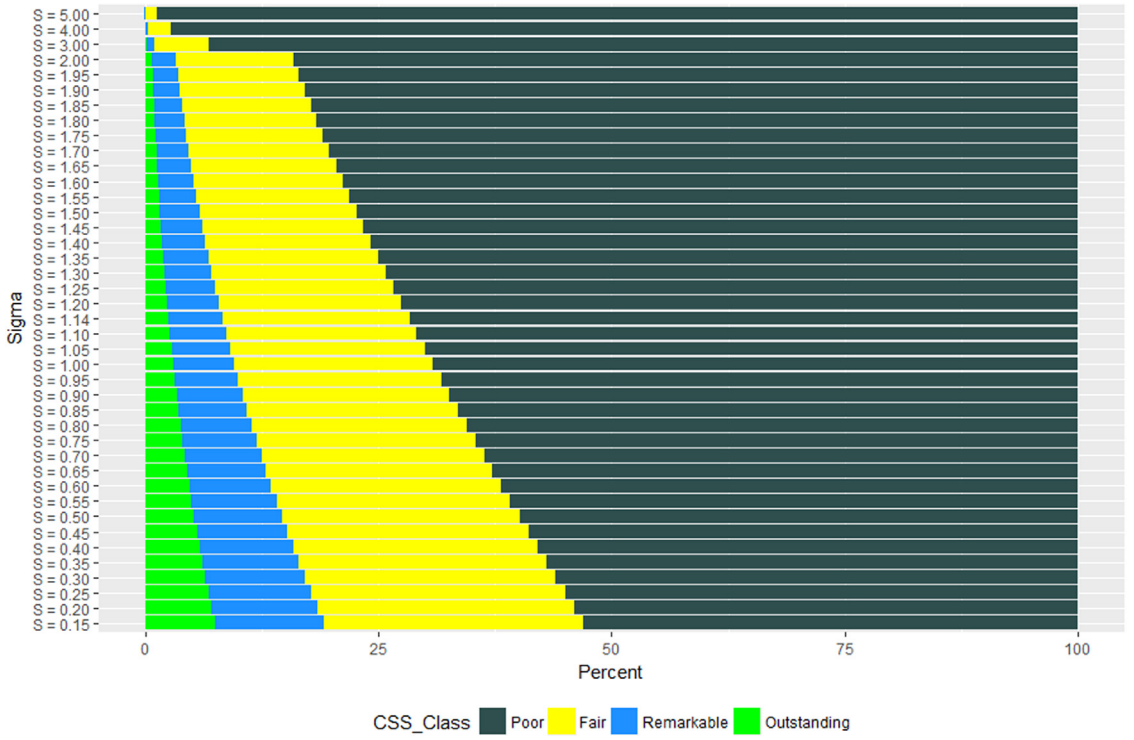

Fig. 1. CSS class composition across different parametrizations of continuous lognormal distribution.

For the case where $\sigma_\lambda = 1.14$ – which corresponds to the strong version of the universality thesis – Table 1 shows the C_p, C_f, C_r and C_o values to be about 71.57, 20.05, 5.84 and 2.55%. Overall this is in good agreement with the 70–21–6–3% CSS pattern but an even closer fit to the pattern emerges for the case when $\sigma_\lambda = 1.05$ and a 70.02–20.85–6.29–2.85% CSS class configuration is obtained. Note that in both these cases the CV values associated with each individual quantity of interest is negligible (at most 4%) meaning the variability of the results over the 10000 applications of the CSS algorithm is very limited. This is in fact the case for all the four quantities across nearly all of the parametrizations considered: with the exception of the more extreme cases ($\sigma_\lambda = 3$–5) the CV for C_p, C_f and C_r is below 5%; more variability is present in the case of C_o values but even here it is only for σ_λ values greater than 1.3 that the CV moves beyond 5% and only for σ_λ values greater than 1.8 that the CV moves beyond 10%.

The different σ_λ values considered yield diverse CSS class configurations but from the selected parametrizations presented in Table 1 a clear trend is visible: as the value of the σ_λ parameter increases the size of C_p also increases and this leads to the proportional decrease in C_f, C_r and C_o values. A visual rendition of this trend across the 41 parametrizations considered is given in Fig. 1. With each 0.05 incremental increase of the σ_λ parameter the C_p value also increases, moving from about 53% when $\sigma_\lambda = 0.15$ to almost 84% when $\sigma_\lambda = 2$; for more extreme values ($\sigma_\lambda = 3$–5) the percent of observations placed by the CSS algorithm in the class of poor performers increases to the point of almost encompassing the entire set of observations ($C_p \approx 93, 97, 99\%$). In the case of C_f values between about 28% (when $\sigma_\lambda = 0.15$) and 12.5% (when $\sigma_\lambda = 2$) are obtained, with extreme σ_λ values leading to ever decreasing C_f shares: about 6% when $\sigma_\lambda = 3$ and 1% when $\sigma_\lambda = 5$. For C_r values of about 11% are obtained in the lower spectrum of σ_λ and values of about 2.5% are obtained when $\sigma_\lambda = 2$, while extreme σ_λ values make the C_r shares drop below 1%. Finally, in the case of C_o values between about 7.5% (when $\sigma_\lambda = 0.15$) and 0.8% (when $\sigma_\lambda = 2$) are obtained; extreme σ_λ values of 3, 4 and 5 also push C_o near 0%.

The analysis of the distribution of the C_p, C_f, C_r and C_o values – see the density graphs in Supplementary material 2 for a better illustration – reveals two additional facts: first, the 41 parametrizations considered yield more variability in the two lower CSS classes (captured by C_p, C_f) while the two higher classes (C_r, C_o) tend to be more compact; second, despite the wide range of values considered for the σ_λ parameter, values for all four quantities tend to be clustered around some typical values which correspond to the general CSS pattern: most C_p values are concentrated between 60 and 80%, i.e. in the vicinity of 70%, most C_f values are concentrated around 21%, most C_r values are concentrated around 6% and most C_o values around 3%.

Having presented these overall results – recall that they are based on continuous lognormal distributions – it is useful to also perform a reliability check of the large scale data simulations given the mathematical considerations from the previous section. Specifically, it is important to establish whether or not the C_p values obtained with the aid of computer simulations are a good match for the more precise C_p values which can be obtained through integration of the lognormal probability

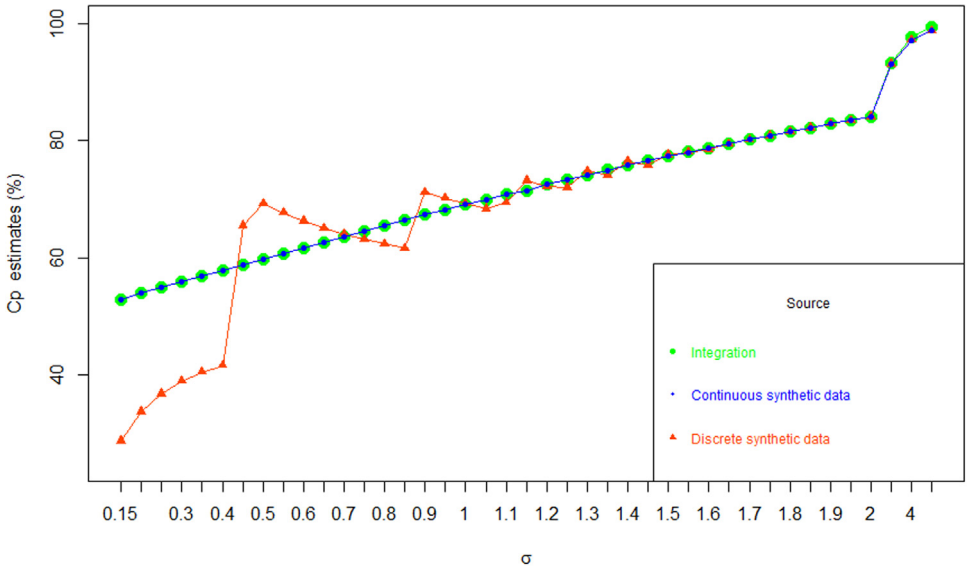

Fig. 2. Comparison of C_p estimates over 41 lognormal parametrizations.

density function, as illustrated in Section 3.1. Fig. 2 shows that there is a near-perfect overlap of C_p estimates obtained through integration with the C_p estimates obtained with computer simulations based on the continuous lognormal scenario. However, this figure also shows that for the scenario involving discretized lognormal data the C_p estimates obtained tend to deviate from the precise, integration-based values; the deviations occur for σ_λ values below 1.5 and are more pronounced for σ_λ values close to 0.5 and especially for those below 0.5.

The CSS class configurations based on discretized synthetic data are in essence similar to those presented above for continuous data (see Supplementary material 2 for detailed data and graphical illustrations). While the results for the discretized scenario lose the smoothness that characterizes continuous data (they are noisier, especially in the lower spectrum of σ_λ values where the effect of rounding in order to obtain discrete values has a significant impact on the resulting distribution of observations) the same essential findings discussed for the continuous case also hold. First, under the strong lognormal universality hypothesis ($\sigma_\lambda = 1.14$) C_p, C_f, C_r and C_o values close to the typical CSS pattern are obtained, namely 73.18%, 18.84%, 5.55% and 2.43%. Second, as the value of the σ_λ parameter increases the size of C_p also tends to increase and this leads to proportional decreases in C_f, C_r and C_o values. Third, despite variation in the σ_λ parameter the C_p, C_f, C_r and C_o values tend to be clustered around the typical CSS pattern.

4.2. Proportion of total accounted for by each CSS class

In addition to the way individual observations are allocated to specific CSS classes another type of analysis usually presented when using the CSS method is the percent of total citations accounted for by each class. For example Albarrán and Ruiz-Castillo (2011) report that on average class I accounts for 22.7% of total citations, class II for 33.3% and classes III and IV for the remaining 44%. Very similar results are reported by Ruiz-Castillo and Waltman (2015). It is worth exploring what shares of a total stock are accounted for by each CSS class when the underlying distribution the observations are derived from is a lognormal one. These shares were computed for all computer simulations ran in the R software for the 41 lognormal parametrizations and are reported in detail in Supplementary material 2. Table 2 and Fig. 3 offer a representation of the results obtained for the continuous case (again, using mean results across the 10000 synthetic samples based on 30000 observations each). A complementary trend to the one affecting the percent of observations in each CSS class is present.

For the continuous case, as the σ_λ parameter increases from 0.15 to 2 – and as the upper tail of the distribution becomes increasingly heavy – the class of poor performers accounts for an ever decreasing share of the total (falling from about 47% when $\sigma_\lambda = 0.15$ to about 16% when $\sigma_\lambda = 2$ and 1% when $\sigma_\lambda = 5$) while the class of outstanding performers accounts for an ever increasing share (rising from about 10% when $\sigma_\lambda = 0.15$ to about 34% when $\sigma_\lambda = 2$ and 70% when $\sigma_\lambda = 5$). The intermediate classes of fair and remarkable observations are substantially more stable than the extreme ones: the first accounts for about 30% of the total across most lognormal parametrizations, while the second accounts for about 20% of the total. For the specific case of $\sigma_\lambda = 1.14$ the first CSS class accounts for about 28% of the total, the second for 31%, the third for 20% and the fourth for about 21%. Note that from Table 2 it is possible to see that for σ_λ values close to 1 the results deviate from the empirical findings cited in the previous paragraph: class I accounts for a greater share of the total than would be expected (roughly 30% instead of 23%) whereas classes III and IV, taken together, account for a lesser share than would be expected (roughly 38% instead of 44%). Note also that for all the parametrizations considered the CV associated with the per cent share of

Table 2
Percentage of total accounted for by each CSS class and corresponding coefficient of variation across selected parametrizations of continuous lognormal distribution.

		Poor	Fair	Remarkable	Outstanding
$\sigma_\lambda = 0.50$	%	40.13	30.76	15.95	13.16
	CV	0.00	0.01	0.01	0.01
$\sigma_\lambda = 0.55$	%	39.17	30.88	16.27	13.69
	CV	0.00	0.01	0.01	0.01
$\sigma_\lambda = 0.60$	%	38.21	30.97	16.59	14.22
	CV	0.00	0.01	0.01	0.01
$\sigma_\lambda = 0.65$	%	37.26	31.06	16.90	14.78
	CV	0.00	0.01	0.01	0.01
$\sigma_\lambda = 0.70$	%	36.32	31.13	17.21	15.34
	CV	0.00	0.01	0.01	0.01
$\sigma_\lambda = 0.75$	%	35.38	31.18	17.51	15.92
	CV	0.00	0.01	0.01	0.01
$\sigma_\lambda = 0.80$	%	34.46	31.22	17.81	16.51
	CV	0.00	0.01	0.01	0.01
$\sigma_\lambda = 0.85$	%	33.54	31.24	18.11	17.11
	CV	0.00	0.01	0.01	0.01
$\sigma_\lambda = 0.90$	%	32.64	31.24	18.40	17.72
	CV	0.00	0.01	0.01	0.01
$\sigma_\lambda = 0.95$	%	31.74	31.23	18.68	18.35
	CV	0.01	0.01	0.01	0.01
$\sigma_\lambda = 1$	%	30.85	31.20	18.95	18.99
	CV	0.01	0.01	0.01	0.01
$\sigma_\lambda = 1.05$	%	29.98	31.15	19.22	19.64
	CV	0.01	0.01	0.01	0.01
$\sigma_\lambda = 1.10$	%	29.12	31.10	19.48	20.30
	CV	0.01	0.01	0.01	0.01
$\sigma_\lambda = 1.14$	%	28.44	31.04	19.68	20.84
	CV	0.01	0.01	0.01	0.01
$\sigma_\lambda = 1.20$	%	27.42	30.93	19.99	21.66
	CV	0.01	0.01	0.01	0.01
$\sigma_\lambda = 1.25$	%	26.60	30.82	20.23	22.35
	CV	0.01	0.01	0.01	0.01
$\sigma_\lambda = 1.30$	%	25.79	30.69	20.47	23.06
	CV	0.01	0.01	0.01	0.01
$\sigma_\lambda = 1.35$	%	24.98	30.56	20.68	23.77
	CV	0.01	0.01	0.02	0.01
$\sigma_\lambda = 1.40$	%	24.20	30.41	20.89	24.50
	CV	0.01	0.01	0.02	0.01
$\sigma_\lambda = 1.45$	%	23.42	30.24	21.10	25.24
	CV	0.01	0.01	0.02	0.01
$\sigma_\lambda = 1.50$	%	22.66	30.06	21.29	25.99
	CV	0.01	0.01	0.02	0.01

observations accounted for by each CSS class is very low (at most 1%, even for the more extreme scenarios) meaning there is a near complete homogeneity of these quantities across the 10000 applications of the CSS algorithm.

For the computer simulations involving discretized lognormal data the same general trends hold but, once again, with the qualification that the overall results reflect the noise induced by the discretization process, especially in the lower end of σ_λ values.

4.3. Discussion

Returning to the first research question formulated in Section 2.3 it seems that a pattern very close to the 70–21–9% rule does actually arise under the strong lognormal universality claim. With regard to the second question, which starts from the premise of the weak version of the universality claim, it seems that the CSS pattern becomes manifest when the σ_λ

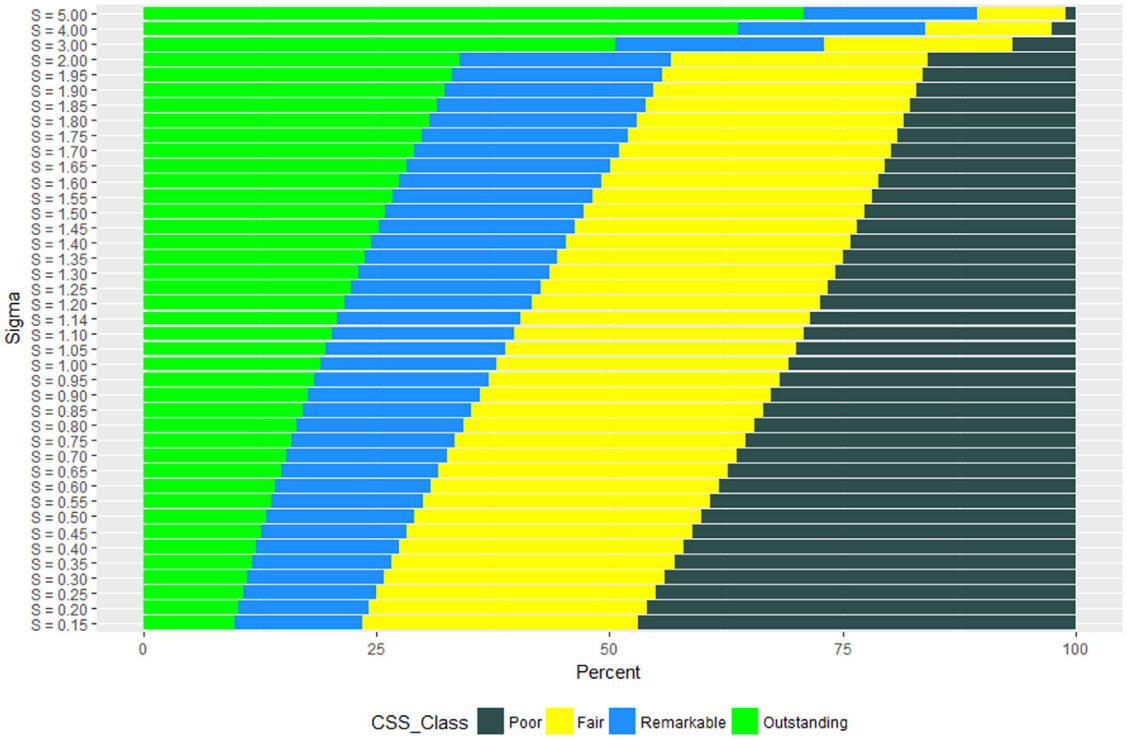

Fig. 3. Proportion of total accounted for by each CSS class under various lognormal parametrizations (continuous data).

parameter takes on values close to 1. Specifically, it seems that the CSS pattern or patterns very close to the most typical one appear when the σ_λ parameter takes values between about 0.8 (this leads to a 65.6–23–7.6–3.8% configuration for continuous data) and 1.3 (74.2–18.6–5.1–2.1% configuration); σ_λ values outside the 0.8–1.3 interval coincide with more atypical CSS class configurations. One should bear in mind in this context that while the 70–21–6–3% pattern is the most representative outcome empirically obtained through application of the CSS method, it is not a strictly ubiquitous one. In fact, the studies mentioned in Section 2.1, beginning with Schubert et al. (1987), have found some scientific fields where C_p can be as low as 60% or as high as 80%, meaning that for those fields C_f, C_r and C_o also deviate from their typical values. However, as the previous sections have shown, it is entirely possible to find parametrizations of the lognormal distribution for which the application of the CSS method also yields atypical class configurations.

A direct implication of the current work is that in general it could be argued that virtually all field-level empirical results obtained in previous studies with the CSS method – not only the typical 70–21–6–3% pattern but atypical class configurations as well – could stem from particular realizations of a lognormal distribution. However, none of the studies making use of the CSS method have also tested for conformity of the empirical citation counts to a lognormal distribution and, conversely, no study demonstrating the fact that citation counts can accurately be modelled within the lognormal framework has also analyzed the empirical data with the CSS method and presented the resulting patterns. This is an important knowledge gap that should be addressed within future studies.

A critical aspect regarding the results obtained for the two research questions addressed in the present paper is the connection between these results and the lognormal universality claims to which they relate. In essence, the results outlined in the current work cannot be taken as confirmation of either the strong or of the weak universality claim. In fact, the opposite may be argued. First, the fact that the typical CSS pattern arises under an entire range (0.8 through 1.3) of σ_λ parameter values instead of only for the rigid 1.14 specification is a clear basis to reject the strong universality claim. Second, although the results of the present work do not directly refute it, given the problem of indiscernibility of several heavy-tailed distributions it is not possible to support even the weak universality claim. Several studies dealing with statistical distributions demonstrate the fact that the lognormal distribution which is at the heart of the universality claims discussed in this study is often indistinguishable from the power law distribution: Mitzenmacher (2003, p. 227) argues that "very similar basic generative models can lead to either power law or lognormal distributions, depending on seemingly trivial variations" and this often leads to debates about which model is more accurate in many fields of science; in a study focusing on citation counts registered in Scopus Brzezinski (2015) notes that when the power law is a plausible model it is indiscernible from alternatives such as the lognormal, Yule and power law with exponential cut-off distributions; Thelwall and Wilson (2014, p. 837) also find that for articles published in one subject and one year (within the limits of a ten-year citation window) "the hooked power law and lognormal distributions are approximately equivalent in their fit to citation data". Consequently, while

citation counts are generally well modelled by heavy-tailed distributions, these need not necessarily be of the lognormal functional form (which is what the weak version of universality would suggest), nor must they necessarily be characterized by a single, rigid parametrization (as the strong version of universality would have us believe).

Given the problem of indiscernibility of several heavy-tailed distributions the truly remarkable scientometric fact may not be that the application of CSS consistently yields the 70–21–6–3% pattern, but rather that citation distributions in general are severely skewed and amenable to modeling within the lognormal framework or with the closely related power law distributions. Historically however, this latter issue has already been addressed. Its underlying explanation can be traced to a phenomenon which has been discussed in the scientometric literature for decades under the guise of several avatars: "Matthew effect" (Merton, 1968), "cumulative advantage" (Price, 1976), "preferential attachment" (Barabási & Albert, 1999). All of these are different expressions for a principle better known as "success breeds success" which, in the case of citation counts, simply means that papers which already have a high number of citations will gain even more citations, while those with few (if any) citations will tend to retain this status. It has been shown (Redner, 2005) that a linear preferential attachment process can yield citation counts that are best accounted for by a lognormal model and much earlier work (Egghe & Rao, 1992) has explained in a similar vein that lognormal distributions are in fact a logical consequence to be expected whenever processes obeying a law of proportionate effect are in operation. Thus, while the CSS pattern may be somewhat new, the explanation for its underlying cause – i.e. for the phenomenon of skewed citation counts – seems to be relatively long-standing.

5. Summary and concluding remarks

Previous empirical works leveraging the method of CSS at high levels of aggregation have documented a remarkable pattern in citation analysis: evaluating in terms of the four CSS hierarchical classes of citedness the scholarly output produced across most fields of science leads to a recurring 70–21–6–3% empirical pattern which indicates that despite their many inherent differences most scientific fields are fundamentally similar in shape. This article has investigated whether or not the CSS pattern can arise when citation counts are assumed to follow a lognormal distribution. This specific distribution was singled out for analysis because, on one hand, it is the most successful distribution used to date for modelling citation data and, on the other hand, it is at the heart of several contentious universality claims in scientometrics. The results of the present article indicate that whenever citation counts are consistent with a lognormal model having a standard deviation parameter close to a value of 1we can expect the application of the CSS method to produce the approximate 70–21–9% pattern. If, however, citation counts are better captured by lognormal distributions with standard deviation parameters further away from 1 we can expect the application of the CSS method to produce class configurations that diverge from the typical pattern but even these more atypical configurations could still be representative for some scientific fields.

While it seems that in essence the CSS pattern is indeed explainable in the framework of the lognormal distribution of citation counts across the sciences, this cannot in itself be taken as evidence in support of either the strong or of the weak universality claim. An important question which should be addressed by future work is whether or not the CSS pattern can also emerge from other distributions that could be used to model the full spectrum of citation counts. Some of the alternative distributions mentioned in Section 2.2 are certainly worth a detailed comparative investigation.

As a final note, the answer to why virtually all fields of science are shown by CSS to be fundamentally similar may be a simple one: regardless of their technical or conceptual specificities, in all scientific fields the same cumulative advantage processes are at work at the level of scholars and their scientific outputs. These cumulative advantage processes lead to skewed productivity and skewed citation counts which the lognormal and perhaps other similarly heavy-tailed distributions capture and which the CSS method simply translates to the specific four-point scale of poor–fair–remarkable–outstanding performance. Following this account it is not the CSS pattern that is ultimately remarkable, but rather the skewness of science that underpins it. The prominent skewness of science – manifest not in one, but in a plurality of functional forms – is in fact the only characteristic of scientometric distributions that may empirically be reasoned to be universal.

Author contributions

Gabriel-Alexandru Vîiu: Conceived and designed the analysis, Collected the data, Contributed data or analysis tools, Performed the analysis, Wrote the paper.

Acknowledgements

The author wishes to acknowledge the importance of several comments made by an anonymous reviewer on a separate article manuscript submitted to the *Journal of Informetrics* (Vîiu, 2017) on a topic related to that of the present work: the reviewer's insistence on the remarkable nature of the pattern detected with the aid of CSS prompted the investigation that ultimately led to the present article. The author also gratefully acknowledges the contribution of the two reviewers of the present article who made valuable suggestions that helped to improve the manuscript. This research was funded by the Research Institute of the University of Bucharest.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.joi.2018.02.002.

References

Abramo, G., D'Angelo, C. A., & Soldatenkova, A. (2017). An investigation on the skewness patterns and fractal nature of research productivity distributions at field and discipline level. *Journal of Informetrics*, *11*(1), 324–335. http://dx.doi.org/10.1016/j.joi.2017.02.001

Albarrán, P., & Ruiz-Castillo, J. (2011). References made and citations received by scientific articles. *Journal of the American Society for Information Science and Technology*, *62*(1), 40–49. http://dx.doi.org/10.1002/asi

Albarrán, P., Crespo, J. A., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, *88*(2), 385–397. http://dx.doi.org/10.1007/s11192-011-0407-9

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. http://dx.doi.org/10.1126/science.286.5439.509

Bonaccorsi, A., Daraio, C., Fantoni, S., Folli, V., Leonetti, M., & Ruocco, G. (2017). Do social sciences and humanities behave like life and hard sciences? *Scientometrics*, *112*(1), 607–653. http://dx.doi.org/10.1007/s11192-017-2384-0

Bornmann, L., & Daniel, H.-D. (2009). Universality of citation distributions–A validation of Radicchi et al.'s relative indicator c f = c/c 0 at the micro level using data from chemistry. *Journal of the American Society for Information Science and Technology*, *60*(8), 1664–1670. http://dx.doi.org/10.1002/asi.21076

Brzezinski, M. (2015). Power laws in citation distributions: Evidence from Scopus. *Scientometrics*, *103*(1), 213–228. http://dx.doi.org/10.1007/s11192-014-1524-z

Burrell, Q. L. (2002). Modelling citation age data: Simple graphical methods from reliability theory. *Scientometrics*, *55*(2), 273–285. http://dx.doi.org/10.1023/A:1019671808921

Castellano, C., & Radicchi, F. (2009). On the fairness of using relative indicators for comparing citation performance in different disciplines. *Archivum Immunologiae et Therapiae Experimentalis*, *57*(2), 85–90. http://dx.doi.org/10.1007/s00005-009-0014-0

Chatterjee, A., Ghosh, A., & Chakrabarti, B. K. (2016). Universality of citation distributions for academic institutions and journals. *Public Library of Science*, 1–11. http://dx.doi.org/10.1371/journal.pone.0146762

Costas, R., Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2016). Currencies of Science: Discussing disciplinary exchange rates for citations and Mendeley readership. In I. Ràfols, J. Molas-Gallart, E. Castro-Martínez, & R. Woolley (Eds.), *Proceedings of the 21st international conference on science and technology indicators* (pp. 1173–1182). Valencia: Universitat Politècnica de València. http://dx.doi.org/10.4995/STI2016.2016.4543

Egghe, L., & Rao, I. K. R. (1992). Citation age data and the obsolescence function: Fits and explanations. *Information Processing and Management*, *28*(2), 201–217. http://dx.doi.org/10.1016/0306-4573(92)90046-3

Egghe, L., & Rao, R. (2002). Theory and experimentation on the most-recent-reference distribution. *Scientometrics*, *53*(3), 371–387. http://dx.doi.org/10.1023/A:1014825113328

Evans, T. S., Hopkins, N., & Kaube, B. S. (2012). Universality of performance indicators based on citation and reference counts. *Scientometrics*, *93*(2), 473–495. http://dx.doi.org/10.1007/s11192-012-0694-9

Glänzel, W., & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. *Journal of Information Science*, *14*(2), 123–127. http://dx.doi.org/10.1177/016555158801400208

Glänzel, W., Thijs, B., & Debackere, K. (2014). The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparison and institutional research assessment. *Scientometrics*, *101*(2), 939–952. http://dx.doi.org/10.1007/s11192-014-1247-1

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. http://dx.doi.org/10.1016/j.joi.2006.10.001

Glänzel, W. (2009). The multi-dimensionality of journal impact. *Scientometrics*, *78*(2), 355–374. http://dx.doi.org/10.1007/s11192-008-2166-9

Glänzel, W. (2010). The role of the h-index and the characteristic scores and scales in testing the tail properties of scientometric distributions. *Scientometrics*, *83*(3), 697–709. http://dx.doi.org/10.1007/s11192-009-0124-9

Glänzel, W. (2011). The application of characteristic scores and scales to the evaluation and ranking of scientific journals. *Journal of Information Science*, *37*(1), 40–48. http://dx.doi.org/10.1177/0165551510392316

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). The Leiden Manifesto for research metrics. Use these ten principles to guide research evaluation. . .. *Nature*, *520*(7548), 9–11. http://dx.doi.org/10.1038/520429a

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

Kurtz, M. J., & Henneken, E. A. (2017). Measuring metrics – A 40-year longitudinal cross-validation of citations, downloads, and peer review in astrophysics. *Journal of the Association for Information Science and Technology*, *68*(3), 695–708. http://dx.doi.org/10.1002/asi.23689

Li, Y., Radicchi, F., Castellano, C., & Ruiz-Castillo, J. (2013). Quantitative evaluation of alternative field normalization procedures. *Journal of Informetrics*, *7*(3), 746–755. http://dx.doi.org/10.1016/j.joi.2013.06.001

Limpert, E., Stahel, W. A., & Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience*, *51*(5), 341. http://dx.doi.org/10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

Low, W. J., Wilson, P., & Thelwall, M. (2016a). Stopped sum models and proposed variants for citation data. *Scientometrics*, *107*(2), 369–384. http://dx.doi.org/10.1007/s11192-016-1847-z

Matricciani, E. (1991). The probability distribution of the age of references in engineering papers. *IEEE Transactions on Professional Communication*, *34*(1), 7–12. http://dx.doi.org/10.1109/47.68421

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63. http://dx.doi.org/10.1126/science.159.3810.56

Mitzenmacher, M. (2003). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, *1*(2), 226–251. http://dx.doi.org/10.1080/15427951.2004.10129088

Moed, H. F., & Halevi, G. (2015). Multidimensional assessment of scholarly research impact. *Journal of the Association for Information Science and Technology*, *66*(10), 1988–2002. http://dx.doi.org/10.1002/asi.23314

Moreira, J. A. G., Zeng, X. H. T., & Amaral, L. A. N. (2015). The distribution of the asymptotic number of citations to sets of publications by a researcher or from an academic department are consistent with a discrete lognormal model. *Public Library of Science*, *10*(11), e0143108. http://dx.doi.org/10.1371/journal.pone.0143108

Morris, S. A. (2005). Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology*, *56*(12), 1250–1273. http://dx.doi.org/10.1002/asi.20208

Perc, M. (2010). Zipf's law and log-normal distributions in measures of scientific output across fields and institutions: 40 Years of Slovenia's research as an example. *Journal of Informetrics*, *4*(3), 358–364. http://dx.doi.org/10.1016/j.joi.2010.03.001

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2014). Within- and between-department variability in individual productivity: The case of economics. *Scientometrics*, *102*(2), 1497–1520. http://dx.doi.org/10.1007/s11192-014-1449-6

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2016). University citation distributions. *Journal of the Association for Information Science and Technology*, *67*(11), 2790–2804. http://dx.doi.org/10.1002/asi.23619
Price, D. J. de S. (1976). A general theory of bibiiometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306. http://dx.doi.org/10.1002/asi.4630270505
R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/
Radicchi, F., & Castellano, C. (2012). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *Public Library of Science*, *7*(3), 1–9. http://dx.doi.org/10.1371/journal.pone.0033833
Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Towards an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272. http://dx.doi.org/10.1073/pnas.0806977105
Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today*, *58*(6), 49–54. http://dx.doi.org/10.1063/1.1996475
Ruiz-Castillo, J., & Costas, R. (2014). The skewness of scientific productivity. *Journal of Informetrics*, *8*(4), 917–934. http://dx.doi.org/10.1016/j.joi.2014.09.006
Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, *9*(1), 102–117. http://dx.doi.org/10.1016/j.joi.2014.11.010
Schubert, A., Glänzel, W., & Braun, T. (1987). Subject field characteristic citation scores and scales for assessing research performance. *Scientometrics*, *12*(5), 267–291. http://dx.doi.org/10.1007/BF02016664
Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, *43*(9), 628–638. http://dx.doi.org/10.1002/(SICI)1097-4571(199210)43:9<628:AID-ASI5>3.0.CO;2-0
Sichel, H. S. (1992). Anatomy of the generalized inverse Gaussian-poisson distribution with special applications to bibliometric studies. *Information Processing & Management*, *28*(1), 5–17. http://dx.doi.org/10.1016/0306-4573(92)90088-H
Stringer, M. J., Sales-Pardo, M., & Nunes Amaral, L. A. (2010). Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. *Journal of the American Society for Information Science and Technology*, *61*(7), 1377–1385. http://dx.doi.org/10.1002/asi.21335
Thelwall, M., & Wilson, P. (2014). Distributions for cited articles from individual subjects and years. *Journal of Informetrics*, *8*(4), 824–839. http://dx.doi.org/10.1016/j.joi.2014.08.001
Thelwall, M., & Wilson, P. (2016). Mendeley readership altmetrics for medical articles: An analysis of 45 fields. *Journal of the Association for Information Science and Technology*, *67*(8), 1962–1972. http://dx.doi.org/10.1002/asi.23501
Thelwall, M. (2016a). Are the discretised lognormal and hooked power law distributions plausible for citation data? *Journal of Informetrics*, *10*(2), 454–470. http://dx.doi.org/10.1016/j.joi.2016.03.001
Thelwall, M. (2016b). Citation count distributions for large monodisciplinary journals. *Journal of Informetrics*, *10*(3), 863–874. http://dx.doi.org/10.1016/j.joi.2016.07.006
Thelwall, M. (2016c). The discretised lognormal and hooked power law distributions for complete citation data: Best options for modelling and regression. *Journal of Informetrics*, *10*(2), 336–346. http://dx.doi.org/10.1016/j.joi.2015.12.007
Thelwall, M. (2016d). The precision of the arithmetic mean, geometric mean and percentiles for citation data: An experimental simulation modelling approach. *Journal of Informetrics*, *10*(1), 110–123. http://dx.doi.org/10.1016/j.joi.2015.12.001
Vîiu, G.-A. (2017). Disaggregated research evaluation through median-based characteristic scores and scales: A comparison with the mean-based approach. *Journal of Informetrics*, *11*(3), 748–765. http://dx.doi.org/10.1016/j.joi.2017.04.003
Van Raan, A. F. J. (2001). Competition amongst scientists for publication status: Toward a model of scientific publication and citation distributions. *Scientometrics*, *51*(1), 347–357. http://dx.doi.org/10.1023/A:1010501820393
Vieira, E. S., & Gomes, J. A. N. F. (2010). Citations to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics*, *4*(1), 1–13. http://dx.doi.org/10.1016/j.joi.2009.06.002
Wallace, M. L., Larivière, V., & Gingras, Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics*, *3*(4), 296–303. http://dx.doi.org/10.1016/j.joi.2009.03.010
Waltman, L., van Eck, N. J., & van Raan, A. F. J. (2012). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, *63*(1), 72–77. http://dx.doi.org/10.1002/asi.21671
Wu, J. (2013). Investigating the universal distributions of normalized indicators and developing field-independent index. *Journal of Informetrics*, *7*(1), 63–71. http://dx.doi.org/10.1016/j.joi.2012.08.007

Gabriel-Alexandru Vîiu is a postdoctoral researcher active within the Social Sciences Division of the Research Institute of the University of Bucharest where his work focuses on scientometrics and policy instruments designed for the purpose of research evaluation. Before beginning his activity at the Research Institute of the University of Bucharest in December 2016 Gabriel completed a PhD in Political Science. The PhD thesis was completed under the supervision of professor Adrian Miroiu between 2011–2014 within the National School of Political and Administrative Studies in Bucharest (NSPAS) and focused on the policies of research evaluation in higher education, on the mechanisms of funding and quality assurance and on the theoretical problems of ranking universities and their associated departments. Prior to the doctoral cycle Gabriel graduated a master program in Political Theory and Analysis and a bachelor program in political science, both within the NSPAS. In addition to being engaged in teaching activities (introductory statistics, decision analysis, history of political and social thought) throughout the past years Gabriel has contributed to several policy studies for the Executive Agency for Higher Education, Research, Development and Innovation Funding, he has worked to implement several institutional development projects within the NSPAS and was also involved in a research project concerned with voting rules. Previous publications: Vîiu, G.-A. (2017). Disaggregated research evaluation through median-based characteristic scores and scales : a comparison with the mean-based approach. *Journal of Informetrics*, 11(3), 748–765. https://doi.org/10.1016/j.joi.2017.04.003, Vîiu, G.-A. (2016). A theoretical evaluation of Hirsch-type bibliometric indicators confronted with extreme self-citation. *Journal of Informetrics*, 10(2), 552–566. https://doi.org/10.1016/j.joi.2016.04.010, Vîiu, G.-A., Păunescu, M., & Miroiu, A. (2016). Research-driven classification and ranking in higher education: an empirical appraisal of a Romanian policy experience. *Scientometrics*, 107(2), 785–805. https://doi.org/10.1007/s11192-016-1860-2, Vîiu, G.-A. (2015). Quality-related funding in Romanian higher education throughout 2003–2011: a global assessment. *Romanian Journal of Society and Politics*, 10(2), 26–59. Retrieved from http://rjsp.politice.ro/sites/default/files/pictures/2_viiu.pdf, Miroiu, A., Păunescu, M., & Vîiu, G.-A. (2015). Ranking Romanian academic departments in three fields of study using the g -index. *Quality in Higher Education*, 21(2), 189–212. https://doi.org/10.1080/13538322.2015.1051794, Vîiu, G.-A., & Miroiu, A. (2015). The quest for quality in higher education: Is there any place left for equity and access? In A. Curaj, L. Deca, E. Egron-Polak, & J. Salmi (Eds.), *Higher Education Reforms in Romania: Between the Bologna Process and National Challenges* (pp. 173–189). Springer. https://doi.org/10.1007/978-3-319-08054-3_9, Miroiu, A., & Vîiu, G.-A. (2013). Ierarhiile universitare şi efectul de inversare. *Revista de Politica Ştiinţei şi Scientometrie*, 2(4), 277–285. Retrieved from http://www.rpss.inoe.ro/articles/106/file, Vîiu, G.-A., & Miroiu, A. (2013). Evaluarea cercetării universitare din România. Abordări metodologice alternative. *Revista de Politica Ştiinţei şi Scientometrie*, 2(2), 89–107. Retrieved from http://rpss.inoe.ro/articles/74/file, Vîiu, G.-A., Vlăsceanu, M., & Miroiu, A. (2012). Ranking political science departments: the case of Romania. *Quality Assurance Review for Higher Education*, 4(2), 79–97. Retrieved from http://www.aracis.ro/fileadmin/ARACIS/Revista_QAR/Septembrie_2012/Articol_6.pdf.