# Method for comparison of the number of citations from papers in different databases

Gerson Pech and Catarina Delgado

[1]*pech@uerj.br*
Department of Nuclear Physics and High Energies, Rio de Janeiro State University, Rua São Francisco Xavier, 524, Rio de Janeiro (Brasil)

[2]*cdelgado@fep.up.pt*
LIAAD INESCTEC and Faculty of Economics, University of Porto, Dr. Roberto Frias, 4200-464 Porto (Portugal)

## Abstract
Citation analysis has been used to compare researchers, fields, institutions and countries. However, not much has been done to compare citations of papers belonging to different databases and published in different years. This comparison could play a relevant role in many systematic literature reviews concerned with the growth, development, and changes of a particular scientific subject. This study aims to examine whether we can use the percentile approach to compare the number of citations from papers in different databases. We argue that this method can convert citations from different databases when there are same articles belonging to more than one database. We apply the method on Thomson Reuters' Web of Science and Elsevier's Scopus databases because they are the leading databases of scholarly impact. In this study we use two different Scopus subject area: Engineering – Industrial and Manufacturing Engineering; and Arts and Humanities –Archaeology. The analysis comprises articles published for the time period 1987–2017, of journals in the Scopus top 10%, corresponding to approximately 152,000 papers.

## Introduction
Citation analysis plays a key role in Scientometry. Many researchers had covered a long journey since the arguments stated by Garfield (1972) highlighted that the results of citation analysis have great potential for management of library journal collections. Garfield (1972) also pointed out that data on citation frequency could be correlated with subscription costs, providing a solid basis for cost-benefit analysis in the management of subscription budgets. Besides the fact that the number of citations is the simplest and most direct indicator of a publication impact (Milojević, Radicchi, & Bar-Ilan, 2017), this metric may provide information on the impact and performance of individual publications, research groups, institutions, countries, and journals (Sangwal, 2013; Waltman, 2016). Therefore, citations can be used for grading the importance of research results, because citation counts seem to correlate with expert assessments (Brito & Rodríguez-Navarro, 2018). Consequently, citations are used for formal and informal evaluations of academics (Thelwall & Wilson, 2014) and, Journal Impact Factor (Garfield, 1972) and h-index (Hirsch, 2005) are widely recognized and used citation impact indicators. However, although many authors have focused their studies on how to apply citation analysis to compare researchers, fields, institutions, and countries (e.g., Fairclough & Thelwall, 2015; Radicchi & Castellano, 2012; Rodríguez-Navarro & Brito, 2018; Waltman, 2016; Zhang, Cheng, & Liu, 2014), little work has been done to investigate how to compare citations of papers belonging different databases and published in different years. This comparison method could play a relevant role in many literature systematic analysis concerned with the growth, development, and changes of a particular scientific subject, and so, is the core of our study. Furthermore, in the last years, the papers that confront citations in different databases are mainly focused on two issues: the coverage that each database provides for the scientific disciplines studied (Li, Burnham, Lemley, & Britton, 2010; Martín-Martín et al., 2018; Winter, Zadpoor, & Dodou, 2014); and a longitudinal comparison involving a very limited period (Moed, Bar-Ilan, & Halevi, 2016; Harzing & Alakangas, 2016). Indeed, as manifested freshly by Martín-

Martín et al. (2018) about citation counts, "there is no recent or systematic evidence about the differences between different databases."

In response, this article aims to address that issue, by using 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. However, as we have shown in this paper, the extension of the method to compare papers published in different years and, simultaneously, also in different databases must be analysed through the linear regression coefficients that correlated the percentiles from different databases.

While this article looks at the number of citations from three decades, it also provides some hints as to how this parameter has changed over the years and thus contributes to a better understanding of the complexity of the citation analysis. The intricate meaning of the citations is still an open topic in the scientometric literature, and the question of what citation counts measure must be investigated carefully (Bornmann & Daniel, 2008). Indeed, in recent years, several authors have questioned the conceptual clarity of citation analysis. Specifically, scientific citations can be copied from the lists of references used in other papers, so that the rate of citing a paper is proportional to the number of citations it has already received (Simkin & Roychowdhury, 2007; Waltman, 2016). Additionally, citing certain authors provides support for a paper and persuades the scientific community of the validity of the findings, introducing bias on the analysis (Chan, Guillot, Page, & Torgler, 2015). On the other hand, scientific evaluation based on citation impact indicators may be improved by considering how significant (according to mention frequency) each paper is cited (Pak, Yu, & Wang, 2018). Beyond these points, it is important to keep in mind that the reason why an author cites an article varies from scientist to scientist (Bornmann & Daniel, 2008).

This paper is organized as follows: in the next section, we present the common approaches in citation analysis and, in the following sections, the percentile approach and numerical examples. Finally, in the last section, the conclusions are presented.

## Citation analysis

The number of citations of a scientific article is a very common measure of the acceptance of that academic publication (Lu & Liu, 2014), and ultimately of the researcher(s), the research group, the institution and the country. The comparison of these (researchers, research groups, institutions, or countries) publishing in different disciplines and periods is only possible with normalized citation scores (Haunschild & Bornmann, 2016). Some popular indicators follow the same formula: $C_{subset}/C_{set}$, where $C_{set}$ is the average number of citations of all publications in a dataset (for example, a scientific area) and $C_{subset}$ is the average number of citations of all publications of a subset. For instance, in the normalized citation impact value indicator, the subset is a country's set of publications on a specific scientific area. If a country has a normalized citation impact value of 1 in a specific subject area, that indicates that the citation impact of papers published by researchers in this country is no more and no less than the average impact of papers in this subject area (Bornmann & Leydesdorff, 2013). In the source normalized impact per paper indicator (SNIP), however, the subset is a journal's set of publications on a specific scientific area (Moed, 2016). A similar approach is followed by the scaled citation count indicator. This is a normalized indicator in which the number of citations of a publication is divided by the average number of citations of the papers published in the same year of the paper being analysed. A value of 1 for a specific paper indicates that the citation impact of this paper is no more and no less than the average impact of papers, in this scientific field, published in the same year. That is, the same normalization concept used to evaluate a set of papers, has been also applied to evaluate the number of citations of a specific paper. In this case, the normalization follows the relation: $C\text{i} / C\text{set}$, where $C\text{i}$ is the citations of the paper i.

Other indicators try to consider the "exposure time" of publications:

- the citation rate per year (also called citation count per year since publication, or adjusted citation index) is the total number of citations of a paper divided by 1/12 of the number of months since the initial publication up to the month of data collection, which gives the average number of citations that a paper has received each year since it was published (Wilcox et al., 2013);
- the citation density is a normalized citation-based indicator which captures the citation impact in terms of both citations per paper, and citations per citation year (Ahmed et al., 2017), by dividing the total number of citations of a set of articles published in a certain year by both the number of papers in that subset and the number of years after the publication.

A normalized variant of the average number of citations per publication is obtained by dividing the total number of citations of a given set of publications by the expected total number of citations (the average number of citations of all publications in the same field, same year and same document type). Some authors claim that this ratio provides the "desired universality of citation distributions," but others refute that claim (Waltman, 2016). Other alternatives to ratios, when it comes to the normalization include applying a logarithmic transformation to citation counts and to normalize citation counts by calculating z-scores and the transformation of citation counts by a two-parameter power–law function, which seems to be the best to create normalized citation distributions that are identical across fields (Waltman, 2016). Still, for some authors the ratio between the actual number of citations of a publication and the average number of citations of all publications that are in the same field and that have at least one citation is the best indicator (Waltman, 2016).

Although field- and time-normalization of metrics is currently a standard procedure in bibliometric studies (Leydesdorff *et al.*, 2016), most of the citation indicators are still based on simple non-normalized averages, either weighted (or factional) or not, like the average number of citations, using the sum of total citations received by the publications being analysed divided by the number of papers in the sample (Waltman, 2016).

Still, some issues are yet to be addressed properly in the literature: (i) the fact that citation counts grown up over time with the increase of journals, papers and the amount of references in the papers; (ii) the tendency of the citations to grow/decrease over time; (iii) the fact that citation frequency is highly skewed, with many infrequently cited papers and relatively few highly cited papers, so one should not see citation rates as representing the central tendency of the distribution; (iv) the fact that different databases provide different citation number for the same article, by counting only the number of citations that appear on publications already on that database (at that time). This can be a problem when conducting a longitudinal study, using data from articles of the same journal, but published in different years, therefore with some only obtained in a different database. Alternatively, when trying to capture the publications from highly respected journals or some highly cited publications that are not in the "main" database being used. While collecting a database of highly cited publications, it may be interesting to search in other databases for other important papers in the field. Some studies (e.g., Lu & Liu 2014) use a citation paralleling approach – for instance, after collecting the top 100 most cited articles from the field in the "main database", identify the number of citations of the 100th most cited article in the new database (because the same article has a different number of citations in different databases) and then search in this new database for all articles of that field with the same, or more, citations than the 100th most cited paper. However, what if we want to use the number of citations as a variable to analyse the publication's acceptance? We cannot use, in the same analysis, the number of citations from different databases? At least, we cannot use them in their "raw data" form. In the next section, we introduce a two-stage method to address these problems.

## The percentile-based approach

*Objectives and research questions*

This exploratory study will address the following research questions:

[RQ1] Assuming the same research field and for the same year, how is it possible to compare paper citations that belong to different databases?

[RQ2] In a systematic literature review when a longitudinal study is developed for a given research area, how to know if it is possible to use the percentile approach to convert the number of citations that appear in one database to an equivalent value in another one?

*Comparison method*

To address RQ1 and RQ2 we developed the percentile-based comparison method with two main stages: (i) a conversion of the number of citations of articles published in different years and (ii) a conversion of the number of citations of articles belonging to different databases.

The steps are as follows:

(i) Method to compare the number of citations (received in a particular year) of papers published in different years:

Step 1 - Consider a sub-area or a set of title sources in a specific database.

Step 2 - For each year, develop a cumulative probability function (CPF) for the number of citations. We can use a characteristic probability distribution function (PDF) like Lognormal, or not.

Step 3 - For each paper, set a citation parameter to be the corresponding percentile that was calculated in Step 2.

Step 4 - Use the citation parameter defined in step 3 to rank the articles, published in any year, in terms of the citations received.

(ii) Method to compare the number of citations of papers in a different database:

Step 1 - Consider a sub-area or a set of title sources in a specific database.

Step 2 - For each year, and each database, separately, develop a cumulative probability function (CPF) for the number of citations. We can use a characteristic probability distribution function (PDF) like Lognormal, or not.

Step 3 - For each year, select the papers that belong in the two databases and obtain a linear regression model to describe the relationship between the number of citations of these papers that are in both databases.

Step 4 – use the model obtained in Step 3 to develop, limited by the uncertainties of the model, a function $F(c_i^y)$ that give the citation relationship between the two databases, and the $F(p_i^y)$ that represents the same for the percentiles.

Among the different normalization procedures that could have been used, the percentile rank approach has the advantage that, intrinsically, implies the normalization of citation counting data (Brito & Rodríguez-Navarro, 2018). In the approach applied in this research each paper is weighted based on the percentile to which it belongs in the citation distribution of its field and of its year of publication. The percentile approach has been extensively applied lately in citation analysis for bibliometric evaluations (Bornmann, 2013; Bornmann, Leydesdorff, & Wang, 2013; Waltman & Schreiber, 2013) and also to predict citation counts (Kosteas, 2018).

The key point of the model is to investigate whether the method used for stage (ii) keeps invariant the method used in stage (i), or, in other words, whether the method used to find equivalence for the conversion of citations over the years is not destroyed by the method of conversion between databases. For example, assume that for a database Ψ the conversion method between years (stage (i)) implies that a particular article published in 2005 with 150 citations is equivalent to an article published in 2015 with 30 citations, and that one article from 2005 with 130 citations is equivalent to one from 2015 with 25 citations. Now, suppose 150 citations from the base Ψ, using the method of stage (ii), are equivalent to 130 citations on the base Ω, for the year 2015. Then, also using the method of stage (ii) for conversion of the number of citations for articles published in different years, but belonging to the base Ω, we must reach the same value of 25 citations that was determined by the method of stage (i). Figure 1 illustrates this problem.

$$\begin{tikzcd} C^{\Psi}(x) \arrow[r, Leftrightarrow, "\textit{stage (i)}"] \arrow[d, Leftrightarrow, "\textit{stage (ii)}"'] & C^{\Psi}(y) \arrow[d, Leftrightarrow, "\textit{stage (ii)}"] \\ C^{\Omega}(x) \arrow[r, Leftrightarrow, "\textit{stage (i)}"] & C^{\Omega}(y) \end{tikzcd}$$

**Figure 1. The correspondence between the bases Ψ and Ω**

## Numerical examples

### *Data collection*

In order to provide some numerical examples, we collected two databases from two very different scientific fields: Engineering and Arts and Humanities. From each field, we selected a narrower subject: Industrial and Manufacturing Engineering (127.208 papers in the Scopus database, from 1987 to 2017) and Archaeology (25.144 papers in the Scopus database). Each database was built with the top 10% articles, in terms of citations, from 1987 to 2017.

Table 1 gives a list of the journals analysed in this study together with some of their characteristics: the CiteScore measures the average citations received per document published in the serial; the CiteScore Percentile indicates the relative standing of a serial title in its subject field (a title will receive a CiteScore Percentile for each subject area in which it is indexed in Scopus); the number of papers published in the range of this study (1987-2017); the publisher and the Scopus Subject area. Scopus Subject areas are defined by the All Science Journal Classification codes in Scopus. It is important to notice that titles can be indexed in multiple subject areas. Data were obtained from the file *CiteScore_Metrics_2011-2017* downloaded on Scopus.com on May 25, 2018, using the following 2 filters: in the column Scopus Sub-Subject Area we selected Industrial and Manufacturing Engineering and Archaeology, and in the column Top 10% (CiteScore Percentile) we selected Top 10%.

**Table 1 - Journals included in the analysis**

| Title | Cite Score | Percentile | SJR | Publisher | Area |
|---|---|---|---|---|---|
| Additive Manufacturing | 7,73 | 99 | 2,611 | Elsevier | IND |
| IEEE Industrial Electronics Magazine | 7,15 | 99 | 1,978 | IEEE | IND |
| Sustainable Materials and Technologies | 7,14 | 99 | 1,548 | Elsevier | IND |
| Chemical Eng. J. | 7,01 | 98 | 1,863 | Elsevier | IND |
| Manufacturing Letters | 6,83 | 98 | 1,313 | Elsevier | IND |
| J. of Industrial Information Integration | 6,5 | 98 | 0,866 | Elsevier | IND |
| J. of Operations Management | 6,13 | 97 | 5,739 | Elsevier | IND |

| | | | | | |
|---|---|---|---|---|---|
| Int. J. of Machine Tools and Manufacture | 5,92 | 97 | 2,700 | Elsevier | IND |
| J. of Cleaner Production | 5,79 | 97 | 1,467 | Elsevier | IND |
| Energy | 5,6 | 96 | 1,990 | Elsevier | IND |
| Int. J. of Production Economics | 5,42 | 96 | 2,401 | Elsevier | IND |
| Composites Part B: Eng. | 5,41 | 96 | 2,039 | Elsevier | IND |
| Virtual and Physical Prototyping | 5,35 | 96 | 1,438 | Taylor & Francis | IND |
| Critical Reviews in Food Sci and Nutrition | 5,15 | 95 | 1,596 | Taylor & Francis | IND |
| Reliability Eng. and System Safety | 4,65 | 95 | 1,665 | Elsevier | IND |
| Food Eng. Reviews | 4,6 | 95 | 1,639 | Springer Nature | IND |
| Int. J. of Greenhouse Gas Control | 4,34 | 94 | 1,458 | Elsevier | IND |
| Int. J. of Precision Eng. and Manuf. Green Tech | 4,31 | 94 | 1,335 | Springer Nature | IND |
| Int. J. of Robust and Nonlinear Control | 4,26 | 94 | 2,028 | Wiley-Blackwell | IND |
| J. of Manufacturing Systems | 4,15 | 93 | 1,548 | Elsevier | IND |
| J. of Materials Processing Tech | 4,15 | 93 | 1,695 | Elsevier | IND |
| Applied Thermal Eng. | 4,14 | 93 | 1,505 | Elsevier | IND |
| Robotics and Comp-Integrated Manufacturing | 4,11 | 92 | 1,041 | Elsevier | IND |
| CIRP Annals - Manufacturing Tech | 4,09 | 92 | 2,034 | Elsevier | IND |
| IEEE Transactions on Industry Applications | 4,05 | 92 | 1,020 | IEEE | IND |
| J. of Process Control | 3,85 | 91 | 1,108 | Elsevier | IND |
| Advanced Materials Technologies | 3,85 | 91 | 1,241 | Wiley-Blackwell | IND |
| Sustainable Production and Consumption | 3,52 | 91 | 0,739 | Elsevier | IND |
| Chemical Eng. Science | 3,44 | 90 | 1,043 | Elsevier | IND |
| Hydrometallurgy | 3,43 | 90 | 1,208 | Elsevier | IND |
| Industrial Management and Data Systems | 3,43 | 90 | 0,904 | Emerald | IND |
| Industrial & Eng. Chemistry Research | 3,4 | 90 | 0,978 | AMC | IND |
| Quaternary Science Reviews | 4,51 | 99 | 2,668 | Elsevier | ARC |
| J. of Archaeological Research | 4,5 | 99 | 2,159 | Springer Nature | ARC |
| J. of Archaeological Science | 2,96 | 98 | 1,885 | Elsevier | ARC |
| J. of World Prehistory | 2,96 | 98 | 2,022 | Springer Nature | ARC |
| Boreas | 2,65 | 98 | 1,273 | Wiley-Blackwell | ARC |
| J. of Archaeological Method and Theory | 2,53 | 98 | 2,014 | Springer Nature | ARC |
| Holocene | 2,43 | 97 | 1,202 | SAGE | ARC |
| Current Anthropology | 2,16 | 97 | 1,160 | Chicago Press | ARC |
| J. of Agrarian Change | 2,15 | 96 | 1,403 | Wiley-Blackwell | ARC |
| J. of Cultural Heritage | 2,11 | 96 | 0,562 | Elsevier | ARC |
| Vegetation History and Archaeobotany | 2,05 | 95 | 1,206 | Springer Nature | ARC |
| American Antiquity | 1,95 | 96 | 1,176 | Cambridge | ARC |
| J. of Anthropological Archaeology | 1,84 | 95 | 1,240 | Elsevier | ARC |
| J. of Social Archaeology | 1,81 | 95 | 0,936 | SAGE | ARC |
| Heritage Science | 1,77 | 95 | 0,491 | Springer Nature | ARC |
| World Archaeology | 1,74 | 94 | 1,349 | Taylor & Francis | ARC |
| Digital App in Arc and Cultural Heritage | 1,72 | 94 | 0,412 | Elsevier | ARC |
| Radiocarbon | 1,7 | 93 | 0,959 | Cambridge | ARC |
| Archaeological and Anthropological Sci | 1,63 | 93 | 1,052 | Springer Nature | ARC |
| J. of Island and Coastal Archaeology | 1,54 | 93 | 0,845 | Taylor & Francis | ARC |
| Cambridge Archaeological J. | 1,47 | 92 | 1,121 | Cambridge | ARC |
| Archaeometry | 1,43 | 92 | 0,587 | Wiley-Blackwell | ARC |
| PalArch's J. of Vertebrate Palaeontology | 1,4 | 92 | 0,403 | PalArchFoundation | ARC |
| Archaeological Prospection | 1,34 | 92 | 0,635 | Wiley-Blackwell | ARC |
| Geoarchaeology - An Int. J. | 1,32 | 91 | 0,823 | Wiley-Blackwell | ARC |
| African Archaeological Review | 1,29 | 91 | 0,862 | Springer Nature | ARC |
| Int. J. of Paleopathology | 1,22 | 90 | 0,618 | Elsevier | ARC |
| Antiquity | 1,21 | 90 | 0,887 | Cambridge | ARC |
| J. of Archaeological Science: Reports | 1,21 | 90 | 0,659 | Elsevier | ARC |
| Frontiers of Architectural Research | 1,2 | 90 | 0,404 | Elsevier | ARC |
| Int. J. of Osteoarchaeology | 1,15 | 90 | 0,652 | Wiley-Blackwell | ARC |

In figure 2 we have the evolution of the number of papers, and the number of citations in the top 10 journals of two analysed Scopus subject areas, from 1987 to 2017.

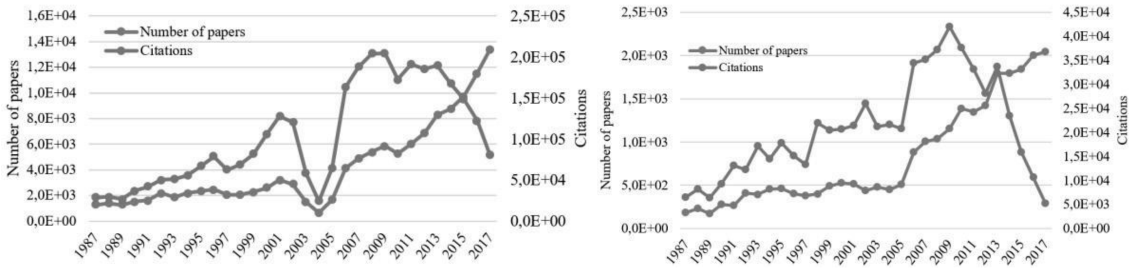


**Figure 2. Number of papers and citations in the top 10 journals of the Industrial and Manufacturing Engineering (left) and Archaeology (right) fields.**

The search of the articles and their corresponding citation numbers was conducted between August and September 2018, on Scopus and WoS sites. For this, we used the Print-ISSN and e-ISSN codes of each journal listed in Table 1, instead of the title name, to avoid collection errors. We collected the results for the period 1987-2017. On the Scopus website, we used the link View Citation Overview. The citation overview is available as a comma separated file (.csv) with the first 20,000 documents included, that we downloaded separately for each year of the interval. For WoS, after performing the search, we used the Create Citation Report functionality, and we downloaded it using the available export data that only allows 500 records to be downloaded at once. The Scopus database of all 31 years was used for the percentiles analysis. For the analysis of the comparison between Scopus and WoS, the two databases were used for the following years: 1987, 1997, 2005 and 2010. Table 2 shows the number of papers of each database, published in each year, and the number of papers that belong to both databases and therefore were analysed.

**Table 2 - Number of papers of each database and number of papers that belong to the two and therefore participated directly in the analysis.**

| Subject area | Number of papers | 1987 | 1997 | 2005 | 2010 |
|---|---|---|---|---|---|
| | Scopus | 1270 | 2054 | 1691 | 5255 |
| IND | WoS | 1000 | 2720 | 4310 | 5608 |
| | Scopus & WoS | 825 | 1893 | 1524 | 4957 |
| | Scopus | 184 | 378 | 506 | 1387 |
| ARC | WoS | 212 | 498 | 788 | 1372 |
| | Scopus & WoS | 140 | 329 | 464 | 1334 |

*Results*

In figure 3, the cumulative probability distribution of citations is shown, in four different periods: (a) 1987 – 1993; (b) 1994 – 2001; (c) 2002 – 2009; and (d) 2010 – 2017. For comparison purposes, the distribution for the first year of each interval – (a); (b); (c) and (d) – is represented by the same symbol and color. The same happens for the distribution of the second, third and subsequent years of each interval.

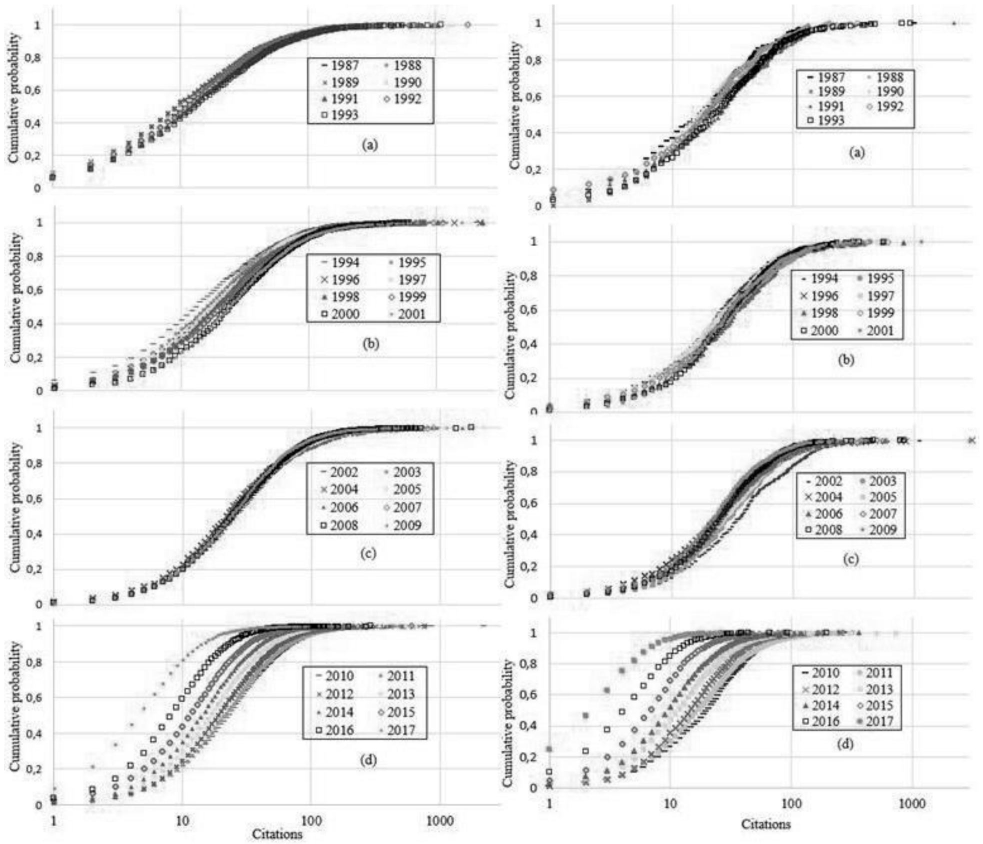


**Figure 3. Cumulative probability distribution of citations for 4 intervals of years: (a) 1987 – 1993; (b) 1994 – 2001; (c) 2002 – 2009; and (d) 2010 – 2017 for Scopus subject areas Industrial and Manufacturing Engineering (left) and Archaeology (right).**

In figure 4, the number of citations across 31 years is shown for the 10th, the 30th, the 50th, the 60th, the 70th, the 80th, the 90th and the 95th percentiles (respectively, P10, P30, P50, P60, P70, P80, P90, P95) in the Industrial and Manufacturing Engineering and Archaeology fields. The equivalence of the number of citations can be obtained by following each of the curves defined by the points of each percentile. For example, a paper with 15 citations that was published in 1989 is equivalent to a paper published in 2003 today with 30 citations (P60). On the other hand, a paper of 2004 with 150 citations is equivalent to a paper of 2015 today with 45 citations (P95).

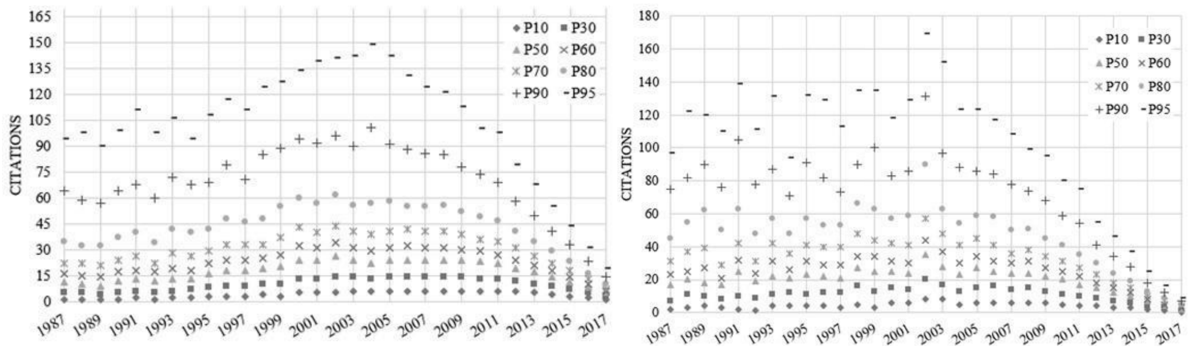


**Figure 4. Number of citations across 31 years for the P10, P30, P50, P60, P70, P80, P90 and P95 percentiles in the Industrial and Manufacturing Engineering (left) and Archaeology (right).**

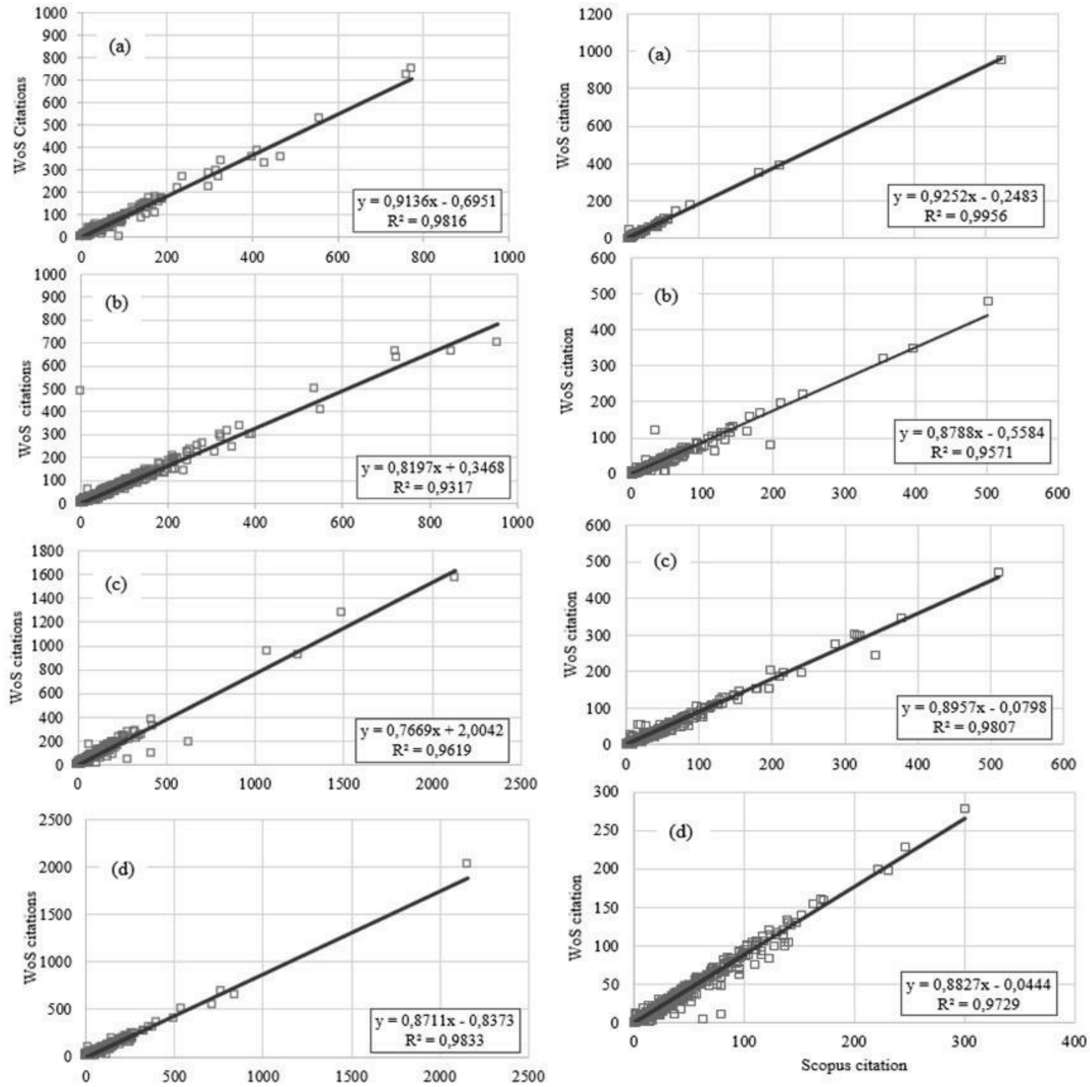


**Figure 5. Linear regression for the number of citations of articles belonging to both the Scopus and the WoS databases (Industrial and Manufacturing Engineering (left) and Archaeology (right)), published in the following years: (a) 1987; (b) 1997; (c) 2005; and (d) 2010.**

A linear regression model was obtained (figure 5) for the number of citations of articles belonging to both the Scopus and the WoS databases, published in the following years: (a) 1987; (b) 1997; (c) 2005; and (d) 2010. The line describes a model able for converting the number of citations from one database into another for each year separately.

A linear regression model was also obtained (figure 6) for the percentiles of citations of articles belonging to both the Scopus and the WoS, published in the following years: (a) 1987; (b) 1997; (c) 2005; and (d) 2010. The size of the points represents the number of papers that have the same values of $x$ and $y$ in the graph. The angular coefficient close to 1 and the linear coefficient close to zero show that, for these examples, even though the number of citations in the two databases is different, the percentiles are seemingly invariants between the Scopus and WoS databases. This invariancy is being investigated further by us and will be the subject of an upcoming paper.

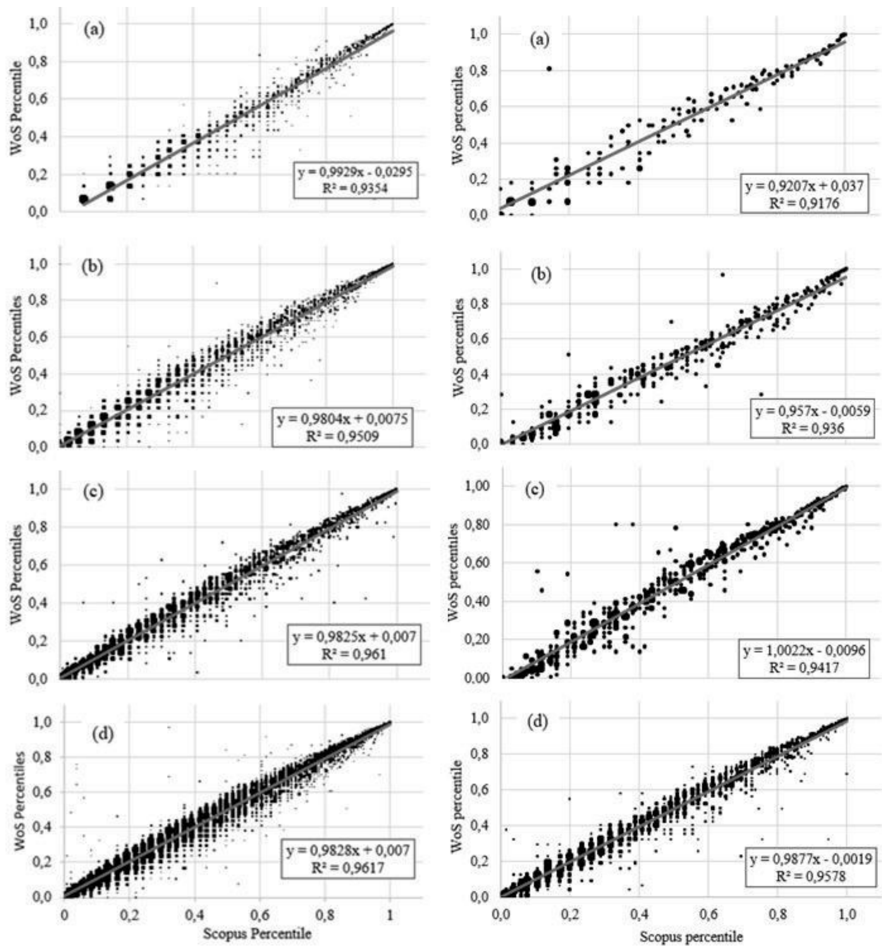


**Figure 6. Linear regression for the percentiles of citations of articles belonging to both the Scopus and the WoS (Engineering (left), Archaeology (right)), in (a) 1987; (b) 1997; (c) 2005; and (d) 2010.**

## Conclusions

In this paper, a percentile-based technique has been introduced to address the problem of having to use, in bibliometric analysis or the data collection stages in systematic literature reviews, citation numbers for publications belonging to different databases (e.g., WoS, Scopus, Google Scholar). When a publication is in different databases, it usually presents a different citation number in each database. We propose a percentile-based method to establish a comparison between the citation numbers of those articles common to both databases, in order to obtain a model that could help us to predict the citation number of the articles that cannot be found in one of the databases. The evidence from the two fields selected (Industrial and Manufacturing Engineering and Archaeology) show that such a model can be derived. However, this is still an exploratory study, and although the results cannot be generalized, they confirm findings from some earlier studies and support the presented technique for comparing and converting citation numbers between different databases. Another contribution of this study is the comparison of models in different fields and different years, suggesting the possibility of a unified conversion model, by including field-related and year-related variables, to capture those influences.

## References


Ahmed, K. K. M., Dhawan, S. M., Gupta, B. M. and Bansal, M. (2017). Highly cited publications output by India in clinical pharmacology during 2000-14: A scientometric assessment. *J. Young Pharm.*, *9*(2), 145–157. DOI: 10.5530/jyp.2017.9.30.

Bornmann, L., & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80.
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. and Leydesdorff, L. (2013). Macro-Indicators of Citation Impacts of Six Prolific Countries: InCites Data and the Statistical Significance of Trends. *PLoS ONE*, *8*(2), e56768. https://doi.org/10.1371/journal.pone.0056768
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*, 933– 944.
Brito, R., & Rodríguez-Navarro, A. (2018). Research assessment by percentile-based double rank analysis. *Journal of Informetrics, 12*(1), 315–329.
Chan, H. F., Guillot, M., Page, L., & Torgler, B. (2015). The inner quality of an article: Will time tell?, *Scientometrics, 104*(1), 19–41.
Winter, J. C. F. de, Zadpoor, A. A., & Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: a longitudinal study. *Scientometrics, 98*(2), 1547–1565.
Fairclough, R., & Thelwall, M. (2015). More precise methods for national research citation impact comparisons. *Journal of Informetrics, 9*(4), 895–906.
Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(4060), 471–479.
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.
Haunschild, R., & Bornmann, L. (2016). Normalization of Mendeley reader counts for impact assessment. *Journal of Informetrics*, *10*(1), 62-73.
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.
Kosteas, V. D. (2018). Predicting long-run citation counts for articles in top economics journals. *Scientometrics,115*, 1395–1412.
Leydesdorff, L., Wouters, P. and Bornmann, L. (2016). Professional and citizen bibliometrics: complementarities and ambivalences in the development and use of indicators—a state-of-the-art report. *Scientometrics*, *109*, 2129–2150. DOI 10.1007/s11192-016-2150-8.
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.
Lu, L. Y. and Liu, J. S. (2014), The Knowledge Diffusion Paths of Corporate Social Responsibility – From 1970 to 2011. *Corporate Social Responsibility and Environmental Management*, 21(2), 113-128.
Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E.D. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics, 12*(4), 1160-1177.
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.
Moed, H. F. (2016). Comprehensive indicator comparisons intelligible to non-experts: the case of two SNIP versions. *Scientometrics*, *106*(1), 51–65. DOI: 10.1007/s11192-015-1781-5.

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.
Rodríguez-Navarro, A., & Brito, R. (2018). Double rank analysis for research assessment. *Journal of Informetrics, 12*(1), 31–41.
Pak, C. M., Yu, G., & Wang, W. (2018). A study on the citation situation within the citing paper: citation distribution of references according to mention frequency. *Scientometrics, 114*(3), 905–918.
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. DOI:10.1371/journal.pone.0033833.
Sangwal, K. (2013). Citation and impact factor distributions of scientific journals published in individual countries. *Journal of Informetrics, 7*(2), 487–504.
Simkin, M. V., & Roychowdhury, Vwani P. (2007). A mathematical theory of citing. *Journal of the American Society for Information Science and Technology, 58*(11), 1661–1673.
Thelwall, M., & Wilson, P. (2014). Distributions for cited articles from individual subjects and years. *Journal of Informetrics, 8*(4), 824–839.
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.
Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*, 365–391.
Wilcox, M.A., Khan, N.R., McAbee, J.H., Boop, F.A. and Klimo, P. Jr. (2013). Highly cited publications in pediatric neurosurgery. *Childs Nerv Syst*, *29*, 2201–2213.
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.