Scientometrics (2016) 107:321–335
DOI 10.1007/s11192-016-1867-8

Do Scopus and WoS correct "old" omitted citations?

Fiorenzo Franceschini[1] · Domenico Maisano[1] · Luca Mastrogiacomo[1]

Received: 11 June 2015 / Published online: 1 February 2016

Abstract Omitted citations—i.e., missing links between a cited paper and the corresponding citing papers—are a consequence of several bibliometric-database errors. To reduce these errors, databases may undertake two actions: (1) improving the control of the (new) papers to be indexed, i.e., limiting the introduction of "new" dirty data, and (2) detecting and correcting errors in the papers already indexed by the database, i.e., cleaning "old" dirty data. The latter action is probably more complicated, as it requires the application of suitable error-detection procedures to a huge amount of data. Based on an extensive sample of scientific papers in the Engineering-Manufacturing field, this study focuses on old dirty data in the Scopus and WoS databases. To this purpose, a recent automated algorithm for estimating the omitted-citation rate of databases is applied to the same sample of papers, but in three different-time sessions. A database's ability to clean the old dirty data is evaluated considering the variations in the omitted-citation rate from session to session. The major outcomes of this study are that: (1) both databases slowly correct old omitted citations, and (2) a small portion of initially corrected citations can surprisingly come off from databases over time.

Keywords Bibliometric database · Database errors · Omitted citations · Old dirty data · Error correction · Engineering-Manufacturing journals

✉ Domenico Maisano
domenico.maisano@polito.it

Fiorenzo Franceschini
fiorenzo.franceschini@polito.it

Luca Mastrogiacomo
luca.mastrogiacomo@polito.it

1 DIGEP (Department of Management and Production Engineering), Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy

Introduction

A sizeable part of the bibliometric literature examines errors in bibliometric databases. A synthetic classification of the major errors is reported in Table 1, distinguishing between author and database mapping errors.

Several studies show that one of the major consequences of these errors is represented by omitted citations, i.e., citations that should be ascribed to a certain (cited) paper but, for some reason, are lost (Moed 2006; Buchanan 2006; Jacsó 2006; Li et al. 2010; Olensky 2013).

Franceschini et al. (2013) proposed an automated algorithm for estimating the omitted-citation rate of bibliometric databases. This algorithm requires the combined use of two or more bibliometric databases and is based upon the idea that the mismatch between the citations occurring in one database and another one is evidence of possible errors/omissions.

In a further study by Franceschini et al. (2014), this algorithm was applied to a relatively large set of publications, showing that, depending on the bibliometric database in use (Scopus or WoS), omitted citations are not distributed uniformly among publishers; e.g., regarding the publications in the Engineering-Manufacturing field, citations from papers published by Wiley-Blackwell are more likely to be omitted by Scopus, while those from papers published by ASME (American Society of Mechanical Engineers) are more likely to be omitted by WoS. A reason behind this result is that some editorial styles imposed by certain publishers can probably hamper the correct identification of the cited papers by some databases.

The presence of database errors, as well as journal coverage or author disambiguation, is probably one of the major concerns of database administrators. In the authors' opinion, database administrators may undertake two actions for reducing database errors (see Fig. 1):

1. Limiting the introduction of "new" dirty data in a database, i.e., errors concerning new papers to be indexed;
2. Cleaning "old" dirty data, i.e., errors concerning papers/journals already indexed by a database.

The recent effort by reviewers, publishers and database administrators in checking the cited-article lists of new papers probably contributes to reducing "new" dirty data. This hypothesis is corroborated by a recent study by Franceschini et al. (2015a), which shows

Table 1 Classification of bibliometric database errors according to Buchanan (2006)

Error type	Author errors	Database mapping errors
Definition	Errors made by authors when creating the list of cited articles for their publication	Failure to establish an electronic link between a cited article and the corresponding citing articles that can be attributed to a data-entry error
Examples	Errors in name and initials of the first author Errors in publication title Errors in publication year Errors in volume number Errors in pagination	Transcription errors Target-source article record errors Cited article omitted from a cited-article list Reason unknown

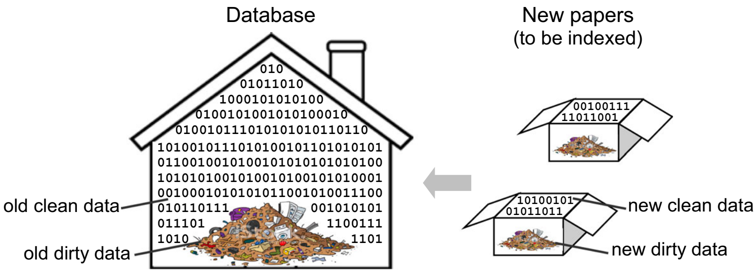


Fig. 1 Intuitive representation of "new" and "old" dirty data in a database

that the databases' propensity to omit newer citations is generally lower than that to omit older citations (see Fig. 2).

Improving the accuracy of old data is certainly much more complicated because it requires the systematic application of suitable error-detection procedures to a huge amount of data. However, this effort would be essential for improving the quality of a database significantly.

This paper focuses on the ability of the major multidisciplinary bibliometric databases, i.e., Scopus and WoS, to clean up old dirty data. For this evaluation, we use a new procedure, derived from the automated algorithm by Franceschini et al. (2013). This procedure consists in (1) repeating the omitted-citation-rate analysis on the same sample of (cited and citing) articles, but in different-time sessions, and (2) observing any variation in the results. A database's ability to clean old dirty data will be evaluated considering the variation in the omitted-citation rate from one session to another one.

The remainder of this paper is organized into four sections. The section "Automated algorithm for examining the omitted citations" briefly recalls the algorithm by Franceschini et al. (2013). The section "Methodology" describes the methodology used in our study, focusing on data collection and analysis. The section "Results" illustrates the results of the analysis, investigating similarities and differences between the two databases examined. Finally, the section "Conclusions" summarizes the original contributions of this paper, highlighting the major results, limitations and suggestions for future research.

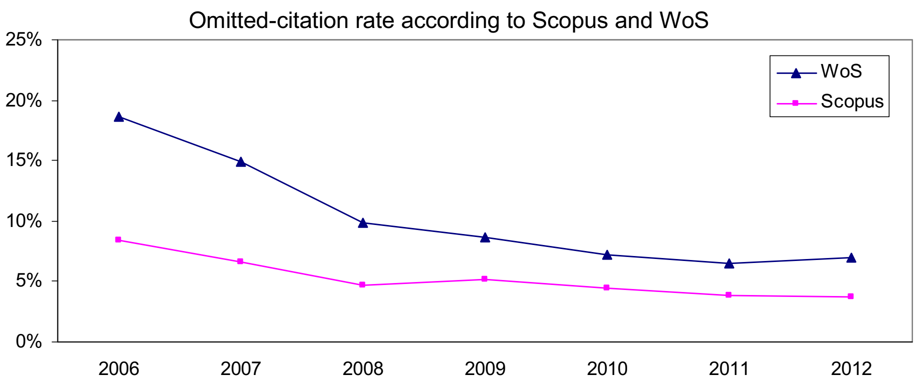


Fig. 2 Omitted-citation rate according to Scopus and WoS, depending on the issue year of citing papers; for details, see (Franceschini et al. 2015a)

This paper is the extended version of the paper (Franceschini et al. 2015b), presented at ISSI'15 (15th International Society of Scientometrics and Informetrics Conference) in Istanbul, Turkey, June–July 2015.

Automated algorithm for analysing the omitted citations

Before recalling the algorithm, we present an introductory example to illustrate how it works. Let us consider a fictitious paper of interest, indexed by Scopus and WoS. The number of citations received by this paper is four in Scopus and six in WoS (see Table 2).

The union of the citations recorded by the two databases is a total of eight citations. Among the citations, only five come from sources (i.e., journals or conference proceedings) officially covered by both databases (highlighted in grey in Table 2). Focusing on these five "theoretically overlapping" (TO) citations, two are omitted by Scopus (but not by WoS) and one is omitted by WoS (but not by Scopus). Therefore, from the perspective of the paper of interest, a rough estimate of the omitted-citation rate is 2/5 $\approx$ 40 % in Scopus and 1/5 $\approx$ 10 % in WoS. The same reasoning can be extended to multiple papers of interest and more than two bibliometric databases.

The automated algorithm, which is based on the combined use of two bibliometric databases (Scopus and WoS in this case), can be summarised in three steps:

1. Identify a set of (P) papers of interest, indexed by both the databases.
2. For each (i-th) paper of the set, identify the TO citations, defined as the portion of documents issued by journals officially covered by Scopus and WoS. The number of TO citations concerning the i-th paper of interest will be denoted as γ_i.
3. For each (i-th) paper of the set and for each database, determine the number (ω_i) of TO citations that do not occur in it and classify them as omitted citations. The omitted-citation rate (p) relating to the P papers of interest, according to a database, can be estimated as:

$$\hat{p} = \sum_{i=1}^{P} \omega_i \bigg/ \sum_{i=1}^{P} \gamma_i. \tag{1}$$

The afore-described algorithm has the great advantage of being automated, i.e., it does not require manual analysis of cited/citing papers. For this reason, it allows estimating the p value of relatively large sets of publications, in a simple and fast way. The price to pay for this advantage is that the algorithm relies on some (potentially questionable) simplifying assumptions:

Table 2 Citation data relating to a fictitious article, according to Scopus and WoS

Citation no	Scopus	WoS
1	✔	Source not covered
2	Source not covered	✔
3	*Omitted*	✔
4	✔	✔
5	✔	✔
6	*Omitted*	✔
7	Source not covered	✔
8	✔	*Omitted*
Total	4	6

The union of the citations recorded by the two databases (see the first column) is a total of eight citations. Among the citations, only five come from sources officially covered by both databases (highlighted in italic)

- It is assumed that the omitted citations of different databases are statistically independent. Actually, to identify a citing paper omitted by one database, it is necessary that the same citing paper occurs in the other database. Of course, the concurrent omission of a citing paper by both databases will prevent its detection, leading to an underestimation of p.
- It is assumed that the incidence of "phantom citations"—that is, erratic citations from papers that did not actually cite the target paper (Garcia-Pérez 2010)—is negligible. According to our algorithm, a phantom citation of one database—if it is (mistakenly) assigned to a paper that is supposed to be covered by other databases—may lead to an incorrect notification of omitted citation for the other database. In the "Appendix" we present a practical analytical model for correcting the $\hat{p}$ values (calculated using Eq. 1), taking account of the—albeit small—distortion produced by the phantom citations of the databases in use.
- The estimation of p is performed on the basis of (1) a set of papers of interests and (2) a portion of the total citations that they obtained (i.e., that ones related to citing articles purportedly covered by both the databases). The results can be extended to the rest of the citations, upon the assumption that the incidence of omissions is similar.
- The algorithm can be readily applied to journal articles, but not as easily to other publication types—for example, book chapters, conference proceedings, monographs, etc.—for two basic reasons: (a) some of these publication types are not covered by databases, (b) lack of exhaustive official lists concerning the coverage of these document types.

For a more detailed description of the algorithm, we refer the reader to Franceschini et al. (2013).

The ability of bibliometric databases to clean old dirty data will be evaluated by applying this algorithm to the same sample of TO citations, in three different-time sessions.

Methodology

The study is based on the analysis of the citations obtained from a relatively large sample of papers of interest. The papers were issued by 33 scientific journals (1) included in the ISI Subject Category of Engineering-Manufacturing (by WoS) and (2) covered by Scopus; Table 3 reports the list of these journals. For each journal, we considered the papers published in the time-window from 2006 to 2012 and the citations that they obtained from papers issued in the same period.

Data collection was repeated in three different-time sessions, spaced about 7 months apart: i.e., session I on August 2013, session II on March 2014 and session III on October 2014. We remark that the duration of each data-collection session (i.e., a few days) is negligible with respect to the time period between two consecutive sessions.

To enable comparisons between data collected in different sessions, we adopted two measures:

1. Among the papers of interest (or cited papers)—i.e., those issued by the 33 Engineering-Manufacturing journals—we selected those indexed in each of the three sessions, by both the (Scopus and WoS) databases; in formal terms:

$$A = A^{(I)} \cap A^{(II)} \cap A^{(III)}, \tag{2}$$

Table 3 List of the Engineering-Manufacturing journals examined

Journal title	ISSN
AI EDAM—Artificial Intelligence for Engineering Design Analysis and Manufacturing	0890-0604
Assembly Automation	0144-5154
CIRP Annals—Manufacturing Technology	0007-8506
Composites Part A—Applied Science and Manufacturing	1359-835X
Concurrent Engineering—Research and Applications	1063-293X
Design Studies	0142-694X
Flexible Services and Manufacturing Journal	1936-6582
Human Factors and Ergonomics in Manufacturing & Service Industries	1090-8471
IEEE Trasaction on Components Packaging and Manufacturing Technology	2156-3950
IEEE Transactions on Semiconductor Manufacturing	0894-6507
IEEE-ASME Transactions on Mechatronics	1083-4435
International Journal of Advanced Manufacturing Technology	0268-3768
International Journal of Computer Integrated Manufacturing	0951-192X
International Journal of Crashworthiness	1358-8265
International Journal of Machine Tools & Manufacture	0890-6955
International Journal of Production Economics	0925-5273
Journal of Advances Mechanical Design Systems and Manufacturing	1881-3054
Journal of Computing and Information Science in Engineering—Transactions of the ASME	1530-9827
Journal of Intelligent Manufacturing	0956-5515
Journal of Manufacturing Science and Engineering—Transactions of the ASME	1087-1357
Journal of Manufacturing Systems	0278-6125
Journal of Materials Processing Technology	0924-0136
Journal of Scheduling	1094-6136
Machining Science and Technology	1091-0344
Materials and Manufacturing Processes	1042-6914
Proceedings of the Institution of Mechanical Engineers Part B—Journal of Engineering Manufacture	0954-4054
Packaging Technology and Science	0894-3214
Precision Engineering—Journal of the International Societies for Precision Engineering and Nanotechnology	0141-6359
Production and Operations Management	1059-1478
Production Planning and Control	0953-7287
Research in Engineering Design	0934-9839
Robotics and Computer-Integrated Manufacturing	0736-5845
Soldering and Surface Mount Technology	0954-0911

For each journal, it is reported its title and ISSN code. Journals are sorted alphabetically according to their title

A being the set of cited papers selected for our analysis and $A^{(I)}$, $A^{(II)}$ and $A^{(III)}$ the sets of papers indexed by both the databases, at the moment of session I, II and III respectively.Also, we excluded articles without DOI code or whose DOI code is not indexed by both databases, as they would be difficult to disambiguate.

2. Among the citations, we selected the so-called TO citations, i.e., those obtained from journals purportedly covered by both databases and issued in the 2006–2012 time-window. To avoid any misunderstanding, we excluded citations from journals covered in the 2006–2012 time-window, but later banned from the database.[1] The official lists of documents covered by the databases in use—which are essential for determining the TO citations—were retrieved from the databases' websites (Scopus Elsevier 2015; Thomson Reuters 2015).

The sample of TO citations used in the analysis is the union of the TO citations (that meet the above requirements), collected in each of the three sessions. In formal terms, this sample of TO citations is:

$$B = B^{(I)} \cup B^{(II)} \cup B^{(III)}, \tag{3}$$

$B^{(I)}$, $B^{(II)}$ and $B^{(III)}$ being the TO citations collected during session I, II and III respectively.

This sample of TO citations will be used for estimating the omitted-citations rate of a certain database, in a certain session; the relationship in Eq. 1 can be used, being:$\hat{p}$ the estimate of the omitted-citation rate related to a certain session and a specific database;P the number of (cited) articles of interest;γ_i the number of TO citations relating to the i-th of the P articles of interest;ω_i the portion of the TO citations, collected in a certain session, which are omitted by a specific database.

Being $\hat{p}$ just an estimate of p—albeit the best possible—a relevant symmetrical $(1 - \alpha)$ confidence interval (CI) can be constructed as[2]:

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{\sum_{i=1}^{P}\gamma_i}}, \tag{4}$$

being: α the type-I error; $z_{1-\alpha/2}$ the unit normal deviate corresponding to $1 - \alpha/2$.

In this case, we consider a symmetrical 95 % CI, therefore $\alpha = 5\ \%$ and $z_{97.5\ \%} \approx 2$.

By adopting this procedure, we will obtain six different estimates of the omitted-citation rate, i.e., one for each of the three sessions and each of the two databases in use. The comparison of these estimates will tell us whether the databases examined are able to correct old omitted citations. Figure 3 provides an intuitive representation of a database, which is "diligent" in gradually correcting old dirty data; will Scopus and WoS be like this?

[1] A possible misunderstanding arises from the fact that, in some cases (mostly on Scopus), the expulsion of a journal from a database entails the entire removal of previously indexed papers, while in other cases (mostly on WoS), previously indexed papers are not necessarily removed.

[2] The CI construction in Eq. 4 is grounded on the following considerations:

- For a generic sample consisting of $n = \Sigma\gamma_i$ TO citations, the number of omitted citations will be a binomially distributed variable with mean value $n\bullet\ p$ and variance $n\bullet\ p\bullet(1 - p)$;
- The aforesaid binomial distribution can be approximated by a normal distribution with the same mean value and variance. This approximation is acceptable in the case $n\bullet\ p \geq 5$ (Ross, 2009), which is generally satisfied when considering relatively large sets of TO citations.
- Based on the previous approximation, the percentage of omitted citations for a sample of n TO citations will be a normally distributed variable with mean value p and variance $p\bullet(1 - p)/n$. Since p is not known, it can be replaced by its best estimate $\hat{p}$.

In conclusion, Eq. 4 defines a symmetric CI around $\hat{p}$, which—with a probability $(1 - \alpha)$—will include the "true" p value.

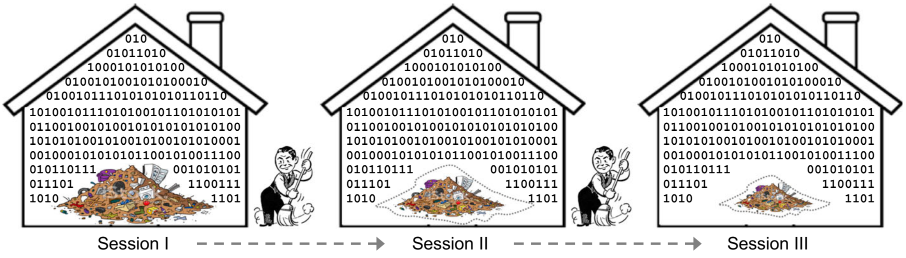


Fig. 3 Intuitive representation of a database able to gradually correct old dirty data

Table 4 Main results of the (repeated) analysis of the omitted-citation rate of databases

Session	$\sum_{i=1}^{P} \gamma_i$	(a) Scopus				(b) WoS			
		$\sum_{i=1}^{P} \omega_i$	$\hat{p}$ (%)	95 % CI		$\sum_{i=1}^{P} \omega_i$	$\hat{p}$ (%)	95 % CI	
I (August 2013)	97,698	5183	5.31	5.16	5.45	7370	7.54	7.37	7.71
II (March 2014)	97,698	4607	4.72	4.58	4.85	6376	6.53	6.37	6.68
III (October 2014)	97,698	4473	4.58	4.44	4.71	6404	6.55	6.40	6.71

Citing and cited articles were issued from 2006 to 2012. Statistics concern each of the three sessions (i.e., session I, II and III) for Scopus and WoS respectively

$P = 23{,}806$ is the total number of (cited) articles, published by 33 Engineering-Manufacturing journals

$\sum \gamma_i$ is the total number of TO citations (which is independent on the session)

$\sum \omega_i$ is the total number of omitted citations relating to each session and each database

$\hat{p}$ is the estimate of the omitted-citation rate relating to each session and each database

The 95 % CI around $\hat{p}$ is obtained applying the approximated relationship in Eq. 4

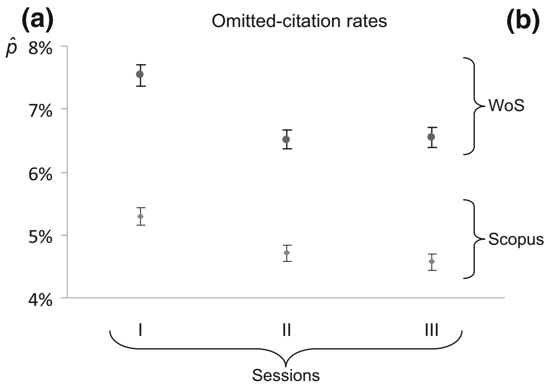
Results

The total number of papers of interest, i.e., those issued by the Engineering-Manufacturing journals examined, is $P = 23{,}806$. The corresponding TO citations are $\Sigma\gamma_i = 97{,}698$. Table 4 contains the $\hat{p}$ values and the relevant 95 % CIs, relating to the three sessions and the two databases examined.

The $\hat{p}$ values of both databases tend to decrease over time, denoting that dirty data have been partially cleaned. Interestingly, the major reduction in the $\hat{p}$ values is between the session I and II for both databases; on the other hand, variations between session II and III are not significant, since the 95 % CIs are partially overlapped (see Fig. 4a); as regards WoS, we can even notice a slight increase in the $\hat{p}$ value between session II and III.

The overall reduction in the number of omitted TO citations ($\Sigma\omega_i$) for WoS is greater than that for Scopus (i.e., $7370-6404 = 966$ against $5183 - 4473 = 710$); however, consistently with what observed in other studies (Franceschini et al. 2014; 2015a), we note that the omitted-citation rates in Scopus are generally lower than those in WoS. Figure 4b shows that the overall percent variations in the $\hat{p}$ values between session I and III are very similar (i.e., -13.7 and -13.1 %, for Scopus and WoS respectively).

Having verified that both databases tend to slowly correct old omitted citations, we now investigate the possible differences in the indexing of individual TO citations, from one



(b) Relevant percent variations

$$\Delta p = \left(\frac{\hat{p}_{final} - \hat{p}_{initial}}{\hat{p}_{initial}} \right)$$

	Δp I to II	Δp II to III	Δp I to III
Scopus	-11.1%	-2.9%	-13.7%
WoS	-13.5%	0.4%	-13.1%

Fig. 4 **a** Graphical representation of the omitted-citation rate in the three sessions, for Scopus and WoS, and **b** relevant percent variations

session to another one. Table 5 summarizes the eight possible events concerning the correct/missing indexing of individual TO citations. Since there are two possible indexing states (i.e., correct or missing indexing) for each of the three sessions, the total number of possible events is $2^3 = 8$; the file containing the complete list of individual TO citations, with the relevant cited papers and their session-by-session indexing by the databases, is available under request to authors.

Not surprisingly, the most frequent events are those with no variation (i.e., the type 1 and 2 events in Table 5), in which the TO citations are indexed correctly ("✔") or incorrectly ("✘") in all the three sessions; the portion of TO citations with no variation is 98.7 % for Scopus and 98.5 % for WoS. The type 3 and 4 events represent corrections in the TO-citation indexing, in session II and III respectively. The total number of corrections in WoS is basically larger than that in Scopus, probably due to the larger level of "initial dirt" in the former database, compared to that one in the latter. Moreover, we note that almost all of the corrections by WoS are concentrated in session II (i.e., 1193 out of 1215).

Despite these differences, the percentage of TO citations corrected by Scopus and WoS are pretty close to each other (i.e., roughly 1 and 1.2 % respectively). This similarity is even more interesting if we consider the fact that, among the set of corrected TO citations, a relatively small subset is shared between the two databases (i.e., 392 citations out of $(997 + 1215 - 392) = 1820$, corresponding to about 21.5 % of the set of corrected TO citations).

The type 5–8 events are characterized by anomalous variations, in which some TO citations, which are correctly indexed in a certain session, are omitted in one (or more) subsequent sessions. It is surprising how citations, which were initially indexed correctly, can come off from a database over time; in other words, these events represent a form of generation of dirty data, which is independent of the introduction of new data in the database. Fortunately, the incidence of these abnormalities is rather low (coincidentally, about 0.3 % for both Scopus and for WoS). In the future, we may conduct a thorough analysis of these anomalies, based on their manual examination; preliminary results show that several of these anomalies involve the so-called "Online-First" articles, whose citations are initially indexed properly but may "disappear" when the final version of an article replaces the Online-First one (see the example in Fig. 5). This is confirmed by other recent studies, i.e., (Haustein et al. 2015; Valderrama-Zurián et al. 2015; Franceschini et al. 2016).

Table 5 Overall statistics concerning the indexing of the individual TO citations, in each session

Type of event		Session			(a) Scopus				(b) WoS			
		I	II	III	Single event		Aggregated events		Single event		Aggregated events	
					TO citations	%	TO citations	%	TO citations	Percent	TO citations	%
No variation	1	✔	✔	✔	92,296	94.5	96,411	98.7	90,195	92.3	96,214	98.5
	2	✘	✘	✘	4115	4.2			6019	6.2		
Correction	3	✘	✔	✔	765	0.8	997	1.0	1193	1.2	1215	1.2
	4	✘	✘	✔	232	0.2			22	0.0		
Anomalous variation	5	✔	✘	✘	102	0.1	290	0.3	164	0.2	269	0.3
	6	✔	✔	✘	112	0.1			77	0.1		
	7	✘	✔	✘	0	0.0			0	0.0		
	8	✔	✘	✔	76	0.1			28	0.0		
				Total	97,698	100	97,698	100	97,698	100	97,698	100

Symbols “✔” and “✘” respectively identify the TO citations correctly indexed or omitted in a certain session

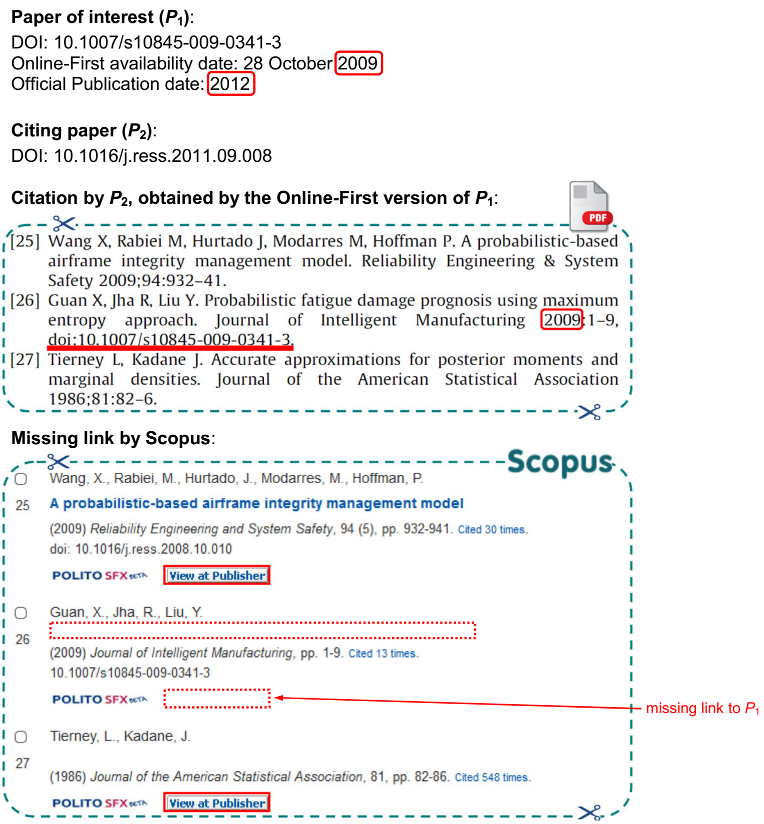


Fig. 5 Example of "disappearance" of a citation obtained by the Online-First version of a paper (i.e., P_1, issued in October 2009), after the publication of the relevant official version (in 2012). The Scopus database was queried in August 2015

Conclusions

The analysis presented in this paper shows that the two bibliometric databases examined tend to gradually reduce the number of old omitted citations, although this reduction is relatively slow for both. It would be interesting to see to what extent these cleanings were due to error-correction campaigns structured by database administrators, or simply due to impromptu database-inaccuracy reports by authors and/or database users (even checking and cleaning up bibliometric data in personal research profiles, such as ResearcherID, Scopus Author ID, ORCID, etc.).

Results of this study show other interesting similarities/coincidences between the two databases examined:

1. Comparing the results related to session I and III (spaced about 14 months apart), we noticed a 13–14 % reduction in the p values for both Scopus and WoS.
2. For both databases, the greatest reduction in the omitted-citations rate was registered in session II and not in session III. This could be just a coincidence or it could denote a sort of "seasonality" of the two databases in cleaning up old dirty data.

3. The portion of TO citations whose indexing varies in the three sessions is roughly the same for both databases, i.e., roughly 1–1.5 %. Apart from the previously omitted TO citations that have been justly corrected, they include a small portion of abnormal variations, i.e., TO citations correctly indexed in some session and subsequently omitted. Coincidentally, the percentage of abnormal variations is 0.3 % for both databases.

The proposed analysis has several limitations:

- Being based on the use of the automated algorithm for estimating the omitted-citation rate, described in (Franceschini et al. 2013), this analysis "inherits" its *pros* and *contras*;
- Even though the set of TO citations includes almost one-hundred thousands citations, the relevant cited papers are all confined within the Engineering-Manufacturing field.
- The analysis was repeated in three sessions over a total period of about 14 months; therefore, it reflects a database's ability to correct errors in short/middle-term period, but not in the long-term period.
- We only examined papers with DOI code in both Scopus and WoS. Since there is no proof that papers with DOI behave in the same way as those without DOI, analysis results are limited to the former type of papers.

In the future, we plan to extend the study to a longer time-scale (e.g., 2 or 3 years) and/or to scientific articles in other disciplines. Furthermore, the study will be expanded for investigating possible links between the omitted citations' propensity to be corrected and the publishers of the relevant citing papers.

Appendix: Model for correcting the omitted-citation rate considering phantom citations

This section presents an analytical model for correcting the $\hat{p}$ values, taking account of the—albeit small—distortion produced by the phantom citations of the databases in use.

The schematic representation in Fig. 6 shows that the phantom citations generated by a certain database can contribute to generating *false TO-citations* and, consequently, *false omitted citations* (δ) related to the competing database. For example, among the (ω_{Scopus}) presumed[3] omitted citations by Scopus, δ_{Scopus} are false due to phantom citations by WoS.

Considering Scopus and WoS, Eq. 1 can be expressed in a compact way as:

$$\begin{aligned} \hat{p}_{\text{Scopus}} &= \sum_{i=1}^{P} (\omega_i)_{\text{Scopus}} \Big/ \sum_{i=1}^{P} \gamma_i = \omega_{\text{Scopus}}/\gamma \\ \hat{p}_{\text{WoS}} &= \sum_{i=1}^{P} (\omega_i)_{\text{WoS}} \Big/ \sum_{i=1}^{P} \gamma_i = \omega_{\text{WoS}}/\gamma \end{aligned}, \tag{5}$$

where ω_{Scopus} and ω_{WoS} are respectively the total number of (presumed) omitted citations related to the Scopus and WoS database;γ is the total number of (presumed) TO-citations available.

[3] The adjective "presumed" indicates that a portion of the omitted citations by one database (e.g., Scopus) can be *false*, due to phantom citations generated by the competing database (i.e., WoS in this case).

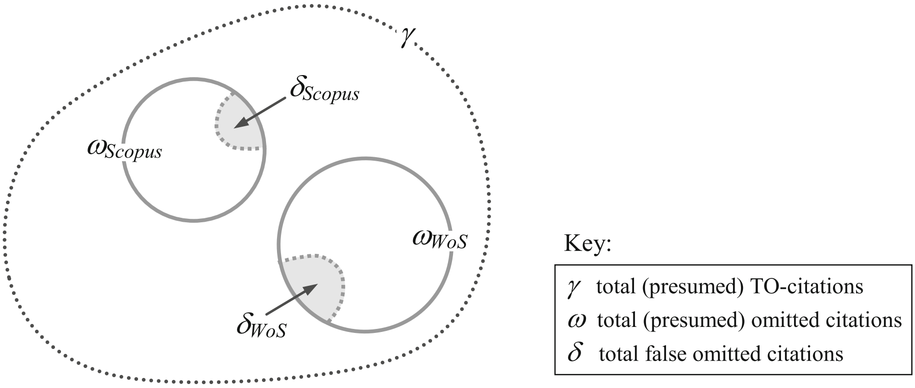


Fig. 6 Schematic representation of the false omitted citations (i.e., δ) related to a database, due to phantom citations by the competing database

Equation 5 provides an estimate of one database's omitted citation rate, which can be distorted by the presence of phantom citations by the competing database.

We define the phantom-citation rate (α) of one database, as the ratio of the number of phantom-citations generated by that database—which coincides with the number of false omitted citations related to the competing database (δ)—and the number of (presumed) TO-citations available (γ):

$$\begin{aligned} \alpha_{\text{Scopus}} &= \delta_{\text{WoS}}/\gamma \\ \alpha_{\text{WoS}} &= \delta_{\text{Scopus}}/\gamma \end{aligned}. \tag{6}$$

From Eq. 1, we obtain:

$$\begin{aligned} \delta_{\text{WoS}} &= \gamma \cdot \alpha_{\text{Scopus}} \\ \delta_{\text{Scopus}} &= \gamma \cdot \alpha_{\text{WoS}} \end{aligned}. \tag{7}$$

The *corrected* number of TO-citations (γ')—i.e., excluding the *false* ones, that is to say that ones produced by phantom citations by Scopus (δ_{WoS}) and WoS (δ_{Scopus})—will be:

$$\gamma' = \gamma - \left(\delta_{\text{WoS}} + \delta_{\text{Scopus}}\right) = \gamma \cdot \left[1 - \left(\alpha_{\text{Scopus}} + \alpha_{\text{WoS}}\right)\right]. \tag{8}$$

The *corrected* number of omitted citations (i.e., excluding the *false* ones) of the two databases will be:

$$\begin{aligned} \omega'_{\text{Scopus}} &= \omega_{\text{Scopus}} - \delta_{\text{Scopus}} = \omega_{\text{Scopus}} - \alpha_{\text{Scopus}} \cdot \gamma \\ \omega'_{\text{WoS}} &= \omega_{\text{WoS}} - \delta_{\text{WoS}} = \omega_{\text{WoS}} - \alpha_{\text{WoS}} \cdot \gamma \end{aligned}. \tag{9}$$

We define the *corrected* omitted-citation rate (p') for both databases as:

$$\begin{aligned} p'_{Scopus} &= \frac{\omega'_{\text{Scopus}}}{\gamma'} = \frac{\omega_{\text{Scopus}} - \alpha_{\text{Scopus}} \cdot \gamma}{\gamma \cdot \left[1 - \left(\alpha_{\text{Scopus}} + \alpha_{\text{WoS}}\right)\right]} = \frac{\hat{p}_{\text{Scopus}} - \alpha_{\text{Scopus}}}{1 - \left(\alpha_{\text{Scopus}} + \alpha_{\text{WoS}}\right)} \\ p'_{\text{WoS}} &= \frac{\omega'_{\text{WoS}}}{\gamma'} = \frac{\hat{p}_{WoS} - \alpha_{\text{WoS}}}{1 - \left(\alpha_{\text{Scopus}} + \alpha_{\text{WoS}}\right)} \end{aligned}. \tag{10}$$

We remark that, having estimated the phantom-citation rate (α) of the databases in use, the formulae in Eq. 10 can be used to correct the $\hat{p}$ values resulting from the application of the automated algorithm, taking account of the distortions produced by phantom citations.

Recent researches indicate that the phantom citation rate related to WoS is $\alpha_{\mathrm{WoS}} \approx 0.5\ \%$ (Garcia-Pérez 2010; Olensky et al. 2016). Assuming that the one related to Scopus (α_{Scopus}) is of the same order of magnitude, we note that α_{Scopus} and α_{WoS} are roughly one order of magnitude lower than the typical $\hat{p}$ values (see Table 4). For this reason, they can be neglected and Eq. 10 becomes:

$$\begin{array}{c} p'_{\mathrm{Scopus}} \approx \hat{p}_{\mathrm{Scopus}} \\ p'_{\mathrm{WoS}} \approx \hat{p}_{\mathrm{WoS}} \end{array}, \tag{11}$$

confirming the simplifying assumption that the incidence of "phantom citations" is generally negligible.

References

Buchanan, R. A. (2006). Accuracy of cited references: The role of citation databases. *College and Research Libraries, 67*(4), 292–303.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2013). A novel approach for estimating the omitted-citation rate of bibliometric databases. *Journal of the American Society for Information Science and Technology, 64*(10), 2149–2156.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2014). Scientific journal publishers and omitted citations in bibliometric databases: Any relationship? *Journal of Informetrics, 8*(3), 751–765.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2015a). Influence of omitted citations on the bibliometric statistics of the major Manufacturing journals. *Scientometrics, 103*(3), 1083–1122.

Franceschini, F., Maisano, D., Mastrogiacomo, L. (2015b). On the correction of "old" omitted citations by bibliometric databases. In *Proceedings of the 15th international society of scientometrics and informetrics (ISSI) conference*, pp. 1200–1207, 29 June–3 September 2015, Istanbul, Turkey, ISBN: 978-975-518-381-7.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). The museum of errors/horrors in Scopus. *To appear in Journal of Informetrics, 10*, 174–182.

Garcia-Pérez, M. A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h-indices in psychology. *Journal of the American Society for Information Science and Technology, 61*(10), 2070–2085.

Haustein, S., Bowman, T. D., Costas, R. (2015). When is an article actually published? An analysis of online availability, publication, and indexation dates. In *Proceedings of the 15th international society of scientometrics and informetrics (ISSI) conference*, pp. 1170–1179, 29 June–3 September 2015, Istanbul, Turkey, ISBN: 978-975-518-381-7.

Jacsó, P. (2006). Deflated, inflated and phantom citation counts. *Online Information Review, 30*(3), 297–309.

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.

Moed, H. F. (2006). *Citation analysis in research evaluation* (Vol. 9). New York: Springer.

Olensky, M. (2013). Accuracy Assessment for Bibliographic Data. In *Proceedings of the 13th international conference of the international society for scientometrics and informetrics (ISSI)* (Vol. 2, pp. 1850–1851). Vienna, Austria.

Olensky, M., Schmidt, M., & van Eck, N. J. (2016). Evaluation of the citation matching algorithms of CWTS and iFQ in comparison to the Web of science. *To appear in Journal of the Association for Information Science and Technology,*. doi:10.1002/asi.23590.

Ross, S. M. (2009). *Introduction to probability and statistics for engineers and scientists*. New York: Academic Press.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician, 55*(3), 182–186.

Scopus Elsevier (2015). *Scopus Content Coverage*. Available at http://www.scopus.com. Retrieved August 2013, March 2014 and October 2014.

Thomson Reuters. (2015). *Master journal list*, http://ip-science.thomsonreuters.com/mjl/[retrieved. Retrieved August 2013, March 2014 and October 2014.

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.