Journal of Library and Information Studies 20:2 (December 2022) pp.75-93
https://doi.org/10.6182/jlis.202212_20(2).075

Drawbacks of Normalization by Percentile Ranks in Citation Impact Studies

Paul Donner[1]

Abstract

This paper discusses drawbacks of the percentile rank method for citation impact normalization which have hitherto been neglected in the bibliometrics literature. The transformation of citation counts to percentile ranks changes ratio scale data into ordinal scale data, for which the notions of the ratio between two values and of the magnitude of a difference between two values are not defined – a substantial loss of information. This distorts citation data particularly severely because the differences between citation counts adjacent in order in publication sets are greater for more highly cited publications and because highly cited publications are more scarce than non-highly cited ones. Moreover, arithmetic operations on ordinal scale data are not meaningful, which rules out arithmetic aggregations such as sums or averages for percentile rank data which are sometimes recommended in the literature. Distortion of citation data by aggregating percentile ranks for average impact indicators is demonstrated with several examples.

Keywords: Citation Normalization; Field Normalization; Percentile Ranks; Ordinal Data

1. Introduction

The Leiden Manifesto (LM) proposes ten guiding principles for responsible research evaluation (Hicks et al., 2015). Since its publication as a comment in *Nature* in 2015, the LM has received considerable attention in academia, evidenced by the fact that, seven years on, it has been cited around 2,000 times in Google Scholar and around 1,000 times in Dimensions. Motivated by "the pervasive misapplication of indicators to the evaluation of scientific performance", the authors "offer this distillation of best practice in metrics-based research assessment so that researchers can hold evaluators to account, and evaluators can hold their indicators to account" (p. 430). The LM does not advocate against metrics-based assessment, but for its responsible use in a supportive role to peer review.

With respect to citation analysis in particular, the LM, in principle 6, points out that different research fields have very different publication and citation practices, leading to typical publication and citation counts that cannot be compared directly across fields. A specialized research topic in bibliometrics has developed which is concerned with the issue of how to calculate bibliometric scores that take such field effects explicitly into account in order to obtain scores that can be compared across fields (Waltman & van Eck, 2019). This method is often called field normalization. The LM (principle 6) claims succinctly:

[1] German Centre for Higher Education Research and Science Studies (DZHW), Berlin, Germany
E-mail: donner@dzhw.eu

> Normalized indicators are required, and the most robust normalization method is based on percentiles: each paper is weighted on the basis of the percentile to which it belongs in the citation distribution of its field (the top 1%, 10% or 20%, for example) (Hicks et al, 2015, p. 430).

In support of this, the LM points to the case of a single paper's seemingly disproportionate influence on a university's ranking, which would not have occurred with percentile normalization. While it is clear that percentile normalization is robust to very high values, we can nonetheless ask, in line with principle 10, "Scrutinize indicators regularly and update them", whether percentile rank (Note 1) normalization, merely by virtue of its robustness, is the most appropriate method for field normalization in general. We scrutinize in this paper if percentile rank transformation is, or ever was, best practice for normalization of citation counts.

The standard approach to citation normalization involves reference sets, that is, sets of publications which are homogeneous with respect to discipline, age and type. While the LM mentions only normalization by field, normalization by publication year and document type is also often required. Usually, sorting of publications by discipline is accomplished by using a research classification system. However, there are citation normalization methods that construct reference sets specific to each individual paper without recourse to classifications and methods which are not based on reference sets (cf. Waltman & van Eck, 2019, section 4.3). Such approaches are not the topic of this contribution. Furthermore, we restrict the analyses to item-level normalization, as the most practical approach to normalization is to calculate normalized citation scores for each item first and only work with aggregates of these scores subsequently, as opposed to calculating normalized scores for publication sets as a whole.

Basic percentile-based citation impact indicators are widely adopted in bibliometrics. The use of percentile values for determining thresholds for highly cited papers indicators (excellence indicators) is common. For example, to calculate the share of papers of units among the 10% most highly cited papers in a scientific discipline, the 90th percentile of the citation distribution is used as a threshold value. Methods for such calculations are available in the commercial research evaluation platforms SciVal and InCites. The proportion of highly cited papers is also a central indicator in CWTS's Leiden Ranking of research institutions (there called PP(top *x*%) with as of the 2020 edition). Percentile rank classes have also been used in a number of studies. For example, the six biennial US National Science Board's *Science and Engineering Indicators* reports from 2008 to 2018 have presented percentile rank class distributions for the US and non-US regions and states for six classes (99th, 95th, 90th, 75th, 50th, and <50th percentile). In these reports, class membership figures were never arithmetically aggregated. The use of arithmetically aggregated percentile rank score indicators is relatively limited. One example is the Integrated Impact Indicator (I3) (Leydesdorff & Bornmann, 2011; Leydesdorff et al., 2019). I3-type indicators are calculated as sums of weighted percentile rank scores. The growing application of percentile-based metrics warrants closer

consideration of which specific kinds of their use are appropriate and which are not.

The remainder of this paper is organized as follows. We first give a brief overview of the topic of citation count normalization, followed by a discussion of the percentile rank approach to normalization in which we also state the main drawbacks of this method which have so far not received adequate consideration. We proceed by discussing some alternative normalization methods. Next, we illustrate the inadequacy of percentile ranks for citation count normalization by, first, a simple example data set, second, a simulation study of units active in a single homogeneous field and, third, a simulation study of units active in two heterogeneous fields. We then consider if the arguments brought forward against percentile ranks also invalidate the application of highly cited rate indicators. We conclude the paper by a discussion of the arguments and results.

2. Citation Counts and Their Normalization

The necessity of citation count normalization arises because different groups of publications exhibit different citation characteristics unrelated to the concept of their scientific impact (for an introduction, see Ioannidis et al., 2016). Usually, the research field, publication year, and document type are considered as factors whose distorting influences need to be corrected for, because otherwise some publications or units would be given an unfair advantage. Citation distributions vary systematically according to these properties, but citation distributions of homogeneous publication sets are assumed to be free of such distorting influences. Hence for homogeneous sets of publications with respect to these properties, no normalization is required, as one is comparing inherently comparable objects, as far as citation analysis is concerned.

If the citation distributions were invariant across fields, field normalization would be unnecessary. If the average numbers of citations per year for publications would not follow a curve that increases from the year of publication for some years, reaches a peak, and then decreases slowly, then a normalization by the specific publication year (or a year stratification) would not be required. A division by the number of years since publication would suffice to produce publication year-normalized citation counts. Within a homogeneous set of publications, one can directly use the unmodified citation counts as a scientific research impact indicator on the level of individual publications. In such a scenario then, we can state that a publication A, which is cited x times as often as another publication B, has a citation impact x times that of B. For example, A is cited 20 times and B is cited 10 times in the same period. Publication A then has two times the citation impact of B. Such a statement is possible because citation data are on a ratio scale (Stevens, 1946). That means that the scale of citations has a zero point and an intrinsic unit magnitude such that ratios between quantities can be formed.

3. Properties of Citation Count Distributions

It is a well-known empirical regularity that citation distributions are highly skewed, more specifically heavy-tailed on the right (positive skewness) (Seglen, 1992). Citation counts do not

follow a Gaussian distribution nor are citation distributions symmetric around their means. This means that publication sets contain a relatively small number of very highly cited publications – high compared to values of citation counts for most of the publications in the set. The concentration of papers over citation count values is typically high for low values and low for high values. Because of these properties, the arithmetic mean is not a very informative descriptive summary statistic of citation distributions and the standard deviation is also not informative of the spread of the distribution. It is a matter of debate which statistical distribution best describes citation counts. In practice, however, reasonably good fits are obtained for power law and log-normal distributions, among others (Thelwall, 2016).

4. Percentile Ranks

Percentile ranks for citation counts are appealing because they indicate the performance level of a publication relative to the chosen reference group of papers by stating the relative share of publications that are less or equally often cited. For example, a publication with a percentile rank citation score of 80 is cited more or as often as 80% of the publications in the reference set. This is a useful way of expressing the relative position of individual publications in their reference groups according to citation counts, which untransformed citation counts do not afford. Furthermore, percentile ranks are also appealing because applying percentile rank transformation to sets of values of very different distributions of raw data normalizes all values to one simple common scale. While there is no single agreed upon calculation method for percentiles, this is of no concern for the present issue. The caveats presented next apply to any and all calculation variants.

4.1 Drawbacks of percentile ranks

4.1.1 Information loss and penalizing high performance

By design and by definition, percentile ranks (PRs) are a transformation of data to an ordinal scale of measurement and therefore incur information loss when applied to ratio or interval scale data. All that can be known about two different percentile rank values is that one is greater, but not by how much. By the transformation from ratio scale to ordinal scale, the information on the magnitudes of differences in the data is irrevocably lost. If there are two publications A and B with citation count percentile ranks 10 and 20, we can not say at all how much more often B has been cited. We certainly cannot say B was cited twice as often as A, although the numbers might suggest so. What we can say is that B has been cited comparatively more often with respect to the proportions of less cited papers in the reference set(s). We cannot say by how much relative to the typical publications in the sets B has been cited more often than A, which is something that is possible with other normalization approaches. Rounding of percentile rank values to integers, as is sometimes advocated (Leydesdorff et al., 2011, p. 1372), exacerbates the information loss. Because of the highly skewed distributions involved, in situations with many publications in a reference set and with high citation levels, this can lead to cases in which publications with high but very different citation values will be assigned the same rounded percentile score, obscuring possibly important differences.

If one is given three different percentile rank values for papers' citation counts, x < y < z, what we can know is their order and that the original scale difference between x and z must be greater than the difference between x and y and that between y and z. But we can not know if the difference on the original scale between x and y is greater or lesser than that between y and z. This holds for any difference between adjacent percentile rank values and in consequence any different values. This poses a problem for citation distributions in particular, as the difference between ordered adjacent values usually becomes greater as one moves towards the high end of the citation distribution. Thus, the higher the real citation counts are, the more severe is the compression imposed by PR transformation, i.e., it is progressive. Needless to say, the upper end of the distribution is where the really important papers are to be found.

The information of whether the difference on the original scale between x and z was huge or minute is lost through percentile rank transformation. The notion of the magnitude of difference between values does not exist on the ordinal scale (Agresti, 2006; Stevens, 1946). The difference on the original scale between 0 and 1 citation on the low end, and between 500 and 5000 citations on the high end of an empirical citation distribution can become one point "difference" or less on the percentile rank scale if the two values in the pairs were adjacent in the order of the values of the empirical data (see Zhou and Zhong (2012) and D'Agostino et al. (2017) for similar arguments).

Because of the positive skewness of citation distributions and the reduction to an ordinal scale by percentile rank calculation, the percentile rank approach to normalization penalizes higher performing items or units by compressing the values of citation counts. This is not a mere theoretical point because percentile rank values are being used as if the differences between such values were meaningful.

When making comparisons of citation impact of publications from different reference sets, the information loss from normalization by PRs precludes answering relevant and important evaluative questions. Since PRs are bounded from above at a value of 100 one cannot state which among the publications with the highest PRs values of different reference sets has had the most impact relative to its set as they must all have the same maximum PR value. Even for two publication sets of the same document type and the same scientific discipline which differ only in that the first set covers some publication year and the second set covers the following year, the two respective most cited publications will both have an equal PR value of 100 while they might have actual citation counts of 50 and 500.

To sum up, the transformation of citation count values into percentile ranks discards crucial information. It obscures how much better or worse items or units really are, compared to others, and this effect is stronger for higher citation values.

4.1.2 Percentile ranks can not meaningfully be aggregated

According to the theory of measurement scales, comparisons of two values of ordinal data are restricted to tests of equality and tests of inequality, i.e., the relations *greater than* and *less than* (Stevens, 1946). Arithmetic operations are not meaningfully defined for the ordinal scale. In the context of citation counts and their

normalization, this has been pointed out earlier by Zhang et al. (2015):

> it should be noted that [the percentile rank method] is a nonlinear transformation which maps the theoretically unbounded citation range [0+ ∞) to the bounded range (0, 1] for percentile ranks. Percentile rank is an unequally spaced measurement and it's inappropriate to calculate the sum or average of percentile ranks […]. So it may be improper to calculate the normalized citation performance at the aggregate level based on summing or averaging percentile ranks of individual publications (Zhang et al., 2015, p. 590).

These concerns have not found any resonance so far. Elaborating on their argument and returning to the domain of percentile ranks in the interval 1 to 100, already by definition the summation of percentile ranks does not work, as for example PR 90 + PR 90 cannot equal PR 180 as percentile ranks only go as high as 100. The standard 26-letter alphabet has a fixed conventional order and is therefore ordinal data, we can associate each letter with a rank number from 1 for A to 26 for Z. Yet it would be clearly nonsense to claim that, since 5 + 6 = 11, therefore E + F = K. And while the arithmetic mean of the sequence of whole numbers from 1 to 26 is 13.5 that does not mean it is valid to conclude that the average letter of the alphabet is halfway between M and N. Numerical values (rank numbers and PRs) associated with ordinal data are just a useful auxiliary tool to help easily keep track of the correct order of the real data, they are not an essential part of any ordinal data and should not be confused for the ordinal data.

Extending their argument, we remark that just as for the results of a sports competition, stating that rank 3 minus rank 1 equals rank 2 is not meaningful, so it is in general with ordinal data. Rank 1 plus rank 1 does not equal rank 2, and neither does percentile rank 1 plus percentile rank 1 equal percentile rank 2. No arithmetic operations can meaningfully be applied to ordinal scale data including aggregations such as sums and averages. Nevertheless, just that has been recommended (Bornmann & Williams, 2020, p. 1471; Leydesdorff et al., 2019, p. 1676; Leydesdorff et al., 2011, p. 1373; Mcallister et al., 1983, p. 208; Mutz & Daniel, 2012).

Proponents of the percentile rank approach have claimed that transformation of citations to percentile ranks creates data on the interval scale, while only the grouping into percentile rank classes and use of weights for class member scores creates ordinal scale data (Leydesdorff et al., 2011, p. 1373). This is not correct. By transformation from raw citations to percentile ranks, only the order of values on the raw scale is preserved, but not the differences between values, hence the result is on the ordinal scale. The grouping into classes and weighting is a further information-lossy transformation but the resulting data is also on the ordinal scale. Treating the resulting values as if they were numerical scores does not change that.

From the above it follows that it is not objectionable to construct and use percentiles and percentile ranks as such (limiting comparisons to equality and inequality) and form classes and compare class sizes across units. Objectionable use starts precisely with performing arithmetic on percentile rank values or class weights.

5. Alternatives

There are a number of well-established alternatives to the use of percentile normalization, Waltman and van Eck (2019) gives an overview. In the classic method of normalization an expected value for the citation count of the reference publication set is calculated, for which the arithmetic mean of the citation counts is used. The normalized citation score for an item in the set is calculated as the ratio of its observed citation count and the expected citation count. From early proposals (Schubert & Braun, 1986), based on journal average citations, incremental developments led to the revised mean normalized citation score (MNCS; Waltman et al., 2011). Especially the criticism of Lundberg (2007) about calculating aggregate observed and expected values first and then calculating the ratio, rather than calculating individual item-level scores first and then aggregating them to unit scores, was a crucial improvement. Prior to this change, the implicit weights of the individual papers in units' publication sets were not uniform, in particular, the weight depended on the expected citation score (Waltman et al., 2011, p. 39).

Such arithmetic mean based methods have been found to be quite sensitive to the presence of high citation counts (Aksnes & Sivertsen, 2004; Antonoyiannakis, 2020), stimulating the search for more robust alternatives. Waltman et al. (2012) justified phasing out the revised MNCS in favor of the share of highly cited papers because of this volatility. Specifically, the authors describe one particular case, which is also alluded to in the LM, thus:

> The MNCS indicator for University of Göttingen turns out to have been strongly influenced by a single extremely highly cited publication. This publication (Sheldrick, 2008) was published in January 2008 and had been cited over 16,000 times by the end of 2010. Without this single publication, the MNCS indicator for University of Göttingen would have been equal to 1.09 instead of 2.04, and University of Göttingen would have been ranked 219th instead of 2nd. Unlike the MNCS indicator, the $PP_{top\ 10\%}$ indicator is hardly influenced by a single very highly cited publication. This is because the $PP_{top\ 10\%}$ indicator only takes into account whether a publication belongs to the top 10% of its field or not. The indicator is insensitive to the exact number of citations of a publication (Waltman et al., 2012, p. 2425).

This position is not entirely convincing because one can just as well argue for what is almost the opposite view: It is not in the least undesirable that a single publication cited 16,000 times in three years should be reflected in a high citation impact indicator value. On the contrary, the fact that the $PP_{top\ 10\%}$ indicator hardly registers this exceptional paper could be seen as a weakness of that indicator. However, such a consideration depends on the type of citation impact performance indicator with respect to intended purpose under discussion. One can at least distinguish between indicators of typical (or average) performance, compound (or total) performance, and high performance (or excellence) (Note 2). The MNCS is an indicator of typical performance while $PP_{top\ 10\%}$ is one of high performance. The usefulness of an indicator of high citation impact that reduces a 16,000 citations paper to irrelevancy seems questionable

– there is no reason why high performance should not be highly concentrated. Given the purpose of an indicator of typical impact, the sensitivity of the MNCS to one very highly cited paper can be judged as undesirable. The replacement of MNCS by $PP_{top\ 10\%}$ in this case then is not just a change of one more volatile impact indicator for a more robust one but also a change of type of indicator from one of average impact to one of high performance.

More complex methods of citation impact normalization have been proposed, of which we just mention one example. For more, the reader is referred to Waltman (2016) and Waltman and van Eck (2019). Lundberg (2007) suggested *z*-score standardization of citation counts of individual papers using the log+1-transformed citation scores instead of raw citation counts, yielding the citation *z*-score $c_{fz[ln]}$. The author justifies the use of the log-transformation by the skewed nature of citation distributions. The advantage of the *z*-score is that, in addition to correcting for different means across reference sets, it also corrects for different standard deviations of citation distributions. Differences between scores of items in this scheme are always expressed in terms of standard deviations on the log-scale – a complete change of the frame of reference from the original data, which may be considered a drawback. For instance, for the two articles A and B from Lundberg (2007)'s Table 1, with original citation counts 12 and 5 and $c_{fz[ln]}$ of 2.2 and 1.2, it can be said that A's score is 1 log-scale standard deviation greater than B's. On the other hand, it makes no sense to say that article A has 1.9 times as much impact as B, as one is no longer reckoning with absolute magnitudes but with relative magnitudes because citation *z*-score is a transformation to the interval scale. The calculation of *z*-scores is also subject to information loss, but less severe than percentile rank calculation (interval scale, rather than ordinal).

Similar to percentile ranks, the citation *z*-score approach also compresses the upper end of the citation distribution, in effect concealing exceptional performance. Unlike the percentile rank approach, this compression is systematic, as taking the logarithm is a deterministic operation that can be reversed to recover the original values. Given the mean and standard deviation, *z*-score values could be turned back into the input citation counts. Percentile ranks can not be transformed back to the original values without knowledge of the entire particular citation distribution.

This compression (or robustness) property of the percentile rank and *z*-score methods suggests a question of a more fundamental nature: Should the citations of a publication be weighted based on the publication's total citation count? Percentile ranks, geometric mean of citations, and citation *z*-score implicitly answer in the affirmative. For instance, with the citation *z*-score method, the example papers A and B with 12 and 5 citations have received intermediate scores calculated as ln(c+1) of 2.6 and 1.8, respectively. One citation for A is worth $2.6 / 12 \approx 0.22$ score points while one citation for B is worth $1.8 / 5 \approx 0.36$ score points.

The specific methods differ and therefore the degree to which exceptional performance is levelled and penalized. In choosing between these methods one is hence confronted with the question of how much less one would like to value highly cited publications. On the other hand, the average of raw (unnormalized) citation counts and MNCS implicitly state that each citation within

a reference set has equal weight, regardless of its magnitude. Absent convincing arguments to the alternative, such a non-discriminative approach seems more intuitive.

6. A Simple Example

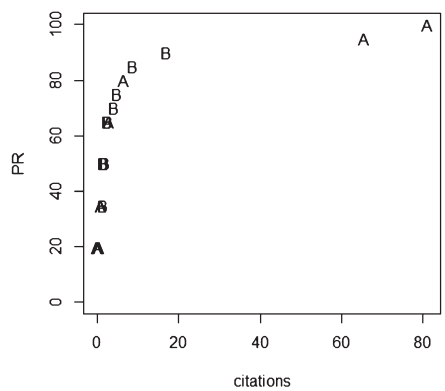
To make the observations discussed above more tangible we present a short analysis of specific data to show an example of the counterintuitive results caused by percentile rank transformation and the arithmetic aggregation of PR values. As a first illustration, let us consider a small example data set. It consists of two groups of papers, A and B, of 10 publications each. Together these 20 publications fashion the whole reference publication set and no cross-field normalization is involved. Citation counts and percentile ranks are given in Table 1. Figure 1 shows the example data in a scatterplot according to citations and percentile rank. Percentile ranks have been computed in the R statistical programming language with the ecdf() empirical cumulative distribution function in this and the following examples.

Given a specific set of values $X = \{x_1, x_2, ..., x_n\}$ and one value $t \in X$ the empirical cumulative distribution function gives the value $F_n(t) = |\{x_i \leq t\}| \ / \ n$. The value is multiplied by 100 to get percentile ranks in the interval (1, 100]. Thus, the formula used here for the calculation or a PR value is $PR_{(t)} = |\{x_i \leq t\}| \ / \ n \times 100$. As an example for a value in the data set of Table 1 the PR value for the paper with citation count 4 is then $PR_{(4)} = |\{0,0,0,0,1,1,1,2,2,2,3,3,3,4\}| \ / \ 20 \times 100 = 14 \ / \ 20 \times 100 = 70$. Other calculation procedures could be used and would lead to differences in the values of the results, but, despite much emphasis in the

Table 1. Simple Example Data Set

Set	Citations	Percentile rank
A	0	20
A	0	20
A	0	20
A	0	20
A	1	35
A	1	35
A	3	65
A	6	80
A	66	95
A	81	100
B	1	35
B	2	50
B	2	50
B	2	50
B	3	65
B	3	65
B	4	70
B	5	75
B	9	85
B	17	90

Figure 1. Scatterplot of Citations and Percentile Ranks for Simple Example Data Set



literature on the optimal percentile calculation method, the differences would be minor and would not touch on the arguments of this paper, which are independent of the specific algorithm.

Set A has two relatively highly cited papers, whereas B does not have any. In this and the following examples we are interested in values of indicators of typical citation impact. Table 2 presents summary statistics and aggregate citation indicator values. Note that set A has around three times as many citations as B, and this is reflected in average citations and the MNCS. Nevertheless, citation *z*-score, average percentile rank and median percentile rank would suggest that the performance of set B is higher – clearly a very contradictory result. This discrepancy is due to the aforementioned compression of the citation distribution tail. The arithmetic average percentile rank, while sometimes recommended (Bornmann & Williams, 2020; Leydesdorff et al., 2019; Leydesdorff et al., 2011; Mutz & Daniel, 2012) is calculated for demonstration only, as the arithmetic operations are invalid for ordinal scale data. But even if we disregard that one result, consider that with an equal number of publications and unit A having more than three times as many citations as unit B, it is unit B which has the higher value in average citation *z*-score and median percentile rank (Note 3).

7. Simulation Studies

Some readers may feel that the above example was too small and contrived to be convincing. To substantiate the above discussion with data of a more realistic size we proceed by performing two simulation studies using synthetic data, run in the R programming language. The first experiment is again restricted to a single homogeneous publication set to illustrate the percentile rank distortion effect in isolation. In the next subsection a scenario with two heterogeneous fields will be considered.

7.1 Single homogeneous publication set scenario

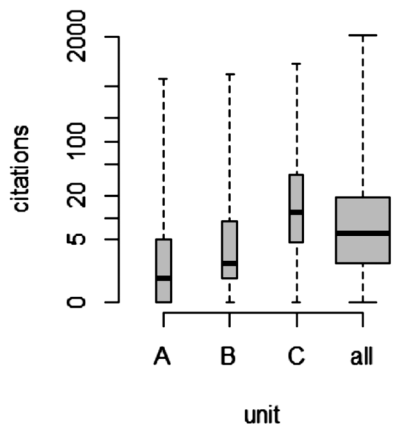
Let A, B, and C be three research units which have the same level of publication activity in terms of number of papers in a single field and year. All three units produce 1,000 items. Not only does using equally sized publication sets make comparisons simple, it is advantageous that they be the same size because the size of publication sets can have a distorting effect even for apparently size-independent indicators (Antonoyiannakis, 2018). The units differ in the typical level of citations their publications have received. All three units' citation counts are drawn from discretized lognormal distributions with mean parameters of A: ln(2), B: ln(3), C: ln(13) on the log-scale. To complete the reference set, there is a group

Table 2. Aggregate Citation Indicator Values for Simple Example Data Set

Set	Citation sum	Average citations	MNCS	Average citation *z*-score	Average PR	Median PR
A	158	16	1.5	-0.081	49	35
B	48	4.8	0.47	0.081	64	65

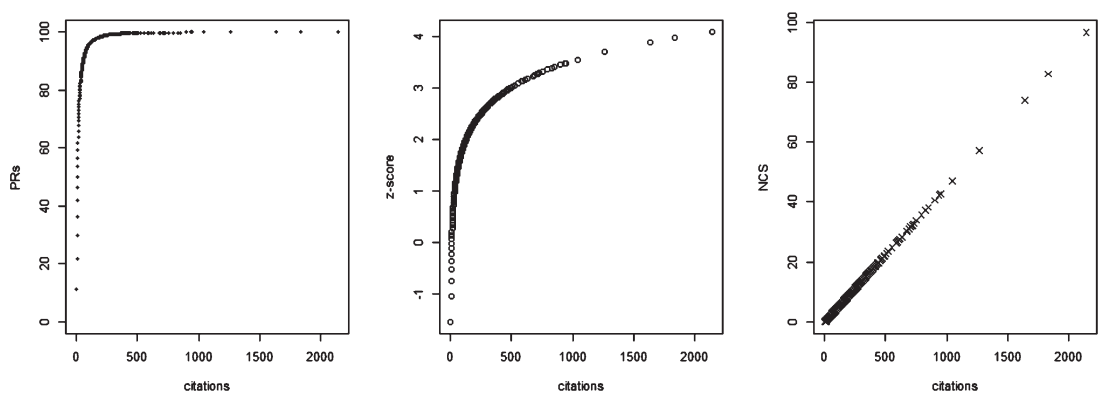
of 10,000 other publications with a discretized lognormal citation distribution with mean ln(8). All standard deviation parameters are set to 1.5 on the log-scale. These input parameters put the average citation impact performance of units A and B below the reference set average and that of unit C above. Figure 2 shows a summarization of the simulated citation distributions of the units and of the whole reference set and Figure 3 shows the scatterplot of all different occurring citation count values and their percentile ranks, *z*-score values, and normalized citation scores.

Figure 2. Simulated Citation Distributions for Three Units and Other Publications Single Homogeneous Publication Set with Three Units Scenario



Note. Box widths proportional to $\sqrt{n}$.

Figure 3. Scatterplots of Different Normalized Citation Scores over Citations. Single Homogeneous Publication Set with Three Units Scenario



We calculate for each unit the sum of citations, the (unnormalized) average number of citations per paper and averages of some of the discussed normalized citation counts, viz. the mean normalized citation score, the average citation *z*-score and the average and median percentile rank. The average percentile rank is calculated only for demonstration purposes, because, as stated above, averages of ordinal data are not meaningful statistics. Table 3 shows the results for this scenario and the third column confirms that the indicator values are not driven by particularly high maximum citation counts for any of the three units in this experiment. All indicators seem to confirm that the typical performance of units A and B are below average while that of unit C is above average, in line with the parameter specifications reported above. Because all units produced the same number of papers and the ratio of the sums of citations between units C and A is 6.5 we can state that both C's typical impact and its compound impact is 6.5 times that of A. Only the unnormalized average citations per paper and the MNCS preserve this ratio. As for *z*-score, no ratio can be formed since the data are transformed to interval scale and the information about the unit magnitude is lost. According to average percentile rank, the ratio would be 1.9. In these cases, direct comparison of different units' indicator values does not give results that correspond to the real factors of difference. That no ratio can be obtained from the *z*-score values is immediately obvious. What is possible is to state that the impact difference between A and C is 1.1 log-transformed standard deviations. Relative differences are not preserved by (log) *z*-score calculation because taking the logarithm is a nonlinear operation. Relative differences are also lost by percentile rank calculation and unlike *z*-score, no meaningful value at all can be expressed for differences although the calculated scores might suggest otherwise.

Let us consider a few actual individual values. While there are 13,000 observations, there are only 341 different observed citation count values in the data set. The lowest is 0 and the highest is 2,146. Table 4 shows more of the two ends of the range, that is, the low and high citation counts. The actual difference between 0 and 1 citations is 1 citation. On the percentile rank scale, it is 10.49 percentile rank points. At the other end of the range, the two most highly cited publications have citation counts of 1,837 and 2,146, hence a difference of 309 citations. Yet their percentile rank "difference" is 0.0077. In this data set the single citation of a

Table 3. Calculated Citation Indicators in Single Homogeneous Publication Set Scenario

Unit	Citations	Max. citations	Average cit. per paper	MNCS	Average cit. *z*-score	Average PR	Median PR
A	6,113	607	6.11	0.28	-0.71	32.91	21.64
B	9,470	704	9.47	0.43	-0.50	38.65	29.95
C	39,711	951	39.71	1.79	0.43	63.44	65.91
all	288,675	2,146	22.21	1.00	0.00	52.38	50.22

Note. Each unit with 1,000 publications.

Table 4. Excerpt from the Citation Values, Percentile Ranks, and the Frequency of Their Occurrence in Single Homogeneous Publication Set Scenario

Citations	Percentile rank	Occurrence
0	11.15	1,449
1	21.64	1,364
2	29.95	1,080
3	36.52	854
4	42.03	717
1,044	99.97	1
1,267	99.98	1
1,641	99.98	1
1,837	99.99	1
2,146	100	1

paper cited once is worth 21.64 PR points while one citation of the most highly cited paper with 2,146 citations is worth 0.05 PR points. The "sum" of the PR points of 9 uncited papers is as much as that of the most highly cited paper. Again, these arithmetic calculations on ordinal data are not meaningful but they illustrate how extreme the compression of the percentile rank calculation is at the upper end of the citation distribution.

7.2 Cross-field normalization scenario

We now come to the case for which field normalization is necessary, a scenario of two publishing research units, A and B, that are equally active in two fields with different citation distributions. Both units have published 100 papers each in each of the two fields. Besides these 200 publications, for each field there are a further 1,000 other publications. The two reference sets corresponding to the fields, X and Y, thus have 1,200 publications each. They have been created with draws from discretized lognormal distributions with parameters of mean equal to log(2) and log(1) on the log scale and the same standard deviation of 2 on the log scale. The R code for all examples is available at https://zenodo.org/record/7313246.

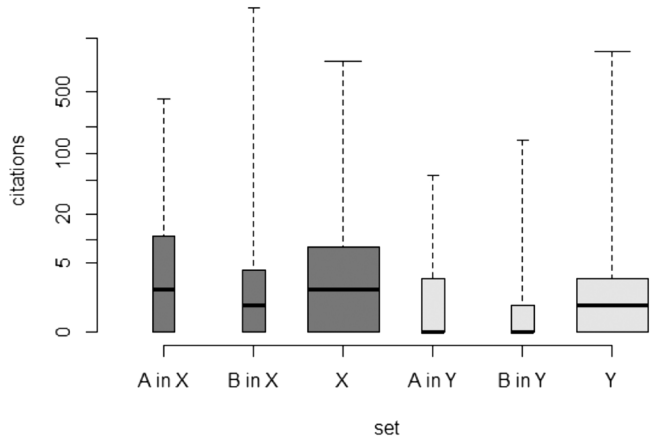
The citation distributions of the publication sets are summarized in Table 5 and plotted in Figure 4. The calculated citation indicators are shown in Table 6. From these tables it can be seen that for the publications in reference set Y, the performance of units A and B is almost even. However, in reference set X, the performance of unit B is much higher than that of A, in this instance due to one very highly cited paper. We would expect an aggregate normalized citation impact indicator to reflect this greater overall performance of unit B. Accordingly, the MNCS value of B is far in excess of that of A. However, the values of both the *z*-score and the average percentile rank, to the contrary, suggest that the performance of unit A is greater.

These results demonstrate that, besides computing the average of percentile ranks being an inadmissible operation, the results of the percentile rank method are misleading. The values of the *z*-score method also produce unintuitive results due to progressive compression of high values.

8. Are Highly Cited Percentage Indicators Affected?

In the preceding sections the disadvantages of percentile rank-based indicators for citation impact have been demonstrated. The number and share of highly cited publications of research units are very commonly used bibliometric

Figure 4. Simulated Citation Distributions for Two Units in Cross-field Normalization Scenario



Note. Box widths proportional to $\sqrt{n}$.

Table 5. Summary of Citation Distributions for Simulated Data Set in Cross-field Normalization Scenario

Set	Publications	Median cit.	Mean cit.	Max. cit.	Sum cit.
unit A in set X	100	2	20.95	414	2,095
unit A in set Y	100	0	4.46	56	446
unit B in set X	100	1	54.01	4,333	5,401
unit B in set Y	100	0	4.60	141	460
set X	1,200	2	14.34	1,097	14,335
set Y	1,200	1	8.97	1,410	8,970

Table 6. Calculated Citation Indicators for Two Units in Cross-field Normalization Scenario

Unit	Citations	Average cit. per paper	MNCS	Average cit. *z*-score	Average PR	Median PR
A	2,541	12.7	0.85	0.07	62.64	56.50
B	5,861	29.3	1.76	-0.19	57.15	51.17

indicators reflecting the performance in the high end of the impact distribution. These indicators are calculated by finding the threshold value of a citation distribution such that publications with citation counts at or above that value belong to the x % most highly cited publications in the reference set. The value of x may be 10 or another, typically small, value. The threshold value is clearly a

percentile. It has been pointed out that the indicator family of percentages of highly cited papers can be defined as a variant of a generalized percentile rank classes indicator family (Bornmann, 2013). If the arguments of the preceding sections are taken seriously and the use of the percentile rank method for aggregate citation impact indicators is to be rejected, does that also mean that highly cited indicators need to be rejected? After all, these indicators use percentile thresholds. What is more, one way of calculating the proportion of highly cited publications of a unit is to define the two percentile rank classes for highly cited and non-highly cited publications with the threshold and assign the class weights 1 and 0. One can then obtain the highly cited percentage of any unit by summing over the class weights of their publications. This would seem to be a case of applying arithmetic to percentile rank-based weights which has been argued as being not meaningful.

The use of the numerical values 0 and 1 is just a convenience, they are not needed at all, as one is not interested in those surrogate values but in the number of papers assigned to the two classes. For demonstration, let H be the set of highly cited papers of a unit and N be the set of non-highly cited papers. Ignoring the issue of ties from papers with citations exactly equal the threshold value, we are only interested in which of these sets each paper belongs to. Mathematically, the operation of calculating the share of highly cited papers for a unit is then accomplished by counting the papers in both sets (set cardinality) and calculating the appropriate ratio: $\frac{|H|}{|H|+|N|}$. Weighted percentile rank classes are unnecessary for calculating shares of highly cited papers and thus the arguments against the percentile rank method do not apply.

Nonetheless, highly cited percentage indicators are also affected by information loss, but this is explicit in the method and intentional. As the name suggests, rather than characterizing the citation distribution as a whole or attempting to locate its central tendency, these indicators are only concerned with the tail end of the distribution. As for using a percentile to define a threshold, as has been mentioned earlier, the criticism in this paper is restricted to percentile ranks when used as substitutes for citation values and does not extend to percentiles as relative positions of particular values in an empirical data set.

9. Conclusion

We have shown by argument and example that percentile rank calculation for normalization of citation counts has severe drawbacks that have hitherto hardly been appreciated in the literature. The percentile rank method discards crucial information and can suggest interpretations that are unjustified. The reduction of citation counts to ordinal data is not needed for normalization and for obtaining robust indicators. Other methods exist that have these properties. However, the citation z-score does so at the cost of relativizing exceptionally high performance, as does percentile rank calculation. We have turned around the argument that robustness to high values is an advantageous property for citation indicators by reasoning that such robust indicators are in fact obscuring exceptional performance and that their robustness can be seen as undesirable insensitivity. However, this particular argument must be considered in the context of the purpose or type of a citation impact indicator (average, compound, or exceptional performance indicator). Most

importantly, arithmetic aggregation of percentile rank scores is inappropriate, as these are ordinal scale data. Arithmetic operations on ordinal data are not meaningful. Contrary to the position articulated in the Leiden Manifesto, percentile rank calculation for citation normalization should not be considered unreservedly as best practice in bibliometrics as it produces misleading results due to an extreme distortion of citation count values. Normalization methods are meant to remove distortion, not introduce more of it.

Notes

Note 1 The LM does not mention percentile ranks but strictly speaking percentiles are specific values of the original data defined by their order position. In bibliometrics, percentile *ranks* are used for normalization, which is also implied in the LM quote by the phrase "weighted on the basis of the percentile". That is to say, the values of citation counts are replaced in analysis by their percentile ranks in the empirical citation distributions.

Note 2 One might in addition also use an indicator of poor citation impact, such as the share of uncited papers, to complement these. It has been argued that instead of using such scalar indicators that reduce the whole impact distribution to one point value one should rather use graphical comparisons of the complete distributions as such, e.g., Adams et al. (2007), Bornmann (2013).

Note 3 It ought to be stated at this point that this reasoning depends on an important proviso. The foregoing holds, provided one is for the purposes of a unit's typical impact indicator value indifferent to the concentration of citations within the set of publications. For example, should two citation distributions from publication sets of equal size and with equal total sum and arithmetic average of citation count, for example the two citation count sets of {0, 0, 15} and {5, 5, 5} have the same value for any typical impact indicator or not? There is no clear consensus in the literature. For the purpose of this article, we side with the position of being indifferent to the internal concentration, that is, the indicator values in the example ought to be equal. The argument in favor is that it is the bulk of the publication set and its citations that are of primary interest. This need not necessarily always be the case. For example, it might be an explicit purpose of indicator design to favor homogeneous over variable performance within units' publication sets.

References

Adams, J., Gurney, K., & Marshall, S. (2007). Profiling citation impact: A new methodology. *Scientometrics, 72*(2), 325-344. https://doi.org/10.1007/s11192-007-1696-x

Agresti, A. (2006). Ordinal data. In S. Kotz, N. Balakrishnan, C. B. Read, & B. Vidakovic (Eds.), *Encyclopedia of statistical sciences* (pp. 5842-5848). John Wiley & Sons. https://doi.org/10.1002/0471667196.ess1881.pub2

Aksnes, D. W., & Sivertsen, G. (2004). The effect of highly cited papers on national citation indicators. *Scientometrics, 59*(2), 213-224. https://doi.org/10.1023/b:scie.0000018529.58334.eb

Antonoyiannakis, M. (2018). Impact factors and the central limit theorem: Why citation averages are scale dependent. *Journal of Informetrics, 12*(4), 1072-1088. https://doi.org/10.1016/j.joi.2018.08.011

Antonoyiannakis, M. (2020). Impact factor volatility due to a single paper: A comprehensive analysis. *Quantitative Science Studies, 1*(2), 639-663. https://doi.org/10.1162/qss_a_00037

Bornmann, L. (2013). How to analyze percentile citation impact data meaningfully in bibliometrics: The statistical analysis of distributions, percentile rank classes, and top-cited papers. *Journal of the American Society for Information Science & Technology, 64*(3), 587-595. https://doi.org/10.1002/asi.22792

Bornmann, L., & Williams, R. (2020). An evaluation of percentile measures of citation impact, and a proposal for making them better. *Scientometrics, 124*, 1457-1478. https://doi.org/10.1007/s11192-020-03512-7

D'Agostino, M., Dardanoni, V., & Ricci, R. G. (2017). How to standardize (if you must). *Scientometrics, 113*(2), 825-843. https://doi.org/10.1007/s11192-017-2495-7

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429-431. https://doi.org/10.1038/520429a

Ioannidis, J. P., Boyack, K., & Wouters, P. F. (2016). Citation metrics: A primer on how (not) to normalize. *PLoS Biology, 14*(9), Article e1002542. https://doi.org/10.1371/journal.pbio.1002542

Leydesdorff, L., & Bornmann, L. (2011). Integrated impact indicators compared with impact factors: An alternative research design with policy implications. *Journal of the American Society for Information Science & Technology, 62*(11), 2133-2146. https://doi.org/10.1002/asi.21609

Leydesdorff, L., Bornmann, L., & Adams, J. (2019). The integrated impact indicator revisited (I3*): A non-parametric alternative to the journal impact factor. *Scientometrics, 119*(3), 1669-1694. https://doi.org/10.1007/s11192-019-03099-8

Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables on citation analysis one more time: Principles for comparing sets of documents. *Journal of the American Society for Information Science & Technology, 62*(7), 1370-1381. https://doi.org/10.1002/asi.21534

Lundberg, J. (2007). Lifting the crown—Citation z-score. *Journal of Informetrics, 1*(2), 145-154. https://doi.org/10.1016/j.joi.2006.09.007

Mcallister, P. R., Narin, F., & Corrigan, J. G. (1983). Programmatic evaluation and comparison based on standardized citation scores. *IEEE Transactions on Engineering*

Management, EM-30(4), 205-211. https://doi.org/10.1109/TEM.1983.6448622

Mutz, R., & Daniel, H.-D. (2012). Skewed citation distributions and bias factors: Solutions to two core problems with the journal impact factor. *Journal of Informetrics, 6*(2), 169-176. https://doi.org/10.1016/j.joi.2011.12.006

Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics, 9*(5/6), 281-291. https://doi.org/10.1007/BF02017249

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science, 43*(9), 628-638. https://doi.org/10.1002/(SICI)1097-4571(199210)43:9<628::AID-ASI5>3.0.CO;2-0

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680. https://doi.org/10.1126/science.103.2684.677

Thelwall, M. (2016). 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. https://doi.org/10.1016/j.joi.2015.12.007

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365-391. https://doi.org/10.1016/j.joi.2016.02.007

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., van Leeuwen, T. N., van Raan, A. F. J., Visser, M. S., & Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science & Technology, 63*(12), 2419-2432. https://doi.org/10.1002/asi.22708

Waltman, L., & van Eck, N. J. (2019). Field normalization of scientometric indicators. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 281-300). https://doi.org/10.1007/978-3-030-02511-3_11

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics, 5*(1), 37-47. https://doi.org/10.1016/j.joi.2010.08.001

Zhang, Z., Cheng, Y., & Liu, N. C. (2015). Improving the normalization effect of mean-based method from the perspective of optimization: Optimization-based linear methods and their performance. *Scientometrics, 102*(1), 587-607. https://doi.org/10.1007/s11192-014-1398-0

Zhou, P., & Zhong, Y. (2012). The citation-based indicator and combined impact indicator—New options for measuring impact. *Journal of Informetrics, 6*(4), 631-638. https://doi.org/10.1016/j.joi.2012.05.004

(Received: 2022/9/1; Accepted: 2022/10/14)

圖書資訊學刊　第20卷第2期　（2022.12）　頁75-93
https://doi.org/10.6182/jlis.202212_20(2).075

引用影響力研究中以百分等級正規化之缺點

Drawbacks of Normalization by Percentile Ranks in Citation Impact Studies

Paul Donner[1]

摘　要

本文探討書目計量文獻中常被忽略之百分等級方法在引用影響力正規化上的缺點。此方法將引用次數轉換為百分等級，使數據由比率尺度轉變成次序尺度。然而，未定義兩值間的比率及兩值間的差異大小易導致重要資訊遺漏。由於在文獻集合中，以引用次數排序時，高被引文獻與其排序相鄰的文獻引用次數落差極大，且高被引文獻相較於非高被引文獻數量更為稀少，因而嚴重地扭曲了引用數據。此外，算術運算在次序尺度資料中是沒有意義的，這也排除了某些文獻所推薦的運算方式，如：用百分等級數據計算總和或是平均。本文以數個案例說明百分等級運算用於影響力指標將扭曲引用數據。

關鍵字：引用正規化、領域正規化、百分等級、次序尺度

[1] 德國高等教育研究暨科學研究中心
German Centre for Higher Education Research and Science Studies (DZHW), Berlin, Germany
E-mail: donner@dzhw.eu

註：本中文摘要由圖書資訊學刊編輯提供。
以APA格式引用本文：Donner, P. (2022). Drawbacks of normalization by percentile ranks in citation impact studies. *Journal of Library and Information Studies, 20*(2), 75-93. https://doi.org/10.6182/jlis.202212_20(2).075
以Chicago格式引用本文：Paul Donner. “Drawbacks of normalization by percentile ranks in citation impact studies.” *Journal of Library and Information Studies* 20, no. 2 (2022): 75-93. https://doi.org/10.6182/jlis.202212_20(2).075