Empirical analysis and classification of database errors in Scopus and Web of Science

Empirical analysis and classification of database errors in Scopus and Web of Science

Fiorenzo Franceschini[1], Domenico Maisano, Luca Mastrogiacomo

[1] *fiorenzo.franceschini@polito.it*
Politecnico di Torino, DIGEP (Department of Management and Production Engineering),
Corso Duca degli Abruzzi 24, 10129, Torino (Italy)

Abstract

In the last decade, a growing number of studies focused on the qualitative/quantitative analysis of bibliometric-database errors. Most of these studies relied on the identification and (manual) examination of relatively limited samples of errors.

Using an automated procedure, we collected a large *corpus* of more than 10,000 errors in the two multidisciplinary databases Scopus and Web of Science (WoS), mainly including articles in the Engineering-Manufacturing field. Based on the manual examination of a portion (of about 10%) of these errors, this paper provides a preliminary analysis and classification, identifying similarities and differences between Scopus and WoS.

The analysis reveals interesting results, such as: (i) although Scopus seems more accurate than WoS, it tends to forget to index more papers, causing the loss of the relevant citations given/obtained, (ii) both databases have relatively serious problems in managing the so-called *Online-First* articles, and (iii) lack of correlation between databases, regarding the distribution of the errors in several error categories.

The description is supported by practical examples concerning a variety of errors in the Scopus and WoS databases.

Keywords: Data accuracy, Database error, Omitted citation, Error classification, Phantom citation, Scopus, Web of Science.

1. Introduction

Bibliometric databases are commonly adopted by individual scientists and research institutions for (i) searching scientific documents, (ii) providing information on the citation impact of the scientific output, and (iii) supporting the selection of the scientific journals where to publish.

The abundance of bibliometric and/or bibliographic disciplinary databases (e.g., PubMed, MathSciNet, PsycINFO, IEEEXplore, EconLit, etc.) contrasts with the relatively limited number of multidisciplinary databases: Google Scholar (GS), Scopus, and Web of Science (WoS). A peculiarity of GS is to automatically index publications/citations through web crawlers, which allows to achieve considerably more coverage than Scopus and WoS. In fact, GS is estimated to contain approximately 160M total documents, while Scopus approximately 13M and WoS

approximately 10M (Orduna-Malea et al., 2015; Mongeon and Paul- Hus, 2016). Unfortunately, the automatic indexing of GS inevitably causes many errors (Labbé, 2010) and (almost) completely disqualifies GS with respect to its two competitors, to the extent that most consider GS simply as a search engine, certainly not a serious bibliometric database. Nevertheless, some recent studies indicate that the GS data quality is gradually improving (Moed, et al., 2016; Prins et al., 2016). Furthermore, the data quality of GS, Scopus and WoS were discussed in a number of comparative studies addressing coverage and overlap (e.g., Meho and Yang, 2007; Archambault et al., 2009; Mikki, 2010; Wildgaard, 2015; Harzing and Alakangas, 2016; Wang and Waltman, 2016).

In the last two years, we have been investigating the Scopus and WoS errors, analyzing the so-called *omitted citations* – i.e., missing links between citing and cited papers – which represent one of the major consequences of database errors (Franceschini et al., 2013). An interesting result – which corroborates the findings of previous studies (Moed and Vriens, 1989; Moed, 2002; Moed, 2005; Buchanan, 2006; Larsen et al., 2007; Hildebrandt and Larsen, 2008; Tunger et al., 2010; Olensky, 2015) – is that the omitted-citation rate of the two databases is far from being negligible: more than 4% for Scopus and more than 6% for WoS (Franceschini et al., 2014). We showed that the editorial style of some publishers can favour database errors and – although Scopus and WoS tend to be more and more careful in indexing new papers – they do little to correct the errors already present in the database (Franceschini et al., 2014; Franceschini et al., 2016a). Also, we came across many weird errors, discussed in a recent "opinion" paper (Franceschini et al., 2016b).

The majority of our past researches relied on the analysis of a relatively large *corpus* of scientific articles, consisting of almost 24,000 cited articles – confined to the Engineering-Manufacturing field – and almost 100,000 corresponding citing articles. Among these articles, thousands of omitted citations were identified using an automated algorithm, which requires the combined use of Scopus and WoS 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 (Franceschini et al., 2013).

In our previous researches (Franceschini et al., 2013, 2014, 2015a, 2016a), we analyzed the Scopus and WoS omitted citations, studying the influence of several factors, such as journal or publisher of cited papers, issue year of citing papers, date of database queries, etc.. However, we did not investigate the causes of these omitted citations – i.e., the nature of database errors – in a detailed and structured way.

Consistently with the categorization suggested by Buchanan (2006), (at least two) types of database errors can be defined:

A. *Pre-existing errors*: errors made by authors/editors/publishers when preparing the list of cited articles for their publication; e.g., errors in the author name(s), article title, issue year, volume number, pagination, etc..

B. *Database mapping errors*: failures to establish an electronic link between a cited article and the corresponding citing articles that can be attributed to data-entry errors in the database; e.g., transcription errors, cited article omitted from a cited-article list, etc..

While the errors in the first category are (at least partly) justifiable, being caused by inaccuracies in the original papers, those in the second one are introduced by databases, in the data-entry process.

The goal of this paper is to delve into the large corpus of omitted citations available from our past research and perform a statistical analysis of the relevant database errors, trying to answer to the following research questions:

- *What are the more frequent errors of Scopus and WoS and the similarities and differences between the two databases?*
- *Are the results of this research in line with those of other researches in the field of bibliometric-database errors?*
- *Does this research provide a representative picture of the Scopus and WoS errors?*
- *In the light of the results obtained, what are the practical implications to users and administrators of the Scopus and WoS databases?*

The proposed statistical analysis requires a thorough manual examination of the database records and the original cited/citing papers, with special attention to the cited-article lists. Due to the relatively large time consumption of this process, it will be limited to the 10% of the (more than 10,000) omitted citations available.

The remainder of the paper is organized into five sections. Sect. 2 recalls the automated algorithm for detecting omitted citations. Sect. 3 illustrates the analysis methodology in detail and presents some indicators for estimating the rate of the so-called *phantom-citations* of the two databases. Sect. 4 describes the analysis results; the description is supported by practical examples concerning various errors in Scopus and WoS. Sect. 5 summarizes the original contributions of this paper, describing its implications and limitations. Additional information is contained in the appendix.

2. 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 Tab. 1).

The union of the citations recorded by the two databases is a total of eight citations. Among these citations, only five come from sources (i.e., journals or conference proceedings) officially covered by both databases (highlighted in grey in Tab. 1). 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 ≈ 40% in Scopus and 1/5 ≈ 20% in WoS. The same reasoning can be extended to multiple papers of interest and more than two bibliometric databases.

Tab. 1. Citation data relating to a fictitious article, according to Scopus and WoS. 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 grey).

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 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 are 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, relating to this database[1]. The omitted-citation rate (p) relating to the P papers of interest, according to a database, can be estimated as:

$$p = \omega / \gamma , \tag{1}$$

where $\gamma = \sum_{i=1}^{P} \gamma_i$ is the total number of TO citations available and $\omega = \sum_{i=1}^{P} \omega_i$ is the corresponding number of omitted TO citations.

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

- 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

[1] We remark that, according to the automated algorithm, the citations omitted by one database are correctly indexed by the other one; the use of the latter database merely represents an expedient to identify these omitted citations automatically.

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*.

- The estimation of *p* is performed on the basis of (i) a set of papers of interests and (ii) 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 omitted citations is uniform.
- It is assumed that the incidence of *phantom citations* – i.e., *false* citations from papers that did not actually cite the target paper, which are generally due to the use of non-sufficiently sophisticated citation-matching algorithms (Garcia-Pérez, 2010) – is negligible. According to our algorithm, a phantom citation of one database may lead to an incorrect notification of omitted citation for the other database. The analysis proposed in this paper will also allow to answer the following additional research question:
  What are the phantom-citation rates of Scopus and WoS and how can they be used to correct the p values estimated through the automated algorithm?
- 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 reasons: (i) some of these publication types are not covered by both the databases in use and (ii) lack of exhaustive official lists concerning the coverage of these publication types.

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

3. Methodology

This study is based on an extended dataset, which was also used for other investigations (Franceschini et al., 2014, 2015a, 2016b). We identified a sample of papers of interest (or cited papers) issued by 33 scientific journals (i) included in the ISI Subject Category of Engineering-Manufacturing (by WoS) and (ii) covered by Scopus; Table A.1 (in the appendix) reports the list of these journals. For each journal, we considered the set of papers published in the time-window from 2006 to 2012 and indexed by both databases, and the citations that they obtained from papers issued in the same period. 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-to-2012 time-window. To avoid any misunderstanding, we excluded citations from journals covered in the 2006-to-2012 time-window, but later banned from the database[2]. The official lists of documents covered by the

[2] 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.

databases in use – which are essential for determining the TO citations – were retrieved from the databases' websites (Scopus Elsevier, 2016; Thomson Reuters, 2016).

The total number of cited papers, i.e., those issued by the journals examined, is $P = 23{,}806$; the corresponding TO citations are $\gamma = 97{,}968$. Tab. 2 contains the relevant number of omitted citations and the estimate of the omitted-citation rates concerning the two databases (i.e., p, determined using the relationship in Eq. 1).

Tab. 2. Number of omitted TO citations resulting from the application of the automated algorithm (the corresponding percentage values in brackets).

	Scopus		WoS	
Total TO citations (γ)	97,698	(100.00%)	*idem*	(*idem*)
Indexed TO citations ($\gamma - \omega$)	93,225	(95.42%)	91,294	(93.45%)
Omitted TO citations (ω)	4,473	($p = 4.58\%$)	6,404	($p = 6.55\%$)

We notice that the total number of omitted TO citations available is relatively large (i.e., 4,473 for Scopus and 6,404 for WoS, corresponding to total 10,877 omitted TO citations). These data (which we make available on request) were collected relatively quickly, using the automated algorithm described in Sect. 2.

Omitted citations are just the ultimate effect of database errors of different nature; the identification and classification of the errors requires several manual activities:

1. Examination of database records;
2. Examination of the original cited/citing papers (e.g., their PDFs), with special attention to the relevant reference lists;
3. Classification of errors into suitable categories.

These manual activities are very time consuming[3], also because of the different editorial styles of scientific journals, which may complicate error detection. For this reasons, we analyzed only the 10% of the omitted TO citations available, i.e., approximately 447 for Scopus and 640 for WoS.

As anticipated in Sect. 1, database errors can be classified into the two categories: *type-A* or *pre-existing errors* and *type-B* or *database mapping errors*. In the case of type-A errors, databases are unable to identify and correct inaccuracies already present in the cited-article list of (citing) papers, using the available information; e.g., in the presence of an error in the author name of a cited article, the corresponding title, volume number and pagination can be used to identify and correct it. On the other hand, type-B errors are far more serious, as they are caused by inaccuracies introduced by one database in data transcription. For each of the above two categories, we will define and describe several sub-categories (see Sect. 4).

The results of the manual analysis can also be used to quantify the phantom-citation rate of Scopus and WoS. The schematic representation in Fig. 1, shows that the phantom citations of one database

[3] Based on our experience, they requires about 15-20 minutes for each error.

– if they are (mistakenly) assigned to papers that are supposed to be covered by the other database – may lead to generate *false TO citations* and, consequently, *false omitted TO citations* (δ) for the other database. For example, among the (ω_{Scopus}) TO citations omitted by Scopus, δ_{Scopus} are false due to phantom citations by WoS. To be rigorous, these omitted TO citations are just *presumed*, as some of them can be false. For the same reason, even the total (γ) TO citations available are just *presumed*, as some of them can be *false*, due to phantom citations generated by both the databases (i.e., δ_{WoS} and δ_{Scopus} for Scopus and WoS respectively).

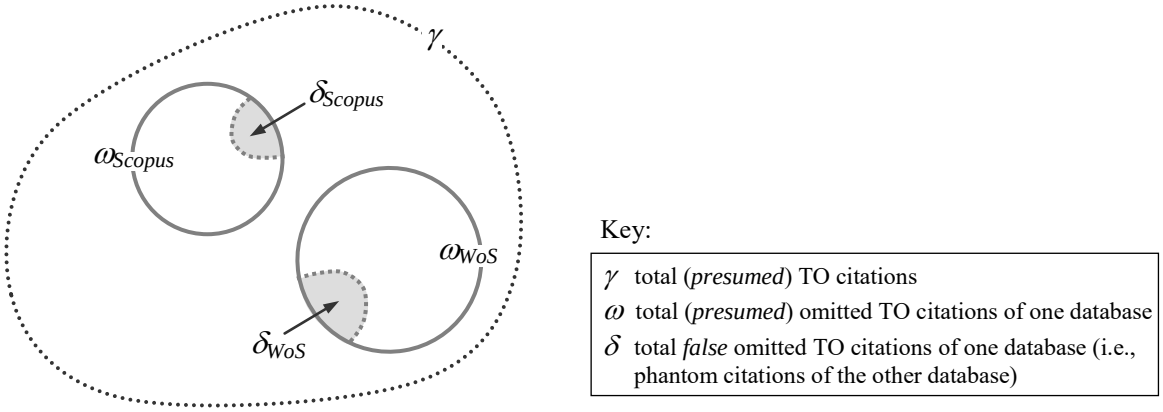


Fig. 1. Schematic representation of the false omitted TO citations (i.e., δ) related to one database, due to phantom citations of the other database.

In Sect. A2 (in the appendix) we go into this point, illustrating a practical way to estimate the phantom-citation rate (α) of databases. The α estimates can be in turn used to correct the omitted-citation rates (p) reported in Tab. 2.

Tab. 3 shows the formulae and concise descriptions of some indicators concerning phantom citations; details on their construction are contained in Sect. A2 (in the appendix).

Tab. 3. Indicators constructed for estimating the influence of phantom citations and correcting the omitted-citation rates. For details on the construction of these indicators, see Sect. A2 (in the appendix).

Indicator description	for Scopus	for WoS
1. Omitted-citation rate (p).	$p_{Scopus} = \omega_{Scopus} / \gamma$	$p_{WoS} = \omega_{WoS} / \gamma$
2. Number of (presumed) omitted citations, which were analyzed manually (o).	$o_{Scopus} \approx 10\% \cdot \omega_{Scopus}$	$o_{WoS} \approx 10\% \cdot \omega_{WoS}$
3. Number of *false* omitted citations of one database (i.e., phantom citations of the other database), detected through the manual analysis (d).	d_{Scopus} (count)	d_{WoS} (count)
4. (Estimated) phantom-citation rate (α).	$\alpha_{Scopus} \approx \frac{d_{WoS}}{o_{WoS}} \cdot p_{WoS}$	$\alpha_{WoS} \approx \frac{d_{Scopus}}{o_{Scopus}} \cdot p_{Scopus}$
5. *Corrected* number of TO citations (i.e., excluding the false ones) (γ').	$\gamma' = \gamma \cdot [1 - (\alpha_{Scopus} + \alpha_{WoS})]$	*idem*
6. *Corrected* number of omitted TO citations (i.e., excluding the false ones) (ω').	$\omega'_{Scopus} = \omega_{Scopus} - \alpha_{Scopus} \cdot \gamma$	$\omega'_{WoS} = \omega_{WoS} - \alpha_{WoS} \cdot \gamma$
7. *Corrected* omitted-citation rate (i.e., excluding the false omitted citations) (p').	$p'_{Scopus} = \frac{p_{Scopus} - \alpha_{Scopus}}{1 - (\alpha_{Scopus} + \alpha_{WoS})}$	$p'_{WoS} = \frac{p_{WoS} - \alpha_{WoS}}{1 - (\alpha_{Scopus} + \alpha_{WoS})}$

4. Analysis results

Before identifying and classifying the errors behind omitted citations, it is appropriate to discriminate between *false* and *authentic* omitted citations. Among the (presumed) omitted TO citations of one database, we estimated the portion of *false* ones, corresponding to phantom citations produced by the other database. Fig. 2 exemplifies a phantom citation produced by WoS, which caused a false omitted TO citation in Scopus. This citation is due to the erroneous substitution of an authentic cited article (P_1) with a false one (P_2) – with same authors, issue year and volume number – in the list of a citing article (P_3). In this case, the error of WoS is twofold: (i) omitted citation related to P_1 and (ii) phantom citation related to P_2.

Tab. 4 contains some indicators concerning the incidence of phantom citations and the correction of the omitted-citation rates. For details, see Sect. A2 (in the appendix).

The incidence of phantom citations in WoS is higher than that in Scopus ($\alpha_{WoS} \approx 0.46\%$ against $\alpha_{Scopus} \approx 0.10\%$). The value of α_{WoS} is in line with that one estimated in other studies – i.e., roughly 0.5% (Garcia-Perez, 2010; Olensky et al., 2016). On the other hand, the estimate of α_{Scopus} represents a novelty in the state of the art.

Authentic cited paper (P_1):

Authors: Wu, Y., Song, Q., Liu, S.
Title: A Normalized Adaptive Training of Recurrent Neural Networks With Augmented Error Gradient
Source: IEEE Transactions on Neural Networks, 19(2): 351-356, 2008
DOI: 10.1109/TNN.2007.908647

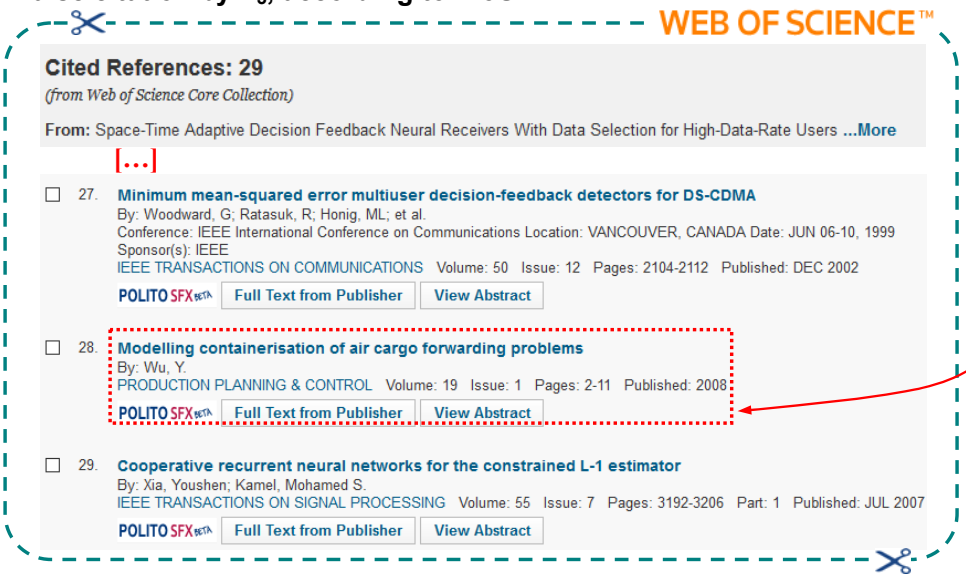
Erroneous cited paper (P_2):

Authors: Wu, Y., Song, Q., Liu, S.
Title: Modelling containerisation of air cargo forwarding problems
Source: Production Planning & Control: The Management of Operations 19(1): 2-11, 2008
DOI: 10.1080/09537280701524168

Citing paper (P_3):

Authors: de Lamare R.C., Sampaio-Neto R.
Title: Space-Time Adaptive Decision Feedback Neural Receivers With Data Selection for High-Data-Rate Users in DS-CDMA Systems
Source: IEEE Transactions on Neural Networks, 19(11):1887-1895, 2008
DOI: 10.1109/TNN.2008.2003286

False citation by P_3, according to WoS:



Original list of P_3:

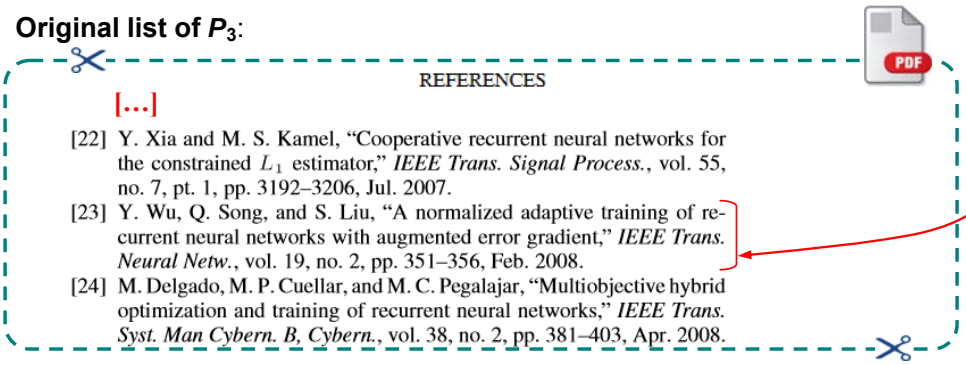


Fig. 2. Example of phantom citation produced by WoS. This citation is due to the erroneous substitution of an authentic cited paper (P_1) with a false one (P_2) – with same authors, issue year and volume number – in the list of a citing paper (P_3). The WoS database was queried in January 2016.

The estimated phantom-citation rates can be used to correct the omitted-citation rates (p) – through the formulae at point 7 in Tab. 3 (for details, see Sect. A2 in the appendix). For both databases, the corrected omitted-citation rates (p') are slightly lower than the initial ones (see Tab. 4).

Tab. 4. Indicators concerning the incidence of phantom citations in the data examined and the correction of the omitted-citation rates.

Parameter	Description	Scopus	WoS
o	No. of omitted TO citations, which have been examined manually	447	640
d	No. of *false* omitted TO citations detected	45	10
$o - d$	No. of *authentic* omitted TO citations	402	630
α	(Estimated) phantom-citation rate	0.10%	0.46%
p	Initial omitted-citation rate	4.58%	6.55%
p'	Corrected omitted-citation rate	4.12%	6.46%

Let us now focus the attention on (i) the *authentic* omitted TO citations, which have been examined manually (i.e., $o - d$) and (ii) the detection and classification of the errors behind them. Tab. 5 summarizes the results of our analysis. It can be seen that the two errors categories (i.e., type-A and type-B) are decomposed into several sub-categories, which depict the specific error causes. These sub-categories are not so different from those identified in other studies (Buchanan, 2006; Olensky, 2015) and their definition is functional to the subsequent description of the more frequent errors detected.

For each (k-th) sub-category, we report the number of errors found and two corresponding frequency indicators, according to the formulae:

$$freq_k^{(1)} = \frac{\text{no. of errors in the } (k^{\text{th}}) \text{ subcategory}}{o - d}, \qquad (2)$$
$$freq_k^{(2)} = freq._k^{(1)} \cdot p'$$

where

$freq_k^{(1)}$ depicts the incidence of a certain (k-th) error sub-category, with respect to the totality of the errors of one database; e.g., for Scopus $freq_{A.1}^{(1)} = 40/402 \approx 10.0\%$.

$freq_k^{(2)}$ estimates the incidence of a certain (k-th) error sub-category, with respect to the totality of the authentic TO citations (i.e., both those indexed and those omitted); e.g., for Scopus the $freq_{A.1}^{(2)} = 10.0\% \cdot 4.12\% \approx 0.41\%$. In other words, this indicator represents the fraction of TO citations omitted due to a certain error sub-category.

Tab. 5. Classification of the errors detected and corresponding frequency indicators (i.e., $freq_k^{(1)}$ and $freq_k^{(2)}$) in the two databases.

		Scopus			WoS		
Type-A or *pre-existing errors*		No.	$freq_k^{(1)}$	$freq_k^{(2)}$	No.	$freq_k^{(1)}$	$freq_k^{(2)}$
A.1	Missing/wrong article title	40	10.0%	0.41%	91	14.4%	0.93%
A.2	Errors in the other fields	18	4.5%	0.18%	99	15.7%	1.01%
	Subtotal	58	14.4%	0.59%	190	30.2%	1.95%
Type-B or *database mapping errors*							
B.1	Errors in the transcription of author name(s) and/or article title	13	3.2%	0.13%	161	25.6%	1.65%
B.2	Incomplete cited-article list	9	2.2%	0.09%	11	1.7%	0.11%
B.3	Omitted cited-article list	8	2.0%	0.08%	14	2.2%	0.14%
B.4	Wrong or missing DOI	9	2.2%	0.09%	14	2.2%	0.14%
B.5	Errors concerning Online-First articles	74	18.4%	0.76%	67	10.6%	0.69%
B.6	Unindexed (citing) articles	127	31.6%	1.30%	16	2.5%	0.16%
B.7	Reasons unknown	104	25.9%	1.07%	157	24.9%	1.61%
	Subtotal	344	85.6%	3.53%	440	69.8%	4.51%
		(*o*–*d*)		*p'*	(*o*–*d*)		*p'*
	Total	402	100.0%	4.12%	630	100.0%	6.46%

Regarding the error contributions, we note that type-B errors predominate over type-A ones, for both databases; two of the possible reasons are:

1. The improved efforts by reviewers/editors/publishers in checking and correcting inaccuracies in the cited-article lists probably contribute to reduce the incidence of pre-existing errors (Franceschini et al., 2016a).
2. The citation matching algorithms of bibliometric databases are probably more and more robust in establishing the correct link between cited and citing articles, even in the presence of type-A errors (Meester et al., 2016). In particular, the citation matching algorithm of Scopus seems more effective than that of WoS, as evidenced by the smaller portion of type-A errors (i.e., 0.59% of the TO citations in Scopus, against 1.95% in WoS).

This result is in partial contradiction with the output of the research by Olensky (2015), showing a higher incidence of type-A errors with respect to type-B ones.

The following two subsections examine the type-A and type-B errors in detail, describing the relevant sub-categories individually. The description is supported by various practical examples.

4.1 Type-A errors

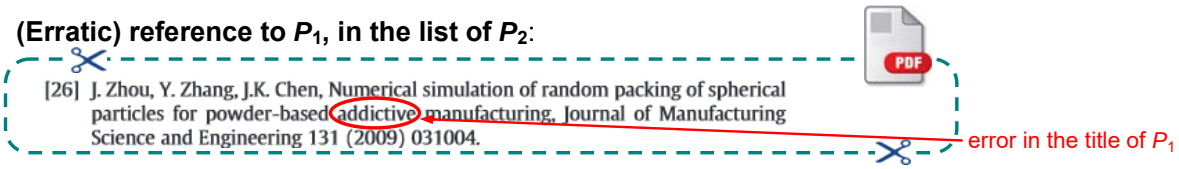
(A.1) *Missing/wrong article title*

For both databases, a very frequent type-A errors concern the missing/wrong title of articles in the reference list of the (citing) papers. See the example in Fig. 3, in which a mistake in the title of a paper (P_1), reported in the list of another paper (P_2), probably compromises the citation match.

Cited paper (P_1):
Authors: J. Zhou, Y. Zhang, J.K. Chen
Title: Numerical simulation of random packing of spherical particles for powder-based additive manufacturing
Source: Journal of Manufacturing Science and Engineering, 131(3): 31004-31012
DOI: 10.1115/1.3123324

Citing paper (P_2):
Authors: T. Jia, Y. Zhang, J.K. Chen, Y.L. He
Title: Dynamic simulation of granular packing of fine cohesive particles with different size distributions
Source: Powder Technology, 218(2012): 76-85
DOI: 10.1016/j.powtec.2011.11.042

(Erratic) reference to P_1, in the list of P_2:



Cited-article list of P_2, according to Scopus:

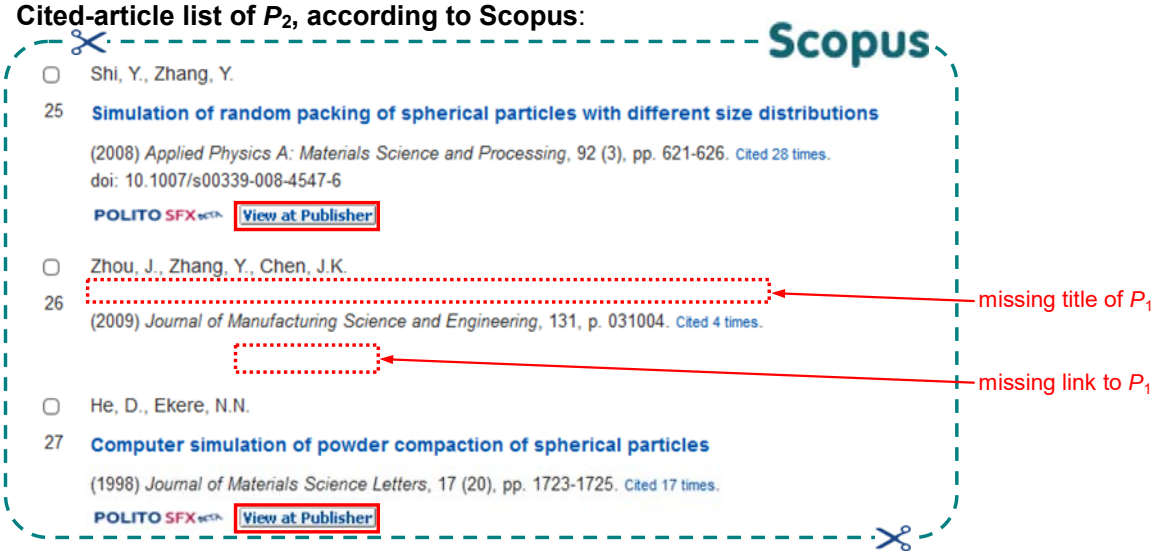


Fig. 3. Example of type-A error in the title of a (cited) paper (P_1), reported in the list of a citing paper (P_2). This error is classified in sub-category A.1-*Missing/wrong article title*. The Scopus database was queried in January 2016.

We also found many references that do not even include the title of the (cited) papers. In some cases journals allow (or even encourage) the use of citation styles in which the title of the cited papers is omitted. This probably increases the risk of generating omitted citations, especially in WoS (see the examples in Fig. 4 and Fig. 5). This result somehow contradicts what inferred by Olensky (2015), i.e., that neither Scopus nor WoS seem to use the article title in the citation-matching process. Our opinion is that, although the presence of (accurate) titles in the cited-article list is not indispensable for the correct citation matching, it probably helps. The only way to dissolve this doubt would be to know the citation matching algorithms Scopus and WoS, which, unfortunately, are not and will probably never be public.

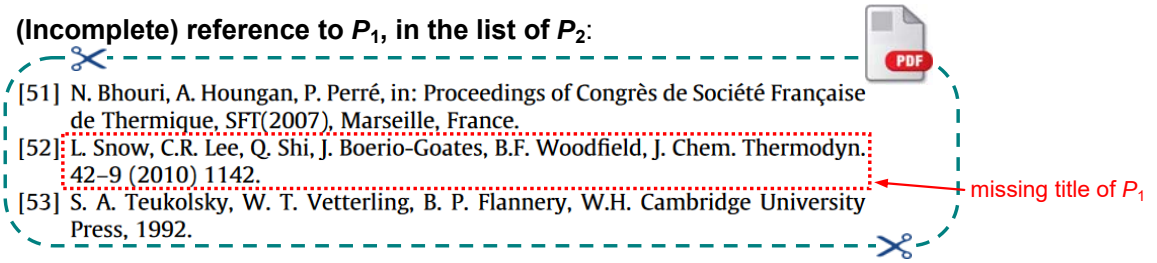
Cited paper (P_1):

Authors: L. Snow, C.R. Lee, Q. Shi, J. Boerio-Goates, B.F. Woodfield
Title: Size-dependence of the heat capacity and thermodynamic properties of hematite (α-Fe_2O_3)
Source: The Journal of Chemical Thermodynamics, 42(9): 1142-1151
DOI: 10.1016/j.jct.2010.04.009

Citing paper (P_2):

Authors: N. Bhouri, S. Bennasrallah, P. Perre
Title: Influence of geometrical structure on sorption isotherms of Jersey and yarns made of cotton at two temperatures
Source: Microporous and Mesoporous Materials, 163(2012): 76-84
DOI: 10.1016/j.micromeso.2012.07.024

(Incomplete) reference to P_1, in the list of P_2:



Cited-article list of P_2, according to WoS:

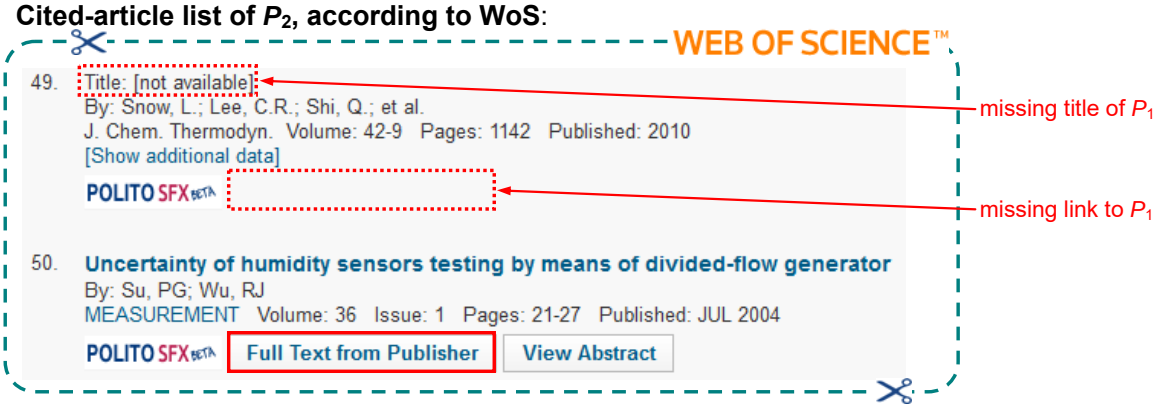


Fig. 4. First example of type-A error due to the missing title of the (cited) paper (P_1), reported in the list of a citing paper (P_2). This error is classified in the sub-category A.1-*Missing/wrong article title*. The WoS database was queried in January 2016.

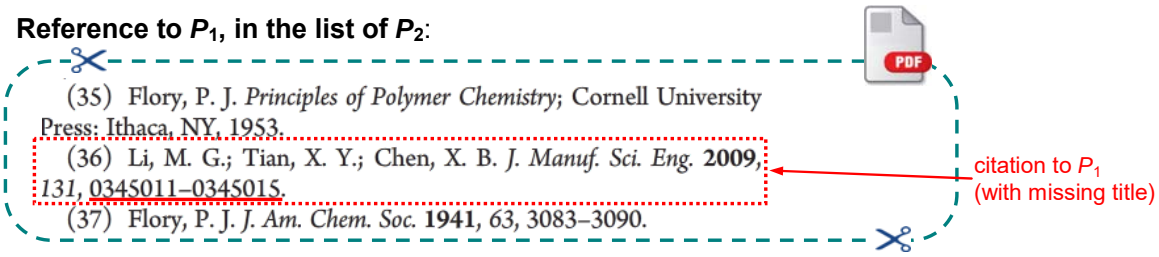
Cited paper (P_1):

Authors: M.G. Li, X.Y. Tian, X.B. Chen
Title: Modeling of Flow Rate, Pore Size, and Porosity for the Dispensing-Based Tissue Scaffolds Fabrication
Source: Journal of Manufacturing Science and Engineering, 131(3): 34501-34505
DOI: 10.1115/1.3123331

Citing paper (P_2):

Authors: L. Pescosolido, W. Schuurman, J. Malda et al.
Title: Hyaluronic Acid and Dextran-Based Semi-IPN Hydrogels as Biomaterials for Bioprinting
Source: Biomacromolecules, 12(5): 1831-1838
DOI: 10.1021/bm200178w

Reference to P_1, in the list of P_2:



List of P_2, according to WoS:

Fig. 5. Second example of type-A error due to the missing title of the (cited) paper (P_1), reported in the list of a citing paper (P_2). This error is classified in the sub-category A.1-*Missing/wrong article title*. We can also notice a pagination error in the original citation by P_2 and in the relevant database transcription. The WoS database was queried in January 2016.

(A.2) *Errors in other fields*

Other type-A errors concern inaccuracies in other fields, such as author name(s), source title, issue year, volume number and pagination. For the purpose of example, Fig. 6 exemplifies an error concerning the author name(s). The incidence of these individual type-A errors is significantly lower than those in sub-category A.1; for this reason, we aggregated them into the same sub-category (A.2). For Scopus, errors in sub-category A.2 are even less numerous than those in sub-category A.1 ($freq^{(2)}_{A.2} \approx 0.18\%$ against $freq^{(2)}_{A.1} \approx 0.41\%$), while for WoS, they have roughly the same incidence ($freq^{(2)}_{A.2} \approx 1.01\%$ against $freq^{(2)}_{A.1} \approx 0.93\%$).

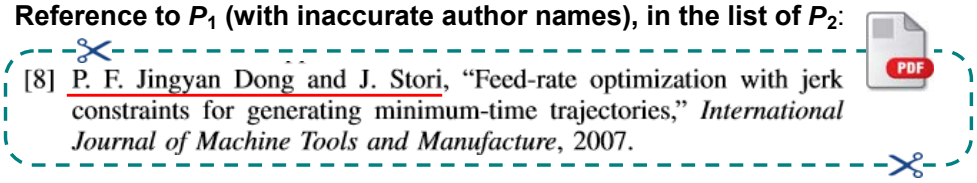
Cited paper (P_1):

Authors: J. Dong, P.M. Ferreira, J.A. Stori
Title: Feed-rate optimization with jerk constraints for generating minimum-time trajectories
Source: International Journal of Machine Tools and Manufacture, 47(12-13): 1941-1955
DOI: 10.1016/j.ijmachtools.2007.03.006

Citing paper (P_2):

Authors: X. Broquere, D. Sidobre, I. Herrera-Aguilar
Title: Soft motion trajectory planner for service manipulator robot
Source: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008. IROS 2008.
DOI: 10.1109/IROS.2008.4650608

Reference to P_1 (with inaccurate author names), in the list of P_2:



Reference to P_1 (with inaccurate author names), according to Scopus:

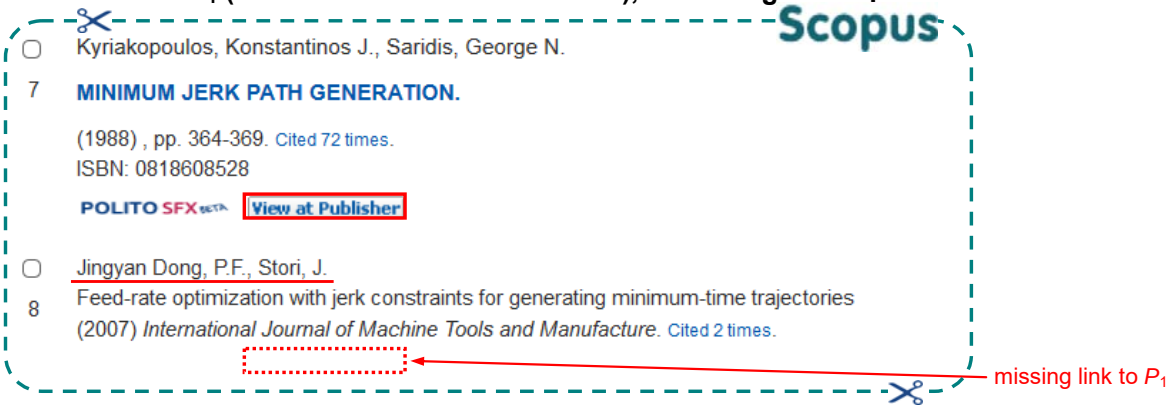


Fig. 6. Example of type-A error due to the inaccurate author names of a (cited) paper (P_1), reported in the list of a citing paper (P_2). This errors is classified in the sub-category A.2-*Errors in other fields*. The Scopus database was queried in January 2016.

4.2 Type-B errors

Type-B errors the database transcription of the (correct) references reported in the list of a (citing) paper. Tab. 5 shows that WoS is slightly weaker than Scopus (3.53% of the TO citations are omitted because of type-B errors for Scopus, against 4.51% for WoS).

(B.1) *Errors in the transcription of author name(s) and/or article title*

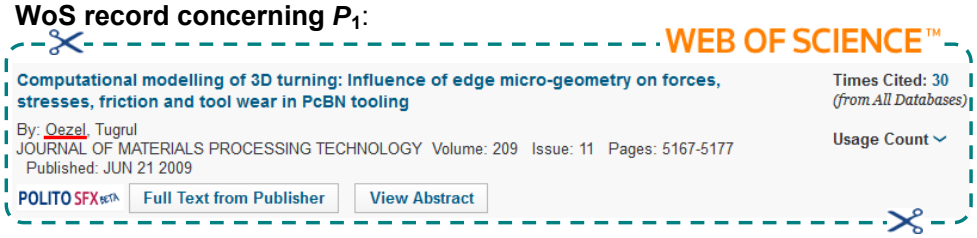
This is the predominant sub-category of type-B errors. See the example in Fig. 7, in which WoS transcribes the author's surname "Özel", related to a (cited) paper (P_1), as "Oezel". Even if this transcription seems legitimate, WoS probably encountered problems in handling the special character "Ö" (German umlaut), failing to establish the citation link with a citing paper (P_2). This type of error is much less frequent in Scopus than in WoS (i.e., $freq_{B.1}^{(2)}$ of 0.13% for Scopus against

1.65% for WoS).

Cited paper (P_1):
Author: T. Özel
Title: Computational modelling of 3D turning: Influence of edge micro-geometry on forces, stresses, friction and tool wear in PcBN tooling
Source: Journal of Materials Processing Technology, 209(11): 5167-5177
DOI: 10.1016/j.jmatprotec.2009.03.002

Citing paper (P_2):
Authors: C. Maranhão, J. Paulo Davim
Title: Finite element modelling of machining of AISI 316 steel: Numerical simulation and experimental validation
Source: Simulation Modelling Practice and Theory, 18(2): 139–156
DOI: 10.1016/j.simpat.2009.10.001

WoS record concerning P_1:



Cited-article list of P_2, according to WoS:

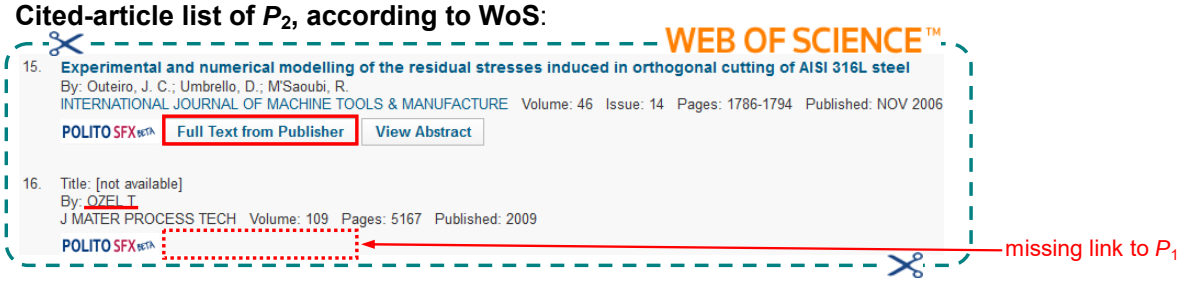


Fig. 7. Example of type-B error of WoS, concerning the name of the author of a paper (P_1), which is reported in the list of a (citing) paper (P_2). This error is classified in sub-category B.1-*Errors in the transcription of author name(s) and/or article title*. The WoS database was queried in January 2016.

(B.2) *Incomplete cited-article list and (B.3) Omitted cited-article list*

Let us now consider two typologies of type-B errors, which are more serious than the previous one, as they involve the incorrect indexing of multiple (cited) articles, causing the omission of many citations. The example in Fig. 8 shows the truncation of part of the list of a (citing) paper in WoS (sub-category B.2), while that in Fig. 9 shows the omission of the entire list of a (citing) paper in Scopus (sub-category B.3).

Paper of interest (P_1):

Authors: L. Sahebdel, S.M. Abbasi, A. Momeni
Title: Microstructural evolution through hot working of the single-phase and two-phase Ti-6Al-4V alloy
Source: International Journal of Materials Research, 102(1): 41-47
DOI: 10.3139/146.110455

Original list of P_1:

(Truncated) list of P_1, according to Scopus:

Fig. 8. Example of cited-article list truncated by the Scopus database: only the first 7 cited articles are properly transcribed, while the remaining (15) are truncated. This error is classified in the sub-category B.2-*Incomplete cited-article list*. The Scopus database was queried in January 2016.

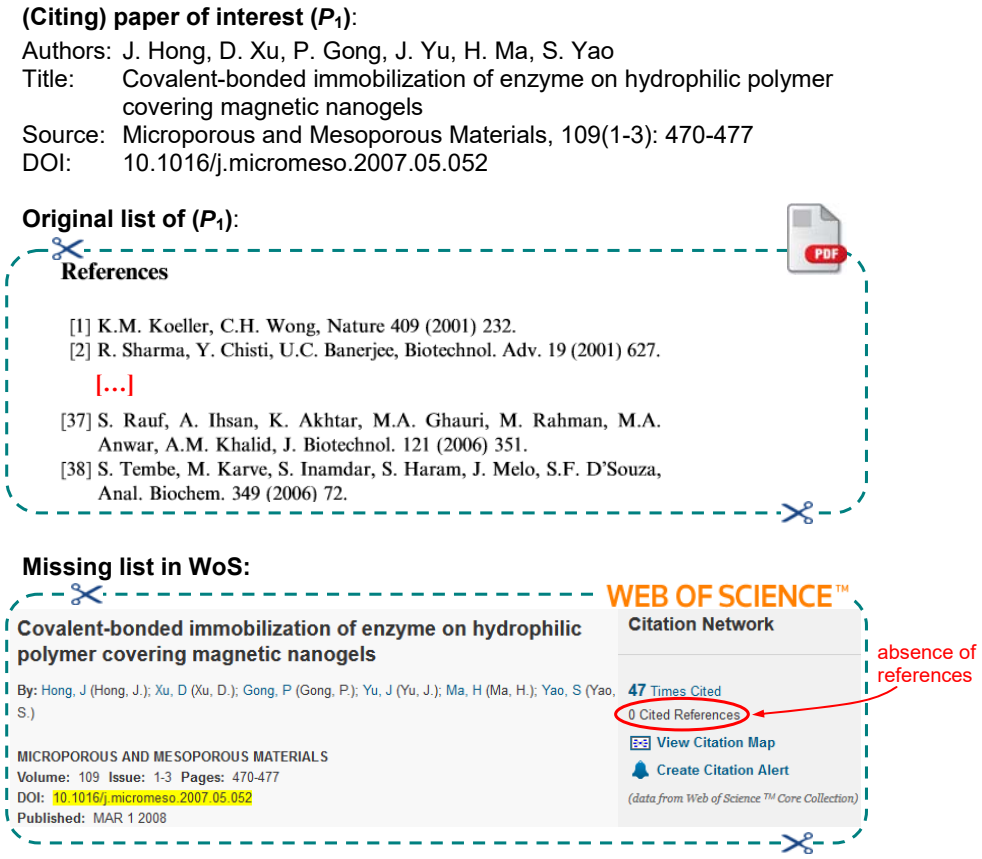


Fig. 9. Example of list omitted by the WoS database. This error is classified in the subcategory B.3-*Omitted cited-article list*. The WoS database was queried in January 2016.

The incidence of errors in sub-categories B.2 and B.3 is not so high for both Scopus and WoS. We also came across some weird variants of these errors, such as authentic cited-article lists replaced with other ones (absolutely irrelevant), anomalous increase in the number of references, etc. – for details, see (Franceschini et al., 2016b).

(B.4) *Wrong or missing DOI*

Other type-B errors concern the missing or incorrect association of an article with the relevant DOI code. We remind that DOI (i.e., Digital Object Identifier) is a character string used to univocally identify entities that are object of intellectual property (Paskin and ID Foundation 2002). Since several years, DOIs are used in bibliometrics for identifying and disambiguating scientific papers, like the "ID card" to a person; therefore, it seems reasonable to expect great attention from bibliometric databases in DOI indexing. Nevertheless, databases sometimes make mistakes.

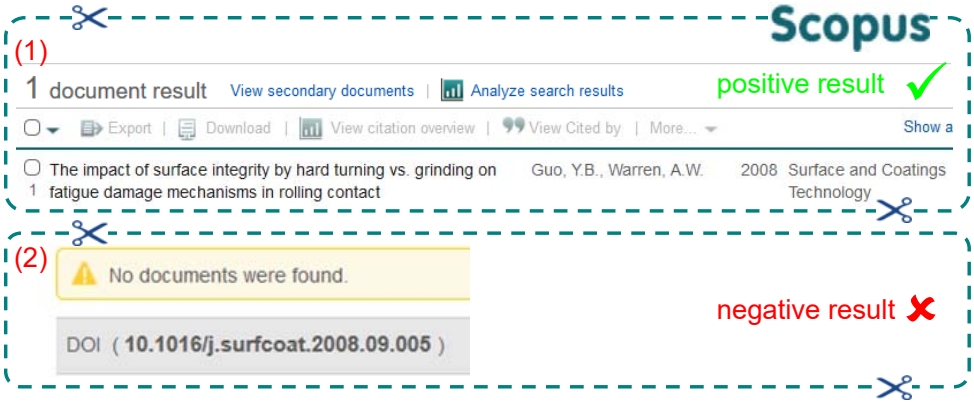
Fig. 10 exemplifies a Scopus error in determining the link between a cited paper (P_1) and a citing one (P_2), probably because of the missing DOI indexing of P_2. To be precise, we cannot be completely sure that the non-match is solely caused by the missing DOI, because of another inaccuracy related to the jumbled author names of paper P_1, in the reference list of P_2 (i.e., "Hashimoto, Warren, Guo" instead of "Hashimoto, Guo, Warren"). The same combination between

missing DOI and other inaccuracies was observed for other database errors. However, we decided to classify these errors in the B.4 sub-category, due to the importance of the DOI code.

Cited paper (P_1):
Authors: F. Hashimoto, Y.B. Guo, A.W. Warren
Title: Surface integrity difference between hard turned and ground surfaces and its impact on fatigue life
Source: CIRP Annals - Manufacturing Technology, 55(1): 81-84
DOI: 10.1016/S0007-8506(07)60371-0

Citing paper (P_2):
Authors: Y.B. Guo, A.W. Warren
Title: The impact of surface integrity by hard turning vs. grinding on fatigue damage mechanisms in rolling contact
Source: Surface and Coatings Technology, 203(3-4): 291-299
DOI: 10.1016/j.surfcoat.2008.09.005

Results of the query of P_2 (1) by title and (2) by DOI, in Scopus:



Reference to P_1 in the list of P_2, according to Scopus:

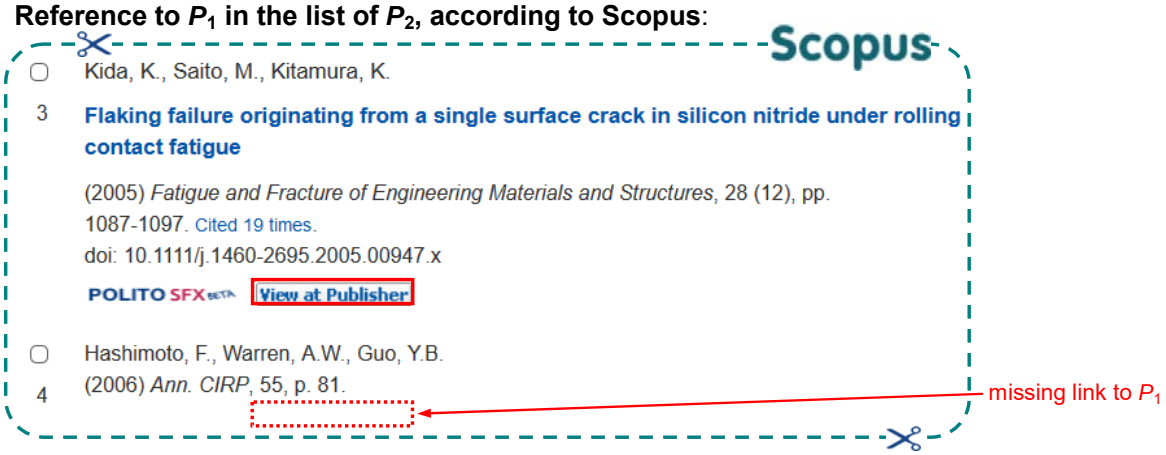


Fig. 10. Missing DOI indexing of a citing article (P_2), which prevents the determination of the citation link with a (cited) paper (P_1). This error is classified in the sub-category B.4-*Wrong or missing DOI*. The Scopus database was queried in January 2016.

In other cases, we observed errors in the DOI transcription or even multiple assignments of the same DOI to several papers – for details, see (Franceschini et al., 2015b).

(B.5) *Errors concerning Online-First papers*

A relatively frequent type-B error concerns the so-called *Online-First* papers, i.e., papers not yet in

the official version, but already available to the scientific community (Haustein et al., 2015).

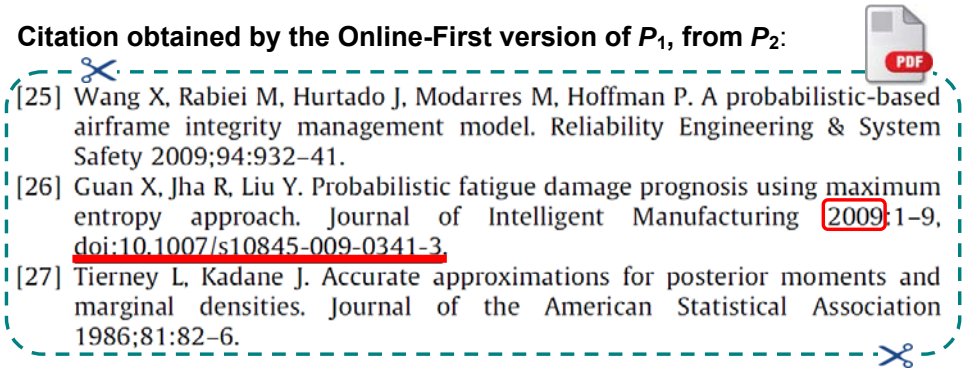
Before getting into the issue, we recall that, for several recent years now, scientific journals have been struggling to include the new-entry papers in their websites as soon as possible, in the form of Online-First papers. Apart from encouraging the spread of new knowledge, this mechanism allows journals to artificially extend the time-window for citation accumulation, resulting in a probable increase of the journal IF and other bibliometric indicators (Falagas and Alexiou, 2008). Bibliometric databases are also struggling to index Online-First papers as soon as possible. Unfortunately, the "double stage" of these papers can favour the generation of database errors; the most common is that of losing the citations obtained by the Online-First version of the paper of interest (P_1), after the publication of the relevant official version (see the example in Fig. 11).

Other authors documented the relatively high incidence of this type of error, both in Scopus and WoS (Haustein et al., 2015; Valderrama-Zurián et al., 2015).

Cited paper (P_1):
Authors: X. Guan, R. Jha, Y. Liu
Title: Probabilistic fatigue damage prognosis using maximum entropy approach
Source: Journal of Intelligent Manufacturing, 23(2): 163-171
DOI: 10.1007/s10845-009-0341-3
Online-First availability date: 28 October 2009
Official Publication date: 2012

Citing paper (P_2):
Authors: X. Guan, J. He, R. Jha, Y. Liu
Title: An efficient analytical Bayesian method for reliability and system response updating based on Laplace and inverse first-order reliability computations
Source: Reliability Engineering & System Safety, 97(1): 1-13
DOI: 10.1016/j.ress.2011.09.008

Citation obtained by the Online-First version of P_1, from P_2:


[25] Wang X, Rabiei M, Hurtado J, Modarres M, Hoffman P. A probabilistic-based airframe integrity management model. Reliability Engineering & System Safety 2009;94:932–41.
[26] Guan X, Jha R, Liu Y. Probabilistic fatigue damage prognosis using maximum entropy approach. Journal of Intelligent Manufacturing 2009;1–9, doi:10.1007/s10845-009-0341-3.
[27] Tierney L, Kadane J. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association 1986;81:82–6.

Missing link in Scopus, in the list of P_2:

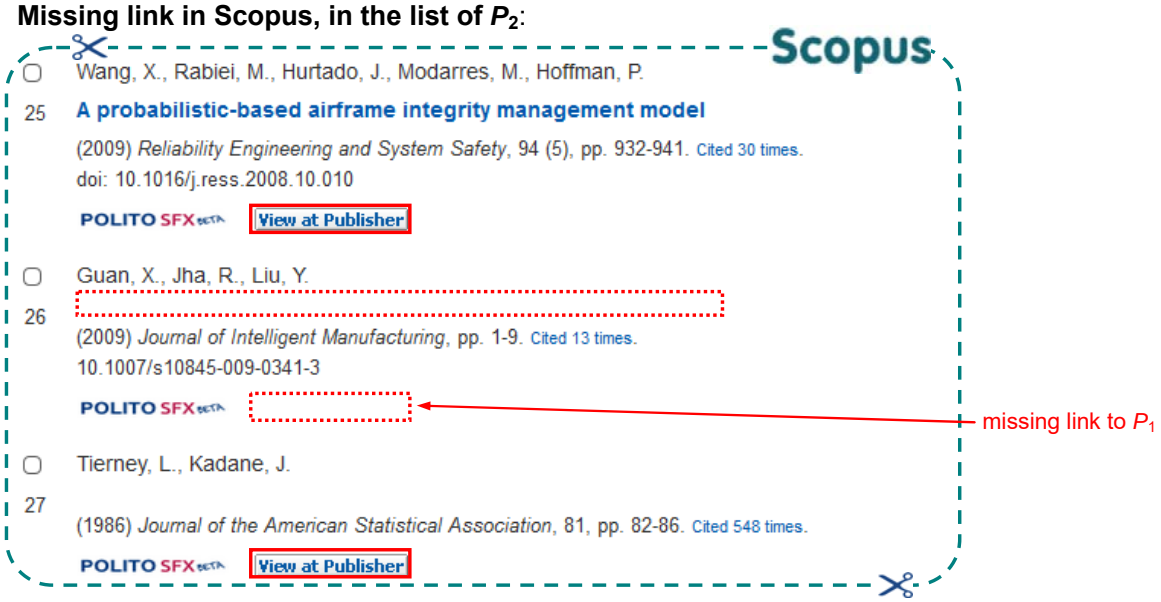


Fig. 11. Example of error classified in the sub-category B.5-*Errors concerning Online-First articles*. A citation obtained by the Online-First version of P_1 (issued in October 2009) is lost after the publication of the relevant official version (in 2012). The Scopus database was queried in January 2016.

(B.6) *Unindexed (citing) articles*

Let us now consider a rather serious type-B error, in which the missing indexing of some (citing) articles caused the omission of their citations. Databases may sometimes forget to index some

(unfortunate) articles, even though they are able to index other articles in the same journal issue (see the example in Fig. 12). This is an extreme form of a database mapping error, in which the citation match fails as some (citing) papers are not even indexed by the database. This error is particularly serious since it causes the omission of multiple citations (i.e., those given by the unindexed citing papers).

Paper of interest (P_1):
Authors: Q. Liang, D. Zhang, Y. Ge, Q. Song
Title: A Novel Miniature Four-Dimensional Force/Torque Sensor With Overload Protection Mechanism
Source: IEEE Sensor Journal, 9(12): 1741-1747
DOI: 10.1109/JSEN.2009.2030975

Results of the query of P_1 (1) by title and (2) by DOI, in Scopus:

Other articles published in the same journal issue of P_1:

Fig. 12. Example of paper mistakenly not indexed by Scopus. This error is classified in the sub-category B.6-*Unindexed (citing) articles*. The Scopus database was queried in January 2016.

This type of error is significantly more frequent in Scopus, than WoS (i.e., $freq_{B.6}^{(2)}$ of 1.30% for Scopus, against 0.16% for WoS).

(B.7) *Reasons unknown*

In this case, a cited article and a relevant citing article are both properly indexed by the database (i.e., without any type-A error); nevertheless, the citation link is not established by the database and

the citation is lost (see the example in Fig. 13). This error sub-category has been denominated as "reasons unknown", since we were unable to identify their possible causes.

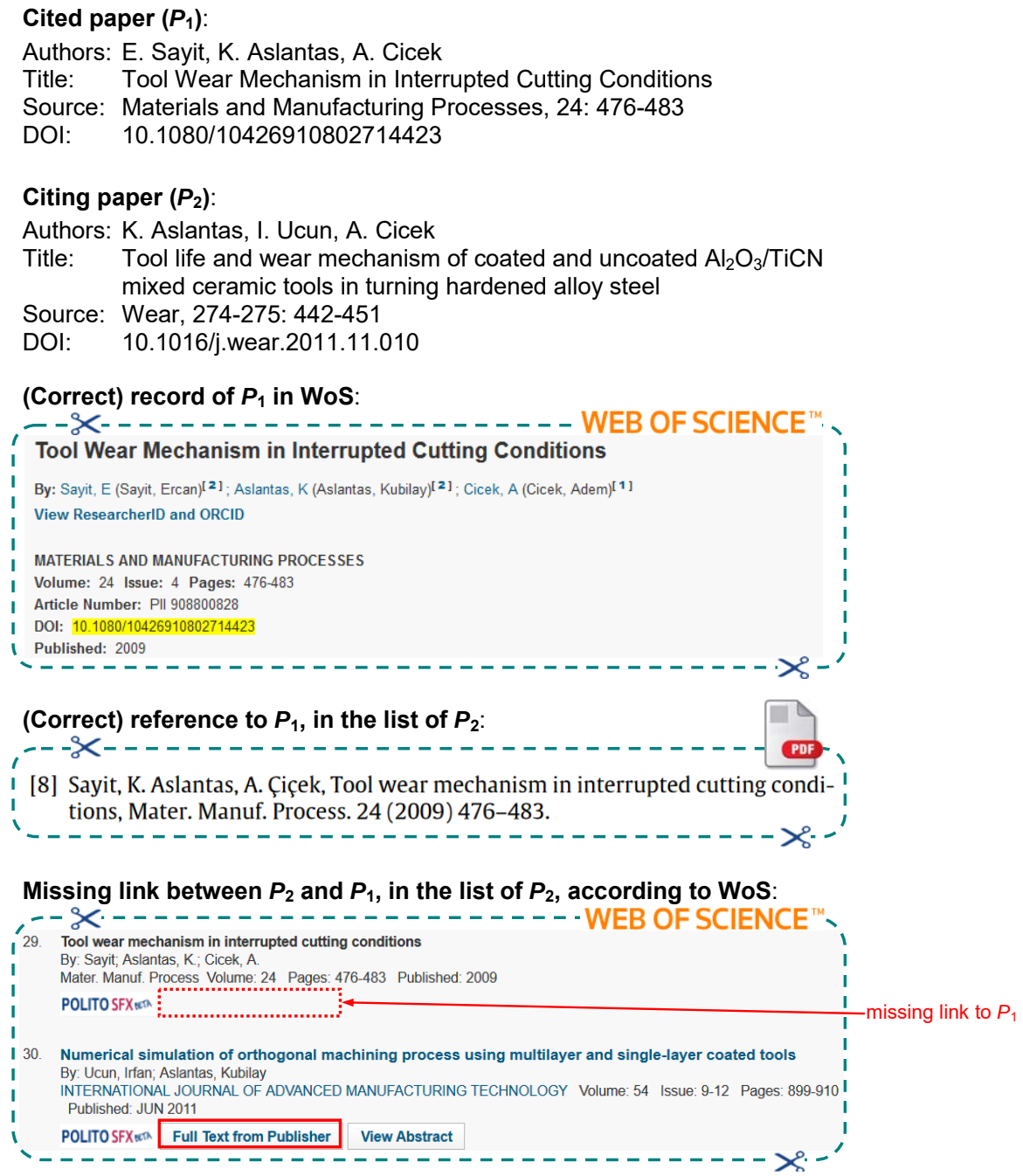


Fig. 13. Example of type-B error, classified in the sub-category B.6-*Reasons unknown*.

4.3 Further remarks on the classification results

Fig. 14 summarizes the classification results, representing the repartition of the errors in the various sub-categories, for both the databases.

At a glance, the predominant error (sub-)categories of the two databases look generally different. This impression is confirmed by a scatter plot in Fig. 15, which denotes the absence of correlation

between the two databases ($R^2 \approx 0.018$). This result is probably due to the use of different citation matching algorithms or metadata, in the indexing process of Scopus and WoS (Olensky et al., 2016).

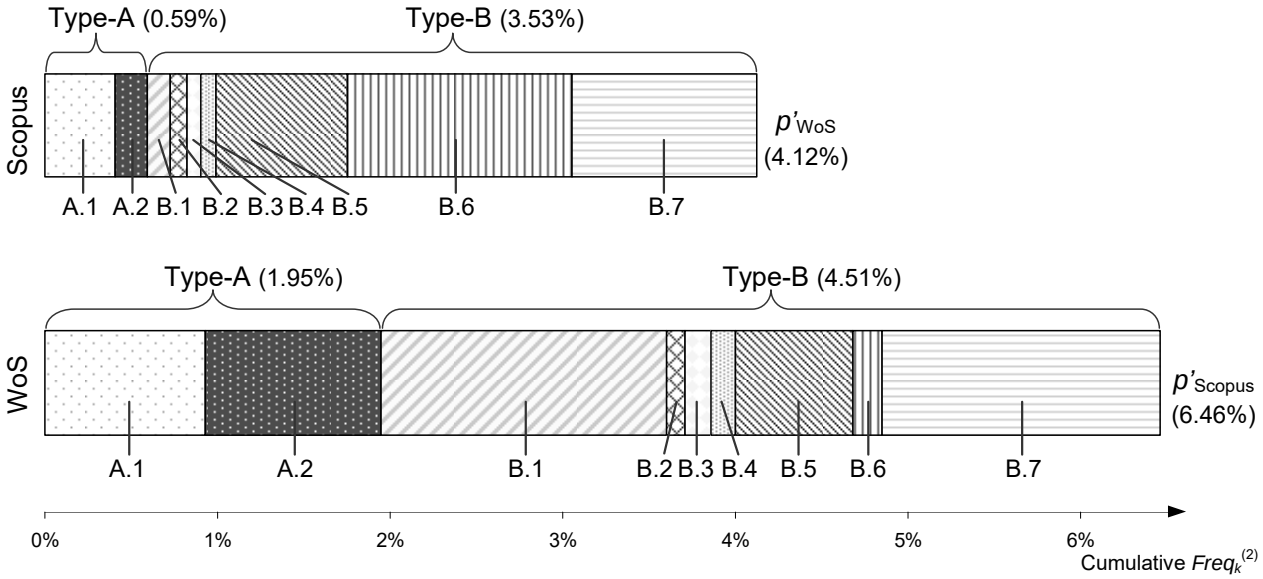


Fig. 14. Graphical representation of the repartition of the errors in the individual sub-categories, for the two databases. Numerical values are reported in Tab. 5.

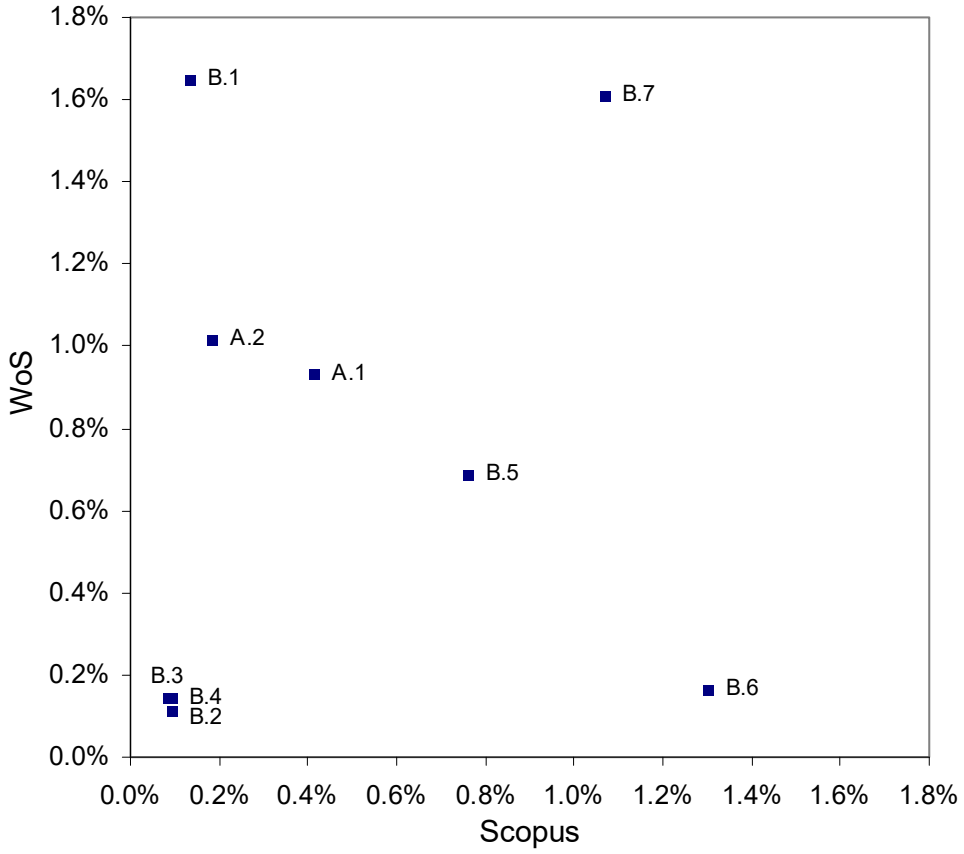


Fig. 15. Scatter plot representing the absolute frequency ($freq_k^{(2)}$) of the error sub-categories for Scopus and WoS.

5. Conclusions

This section sums up and discusses the results of this research from the perspective of the previously formulated research questions.

- *What are the more frequent errors of Scopus and WoS and the similarities and differences between the two databases?*

  Through the manual analysis of a relatively large amount of database errors, we identified several error typologies (some of which are new to the state of the art, e.g., those in the sub-categories B.2 and B.5) and several weaknesses of the Scopus and WoS databases, such as:

  - Regarding type-A errors, WoS seems significantly weaker than Scopus (1.95% of the TO citations are omitted because of type-A errors in WoS, against 0.59% in Scopus). A possible interpretation of this result is that the Scopus citation matching algorithm seems more robust than the WoS one, in the presence of dirty data.
  - Another weakness of WoS with respect to Scopus is represented by the type-B errors concerning the incorrect transcription of the author name(s) and/or title ($freq_{B.1}^{(2)}$ of 1.65% for WoS against 0.13% for Scopus).
  - Although Scopus seems more accurate than WoS, it has a higher propensity to forget to index some papers (error sub-category B.6), losing the citations that they gave/obtained (i.e., $freq_{B.6}^{(2)}$ of 1.30% for Scopus against 0.16% for WoS).
  - Managing the Online-First articles (error sub-category B.5) seems rather problematic for both databases ($freq_{B.5}^{(2)}$ of 0.76% for Scopus against 0.69% for WoS). The typical consequence of these errors is to lose the citations obtained by the Online-First version of a paper of interest, after the publication of the relevant official version.

  The analysis showed the lack of correlation between Scopus and WoS, regarding the distribution of the errors in the different (sub-)categories. This is probably due to the fact that the two databases use different citation matching algorithms and/or metadata, in the indexing process.

- *Are the results of this research in line with those of other researches in the field of bibliometric-database errors?*

  We remark that the relatively large sample of (presumed) database errors is a distinctive element of this research. Having said that, some of the findings presented are in line with those of other studies, e.g., the identification of the more frequent error (sub-)categories, the estimate of the phantom-citation rate of WoS (Garcia-Perez, 2010; Olensky, 2015), the fact that both Scopus and WoS seem to have relatively serious problems in managing the citations obtained/given by the Online-First articles (Haustein et al., 2015; Valderrama- Zurián et al., 2015; Franceschini et al., 2016b), etc..

On the other hand, some inconsistencies emerged; for example, it was shown that type-B errors tend to predominate over type-A ones or that pre-existing inaccuracies concerning the title of the cited articles probably complicate the citation match, contradicting the findings by Olensky (2015). These inconsistencies could be due to several reasons:

- The relatively small sample of papers used in the previous database-error classifications; e.g. the research by Olensky (2015) is based on the manual analysis of 300 cited papers and the relevant citing ones.
- The fact that, among the more than 10,000 database errors available, we manually analyzed just a fraction (i.e., 10%) of them, generally concerning citations in the Engineering-Manufacturing field.
- The relatively strong simplification of associating one-and-only-one error cause (and therefore one-and-only-one error sub-category) with each omitted citation. We are aware that omitted citations are not rarely caused by a combination of more than one typology of inaccuracy. The identification of the error cause that seems more decisive, among the possible ones, is indeed subjective.

- *Does this research provide a representative picture of the Scopus and WoS errors?*
  We would be tempted to answer saying "yes, it does". The reason is that – despite our focus was mainly on publications in the Engineering-Manufacturing field – the error mechanisms identified appear to be independent from this particular scientific field. As a proof, the results obtained are often in line with those of other studies based on publications from other scientific fields.
- *What are the phantom-citation rates of Scopus and WoS and how can they be used to correct the p values estimated through the automated algorithm?*
  The analysis of the presumed omitted citations allowed to identify a certain amount of phantom citations and to estimate the phantom-citation rate of the two databases: $\alpha_{Scopus} \approx 0.10\%$ and $\alpha_{WoS} \approx 0.46$. Using these data, the omitted-citation rates estimated in our previous studies have been slightly adjusted (i.e., $p'_{Scopus} \approx 4.12\%$ against $p_{Scopus} \approx 4.58\%$ and $p'_{WoS} \approx 6.46\%$ against $p_{WoS} \approx 6.55\%$).
- *In the light of the results obtained, what are the practical implications to users and administrators of the Scopus and WoS databases?*
  Although the influence of omitted citations is not very high for both databases – it could lead to significant distortions when considering relatively small sets of cited/citing papers, e.g., those representing the production output of individual scientists. From a practical viewpoint, individual users cannot do much, given the difficulty to identify the possible omitted citations manually. Despite this, our advice is to compare data from different databases as much as possible. In this sense, this research contributed to identify the main weaknesses of Scopus and WoS. Also, the

use of GS may help to identify omitted citations, due to the great coverage.

Once possible database errors are identified, they can be notified to the database staff through dedicated support/feedback mechanisms. We have noticed that Scopus and WoS are both very responsive to these feedbacks (Meester et al., 2016).

As regards database administrators we renew our exhortation to improve in terms of data cleaning. We remark that *all* the database errors analyzed and classified in this research were preventable: in fact, all the citations omitted by one database are, by definition, correctly indexed by the other one.

We are aware that the citation-matching algorithms used by databases will never be infallible, as they struggle to find the optimal balance between (i) the risk of failing to identify authentic citations (*false negatives*) and (ii) that of assigning phantom citations (*false positives*). Nevertheless, we believe that databases could introduce additional (automated) controls on the results of the citation mapping process. This would be much more effective than waiting for the feedbacks from users, with important benefit in terms of database usability and accuracy.

Acknowledgements

Authors gratefully acknowledge the contribution of Silvia Milan B.Sc., in analysing and classifying database errors with great accuracy and patience.

References

Archambault, É., Campbell, D., Gingras, Y., Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. Journal of the American Society for Information Science and Technology, 60(7), 1320-1326.

Buchanan, R.A. (2006). Accuracy of Cited References: The Role of Citation Databases. College & Research Libraries, 67(4), 292-303.

Falagas, M.E., Alexiou, V.G. (2008). The top-ten in journal impact factor manipulation. Archivum immunologiae et therapiae experimentalis, 56(4): 223-226.

Franceschini, F., D. Maisano and L. Mastrogiacomo (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). Errors in DOI indexing by bibliometric databases. Scientometrics, 102(3), 2181-2186.

Franceschini, F., Maisano, D., Mastrogiacomo, L. (2016a). Do Scopus and WoS correct "old" omitted citations? Scientometrics, DOI: 10.1007/s11192-016-1867-8.

Franceschini, F., Maisano, D., Mastrogiacomo, L. (2016b). The museum of errors/horrors in Scopus. Journal of Informetrics, 10(1), 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.

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.

Haustein, S., Bowman, T. D., Costas, R. (2015). When is an article actually published? An analysis of online availability, publication, and indexation dates. Proceedings of the 15th International Society of Scientometrics and Informetrics (ISSI) Conference, 1170–1179, 29 June - 3 September 2015, Istanbul,

Turkey, ISBN: 978-975-518-381-7.
Hildebrandt, A.L., Larsen, B. (2008). Reference and citation errors: A study of three law journals. Presented at the 13th Nordic Workshop on Bibliometrics and Research Policy. 11–12 September 2008, Tampere, Finland.
Labbé, C. (2010). Ike Antkare, one of the great stars in the scientific firmament. ISSI Newsletter, 6(2): 48-52.
Larsen, B., Hytteballe Ibanez, K., Bolling, P. (2007). Error rates and error types for the Web of Science algorithm for automatic identification of citations. Presented at the 12th Nordic Workshop on Bibliometrics and Research Policy. 13–14 September 2007, Copenhagen, Denmark.
Meester, W.J., Colledge, L., Dyas, E.E. (2016). A response to "The museum of errors/horrors in Scopus" by Franceschini et al. Journal of Informetrics, 10(2): 569-570.
Meho, L.I., Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. Journal of the American Society for Information Science and Technology 58(13), 2105-2125.
Mikki, S. (2010). Comparing Google Scholar and ISI Web of Science for earth sciences. Scientometrics, 82(2), 321-331.
Moed, H.F., Vriens, M. (1989). Possible inaccuracies occurring in citation analysis. Journal of Information Science, 15(2), 95–107.
Moed, H. (2002). The impact-factors debate: The ISI's uses and limits. Nature, 415(6873): 731-732.
Moed, H.F. (2005). Citation Analysis in Research Evaluation. Information Science and Knowledge Management (Vol. 9). Dordrecht: Springer.
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.
Olensky, M. (2015). Data accuracy in bibliometric data sources and its impact on citation matching. Doctoral dissertation. Humboldt-Universität zu Berlin (Germany). Retrieved from edoc.huberlin.de/dissertationen/olensky-marlies-2014-12-17/PDF/olensky.pdf
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.
Orduna-Malea, E., Ayllón, J. M., Martín-Martín, A., López-Cózar, E. D. (2015). Methods for estimating the size of Google Scholar. Scientometrics, 104(3): 931-949.
Paskin, N. and I. D. Foundation (2002). The DOI® handbook, IDF-Intern. DOI Foundation.
Prins, A.A., Costas, R., van Leeuwen, T.N., Wouters, P.F. (2016). Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data. To appear in Research Evaluation, DOI: 10.1093/reseval/rvv049.
Scopus Elsevier (2016). *Scopus Content Coverage*. Available at http://www.scopus.com [retrieved on October 2014].
Thomson Reuters (2016). *Master Journal List*, http://ip-science.thomsonreuters.com/mjl/ [retrieved on October 2014].
Tunger, D., Haustein, S., Ruppert, L., Luca, G., Unterhalt, S. (2010). The Delphic Oracle: An analysis of potential error sources in bibliographic databases. In CWTS (Ed.), Proceedings of the 11th International Conference on Science and Technology Indicators (pp. 282–283). Leiden, Netherlands: CWTS.
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.
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.
Wildgaard, L. (2015). A comparison of 17 author-level bibliometric indicators for researchers in Astronomy, Environmental Science, Philosophy and Public Health in Web of Science and Google Scholar. Scientometrics, 104(3), 873-906.

Appendix

A1. List of the Engineering-Manufacturing journals examined

Tab. A.1. List of the Engineering-Manufacturing journals examined. For each journal, it is reported its title and ISSN code. Journals are sorted alphabetically according to their title.

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 & Control	0953-7287
Research in Engineering Design	0934-9839
Robotics and Computer-Integrated Manufacturing	0736-5845
Soldering & Surface Mount Technology	0954-0911

A2. Model for correcting the omitted-citation rate, considering the effect of phantom citations

This section provides a mathematical explanation of the formulae reported in Tab. 3. Considering the representation in Fig. 1, it is easy to deduce that the omitted-citation rates of Scopus and WoS are:

$$\begin{aligned} p_{Scopus} &= \omega_{Scopus} / \gamma \\ p_{WoS} &= \omega_{WoS} / \gamma \end{aligned}, \qquad \text{(A1)}$$

where

ω_{Scopus} and ω_{WoS} are respectively the total number of (presumed) omitted TO citations related to the Scopus and WoS database;

γ is the total number of (presumed) TO citations available.

Eq. A1 provides an estimate of one database's omitted citation rate, which can be distorted by the presence of phantom citations by the other 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 TO citations related to the other database (δ) – and the number of (presumed) TO citations available (γ):

$$\begin{aligned} \alpha_{Scopus} &= \delta_{WoS} / \gamma \\ \alpha_{WoS} &= \delta_{Scopus} / \gamma \end{aligned} . \tag{A2}$$

The apparent reversal of the "Scopus" and "WoS" subscript in Eq. A2 depends on the fact that the false omitted TO citations relating to one database are due to phantom citations by the other database.

From Eq. A2, we obtain:

$$\begin{aligned} \delta_{WoS} &= \gamma \cdot \alpha_{Scopus} \\ \delta_{Scopus} &= \gamma \cdot \alpha_{WoS} \end{aligned} . \tag{A3}$$

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:

$$\gamma' = \gamma - (\delta_{WoS} + \delta_{Scopus}) = \gamma \cdot [1 - (\alpha_{Scopus} + \alpha_{WoS})] . \tag{A4}$$

The *corrected* number of omitted citations – i.e., excluding the *false* ones – of each database will be:

$$\begin{aligned} \omega'_{Scopus} &= \omega_{Scopus} - \delta_{Scopus} = \omega_{Scopus} - \alpha_{Scopus} \cdot \gamma \\ \omega'_{WoS} &= \omega_{WoS} - \delta_{WoS} = \omega_{WoS} - \alpha_{WoS} \cdot \gamma \end{aligned} . \tag{A5}$$

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

$$\begin{aligned} p'_{Scopus} &= \frac{\omega'_{Scopus}}{\gamma'} = \frac{\omega_{Scopus} - \alpha_{Scopus} \cdot \gamma}{\gamma \cdot \left[1 - \left(\alpha_{Scopus} + \alpha_{WoS}\right)\right]} = \frac{p_{Scopus} - \alpha_{Scopus}}{1 - \left(\alpha_{Scopus} + \alpha_{WoS}\right)} \\ p'_{WoS} &= \frac{\omega'_{WoS}}{\gamma'} = \frac{p_{WoS} - \alpha_{WoS}}{1 - \left(\alpha_{Scopus} + \alpha_{WoS}\right)} \end{aligned} . \tag{A6}$$

We remark that, having estimated the phantom-citation rate (α) of the databases in use, the formulae in Eq. A6 can be used to correct the *p* values resulting from the application of the automated algorithm, taking account of the distortions produced by phantom citations.

Since we manually analyzed only a portion of the (presumed) omitted TO citations available (precisely $o_{Scopus} = 447$ for Scopus and $o_{WoS} = 640$ for WoS), α_{Scopus} and α_{WoS} can be estimated as:

$$\alpha_{Scopus} = \frac{\delta_{Scopus}}{\gamma} \approx \frac{\frac{d_{WoS}}{o_{WoS}} \cdot \omega_{WoS}}{\gamma} = \frac{d_{WoS}}{o_{WoS}} \cdot \frac{\omega_{WoS}}{\gamma} = \frac{d_{WoS}}{o_{WoS}} \cdot p_{WoS}, \qquad \text{(A7)}$$

$$\alpha_{Scopus} \approx \frac{d_{Scopus}}{o_{Scopus}} \cdot p_{Scopus}$$

being:

o_{Scopus} and o_{WoS} the number of (presumed) omitted TO citations by Scopus and WoS, which were analyzed manually;

d_{Scopus} and d_{WoS} the number of phantom citations, among the o_{Scopus} and o_{WoS} omitted TO citations, which were analyzed manually.

The α_{Scopus} and α_{WoS} estimated in Eq. A7 are based on the reasonable assumption that false citations are randomly distributed among the total (presumed) omitted TO citations. According to this assumption, the ratio between the d and o values related to a certain database can be considered equal to the ratio between δ and ω (see the representation scheme in Fig. 16). In formal terms:

$$\frac{d_{Scopus}}{o_{Scopus}} \approx \frac{\delta_{Scopus}}{\omega_{Scopus}}, \qquad \text{(A8)}$$

$$\frac{d_{WoS}}{o_{WoS}} \approx \frac{\delta_{WoS}}{\omega_{WoS}}$$

from which we derive the terms δ_{Scopus} and δ_{WoS} (already replaced in Eq. A7):

$$\delta_{Scopus} \approx \frac{d_{Scopus}}{o_{Scopus}} \cdot \omega_{Scopus} \qquad \text{(A9)}$$

$$\delta_{WoS} \approx \frac{d_{WoS}}{o_{WoS}} \cdot \omega_{WoS}$$

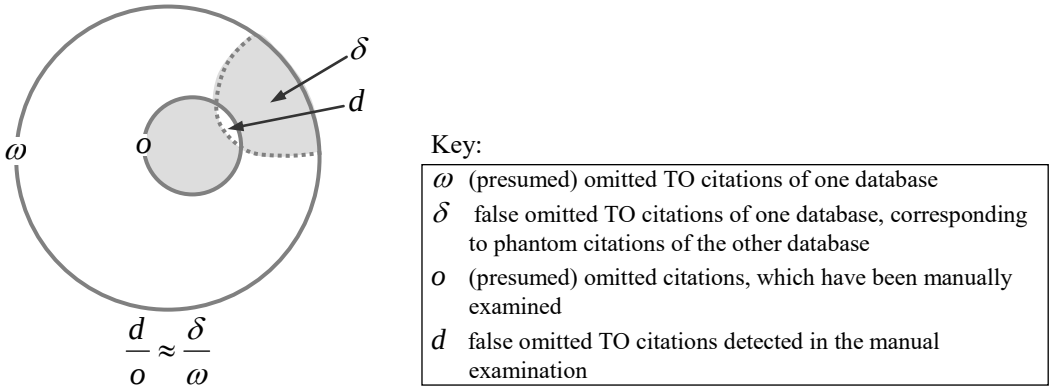


Fig. 16. Schematic representation of the subset of (*o*) presumed omitted TO citations by a database of interest, which were examined manually. These citations were selected from a sample of (*ω*) presumed omitted TO citations, which includes (*δ*) *false* TO citations, due to phantom citations of the other database.

Let us now return to the definition of α_{Scopus} and α_{WoS}. These phantom-citation rates (in Eq. A2) are defined as the ratio between the phantom citations and the (γ) total (presumed) TO citations

available. We remark that the (presumed) TO citations are influenced by the phantom citations produced by both the databases in use (i.e., the one of interest and the other one). For this reason, we cannot say that the phantom-citation rate of one database is completely independent from the behaviour of the other database.

From the perspective of one-and-only-one database, i.e., ignoring the other one and the corresponding phantom citations, the phantom-citation rate can be redefined as the ratio between the phantom citations of this database and the citations that are or should be indexed by the database itself; in formal terms:

$$\begin{aligned} \alpha'_{Scopus} &= \frac{\delta_{WoS}}{\gamma - \delta_{Scopus}} = \frac{\delta_{WoS}}{\gamma \cdot (1 - \alpha_{WoS})} = \frac{\alpha_{Scopus}}{1 - \alpha_{WoS}} \\ \alpha'_{WoS} &\approx \frac{\delta_{Scopus}}{\gamma - \delta_{WoS}} = \frac{\alpha_{WoS}}{1 - \alpha_{Scopus}} \end{aligned} . \quad \text{(A10)}$$

Terms $\gamma - \delta_{Scopus}$ and $\gamma - \delta_{WoS}$ (in the denominator of the previous formulae) represent the TO citations "purified" from the phantom citations produced by the other database. Even though the estimates obtained using α'_{Scopus} and α'_{WoS} are perhaps more rigorous than those obtained using α_{Scopus} and α_{WoS}, their difference is actually negligible, due to the fact that α_{Scopus} and α_{WoS} are much smaller than 1. As a confirmation of this, Tab. 6 reports the numerical values of the parameters discussed in this section, resulting from the analysis. For simplicity, in the rest of the document we just refer to the initial definition of the phantom-citation rate (i.e., α_{Scopus} and α_{WoS}), not the "more rigorous" one (i.e., α'_{Scopus} and α'_{WoS}).

Tab. 6. Synthetic indicators related to the phantom citations examined.

Parameter	Scopus	WoS
γ	97,698	*idem*
ω	4,473	6,404
p	4.58%	6.55%
o	447	640
d	45	10
α	0.1024%	0.4609%
γ'	97147.6	*idem*
ω'	4022.7	6303.9
p'	4.12%	6.46%
α'	0.1029%	0.4614%