Scientometrics (2020) 123:223–252
https://doi.org/10.1007/s11192-020-03386-9

Percentile and stochastic-based approach to the comparison of the number of citations of articles indexed in different bibliographic databases

Gerson Pech[1,3] · **Catarina Delgado**[2,3]

Received: 9 August 2019 / Published online: 22 February 2020

Abstract
Recent studies have shown that the coverage of Scopus and Web of Science (WoS) databases differs substantially. Consequently, the citation counts of a paper are different depending on the database used, making it difficult to apply both together. To address this problem, this paper aims to examine whether the percentile- and stochastic-based approach is effective for converting citation counts between two databases while guaranteeing its time-normalization. For this analysis, we collected a dataset of 326,345 papers, published in 1987–2017 in the top 10% source titles of the following fields: Industrial and Manufacturing Engineering, Aquatic Science, Social Psychology and Archaeology. First, we applied the linear regression model to the citation percentiles of indexed papers in both databases. Secondly, we used the predicted results of this linear dependence, combined with the Monte Carlo simulations, to obtain the probability density function of a percentile from papers in the database in which they are missing. The results indicate that, with the method proposed in this paper, it is possible to convert the citation counts of articles between Scopus and WoS. In addition, it also predicts the citation impact of a missing paper on one of those databases, based on the citation impact on the other database. Tests on subsamples, using Lin's concordance coefficient, suggest substantial agreement between estimated and real citation values. This allows the combined use of the citation counts of two databases, improving the coverage and accuracy of both bibliometric studies and bibliometric indicators.

Keywords Citation analysis · Citation impact · Bibliographic databases · Database normalization · Citation normalization · Percentile-based approach · Probability density function · Monte Carlo simulation · Scopus · Web of science

✉ Gerson Pech
pech@uerj.br; gpech@fep.up.pt

Catarina Delgado
cdelgado@fep.up.pt

[1] Department of Nuclear Physics and High Energies, Rio de Janeiro State University, Rio de Janeiro 20550-900, Brazil

[2] LIAAD/ INESC TEC, University of Porto, 4200-465 Porto, Portugal

[3] Faculty of Economics, University of Porto, 4200-464 Porto, Portugal

Introduction

Citation analysis of papers indicates the contribution level of work in developing a research field and reveals the interest a paper rises in the scientific community. Many authors in scientometrics (e.g., Fairclough and Thelwall 2015; Glänzel 2011; Milojević et al. 2017; Radicchi and Castellano 2012; Rodríguez-Navarro and Brito 2018; Thelwall 2019; Waltman 2016; Zhang et al. 2014) have been focusing their research on how to apply citation analysis to compare the impact and the visibility of articles, journals, researchers, research groups, institutions, and countries, and the level of attractiveness of research fields for literature mapping, performance evaluation, or funding purposes. This work follows this track and examines the problem of how to compare the number of citations of papers indexed in different databases. This comparison method could play an important role when conducting a systematic bibliometric analysis concerned with the growth, development, and changes of a scientific subject.

Indeed, in the last years, the studies comparing Scopus and Web of Science (WoS) databases have shown their coverage differs substantially (Mongeon and Paul-Hus 2016). This discrepancy depends on the research field, country of publication and the period of publication. These differences, in some cases, reach 40%-50% of the analyzed dataset (Harzing and Alakangas 2016; Li et al. 2010; Martín-Martín et al. 2018a, b; Mongeon and Paul-Hus 2016; Wang and Waltman 2016). Consequently, the conclusions of bibliometric analyses may depend on the database used (Mongeon and Paul-Hus 2016), which is unwelcome. This suggests that to develop a more complete and less biased bibliometric study, these two databases need to be used together. In order to do so, three challenges need to be overcome. First, articles in different databases have different numbers of citations, even if they belong to both databases. Therefore, we have to decide which number of citations we are going to use. Second, some articles may only be indexed in one of those databases. Thus, we have to decide how to estimate (if possible) the missing number of citations for those articles. Third, the number of citations might not be the best indicator of an article´s impact, since previous studies (Bornmann and Leydesdorff 2017; Pech and Delgado 2019; Pech et al. 2019; Petersen et al. 2019; Waltman 2016) demonstrated that when comparing articles from different years, the most common normalization methods introduce a bias in the number of citations for either earlier years or later years (depending on the chosen normalization method). For longitudinal studies, Pech et al. (2019) have indicated that using the percentile of the number of citations of an article is more accurate and unbiased than the common methods that have been used in most recent literature analytical reviews. A similar problem happens when comparing articles from different bibliographic databases, such as Scopus and WoS.

The comparison of the number of citations from each of the two bibliographic databases can also play an important role when calculating researchers' performance metrics as it can minimize the dependence between these metrics and the database used (Wildgaard et al. 2014). In addition, as Rousseau (2007) emphasizes, it is common to use, for *h*-index purposes, only articles published in journals indexed in WoS, without considering other types of scientific and peer-reviewed publications (such as conference proceedings or book chapters) or articles indexed in other databases, such as Scopus, that might well be an author's most cited work. Moreover, using different databases to obtain an author's *h*-index results in different *h*-index values for the same author (Bornmann and Leydesdorff 2018)—which database should one follow when assessing the productivity and quality of a researcher (or of a research group or institution) for performance assessment or for funding purposes?

(De Groote and Raszewski 2012). Indeed, with the conversion of citations between databases, the deficiencies caused by the difficulty of using two databases in bibliometric performance indicators could be minimized. In fact, there are other important shortcomings regarding the use of the *h*-index (Bornmann 2014) not related to this study, and therefore outside the scope of this paper.

In response to these issues, we argue, in this paper, that a percentile and stochastic-based approach can effectively address the merge of the two databases and their articles' citation impact information. We examined the use of a percentile-based approach to compare (and convert) the number of citations from papers in different databases when there is a subset of articles in both databases. The percentile approach has been extensively applied lately in citation analysis for bibliometric evaluations (Bornmann 2013; Bornmann and Marx 2014; Bornmann et al. 2013; Waltman and Schreiber 2013; Thelwall 2016; González-Betancor and Dorta-González 2017) and also to predict citation counts (Kosteas 2018). In addition, this technique is the basis of the percentile share method (Jann 2016), that has been used to investigate inequalities (Davies et al. 2017, Mishra 2018, Mishra and Kumar 2018), specifically, the skewness of citation impact data to demonstrate (by an easy visualization) that inequality in impact appears in various disciplines and fields (Bornmann and Leydesdorff 2017).

As we will be presenting, the necessary time- and database-normalization can be adequately done by way of a linear regression analysis of the relationship between the percentiles from the different databases or from different years. With the linear regression coefficients and their standard error, it is possible to obtain, using the Monte Carlo method, the probability distribution function of the percentile of a paper on a database in which it is not indexed. The idea of using the probability distribution function is not a new approach in citation analysis. Stegehuis et al. (2015) used the probability density function to address the citation count behaviour forecasting issue. These authors argue it would be more logical to calculate the probability that a paper will receive a certain range of citations in the future, instead of predicting the average number of forthcoming citations.

We used the citation data from 326,345 papers, published in the period 1987–2017, from the top 10% source titles in four different fields (the Scopus sub-subject areas: Industrial and Manufacturing Engineering; Aquatic Science; Social Psychology and Archaeology). We developed a percentile-based conversion method, with the following steps: (1) data collection from Web of Science (WoS) and Scopus (articles and their citation counts); (2) percentile calculation (percentile of the citation count of each article in each database); (3) database comparison (identification of articles belonging to the two databases); (4) linear regression to obtain regression coefficients and their standard errors; and (5) Monte Carlo simulation to obtain the percentile probability density function of the citations count of a paper on a database, based on the percentile on the other database. This percentile probability density function allows the conversion between the two databases. Details of the process used by this paper are provided in the following sections.

Objectives and research questions

The aim of this paper is to examine whether the percentile- and stochastic-based approach is effective for converting citation counts between two databases while guaranteeing the normalization of citation counts over the years. Thus, this exploratory study will address the following research questions:

[RQ1] What are the differences between percentiles of citation counts, in WoS and Scopus databases, for the same papers?

[RQ2] Given a specific confidence level and the percentile of the citation counts of a paper on a database, what is the confidence interval of the percentile on the other database?

[RQ3] How is it possible to merge the paper indexed in two different databases and still use the paper citation-based impact information in a unified manner?

In the next section, we present the data used in this study and introduce the five-stage method implemented to answer these research questions.

Methods

Data collection

We collected data from Scopus and WoS of four research fields defined by the All Science Journal Classification code (ASJC). ASJC is the Scopus journal categorization, and it is often used in bibliometric studies. The research fields (in Scopus, these fields are defined as the sub-subject areas) used in this paper are the following: Industrial and Manufacturing Engineering (ASJC = 2209, hereafter INDENG), Aquatic Science (ASJC = 1104, hereafter AQUSCI), Social Psychology (ASJC = 3207, hereafter SOCPSY), and Archaeology (ASJC = 1204, hereafter ARCHAE). According to this classification, these fields belong to the following Scopus subject areas, respectively: *Engineering, Agricultural and Biological Sciences, Psychology and, Arts and Humanities*.

Each database was built with the top 10% source titles, in terms of citations, according to Scopus CiteScore Percentiles, i.e., journals with CiteScore Percentiles from the 99th–90th. For the purposes of this work, we considered this filter to align our results with the usual practice of several authors of bibliometric studies who applied in their works,: (1) only the journals with high impact in a studied area (e.g., Alajmi and Alhaji 2018; Darko and Chan 2016; Filardo et al. 2016; Laengle et al. 2017; Pesta 2018; Shang et al. 2015; Zhu and Zhu 2016); or (2) the most highly cited papers of an area (e.g., Adam et al. 2017; Bohl 2017; Dokur and Uysal 2018; Yeung et al. 2018). These papers have also a high probability of being in the top journals of the area. The Print-ISSN and e-ISSN of the titles were obtained from the file *CiteScore_Metrics_2011–2017* downloaded from Scopus.com on May 25, 2018, by selecting the top 10% journals in the four ASJC codes (sub-subject areas) mentioned above.

The selection of these four ASJC codes, INDENG, AQUSCI, SOCPSY, and ARCHAE, made it possible to study data from very different subjects in terms of content and scientific production, without the computational burden of analysing all the research fields available. Each code was selected from one of four core research areas: *Technologies and Exact Sciences; Social Sciences; Education and Humanities;* and *Life and Health Sciences.* This classification is followed in authors' universities (Santiago 2018, p. 10), and in authors' opinion, it can represent scientific knowledge as well as the Scopus four broad subject clusters classification (*Health Science, Life Science, Physical Science, and Social Science*). Since this is an issue in discussion in the literature, and other definitions and delimitations of subject areas have been proposed to improve and update this classification (Gómez-Núñez et al. 2011), authors decided to use the classification they are more used to.

The four ASJC codes were selected, from the 303 ASJC codes in those four major research groups, based on the number of top 10% source titles, by following these rules:

- In order to represent a considerable range of fields with different journal numbers, we calculated the relative frequency distribution of the field top 10% source titles;
- To keep the task of data collection realistic in terms of time spent and effort, we decided to avoid fields with an excessive number of journals, and therefore with an excessive number of papers to process.
- Fields with an insignificant amount of journals were also avoided to guarantee the necessary data variability to support statistical analysis. To achieve this goal, the cumulative probability of the number of top 10% source titles per field was obtained, and the field with [75% + k 5%] was selected in each one of the large core fields (*Technologies and Exact Sciences*; *Social Sciences*; *Education and Humanities* and; *Life and Health Sciences*). To support the requirements mentioned above and to select fields with a different number of journals, we decided to use $k = 0$, 1, 2 and 3, where each value k is randomly assigned to a specific large core field (0 to the *Life and Health Sciences* field; 1 to the *Education and Humanities* field; 2 to the *Social Sciences* field; and 3 to the *Technologies and Exact Sciences* field). Since 75% is equivalent to approximately 20 journals, considering the cumulative probability function, 80% is equivalent to approximately 24, and 90% is equivalent to 32, we selected, respectively, AQUSCI (20 journals), ARCHAE (24 journals) and INDENG (32 journals). For SOCPSY (25 journals), as 85% is equivalent to approximately 26, and there is no field with this number, we selected the field with the closest number of journals, which is SOCPSY with 25 journals.

Considering these four fields, the dataset consists of all papers published in the period 1987–2017, of the document type Article or Review, obtained from the query that excludes: editorial OR biography OR early access OR correction OR unspecified OR retraction OR abstract OR retracted publication OR bibliography OR book OR reference material OR meeting OR letter, and include only papers in English language. We used the ISSN codes of each journal, instead of the title name, to avoid collection errors. This search was conducted between April and May 2019, on Scopus and WoS sites.

In total, the data contains about 326 thousand different papers (of which, about 292 thousand from WoS, 274 thousand from Scopus, and almost 240 thousand indexed in the two databases). Figure 1 presents the evolution over the period 1987–2017 of the number of different papers used in this work considering the Scopus and the WoS databases. As shown by this figure, each field contributes to a different number of articles per year, which makes it possible to explore RQ1, RQ2, and RQ3 in different situations in terms of this characteristic.

Citation conversion process

A fundamental point about citation counts when comparing the article citations from a specific field, but published in different years, is the effect of two relevant factors. First, the evidence that the longer the paper has been exposed since its publication, the greater is the likelihood that it will have a greater citation count. Secondly, the growth of the number of papers published over the past few years and the growth of the number of references in each paper. For a broader view on the dependence between the number of citations and the

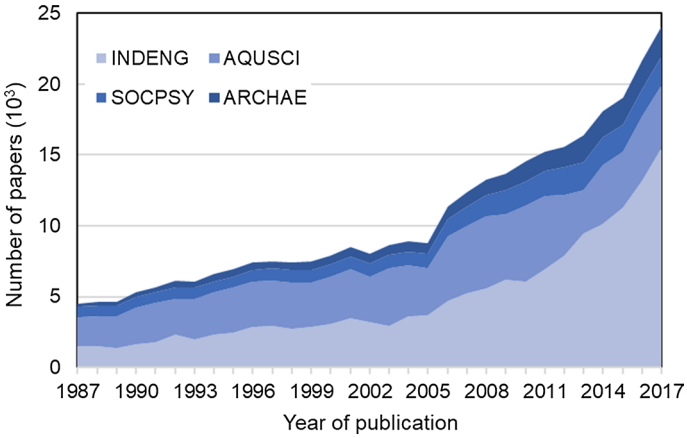


Fig. 1 Number of different papers included in our dataset according to the year of publication of the four research fields studied in this work, taking into account the Scopus and the WoS databases

time, see Petersen et al. (2019). These researchers used the concept of 'citation inflation' to investigate in-depth the two factors mentioned above and propose policies, such as limiting the number of references in papers, to improve the quality of the evaluations that use citation counts as a metric.

The dependence between the citations and the paper exposure time cannot be just considered linear, because the number of papers published, and because the number of citations has grown by different behaviors, depending on the field and period. For a detailed study of field- and time-normalization see Bornmann and Wohlrabe (2019) who recently applied proposed metrics based on arithmetic averages and percentiles in economics and their subareas. For instance, in the subject area Agricultural and Biological Sciences (in part, included in this paper) in 1987 were published close to 38,000 papers; in 1997, 78,000; 2007, 121,000, and in 2017, 209,000, considering only articles and reviews in the Scopus database. In other subject areas, this kind of nonlinear growth is also observed. In addition, the average citations number of papers published in different years, considering one research area, presents a more complex behavior. For example, in the sample of papers from INDENG area, studied by this work, the citation averages of articles published in 1987, 1990, 1997, 2000, 2007, 2010 and 2017 are, respectively, 25, 27, 31, 38, 40, 35 and 10. That is, the value grows for articles published\around 2007 and then begins to decrease. As a method of time-normalization, Wang (2013) analyzed the possibility of including only citations accumulated during a specific period after publication, called "citations window". However, Petersen et al. (2019) showed that the number of publications has been growing at around 4% per year since the 1960s, causing a temporal bias in analyzes using the citations window. To address this, the authors applied a correction based on the percentiles of paper citations.

Another important issue in terms of citations, which is precisely the theme of this study, is that the two databases mostly used in this scientific research, Scopus and WoS (Mongeon and Paul-Hus 2016), despite having a large database of titles in common, have some different and relevant gaps of coverage for many areas (Martín-Martín et al. 2018b). It is not the aim of this article to explore this point quantitatively; however, some examples we found in data used in this work, may help us to better understand the reasons why we have developed the

methods applied in this paper. For example, the paper: "Ignore fishers' knowledge and miss the boat", published by *Fish and Fisheries*, the title that has the highest SNIP and SJR value of the research field Aquatic Science, has 288 citations in WoS but does not appear in Scopus. In Social Psychology, another field included in this study, if we again get the journal with the highest SNIP and SJR of the field, i.e., *Personality and Social Psychology Review*, the paper "A meta-analysis of personality in scientific and artistic creativity" appears in Scopus with 902 citations, but does not appear in WoS at all. The same happens with the paper "Making sense of 3-D printing: Creating a map of additive manufacturing products and services" that has 230 citations in Scopus but it doesn't appear in WoS. This paper was published by *Additive Manufacturing*, again the highest SNIP and SJR of Industrial and Manufacturing Engineering, a field included in this work. Consequently, such kind of differences between the coverage of these two databases also leads to discrepancies in the citation counts. Thus, a complete bibliometric analysis of a research field needs to, at least, use these two databases. Otherwise, some important articles will be left out. In order to do that, it was necessary a conversion method between the citation counts of these two databases, which didn't break normalization over the years.

To introduce the conversion process suggested in this paper, we will begin without still defining their normalization rules. Thus, to illustrate this general process capable of normalizing citations in terms of period and simultaneity, between databases, let us assume two databases but without defining their normalization rules for a while: Ψ and Φ, in this study the Scopus and WoS, respectively. In addition, we will define ψ and φ as a subset of Ψ and Φ used in this research, composed only by the papers from top 10% source titles (see the details of this subset in "Data collection section").

Let us assume that:

- For the database Ψ, the comparison between years implies that a particular article i published in t_1, and with $C_i^{\Psi}(t_1)$ citations, is equivalent, in terms of citations impact, to an article j published in t_2 with $C_j^{\Psi}(t_2)$ citations; and that
- For the database Φ, one article k published in t_1, with $C_k^{\Phi}(t_1)$ citations, is equivalent to one l from t_2 with $C_l^{\Phi}(t_2)$ citations.

If, in t_1, $C_i^{\Psi}(t_1)$ can be converted in $C_k^{\Phi}(t_1)$ by using some normalization parameter which reflects the same citation impact, then, in t_2, $C_j^{\Psi}(t_2)$ must be converted into $C_l^{\Phi}(t_2)$ by the same normalization process, that was used in Ψ. Only in this way, the conversion process will be consistent between the databases and, the equivalence between different papers published in different years may be preserved (Fig. 2). In Fig. 2, the articles i and k published in t_1 have the same impact because the conversion between the two databases generates equal values for the normalized parameter, even with a different citation count. The same happens for j and l, in t_2. In addition, i and j (different papers published in different years) have equal impacts after time- normalization, despite the possible difference between citations. Therefore, applying the appropriate normalization method, we can conclude that the four different articles, with different citation counts, have the same impact since their normalized citation parameters are the same.

Accordingly, if $N_i^{\Psi}(t_1)$, $N_j^{\Psi}(t_2)$, $N_k^{\Phi}(t_1)$ and $N_l^{\Phi}(t_2)$ are, respectively, the normalization parameters of $C_i^{\Psi}(t_1)$, $C_j^{\Psi}(t_2)$, $C_k^{\Phi}(t_1)$ and $C_l^{\Phi}(t_2)$ citations (the normalization rules proposed in this study will be presented below), we will have:

$$N_i^{\Psi}(t_1) = N_j^{\Psi}(t_2) = N_k^{\Phi}(t_1) = N_l^{\Phi}(t_2) \tag{1}$$

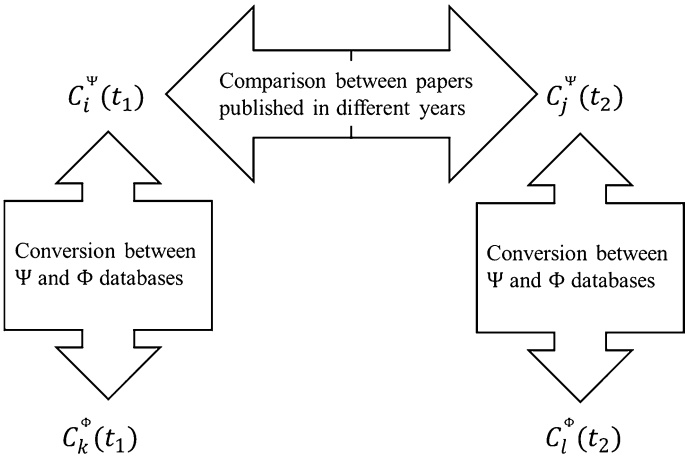


Fig. 2 The correspondence between the citations in two databases (Ψ and Φ) and across the years t_1 and t_2. The horizontal arrow represents the comparison between different papers published in different years, and the vertical arrows represent the conversion between the databases. If in the comparison between t_1 and t_2, in Ψ, $C_i^{\Psi}(t_1)$ is equivalent to $C_j^{\Psi}(t_2)$, and if in the conversion between the databases, $C_i^{\Psi}(t_1)$ is equivalent to $C_k^{\Phi}(t_1)$, so $C_l^{\Phi}(t_2)$ must be the same, if calculated by the longitudinal comparison or if calculated by the process of database conversion

Thus, comparing the articles from any database and published in any year, by means of their normalized values, it is possible to classify them in an appropriate list using the two databases and to develop the analysis that is required for the bibliometric study. However, we still must define the rules of normalization for both the conversion between the databases and the comparison between different papers from different years. This is one of the key ideas of this work. In order to address this problem, we have established the following basic assumption: the same normalization rule should fit both the conversion between Ψ and Φ and the temporal comparison. Thus, the properties represented by the parameters coming from both the time conversion and the database comparison will have the same characteristics in terms of impact.

The adoption of the percentiles method seems to be appropriate as a normalization because, among the different normalization procedures that could have been used (Waltman 2016), the percentile approach has the advantage of intrinsically reflect the normalization of citation counting data, since their weight is independent of external values (Brito and Rodríguez-Navarro 2018). In addition, the use of percentiles to normalize citations may be more effective in comparing papers from different years of publication than the use of averages or statistical distributions that are dominated by few highly cited papers (Bornmann 2013).

To ensure the conversion between the two databases will allow the temporal comparison, the percentiles of an article indexed in the two databases have to be equal in these two databases, within a margin of error with acceptable confidence limits. If this is roughly the scenario for a set of papers in a bibliometric study, the temporal comparison will automatically be possible. To clarify this concept, consider the following examples:

1. The paper *Energy recovery from molten slag and exploitation of the recovered energy* (hereafter, *Energy*) written by G. Bisio, and published in 1997, is indexed in the Scopus

and WoS, and in the same percentile (0.95), despite the different numbers of citations in each database;
2. The paper by Hnaien et al. (2016), *A mixed-integer programming model for integrated production and maintenance* (hereafter, *Mixed-integer*), is also indexed in the two databases and in the same percentile (0.95).

In this sense, considering the percentile approach as a normalization method, since these two articles are in the same percentile in both ψ and in φ, they are also equivalent by temporal comparison, independently of the selected database. Other examples such as the one above can be found in Table 1 for Percentiles 0.95; 0.90; 0.85; 0.80; and 0.75 (we use here five different values of N_i of the 1st quartile, as an example). Table 1 presents the results from 16 articles from the INDENG for four different years. In this table, the articles' titles were truncated due to space constraints.

All the articles presented have the same percentile in the two databases ($N_i^{\psi}(t) = N_i^{\varphi}(t) = N_i$), and a different number of citations, as presented in the Table 1. That is, for each table row, the four articles are equivalent both in the database conversion and in the temporal comparison and, therefore, we have:

$$N_i^{\psi}(t_1) = N_j^{\psi}(t_2) = N_k^{\psi}(t_3) = N_l^{\psi}(t_4) = N_i^{\varphi}(t_1) = N_j^{\varphi}(t_2) = N_k^{\varphi}(t_3) = N_l^{\varphi}(t_4) \quad (2)$$

Note that the events represented by Eq. 2 are just particular cases, whose possibility of generalization will be explored by this paper with the method presented in the next section.

In the cases where a papers' percentile is different in both databases, a temporal comparison with a paper from a different year could be possible only in one of those databases. For instance, if a 2007 paper is in percentile 0.95 in Scopus and 0.80 in WoS, it would be equivalent to the 1997 paper *Energy*, presented above, in temporal comparison in the Scopus database (both are in percentile 0.95) but would not be in WoS (*Energy* is in the 0.95 percentile, not in the 0.80 percentile). This scenario does not reveal any inconsistency, only that the method may not be applied directly in both databases. In this case, it would require corrections that could, for example, depend on the subject field, the number of papers used to calculate the percentile, the year of publication or other variables. However, if the percentiles rank of the papers indexed in Scopus and WoS is approximately equal, as suggested above, papers indexed in at least one of those two databases could be used in a longitudinal study. That is another key idea of this study: the need to examine how precisely and under what conditions the two percentiles can be considered approximately equal.

Percentile-based conversion method

To address our research questions, we developed the percentile-based conversion method (Fig. 3) with two linked purposes. The first purpose is to examine the differences between percentiles, in Scopus and WoS databases, of indexed papers. After that, if, as expected, the percentiles calculated for the same article in the two databases can be considered approximately similar, the second purpose is to estimate the database conversion uncertainty, in terms of confidence level and interval.

The percentile-based conversion method was developed with five main steps (Fig. 3):

1. *Dataset* The data of papers from the top 10% source titles were downloaded considering the four research fields mentioned in the "Data collection section" and for each year

Table 1 Examples of indexed papers in the two databases and that have the same percentile (N_i) in these two databases

N_i	1987 Article title $[C_i^{\psi}(t_1)\|C_i^{\varphi}(t_1)]$	1997 Article title $[C_i^{\psi}\|C_i^{\varphi}(t_2)]$	2007 Article title $[C_i^{\psi}(t_3)\|C_i^{\varphi}(t_3)]$	2017 Article title $[C_i^{\psi}(t_4)\|C_i^{\varphi}(t_4)]$
0.95	Kinetics of pyrolysis of moroccan … [94|90]	Energy recovery from molten slag … [111|94]	A mixed integer programming model … [140|130]	Producing PHAs in the bioeconomy … [32|28]
0.90	Solubility of Co_2 in aqueous … [57|54]	Stochastic simulation approach to … [73|69]	Numerical simulation of periodic bubble … [99|88]	Year-round performance and cost … [23|21]
0.85	A study of micromixing in tee mixers … [43|39]	An experimental investigation of heat … [56|53]	Laser milling of ceramic … [73|66]	Sorption enhanced catalytic Steam … [19|17]
0.80	Performance of a continuous solid … [34|30]	Improved operational policies for batch … [47|41]	Novel thermal-swing sorption-enhanced … [61|55]	Clean production pathways for regional … [16|14]
0.75	Load model for stability studies … [26|25]	The effects of machining settings … [39|35]	Studies on the performance of a … [53|47]	Distortion prediction and compensation … [14|13]

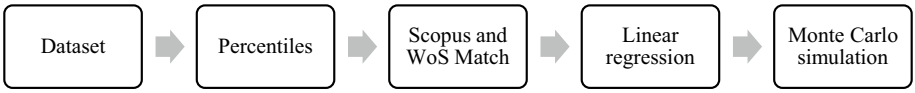


Fig. 3 The five steps method for comparison of the number of citations from papers in different databases

separately, during the period from 1987–2017. The data downloaded refer to the article title, DOI, authors, journal title, and citation counts. This procedure was conducted for both Scopus and WoS. Scopus data was analyzed in Citavi 6 (www.citavi.com)—Swiss Academic Software GmbH—to remove the duplications in the database (Valderrama-Zurián et al. 2015) and therefore prevent some avoidable errors related to the number of citations: (1) identical articles indexed in the same database but with different number of citations; (2) different papers with the same DOI creating a false positive in the next step of the research method.

2. *Percentiles* Considering the dataset of this study, the percentile of each paper was calculated regarding the papers published in the same year within the research field to which the paper belongs. The calculation was done in Excel (Office 365—64bits) using the PERCENTRANK.INC function ranging from 0–1 with 10 decimal places, for accuracy purposes. Calculated this way, if a paper has percentile 0.95 it means there are 95% of the papers with fewer citations than this paper. Thus, N_i can be written by the formula below:

$$N_i = \frac{k-1}{n-1}, \tag{3}$$

where N_i is the percentile of paper i, k is the rank number in ascendant order and n is the number of papers of the sample. As an example, given a sample of $n = 3001$ papers, if 60 papers have no citations, papers that have one citation ($k = 61$) will have $N_i = 0.02$. If some papers have the same number of citations, there is a tie, and they will all have the same percentile, calculated considering k the first rank number in ascendant order. In the previous example, if in the mentioned case five papers ($k = 61,\ldots,65$) had one citation, the five would have the same percentile, $N_i = 0.02$. This percentile rank formula assigns the zero percentile to articles with zero citations ($k = 1$), ensuring that the absence of impact of these works is reflected in the same way in all the sets of papers studied (Bornmann et al. 2013). An analogous concept can be applied to the paper with the highest number of citations of each year, which in this case has $N_i = 1$ ($k = n$). This procedure was performed in both ψ and φ in two separate ways: (1) considering all papers; (2) excluding papers that had no citations (hereafter, NO 0).

3. *Scopus and WoS Match* The goal of this step is to identify the articles indexed in the two databases and to collect their percentiles in each of the databases. Firstly, this procedure was done confronting the DOI of the Scopus database articles with those of the WoS database, taking into account each research field and year, separately. For papers that did not have DOI, or where the corresponding DOI in the WoS database was not found, the equivalence was checked by the paper title. As in the paper titles, some characters are not recognized in the same way in the two databases, we removed the characters " () [] { } / - * $ \ & % _ + ^ > < ! ?. ; —: @. ∞ in all titles. After that, we converted all titles to lowercase and we trimmed the titles to strings of 35 characters to do the match. As a result, the main database for this research was constructed. It is composed of the following data for each paper: research field, year of paper publication, Percentile in

Scopus, Percentile in WoS, Percentile NO 0 in Scopus, Percentile NO 0 in WoS. This dataset contains all the papers with a match in the two databases. Therefore, this dataset was the input for the linear regression developed in the next step of the model.

4. *Linear regression* This step verifies whether the percentiles of the papers indexed in the two databases can be modelled by a linear regression. Linear regression involving citation counts has been studied by other authors. For example, using a set of articles published in 12 journals, Moed et al. (2016) showed that there is a strong linear correlation between citations from Google Scholar and Scopus. On the other hand, Abramo et al. (2019) showed that linear regression can be an efficient technique to forecast long time impact of papers, when using as input a two or three-year citation window. Following these studies, we conducted multiple linear regression analyses using as dependent variable the WoS percentile, and as independent variable the percentile in Scopus. In addition to the dependence between the percentiles of the two databases, other dependencies were also tested (e.g., number of papers, number of articles without citations, year of publication, number of articles in the sample, and their relative percentages). The linear regression between the percentile values of the articles published in the same year and indexed in the two databases can be expressed as follows:

$$N_i^{\varphi} = N_i^{\psi}\left(\beta_1 \pm \varepsilon_1\right) + \beta_0 \pm \varepsilon_0 \tag{4}$$

where $\beta_1, \varepsilon_1, \beta_0$ and ε_0, are the linear regression coefficients and their corresponding statistical errors. One of the main ideas of the model proposed in this paper is that the estimate of these values provides indications about an effective normalization process to convert citations between the databases and over the years.

5. *Monte Carlo simulation* The objective of applying the Monte Carlo simulation in the method developed in this study is to answer the following question: what is the probability density function of N_i^{φ},. For a specific N_i^{ψ}? That is, if paper i is in ψ (and has a N_i^{ψ} value), but not in φ, what is the probability density function of N_i^{φ}? Thus, with the answer to this question, we can estimate the uncertainty of assigning to paper i a certain normalization value, even in a database in which it is not indexed. That is t method to estimate the conversion uncertainty between two databases.

   To do that, we replaced the linear regression coefficients (β_0 and β_1) by Gaussian distributions, using the values found in step 4 for β_0 and β_1 as the μ parameters, and ε_0 and ε_1 as the σ parameters of these functions. The use of the Gaussian distribution in this model follows the results of previous works about the normal linear regression model (Goel and DeGroot 1980; Spanos 1995) and its applications (Chen et al. 2001; Jiang et al. 2013). These studies predict that, in several situations where there is linearity between two variables, the errors related to the linear regression model coefficients can be described by normal distributions. In this case, the mean for these parameters will match the coefficients β_1 and β_0 found. Therefore, Eq. 4 is rewritten as follows:

$$N_i^{\varphi} = N_i^{\psi} F\left(\beta_1, \varepsilon_1\right) + F\left(\beta_0, \varepsilon_0\right) \tag{5}$$

where $F\left(\beta_1, \varepsilon_1\right)$ and $F\left(\beta_0, \varepsilon_0\right)$ are the Gaussian distributions mentioned above which parameters depend on the research field and the year of the paper publication. For each specific field, the Monte Carlo method was used to simulate the $F\left(\beta_1, \varepsilon_1\right)$ and $F\left(\beta_0, \varepsilon_0\right)$ values, based on the probability of simulating a paper published in t_1 (given by the number of papers published in t_1 divided by the total number of papers). As a result,

the described model can produce the probability distribution of N_i^{φ}, for defined values of N_i^{ψ}, and thereby is able to determine the conversion uncertainty.

With the Monte Carlo Method, we are extrapolating the dataset of this study, pursuing generic results that consolidate the domain of applicability of these research findings. Monte Carlo has already been applied in several scientometrics studies. For example, Yamashita and Okubo (2006) used it to estimate some parameters of the Probabilistic Partnership Index, which is an index developed by the authors to study collaborations between countries; Rodriguez and Pepe (2008) used this method to analyze the dependencies in the co-authorship network; by employing Monte Carlo, Demarest et al. (2014) studied resemblances between recommendations received by authors and recommendations given by reviewers in the same social group as they belong; Schulz (2016) used Monte Carlo to understand how the ambiguities of authors' names generate errors that can impact the author rankings; and also using Monte Carlo, Pan et al. (2018) demonstrated how the growth of scientific effort impacts the structure of the scientific citation network. In all these works, the purpose of using Monte Carlo was to broaden the dataset available to obtain more accurate results, which is also one of the purposes of this study.

Results and discussions

In this section, we present the empirical results of the five-step method described previously.

Dataset

Our dataset contains 326,345 different papers, published from 1987 to 2017. 15.9% of the articles are indexed only in WoS, 10.6% are indexed only in Scopus, and 73.5% are indexed in both databases. 46.2% of the papers are from the INDENG, 34.0% from the AQUSCI, 11.4% from the SOCPSY and 8.5% ARCHAE field.

Percentiles

Figure 4 shows the Scopus citations (C_i^{ψ}) by the solid line, and the WoS citations $\left(C_i^{\varphi}\right)$ by a dotted line, over the 31 years, for $N_i = \{0.30, 0.50, 0.75, 0.90, 0.95, 0.99\}$ in the four research fields studied. In Fig. 4, we omitted the ψ and φ superscripts since the different line types refer to the two databases. The equivalence of the number of citations can be obtained by following each of the curves defined by the solid line (for Scopus), or dotted line (for WoS), of each normalization value. For example, in the INDENG field, a paper with 30 citations in Scopus that was published in 1990 is equivalent to a paper published in 2000 with 50 citations and with a paper published in 2015 also with 25 citations ($N_i = 0.75$). In the same field and the same years, following the line of $N_i = 0.95$, we could see that this normalization results in $C_i^{\psi} = 100$, 140 and 56, respectively. The variation of

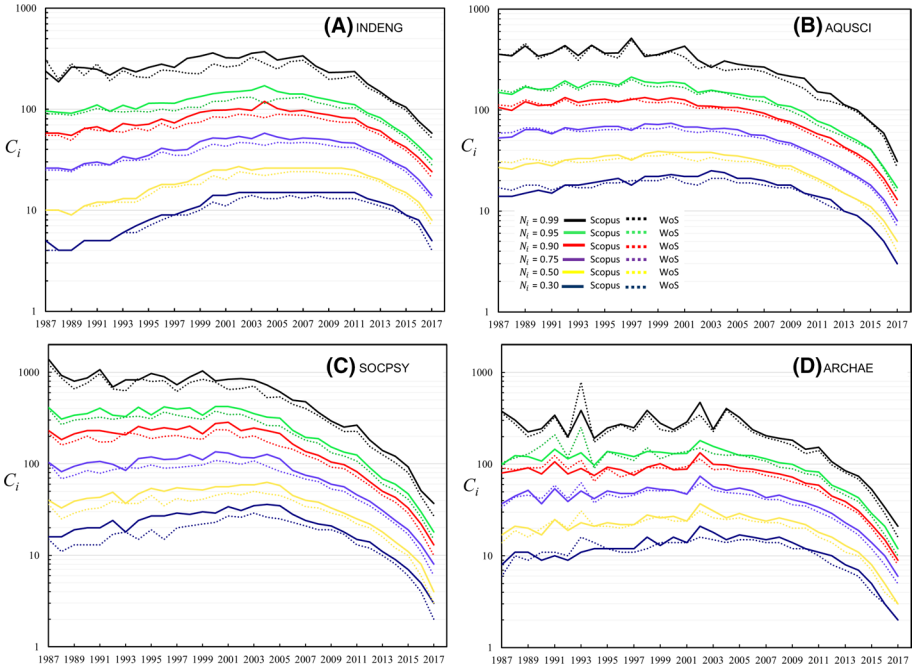


Fig. 4 Number of citations (C_i) across 31 years, for the Percentiles (N_i) equal to 0.30; 0.50, 0.75, 0.90, 0.95 and 0.99 for papers in the research field: **a** industrial and manufacturing engineering—INDENG, **b** aquatic science—AQUSCI, **c** social psychology—SOCPSY and **d** Archaeology—ARCHAE. The vertical axis displays in log scale to visualize easier the citations of different percentiles

the curve in the earlier years may be related to a smaller number of papers than in the later years, especially in the ARCHAE field, which has the least number of papers.

Scopus and WoS match

In the matching procedure, we found 239,780 papers indexed in both Scopus and WoS. Table 2 shows the number of articles in each research field during the period 1987–2017. The total number of papers is shown in the last line. Consequently, we can see that INDENG has almost twice as many papers as AQUSCI, which has about twice as many as SOCPSY. ARCHAE is the one with the lowest number of articles, corresponding to 75% of what SOCPSY has. Considering only the papers that were successful in the match between Scopus and WoS, Fig. 5 presents the percentage of papers from each of the research fields over the total.

It was based on the correspondence found among papers from both databases that it was possible to study the existing correlation among the corresponding citation counts. For this purpose, we calculated the Spearman (ρ) correlation coefficient between the number of citations in Scopus and WoS using the StatTools 7.6 (Palisade 2016a). Figure 6 presents ρ results for each research field, in each year of the studied period, revealing a strong correlation ($\rho > 0.90$) between citation counts from both databases. This analysis allows us to understand there isn't an area in which the correlation is the strongest or the weakest since this association varies from year to year. The minimum value, maximum value and

Table 2 Number of papers that were successful in the correspondence between Scopus and WoS per year of publication and from each of the research fields. The last line corresponds to the total of each field

Year	Match INDENG	Match AQUSCI	Match SOCPSY	Match ARCHAE
1987	1111	918	541	146
1988	1204	1217	577	186
1989	1112	1268	555	138
1990	1168	1422	605	154
1991	1333	1486	568	150
1992	1525	1627	559	167
1993	1507	1771	563	208
1994	1663	1961	516	221
1995	1812	2058	523	296
1996	1753	1667	609	324
1997	1547	1954	613	339
1998	1624	1450	622	340
1999	1589	1434	624	367
2000	2041	1443	650	485
2001	2367	2020	619	343
2002	2545	1681	600	401
2003	1507	1040	424	442
2004	745	1082	487	387
2005	1677	1569	604	479
2006	3515	2947	1058	826
2007	3731	3184	1210	950
2008	4876	3135	1360	1004
2009	5529	2920	1525	1105
2010	5015	3161	1644	1210
2011	6302	3312	1708	1328
2012	7225	2559	1876	1397
2013	8629	2737	1944	1693
2014	9615	2836	1849	1704
2015	10,652	2819	1653	1731
2016	12,309	3374	1749	1939
2017	14,246	3551	1872	1936
Total	121,474	65,603	30,307	22,396

mean of ρ are, respectively: 0.9666, 0.9888 and 0.9804 for INDENG; 0.9474, 0.9910 and 0.9792 for AQUSCI; 0.9116, 0.9956 and 0.9852 for SOCPSY; 0.9202, 0.9886 and 0.9759 for ARCHAE. It is observed a lower (but still strong) correlation for papers published in more recent years, as expected; whereas on average, they have not accrued as many citations to stabilize their position in both databases rankings. Correlations below average for ARCHAE, in 1991 ($\rho = 0.9334$), and for AQUSCI for years as 1999 ($\rho = 0.9524$) and 2002 ($\rho = 0.9476$) are mainly caused by the non-identification of some citations in one of the two databases even by articles indexed journals in the database (see some examples of outliers in the "Linear regression section"). A strong Spearman correlation for the number of

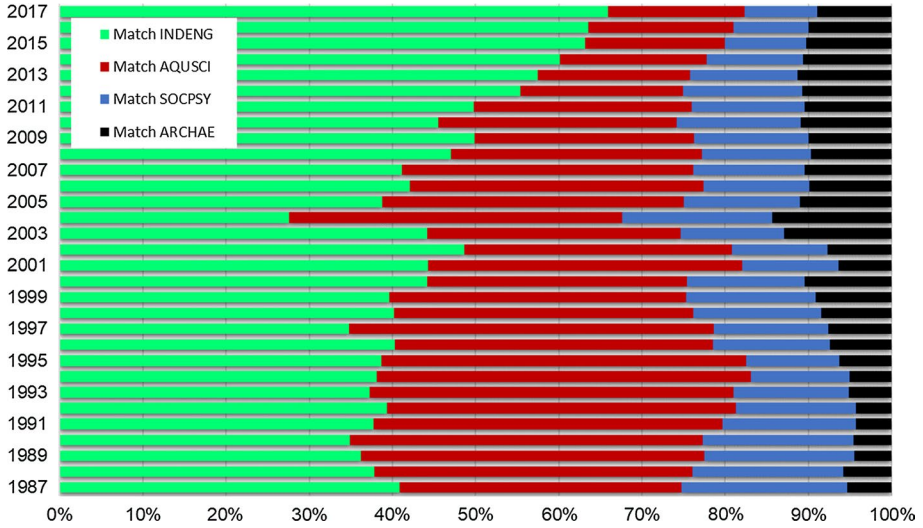


Fig. 5 Percentage of papers from each of the research fields over the total of the year that was successful in the match between Scopus and WoS

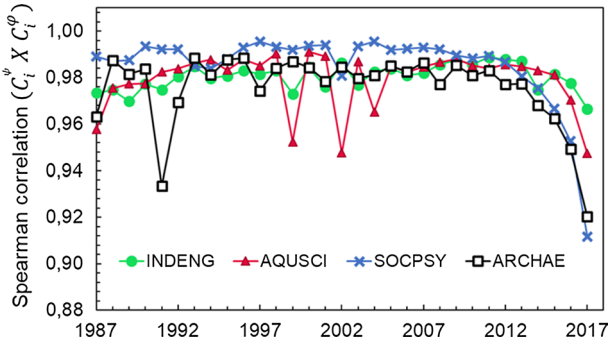


Fig. 6 Spearman correlation coefficient (ρ) between the citation counts of papers belonging in Scopus and WoS for the four fields and for each year in the period 1987–2017

citations was also obtained by Martín-Martín et al. (2018a, b), who studied and compared the coverage of Scopus, WoS, and Google Scholar (GS) databases. These authors investigated the citations from these three databases from 2515 highly cited documents published in 2006 from 8 different large categories (Martín-Martín et al. 2018b) and 2,448,055 citations of 2299 English language highly cited documents from 252 Google Scholar subject categories (Martín-Martín et al. 2018a). The authors were the first to generate systematized results, showing that GS is broader in terms of citations than WoS and Scopus in many areas of research and it is much higher in areas where WoS and Scopus have little coverage, including the Social Science and Humanities (Martín-Martín et al. 2018a). The authors calculated ρ for correlations of citation counts (GS-WoS, GS-Scopus) for several areas including the four, partially studied in this paper. They demonstrated that, for these areas, ρ is in the range 0.90–0.98 which is compatible with the values found in this work.

With a goal to study the linear behavior between C_i^{ψ} and C_i^{φ}, we applied the linear regression model between these two variables, by field and by year studied. From these

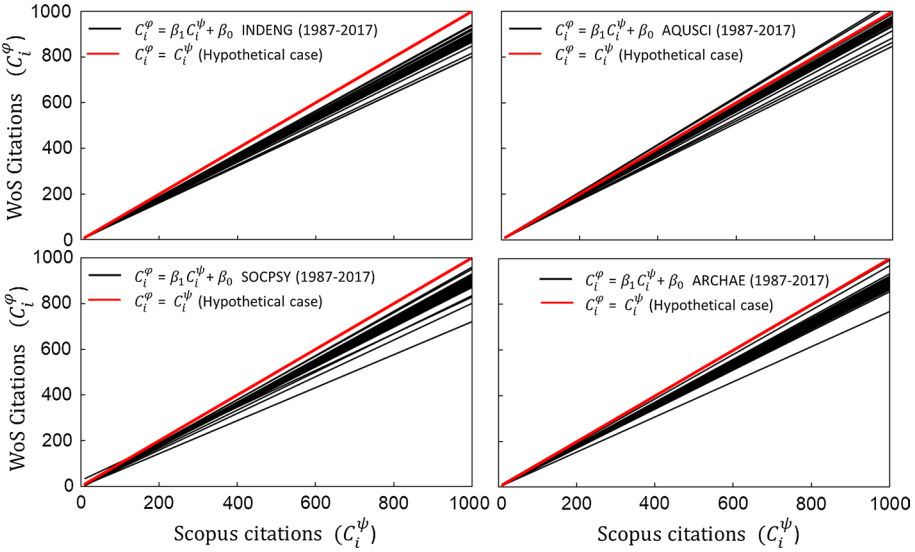


Fig. 7 Linear regressions between the Scopus and WoS citation counts for the four fields studied in the period 1987–2017. The values of β_1 and β_0 depend on the year of publication. Comparing the results obtained with the hypothetical line ($C_i^{\psi} = C_i^{\varphi}$) we see that the number of citations in Scopus tends to be higher than in WoS for almost all situations

linear regressions, it was obtained 31 straight lines corresponding to the 31 years of the 1987–2017 period, for each research field. They are represented in Fig. 7. It is important to notice that the slope of the line depends on the year, demonstrating there is no general pattern for all years, which could establish a unique behavior between these two variables. In other words, the analysis between the citations is not enough to create a method of conversion between the citations of the two databases. As a comparison, Fig. 7 also shows the hypothetical line $C_i^{\psi} = C_i^{\varphi}$ which would represent the case where the number of citations was equal on both databases. Except for two cases corresponding to the first two years of the AQUSCI sample (1987 and 1988), the linear regression analysis showed that in the other 122 datasets, the number of citations in Scopus is systematically higher than WoS. This result is aligned with the study of Martín-Martín et al. (2018a) for highly cited documents in which they showed that Scopus has an average of citations 12.04% higher than WoS. In addition, to perform a comparative analysis between the results obtained in this work and the results obtained by the work of Martín-Martín et al. (2018a), we considered only the data obtained by the authors on the databases we used. That is, only the citations on WoS and Scopus. If we consider only the 2007 citation data, the 12 years between publication and data collection, as in Martín-Martín et al. (2018a), the results of this research would point to a range between 10 and 15% more citations in Scopus compared to WoS for articles with more than 50 citations. This result is compatible with that found by Martín-Martín et al. (2018a) who showed that there is a range of 5% to 22% citations in Scopus more than in WoS for the researched set of papers.

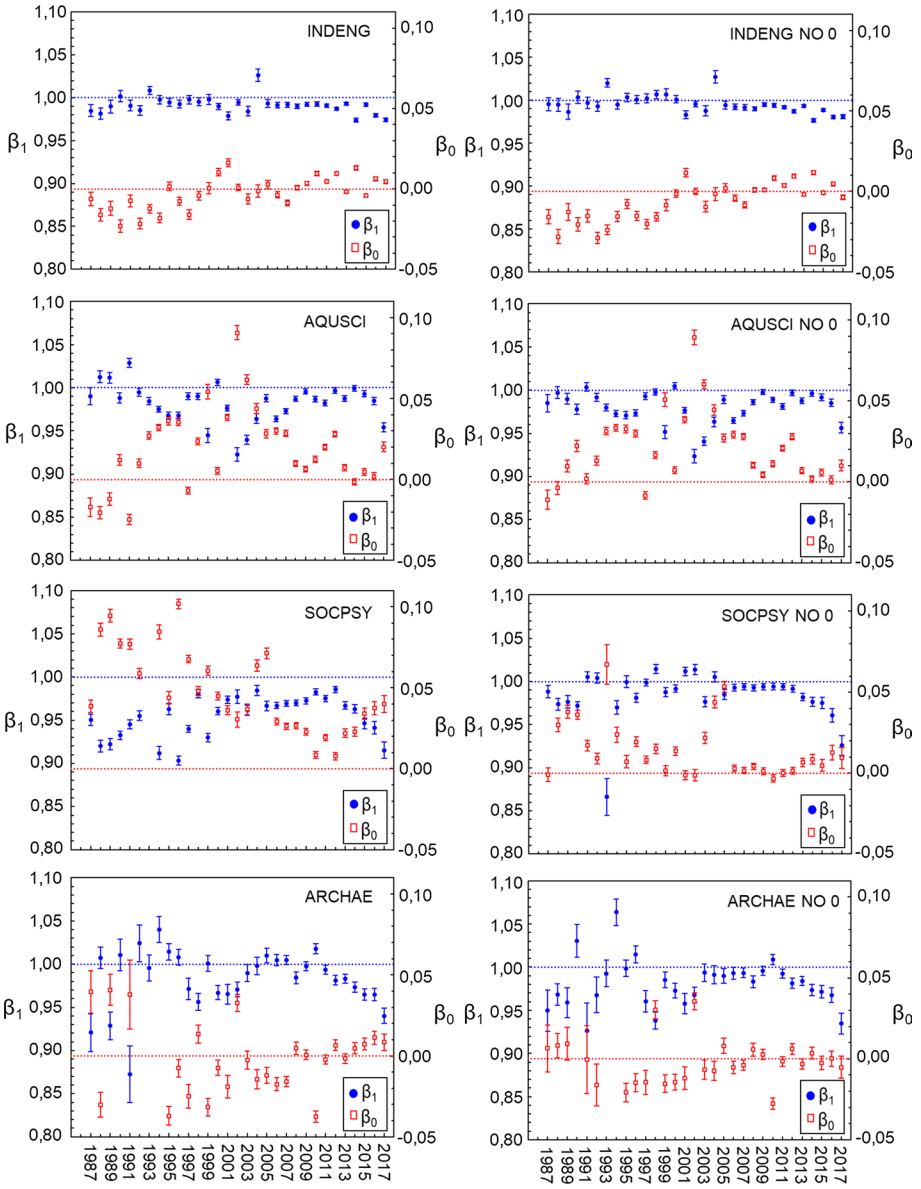


Fig. 8 Coefficients of linear regression over the period 1987–2017 for the citation percentiles for papers indexed in Scopus and WoS databases in the four research fields: industrial and manufacturing engineering—INDENG; aquatic science—AQUSCI, social psychology—SOCPSY and Archaeology—ARCHAE. The four graphics on the right side show the results excluding papers without citations (NO 0), and the four graphics on the left include all the papers. The β_1 values are indicated by the scale at the left axis and the β_0 by the scale on the right axis. The ε_1 and ε_0 values are represented for the error bar

Linear regression

A linear regression model was obtained for the percentiles of articles indexed in both

databases. Figures 8 and 9 show the results of these analyses. Figure 8 shows the results of $\beta_1, \varepsilon_1, \beta_0$ and ε_0 per year and per research field, considering: all the papers, on the left side, and excluding papers with no citations (NO 0), on the right side. Two vertical axis scales are displayed. The right axis corresponds to the values of β_0 represented by the empty squares, and the left axis corresponds to the values of β_1 that is represented by the solid circles. The variations of β_0 and β_1 for INDENG, in more recent years, are lower than in the other fields studied. This can be explained since the number of articles in this field is larger than in the other fields in this period (see Table 2 and Fig. 5).

The same is true for the NO 0 data. In AQUSCI, although the variations of β_0 and β_1 are greater than in INDENG, they go in opposite directions, thus cancelling each other in some measure. In SOCPSY, the results obtained in terms of the closeness of β_0 with 0, and β_1 with 1 are better for NO 0 values. Compared with the other three research fields, in SOCPSY, the number of papers without citations in φ is reasonably different from that in ψ, which may have generated this result. For example, the average in the period 1987–2017 of the percentage of papers without citations (number of papers without citations divided by the total number of papers) considering, respectively, φ and ψ are: 1.03 and 1.19 (in AQUSCI); 3.39 and 3.05 (in INDENG); 4.84 and 1.28 (in SOCPSY); 2.89 and 3.18 (in ARCHAE). Since in ARCHAE the dataset is smaller (see Table 2 and Fig. 5), we expect larger year-to-year variations and, also, larger error bars. These two features are shown in Fig. 8.

Figure 9 provides additional information on the linear regression results showing the strong correlation between the Scopus and the WoS percentile for articles belonging to these two databases. Taking the 124 cases studied (4 fields in 31 years) into account, 95% of R^2 values are above 0.900 and, 55% above 0.965. The R^2 distribution median is 0.966, and the mean R^2 is 0.955. As can be seen from Fig. 9, there are several model outliers. Some outliers are generated by the difference in coverage between the databases. However, this is neither the main reason, nor it generates the largest outliers. In fact, the largest discrepancies in the linear regression of the percentiles have the following reasons: (1) The list of references of one article is different if it is accessed by WoS, or if it accessed by Scopus (Franceschini et al. 2016); (2) Papers appear duplicated on Scopus with different DOIs (Valderrama-Zurián et al. 2015); (3) WoS (or Scopus) did not identify several citations, although the papers citing the article are in its database, with the reference lists citing this article (Franceschini et al. 2016). A quantitative study of these outliers, including an analysis of the best way to handle them in the model proposed by this work, is being conducted by the authors.

In addition to the percentiles of both databases, other variables (e.g., number of papers per year, number of articles without citations, year of publication, number of articles in the sample, and relative percentages) were tested in the multiple linear regression, but the results were not statistically significant.

Monte Carlo simulation

We conducted the Monte Carlo simulation using the software @Risk version 7.6 (Palisade 2016b) to answer the 2nd research question (what is the probability density function of N_i^{φ}, for a specific N_i^{ψ} value?). Figure 10 presents the results, and Table 3 reveals the descriptive statistic of the N_i^{φ} distributions, considering N_i^{ψ} equal to 0.75, 0.80, 0.85, 0.90, and 0.95. We are exposing the results for these five values because, in much bibliometric analysis,

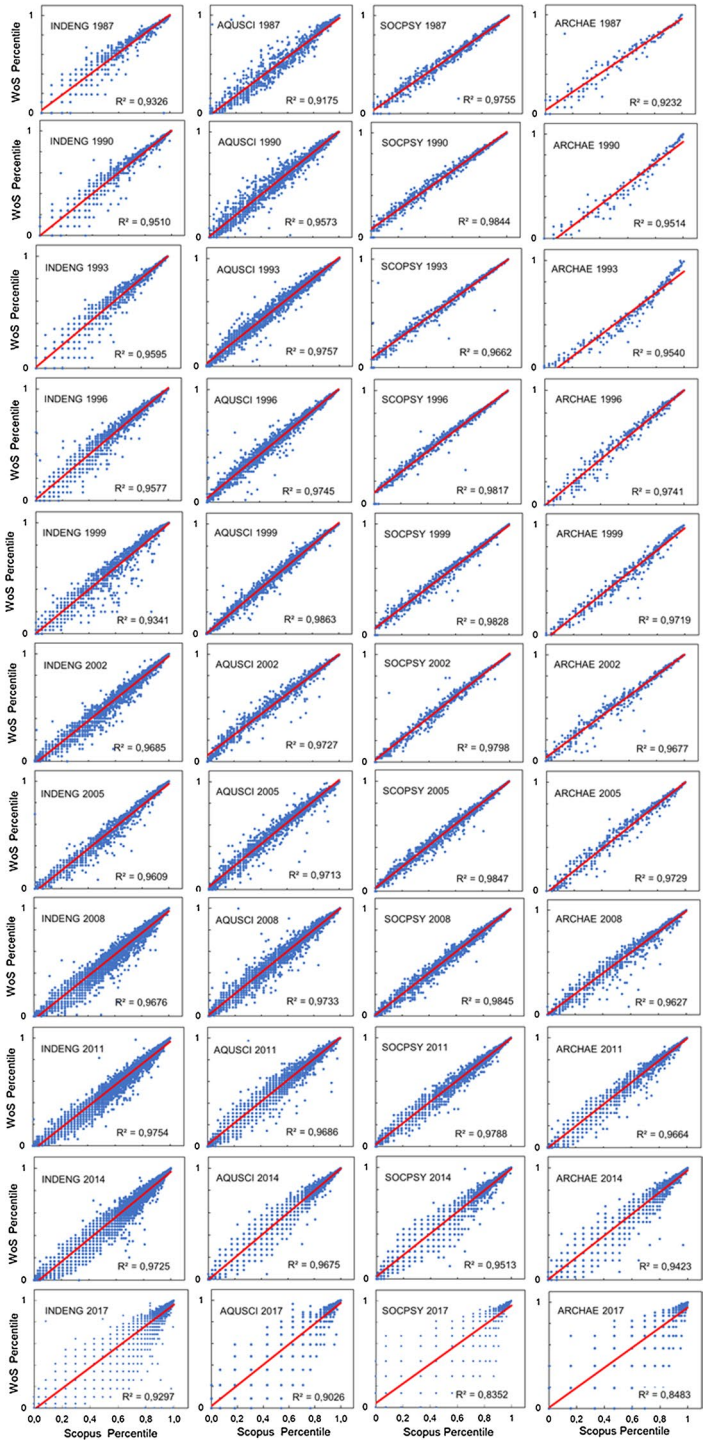
Fig. 9 Linear regressions, uniformly distributed over the period 1987–2017, for the citation percentiles for papers indexed in Scopus and WoS databases in the four research fields: industrial and manufacturing engineering—INDENG, aquatic science—AQUSCI, social psychology—SOCPSY and Archaeology—ARCHAE. Each point in the scatter plot can represent many matches between Scopus and WoS. The R-Squared values are also revealed in the figures

the authors only use the most cited articles, and because that is also the focus of our work. However, similar distributions for smaller values of N_i^{ψ} can be obtained in the same way.

The simulation was conducted for 10,000 iterations for each percentile of each research field, following the assumption that we had 10,000 papers in the $N_i^{\psi} = x$ percentile, and indexed in the two databases. The goal was to find the corresponding percentile values in the other database (N_i^{φ}). The distributions of N_i^{φ} for the INDENG are the narrowest, although the mean is slightly shifted to values smaller than N_i^{ψ}. For the five N_i^{ψ} values, the average of this shift is approximately 1%. That is, (0.75–0.7421)/0.75 = 0.01053; (0.80–0.7915)/0.80 = 0.01063; (0.85–0.8409)/0.85 = 0.0107; (0.90–0.8903)/0.90 = 0.0108; and (0.95–0.9398)/0.95 = 0.0107. For the AQUSCI and SOCPSY the mean, median and mode are very close to the value of N_i^{ψ}, although their distributions are not as narrow as in INDENG. For example, the percentage differences between the mean of the distributions and the N_i^{ψ} values, calculated as above for INDENG, are 0.8%; 0.6%; 0.5%; 0.3%; 0.2%, for AQUSCI, and 0.9%; 0.5%; 0.2%; 0.1%; 0.3%, for SOCPSY. This same value calculated for ARCHAE is around 2.5%. The results of this analysis do not change substantially when we look at the NO 0 curves. When we compare the curves of all papers with NO 0 curves, we can see that the parameter that most changes is the σ of the SOCPSY distributions (Table 3). Indeed, for NO 0, β_0 and β_1 of SOCPSY vary less, year by year, than in the case where the percentile is calculated with all papers (see Fig. 8).

This variation of β_0 and β_1, also generates a secondary peak in the probability density function of SOCPSY. That is, a peak that appears in the curves for percentile values larger than the mode value. Specifically, in some of the early years of the studied period, the β_0 for SOCPSY is in the range 0.08–0.10, i.e. higher than the β_0 for later years (see Fig. 8). Since these high values are not fully compensated by the low values of β_1, it produces a secondary peak that appears in the probability distribution functions of N_i^{ψ}, observed in Fig. 10. The cause is the difference in the number of papers with zero citations between the two databases for the initial years of the studied period in the SOCPSY field. For instance, in 1988 there are 58 papers (7.6%) without citations in WoS and only 8 (1.3%) in Scopus. In 1989, there are 61 (8.2%) papers without citations in WoS and 10 (1.7%) in Scopus. This difference in the number of papers without citations comes, mainly, from existing gaps in the database covers. For example, the journal Political Psychology, with some articles without citations between 1988 and 1989, indexed by WoS in 1985, was only indexed by Scopus in 1996. Similar effects are also present in other field distributions, however, much less markedly.

Table 3 also presents the value of the percentile 5% and 95% for both cases (all papers and NO 0). Finally, to better exemplify the results of this study, Table 4 presents the probability that N_i^{φ} is between N_i^{ψ} - 0.02 and $N_i^{\psi} + 0.02$ for all distributions shown in Fig. 10. For INDENG and AQUSCI, considering all the curves presented, this probability is between 82 and 92%, revealing that, for a deviation of 0.02 in the percentile value, the level of confidence is high. For example, for a paper that has a percentile of 0.85 on the ψ database, the probability of it having a percentile in the range 0.83–0.87 on the φ are 90.00%; 87.80%; 74.4% and 55.9%, for INDENG, AQUSCI, SOCPSY, and ARCHAE, respectively.


INDENG 1987 R² = 0,9326
AQUSCI 1987 R² = 0,9175
SOCPSY 1987 R² = 0,9755
ARCHAE 1987 R² = 0,9232
INDENG 1990 R² = 0,9510
AQUSCI 1990 R² = 0,9573
SOCPSY 1990 R² = 0,9844
ARCHAE 1990 R² = 0,9514
INDENG 1993 R² = 0,9595
AQUSCI 1993 R² = 0,9757
SCOPSY 1993 R² = 0,9662
ARCHAE 1993 R² = 0,9540
INDENG 1996 R² = 0,9577
AQUSCI 1996 R² = 0,9745
SCOPSY 1996 R² = 0,9817
ARCHAE 1996 R² = 0,9741
INDENG 1999 R² = 0,9341
AQUSCI 1999 R² = 0,9863
SOCPSY 1999 R² = 0,9828
ARCHAE 1999 R² = 0,9719
INDENG 2002 R² = 0,9685
AQUSCI 2002 R² = 0,9727
SOCPSY 2002 R² = 0,9798
ARCHAE 2002 R² = 0,9677
INDENG 2005 R² = 0,9609
AQUSCI 2005 R² = 0,9713
SCOPSY 2005 R² = 0,9847
ARCHAE 2005 R² = 0,9729
INDENG 2008 R² = 0,9676
AQUSCI 2008 R² = 0,9733
SOCPSY 2008 R² = 0,9845
ARCHAE 2008 R² = 0,9627
INDENG 2011 R² = 0,9754
AQUSCI 2011 R² = 0,9686
SOCPSY 2011 R² = 0,9788
ARCHAE 2011 R² = 0,9664
INDENG 2014 R² = 0,9725
AQUSCI 2014 R² = 0,9675
SOCPSY 2014 R² = 0,9513
ARCHAE 2014 R² = 0,9423
INDENG 2017 R² = 0,9297
AQUSCI 2017 R² = 0,9026
SOCPSY 2017 R² = 0,8352
ARCHAE 2017 R² = 0,8483
WoS Percentile
Scopus Percentile
0 0,2 0,4 0,6 0,8 1,0

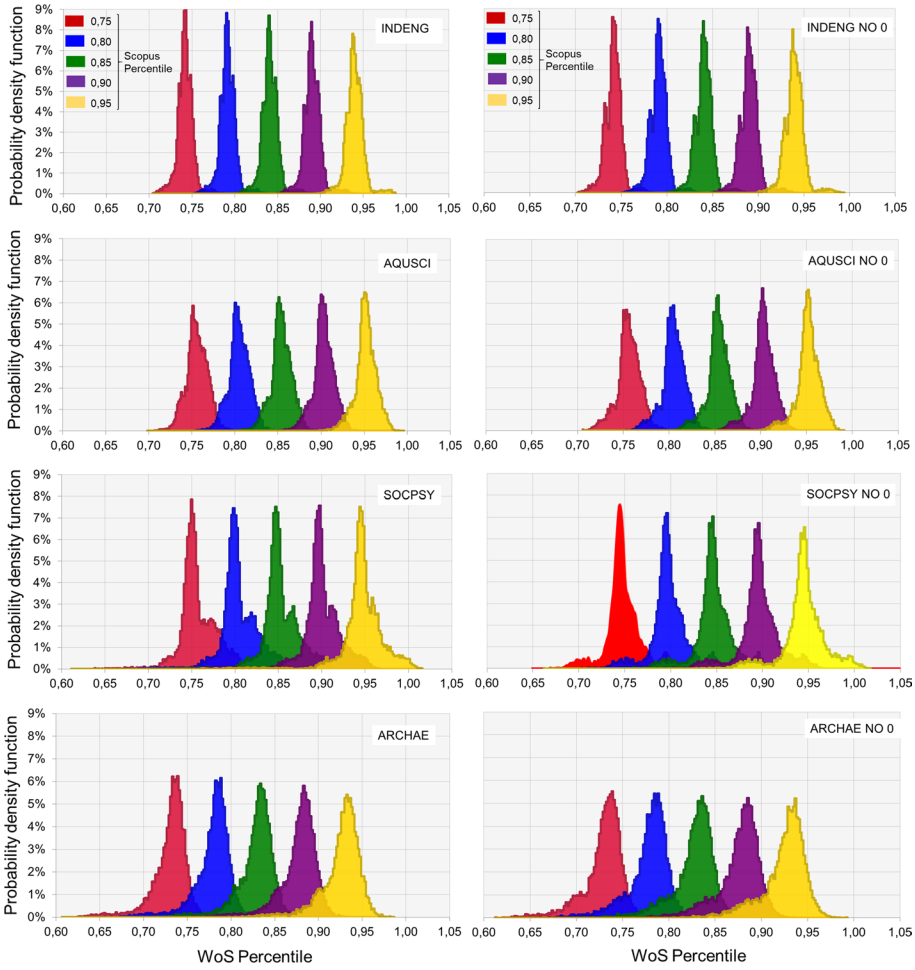


Fig. 10 Probability Density Function of the N_i^{φ} for N_i^{ψ} equal to 0.75, 0.80, 0.85, 0.90 and 0.95

Since the citation counts of missing articles on a database are estimated by the model, we must include these articles on the database to calculate again all percentiles, since previously the percentiles were calculated without these articles. In this way, we will ensure that each article will have associated a normalized indicator that represents its impact on both databases. In order to, approximately, evaluate the effect of this procedure on the previously percentiles, we performed the calculation using, as an example, the data of 2009, following the next steps:

- For a missing paper in WoS, but which belong to Scopus, we estimated the WoS percentile equal to the percentile obtained by this article in Scopus (N_i^{ψ}). Here we consider the results generated by the model presented in this study that show that these percentiles are very close;

Table 3 Statistical parameters of the probability density function of N_i^{φ}

	N_i^{ψ}	Probability density function of N_i^{φ}					Probability density function of N_i^{φ}(NO 0)				
		μ	$\tilde{x}$	σ	5%	95%	μ	$\tilde{x}$	σ	5%	95%
INDENG	0.75	0.7421	0.7420	0.0090	0.7275	0.7550	0.7414	0.7418	0.0093	0.7270	0.7541
	0.80	0.7915	0.7914	0.0093	0.7770	0.8047	0.7909	0.7912	0.0095	0.7764	0.8039
	0.85	0.8409	0.8407	0.0096	0.8264	0.8545	0.8405	0.8406	0.0098	0.8257	0.8538
	0.90	0.8903	0.8900	0.0099	0.8758	0.9042	0.8900	0.8900	0.0100	0.8749	0.9038
	0.95	0.9398	0.9394	0.0102	0.9250	0.9539	0.9396	0.9395	0.0103	0.9242	0.9536
AQUSCI	0.75	0.7557	0.7554	0.0125	0.7342	0.7754	0.7559	0.7561	0.0126	0.7326	0.7754
	0.80	0.8048	0.8047	0.0123	0.7830	0.8244	0.8049	0.8052	0.0125	0.7814	0.8244
	0.85	0.8538	0.8538	0.0121	0.8319	0.8733	0.8539	0.8542	0.0125	0.8303	0.8734
	0.90	0.9029	0.9030	0.0120	0.8808	0.9222	0.9029	0.9032	0.0125	0.8792	0.9223
	0.95	0.9519	0.9520	0.0121	0.9293	0.9711	0.9519	0.9523	0.0126	0.9281	0.9712
SOCPSY	0.75	0.7567	0.7531	0.0205	0.7279	0.7932	0.7488	0.7472	0.0194	0.7124	0.7870
	0.80	0.8042	0.8015	0.0218	0.7731	0.8413	0.7979	0.7967	0.0203	0.7582	0.8365
	0.85	0.8517	0.8499	0.0233	0.8187	0.8893	0.8470	0.8463	0.0212	0.8043	0.8860
	0.90	0.8992	0.8983	0.0250	0.8646	0.9377	0.8962	0.8958	0.0222	0.8503	0.9359
	0.95	0.9467	0.9467	0.0268	0.9103	0.9861	0.9453	0.9453	0.0232	0.8967	0.9853
ARCHAE	0.75	0.7298	0.7338	0.0200	0.6930	0.7531	0.7288	0.7330	0.0210	0.6882	0.7542
	0.80	0.7789	0.7829	0.0204	0.7419	0.8031	0.7778	0.7823	0.0215	0.7351	0.8040
	0.85	0.8280	0.8321	0.0209	0.7899	0.8530	0.8268	0.8315	0.0220	0.7819	0.8539
	0.90	0.8771	0.8814	0.0214	0.8373	0.9030	0.8758	0.8808	0.0226	0.8294	0.9037
	0.95	0.9262	0.9307	0.0221	0.8850	0.9530	0.9249	0.9300	0.0233	0.8770	0.9534

Table 4 Confidence level of N_i^{φ} for N_i^{ψ} equal to 0.75, 0.80, 0.85, 0.90 and 0.95 considering an error of 0.02

N_i^{ψ}	N_i^{φ}	Confidence level							
		INDENG		AQUSCI		SOCPSY		ARCHAE	
		All (%)	NO 0 (%)	All (%)	NO 0 (%)	All (%)	NO 0 (%)	All (%)	NO 0 (%)
0.75	0.75 ± 0.02	92.10	89.80	85.30	83.70	70.00	79.80	60.90	57.20
0.80	0.80 ± 0.02	91.30	88.00	86.90	85.20	72.00	78.60	58.60	55.10
0.85	0.85 ± 0.02	90.00	85.90	87.80	86.30	74.40	77.50	55.90	53.50
0.90	0.90 ± 0.02	87.90	84.00	88.40	87.40	75.60	76.00	53.70	51.70
0.95	0.95 ± 0.02	85.20	82.20	88.70	88.00	75.50	74.70	51.60	49.70

- We searched WoS for the article that has the percentile value, N_i^{φ}., closest to the percentile obtained by the previous step, and assign to the missing article the percentile N_i^{φ};
- We included this article in the WoS database, with C_i^{φ} being the same as the article found in the previous step;
- We performed the prious steps for all Scopus articles that are missing on WoS, and thus completed the WoS database, now defined by φ';
- We recalculated the percentiles ($N_i^{\varphi'}$) considering the values $C_i^{\varphi'}$ of all articles;

Table 5 Percentile recalculation results when we included missing articles in the database. Calculations were performed for papers published in 2009

	INDENG	AQUSCI	SCOPSY	ARCHAE
P^{φ}	5579	3135	1575	1161
ΔP^{φ}	641	213	91	45
$\overline{\Delta N_i^{\varphi}}$	0.0013	0.0016	0.0026	0.0032
$\overline{\Delta N_i^{\varphi}/N_i^{\varphi\prime}}$	1.10%	0.60%	1.63%	2.52%

- For each article, we calculated: $\Delta\ N_i^{\varphi} = \left| N_i^{\varphi'} - N_i^{\varphi} \right|$.

Table 5 presents the recalculation results considering the 2009 publications. It indicates the following values: the number of papers in WoS (P^{φ}); the number of missing papers that were included during that year (ΔP^{φ}); the arithmetic average of $\Delta\ N_i^{\varphi}$ $\left(\overline{\Delta N_i^{\varphi}}\right)$; and the average of percentage differences $\left(\overline{\Delta N_i^{\varphi}/N_i^{\varphi'}}\right)$. These results show that there was a change to the third decimal place of the percentile value. In addition, the recalculation generates less variation in the two fields that have the largest number of articles and the smallest σ Probability Density Function of N_i^{φ}. Indeed, the tendency in these cases is that the inclusion of the missing articles causes fewer changes to the original configuration values.

Model performance assessment

In order to assess the performance of the methodology developed in this study, we divided the matched papers of each field into three parts. Part {A}: approximately 75% of randomly selected papers. In this part, papers are considered belonging to both databases. Part {B}: approximately 50% of the remaining papers also randomly selected. In this part, papers are considered missing out in Scopus. Part {C}: the rest of the papers. This last part represents papers missing out in WoS. The test consists of using the WoS citations of the papers in {B} to estimate their citations in Scopus, and vice-versa for {C}. Therefore, as the real values are known, we can use them to know whether they are in concordance with the estimated values.

The procedure to prepare the sample was the following: (1) we used the data of seven different years with the same time interval between them (1987, 1992, 1997, 2002, 2007, 2012 and 2017); (2) we calculated the percentile of the papers in both databases, classified

Table 6 Lin's concordance correlation coefficient (ρ_C) and its 95% of confidence intervals ($\rho_{C,\text{lower}} - \rho_{C,\text{ upper}}$) for WoS citations estimated by Scopus citations and their real values, ρ_C ($S \rightarrow W$); and Scopus citations estimated by WoS and their real values, $\rho_C(W \rightarrow S)$

	INDENG	AQUSCI	SOCPSY	ARCHAE
$\rho_C(S \rightarrow W)$	0.9884	0.9818	0.9652	0.9851
$\rho_{C,\text{lower}}(S \rightarrow W)$	0.9877	0.9801	0.9619	0.9827
$\rho_{C,\text{upper}}(S \rightarrow W)$	0.9891	0.9833	0.9681	0.9872
$N(S \rightarrow W)$	3990	1935	1089	668
$\rho_C(W \rightarrow S)$	0.9857	0.9844	0.9866	0.9647
$\rho_{C,\text{lower}}(W \rightarrow S)$	0.9848	0.9831	0.9851	0.9600
$\rho_{C,\text{upper}}(W \rightarrow S)$	0.9866	0.9856	0.9880	0.9689
$N(W \rightarrow S)$	3991	1935	1093	668
N(total)	31,930	15,472	7271	5337

by year of publication, considering papers from {B} ({C}) not included in the calculation of Scopus (WoS) percentiles; (3) for the papers of {B} ({C}) we considered the Scopus (WoS) percentile as the closest one to its WoS (Scopus) percentile in the same year of publication; (4) we assigned the number of citations corresponding to this percentile in the database where the paper was missing.

We used Lin's concordance correlation coefficient (ρ_C), and its relative 95% confidence intervals (Lin 1989), to compare the real values with the estimated values for each research field. ρ_C. is a measure to test how well bivariate pairs of observations conform relatively to a gold standard or, in this case, to the real values. Table 6 presents the results of ρ_C and the 95% of confidence intervals defined by the lower bound ($\rho_{C,\text{lower}}$) and the upper bound ($\rho_{C,\text{upper}}$) for two cases: (1) concordance between t WoS citations estimated by Scopus citations and their real values, defined by ρ_C ($S \rightarrow W$) and; (2) Scopus citations estimated by WoS and their real values, defined by $\rho_C(W \rightarrow S)$. In Table 6, N(total) is the number of papers included in the test in each field (number of matched papers given in Table 2); $N(S \rightarrow W)$ is the number of papers in Scopus used to estimate the WoS percentiles and; vice-versa to $N(W \rightarrow S)$. The results reveal a substantial agreement between the estimated and real values. This conclusion is based on the interpretation given by McBride (2005) which suggests the following scale for $\rho_{C,\text{lower}}$: Almost perfect (>0.99); Substantial (0.95–0.99); Moderate (0.90–0.95); and Poor (<0.90). The lowest values of $\rho_{C,\text{lower}}$ came from SOCPSY and ARCHAE, where the confidence level to adopt the relation $N_i^{\psi} \approx N_i^{\varphi}$ is lower than in AQUSCI and INDENG (see Table 3), however still it is >0.95, established, by McBride (2005), as a substantial concordance.

Final remarks

In order to answer the research questions introduced earlier in this article, we collected a dataset of 326,345 papers of four different research fields, using the top 10% source titles of each field, and conducted a longitudinal study of the citation counts in the two widely used scientific bibliometric databases: Scopus and WoS. In this paper, we examine whether the percentile- and stochastic-based approach are effective for converting citation counts between these two databases.

To answer [RQ1] (What are the differences between percentiles of citation counts, in WoS and Scopus databases, for the same papers?), a predictive analysis was conducted using a linear regression model. The results of this study show that the dependence between the citation count percentiles calculated in Scopus and the citation count percentile calculated in WoS can be described by a 2-parameter (β_0. and β_1) linear model. Additionally, the linear regression analysis reveals that $\beta_0 \approx 0$ and $\beta_1 \approx 1$,. which demonstrates that the differences between the percentiles of the citation number, in WoS and Scopus, are close to zero (see Fig. 8). This evidence is compatible with the strong Spearman correlatio (ρ) between citation counts in Scopus and WoS found in this work in the period 1987–2017 and, also, is aligned with the results fou by Martín-Martín et al. (2018b) using highly cited documents published in 2006. The high correlation between GS and WoS, and between GS and Scopus for the citation counts (Martín-Martín et al. 2018b) shows that between Scopus and WoS ρ may also be high and therefore, the estimate that the percentiles are close in these two databases needs to be evaluated. Thus, considering the standard errors associated with the linear coefficients, the differences between them over the years, and the differences for the research fields studied, we examined these estimates in more detail.

The answer to the [RQ2] (Given a specific confidence level and the percentile of the citation counts of a paper on a database, what is the confidence interval of the percentile on the other database?) drive this issue. We conducted the Monte Carlo simulation with this purpose, since applying this method we are expanding the limited dataset used and leading to results that can be statistically analyzed. The outcomes can be summarized by the probability density functions of the percentiles. The results presented in Table 3 show that if $N_i^{\psi} = x$, the mean and the median of the probability distributions of N_i^{φ} are, nearly, equal to x, too. Additionally, for INDENG and AQUSCI, if $N_i^{\psi} = x$, the probability that $N_i^{\varphi} = x \pm 0,02$ is in the range of 82.2–92.1% (see Table 4). For SOCPSY and ARCHAE, where the number of articles is smaller, the distribution is not so narrow, and consequently, the uncertainty of the assumption that the two percentiles are equal is larger. We have demonstrated in this paper that the stochastic comparison between percentiles of different databases is able to predict the percentile of an article on a database to which it does not belong.

This consequence supports the response to [RQ3] (How is it possible to merge the articles indexed in two different databases and still use the article citation-based impact information in a unified manner?). In fact, using this approach as a normalization method, we can combine articles from different databases and, effectively, compare them by this criterion, since it maintains the original impact information of the article. In addition, we use the citation percentile also for time-normalization purposes, along with the database-normalization, thus avoiding the use of different impact measures in the same process of comparison.

The observed data of the time-evolution of the citations over the 31 years are compatible with the evolution of the citation-percentile correspondence presented by Petersen et al. (2019). The authors used a 10-year citation window (C_{10}) because they consider this period to be the most relevant for papers to accumulate citations, and showed that only this normalization is not enough, because for the same percentile, C_{10} increases over the years. In our case, as we consider all the citations, we can expect that the percentile curves grow until around 2008 and then begin to decline because the papers have not yet had 10 years to accumulate citations. Most of the curves presented in Fig. 4 illustrate this behavior.

Conclusions

Scopus and WoS are the most relevant when it comes to scientific research titles. However, these databases index journals using different criteria, so articles that belong to one database may not belong to the other. Consequently, the number of citations of a paper depends on the database used, which hinders the comparison of the impact of papers from different databases. This problem affects bibliometric studies because they need to use both databases to consolidate the knowledge of a given research field, provide state-of-the-art and guide future researches. In addition, this problem also affects the measurement of bibliometric indicators, since the calculation is based on the number of articles published and the number of citations, and both depend on the database used (Bornmann and Leydesdorff 2018). Therefore, with the conversion of citations between databases, the shortcoming caused by the difficulty of using two databases in bibliometric performance indicators can be minimized. In fact, this case is even worse because it influences investments in research institutions and groups, and in the evaluation of the researchers themselves.

To address this problem, we introduce, in this paper, a method that combines the use of percentiles as a normalization tool with a Monte Carlo simulation. With this method, it is possible to convert the citation counts of articles between Scopus and WoS and to predict, within an error, the citation impact of a missing paper on one of those databases, based on the citation impact on the other database. This allows the combined use of the two databases, reducing biases and improving the coverage and accuracy of both bibliometric studies and scientific performance measurements.

The percentiles are found for each article, within each database, field and publication year. The results indicate that the relationships between the percentiles in each database, for each set k of year and research field, can be modeled as linear, through linear regressions with parameters β_{1k} and β_{0k}. In most cases, β_{1k} is close to one, and β_{0k} is close to zero, suggesting the percentiles in both databases are likely the same.

This analysis revealed the existence of outliers caused by errors in Scopus and WoS citation counts such as "cited article omitted from a cited-article list" (Franceschini et al. 2016), whose qualitative and quantitative aspects have been the subject of previous studies and whose aspects related to this model will be analyzed in an upcoming paper by the authors. Although these incorrect citation counts are not significant to change results for studies like ours, that use a large database, it may be relevant for studies using relatively smaller data sets, or even for the specific assessment of researchers.

Based on the linear regression results between percentiles, we applied the Monte Carlo method to calculate the probability of an article being in the percentile x in one database and in percentile between x-Δx and $x+\Delta x$ in the other database. When an article is not found in one of the databases, one can obtain the probability distribution of the article's percentile in that database, based on the simulation results and the (known) percentile in the other database. Having a $[x-\Delta x, x+\Delta x]$ interval for that percentile, it is possible to estimate the corresponding $[C_i-\Delta C_i, C_i+\Delta C_i]$ interval for the missing number of citations of the article. In this study we used the normal distribution to describe the residuals of linear regression of percentiles. However, the application of this model to a broader data set, in terms of fields, could be able to analyze others different probability distribution functions, also appropriated. Finally, in the model performance assessment, Lin's coefficient showed a substantial concordance between the number of citations estimated by this model and the real values.

Our results were obtained using papers in English, published in the top 10% source titles of Scopus, on four fields of research and therefore cannot be directly generalized. In fact, the percentile of citations strongly depends on the characteristics of the journals and on the type and language of the publications (among other characteristics) on the dataset and, therefore, the use of a broader set of publications and journals in each area can generate other results. Furthermore, we use the Scopus and WoS databases, which means that for other databases with different technical features (e.g., Google Scholar) the methods employed need to be adapted.

A relevant issue to be subject of future works is the development of a single database for several fields combining Scopus and WoS in a large period, using the results obtained by this study, and the recalculation method (see "Monte Carlo simulation section") of the normalized indicators when the missing papers are also included in the database with its estimated citations.

Acknowledgements The authors are grateful to two anonymous reviewers for their valuable recommendations to improve the manuscript. We acknowledge the support of ERDF—European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation—COMPETE 2020

Programme and the Portuguese funding agency, FCT—Fundação para a Ciência e a Tecnologia within project POCI-01-0145-FEDER-031821.

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.

Abramo, G., D'Angelo, C. A., & Felici, G. (2019). Predicting publication long-term impact through a combination of early citations and journal impact factor. *Journal of Informetrics,13*(1), 32–49.

Adam, A., Ras, R., Bhattu, A. S., Raman, A., & Perera, M. (2017). "Researching the research" in prostate cancer: A comparative bibliometric analysis of the top 100 cited articles in the field of prostate cancer. *Current Urology,11*(1), 26–35.

Alajmi, B., & Alhaji, T. (2018). Mapping the field of knowledge management: Bibliometric and content analysis of journal of information and knowledge management for the period from 2002 to 2016. *Journal of Information and Knowledge Management,17*(3), 1850027.

Bohl, M. A., Turner, J. D., Little, A. S., Nakaji, P., & Ponce, F. A. (2017). Assessing the relevancy of "Citation Classics" in neurosurgery: Part II foundational papers in neurosurgery. *World Neurosurgery,104*, 939–966.

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 and Technology,64*(3), 587–595.

Bornmann, L. (2014). H-Index research in scientometrics: A summary. *Journal of Informetrics,8*(3), 749–750.

Bornmann, L., & Leydesdorff, L. (2017). Skewness of citation impact data and covariates of citation distributions: A large-scale empirical analysis based on Web of Science data. *Journal of Informetrics,11*(1), 164–175.

Bornmann, L., & Leydesdorff, L. (2018). Count highly-cited papers instead of papers with h citations: use normalized citation counts and compare "like with like"! *Scientometrics,115*(2), 1119–1123.

Bornmann, L., & Marx, W. (2014). How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations. *Scientometrics,98*(1), 487–509.

Bornmann, L., & Wohlrabe, K. (2019). Normalisation of citation impact in economics. *Scientometrics,120*(2), 841–884.

Bornmann, L., Leydesdorff, L., & Wang, J. (2013). Which percentile-based approach should be preferred for calculating normalized citation impact values? An empirical comparison of five approaches including a newly developed citation-rank approach (P100). *Journal of Informetrics,7*(4), 933–944.

Brito, R., & Rodríguez-Navarro, A. (2018). Research assessment by percentile-based double rank analysis. *Journal of Informetrics,12*(1), 315–329.

Chen, J. S., Hubbard, S., & Rubin, Y. (2001). Estimating the hydraulic conductivity at the South Oyster Site from geophysical tomographic data using Bayesian techniques based on the normal linear regression model. *Water Resources Research,37*(6), 1603–1613.

Darko, A., & Chan, A. P. C. (2016). Critical analysis of green building research trend in construction journals. *Habitat International,57*, 53–63.

Davies, J., Fortin, N. M., & Lemieux, T. (2017). Wealth inequality: Theory, measurement and decomposition. *Canadian Journal of Economics,50*(5), 1224–1261.

De Groote, S. L., & Raszewski, R. (2012). Coverage of Google Scholar, Scopus, and Web of Science: A case study of the h-index in nursing. *Nursing Outlook,60*(6), 391–400.

Demarest, B., Freeman, G., & Sugimoto, C. R. (2014). The reviewer in the mirror: examining gendered and ethnicized notions of reciprocity in peer review. *Scientometrics,101*(1), 717–735.

Dokur, M., & Uysal, E. (2018). Top 100 cited articles in traumatology: A bibliometric analysis. *Turkish Journal of Trauma and Emergency Surgery,24*(4), 294–302.

Fairclough, R., & Thelwall, M. (2015). More precise methods for national research citation impact comparisons. *Journal of Informetrics,9*(4), 895–906.

Filardo, G., da Graca, B., Sass, D. M., Pollock, B. D., Smith, E. B., & Martinez, M. A.-M. (2016). Trends and comparison of female first authorship in high impact medical journals: observational study (1994–2014). *BMJ (Clinical Research Ed.),352*, i847.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of Informetrics,10*(4), 933–953.
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.
Goel, P. K., & DeGroot, M. H. (1980). Only normal distributions have linear posterior expectations in linear regression. *Journal of the American Statistical Association,75*(372), 895–900.
Gómez-Núñez, A. J., Vargas-Quesada, B., de Moya-Anegón, F., & Glänzel, W. (2011). Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis. *Scientometrics,89*(3), 741–758.
González-Betancor, S. M., & Dorta-González, P. (2017). An indicator of the impact of journals based on the percentage of their highly cited publications. *Online Information Review,41*(3), 398–411.
Harzing, A.-W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics,106*(2), 787–804.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America,102*(46), 16569–16572.
Hnaien, F., Yalaoui, F., Mhadhbi, A., & Nourelfath, M. (2016). A mixed-integer programming model for integrated production and maintenance. *IFAC-PapersOnLine*, *49*(12), 556–561.
Jann, B. (2016). Assessing inequality using percentile shares. *Stata Journal,16*(2), 264–300.
Jiang, Z., Schrank, C., Mariethoz, G., & Cox, M. (2013). Permeability estimation conditioned to geophysical downhole log data in sandstones of the northern Galilee Basin, Queensland: Methods and application. *Journal of Applied Geophysics,93*, 43–51.
Kosteas, V. D. (2018). Predicting long-run citation counts for articles in top economics journals. *Scientometrics,115*, 1395–1412.
Laengle, S., Merigó, J. M., Miranda, J., Słowiński, R., Bomze, I., Borgonovo, E., et al. (2017). Forty years of the European Journal of Operational Research: A bibliometric overview. *European Journal of Operational Research,262*(3), 803–816.
Li, J., Burnham, J. F., Lemley, T., & Britton, R. M. (2010). Citation analysis: Comparison of Web of Science®, Scopus™, SciFinder®, and Google Scholar. *Journal of Electronic Resources in Medical Libraries,7*(3), 196–217.
Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics,45*(1), 255–268.
Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E. D. (2018a). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics,12*(4), 1160–1177.
Martín-Martín, A., Orduna-Malea, E., & López-Cózar, E. (2018b). Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison. *Scientometrics,116*(3), 2175–2188.
McBride, G. B. (2005). A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. In *NIWA client report, HAM2005-062*.
Milojević, S., Radicchi, F., & Bar-Ilan, J. (2017). Citation success index—An intuitive pair-wise journal comparison metric. *Journal of Informetrics,11*(1), 223–231.
Mishra, A. K. (2018). Household income inequality and income mobility: Implications towards equalizing longer-term incomes in India. *International Economic Journal,32*(2), 271–290.
Mishra, A. K., & Kumar, A. (2018). What lies behind income inequality and income mobility in India? Implications and the way forward. *International Journal of Social Economics,45*(9), 1369–1384.
Moed, H. F., Bar-Ilan, J., & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. *Journal of Informetrics,10*(2), 533–551.
Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics,106*(1), 213–228.
Palisade. (2016a). StatTools 7.6, Ithaca NY: Palisade Corporation. www.palisade.com
Palisade. (2016b). @Risk 7.6, Ithaca NY: Palisade Corporation. www.palisade.com
Pan, R. K., Petersen, A. M., Pammolli, F., & Fortunato, S. (2018). The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics,12*(3), 656–678.
Pech, G., & Delgado, C. (2019). Method for comparison of the number of citations from papers in different databases. In *17th International conference and Proceedings on scientometrics and informetrics, ISSI 2019, No. 2,* (pp. 2419–2429).
Pech, G., Delgado, C., & Vieira, N. (2019). Percentile citation-based method for screening the most highly cited papers in longitudinal bibliometric studies and systematic literature reviews. In *12th Annual conference and Proceedings of the EuroMed Academy of Business, EUROMED 2019*, (pp. 911–923).

Pesta, B. J. (2018). Bibliometric analysis across eight years 2008–2015 of Intelligence articles: An updating of Wicherts (2009). *Intelligence,67*, 26–32.
Petersen, A. M., Pan, R. K., Pammolli, F., & Fortunato, S. (2019). Methods to account for citation inflation in research evaluation. *Research Policy,48*(7), 1855–1865.
Radicchi, F., & Castellano, C. (2012). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS ONE,7*(3), e33833.
Rodriguez, M. A., & Pepe, A. (2008). On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Informetrics,2*(3), 195–201.
Rodríguez-Navarro, A., & Brito, R. (2018). Double rank analysis for research assessment. *Journal of Informetrics,12*(1), 31–41.
Rousseau, R. (2007). The influence of missing publications on the Hirsch index. *Journal of Informetrics,1*(1), 2–7.
Santiago, A. M. A. et al. (2018). Relatório de Autoavaliação Institucional da Universidade do Estado do Rio de Janeiro - Comissão Própria de Avaliação da UERJ—CPA—ano base 2017 (Institutional Self-Evaluation Report of the Rio de Janeiro State University—UERJ Own Evaluation Committee—base year 2017), (p.15).
Schulz, J. (2016). Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analyses. *Scientometrics,107*(3), 1283–1298.
Shang, G., Saladin, B., Fry, T., & Donohue, J. (2015). Twenty-six years of operations management research (1985–2010): Authorship patterns and research constituents in eleven top rated journals. *International Journal of Production Research,53*(20), 6161–6197.
Spanos, A. (1995). On normality and the linear regression model. *Econometric Reviews,14*(2), 195–203.
Stegehuis, C., Litvak, N., & Waltman, L. (2015). Predicting the long-term citation impact of recent publications. *Journal of Informetrics,9*(3), 642–657.
Thelwall, M. (2016). 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.
Thelwall, M. (2019). The influence of highly cited papers on field normalised indicators. *Scientometrics,118*(2), 519–537.
Valderrama-Zurián, J.-C., Aguilar-Moya, R., Melero-Fuentes, D., & Aleixandre-Benavent, R. (2015). A systematic analysis of duplicate records in Scopus. *Journal of Informetrics,9*(3), 570–576.
Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics,10*(2), 365–391.
Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology,64*(2), 372–379.
Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics,94*(3), 851–872.
Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics,10*(2), 347–364.
Wang, Y., Zeng, A., Fan, Y., & Di, Z. (2019). Ranking scientific publications considering the aging characteristics of citations. *Scientometrics,120*(1), 155–166.
Wildgaard, L., Schneider, J. W., & Larsen, B. (2014). A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics,101*(1), 125–158.
Yamashita, Y., & Okubo, Y. (2006). Patterns of scientific collaboration between Japan and France: Intersectoral analysis using Probabilistic Partnership Index (PPI). *Scientometrics,68*(2), 303–324.
Yeung, A. W. K., Heinrich, M., & Atanasov, A. G. (2018). Ethnopharmacology-A bibliometric analysis of a field of research meandering between medicine and food science? *Frontiers in Pharmacology,9*, 215.
Zhang, Z., Cheng, Y., & Liu, N. C. (2014). Comparison of the effect of mean-based method and z-score for field normalization of citations at the level of Web of Science subject categories. *Scientometrics,101*(3), 1679–1693.
Zhu, H., & Zhu, Q. (2016). Mergers and acquisitions by Chinese firms: A review and comparison with other mergers and acquisitions research in the leading journals. *Asia Pacific Journal of Management, 33*(4), 1107–1149.