

Open and overlooked: the penalties for preprint open access papers and translated journals in citation analysis

Paul Donner*

**donner@dzhw.eu*

ORCID: 0000-0001-5737-8483

Department 2 'Research System and Science Dynamics', German Centre for Higher Education Research and Science Studies (DZHW), Germany

The citations to open access preprint versions of papers and citations to papers in journals translated to English are not regularly counted in major proprietary citation index databases or in ordinary bibliometric research assessment even though they arguably reflect a true part of a work's scientific impact. Here we explore the extent of these phenomena using Web of Science data.

1. Introduction

Many scientific publications are published as open access preprints, often a considerable time before any eventual formal publication in a scientific journal or book. These versions may receive substantial numbers of citations which would not be counted in conventional citation analysis because the default mode is that only citations to regularly published items are counted. This could potentially disadvantage authors and institutions committed to sharing their findings early, openly, and widely. Here we investigate the size of this possible distortion which prior research shows as being far from negligibly small.

Aman (2015) found that WoS-indexed references to arXiv alone, for publications from 1991-2013, amount to 900,000 in number, more than enough to motivate an updated analysis including the wider preprint literature. Many studies have tried to identify an open access citation advantage (Langham-Putrow, Bakker, & Riegelman, 2021). Insofar as these studies have not included citations to preprint versions of both open and closed access publications, their results would be questionable. Collecting and analyzing citations to preprint versions of articles also has other applications. Traag (2021) used arXiv preprint citations to published works to formally estimate the causal effect of publication of articles in specific journals on citation impact of articles, i.e. a journal halo effect.

Scopus has included preprints as listed documents on author profiles since 2021 but citations to these are not included. In 2023 Clarivate Analytics launched the Preprint Citation Index, covering documents from five platforms. This index provides citation counts tallying references from preprints to preprints and links cited references to the Web of Science Core Collection. However, citation counts to preprint and final published versions of the same item are also not unified in Web of Science. We experimentally carry out exactly this crucial additional process and analyze its consequences.

A similar situation exists for articles in translated journals. A small number of journals are published in a language other than English and are translated back-to-back. Referencing authors could cite either version but the citation counts should be unified at some point for valid citation analysis. Web of Science may index either version of some of these journals so it would appear most appropriate to also link citing references to original and translated item to the indexed items to fully reflect their citation impact. Aksenteva (2015) raised awareness of this issue, noting that the indexing policy of SCI/WoS has been inconsistent. Figure 1 shows an updated version of the example case she gave. WoS identified and counted 35 of the references to the

Russian version, which is indexed as a source document in WoS, but not the 81 references to the English translated version. Studying one prominent Russian-language journal, Aksenteva (2015) found that on average 67 % of articles' citations were not linked to the indexed items and thus not countable in ordinary citation analysis using WoS. Here this type of analysis is extended to many more translated journal titles and the results from counting references to both language versions are compared to the citation counts in WoS based presumable on one version only.

Figure 1: Screenshot of WoS cited reference search showing separate citation counts of a translated journal item (taken Jan. 2024)

<input type="checkbox"/>	Cited Author	Cited Work	Title	Year	Volume	Issue	Page	Identifier	Citing Articles ↓
<input type="checkbox"/>	Kerner, B.S.; Osipov, V.V.	Soviet Physics - Uspekhi	Self-organization in active distributed media: scenarios for the spontaneous formation and evolution of dissipative structures	1990	33	9	679-719	10.1070/ PU1990v033n09ABEH002627	81
<input type="checkbox"/>	KERNER, BS; OSIPOV, VV	USP FIZ NAUK+	SELF-ORGANIZATION IN ACTIVE DISTRIBUTED MEDIA - SCENARIOS OF SPONTANEOUS FORMATION AND EVOLUTION OF	1990	160	9	1-73	10.3367/JFNr.0160.199009a. 0001	35

These two phenomena have in common that they concern specific subsets of scientific works which exist in more than one version but citation analysis is restricted to one of the versions resulting in negative citation count biases for such works.

2. Data and methods

We carried out this study using a mid-2023 snapshot of the Web of Science Core Collection, publication years 1980 to present, which includes the Science, Social Science, and Arts and Humanities Citation Indexes and the two Proceedings Indexes, but not the regional/national CIs, Book CI, Emerging Sources CI, or the new Preprint CI.

Cited reference strings were text-indexed for easy and fast search queries. For preprint citations, we restricted the search to unmatched references (cited references not linked by WoS to any indexed source item) which had some present author name information. We stored all cited reference strings containing a pattern matching any of 11 preprint platforms or either of the strings “working paper” or “preprint” in the cited source title field. Missing publication years were in some cases automatically replaced with identified year-like pattern in the cited source title field when present. Reference strings with cited title containing “reply”, “comment”, “corrigendum”, “correction”, “vol” were removed, as these were found to lead to false matches in the following steps. The remaining candidate preprint citations were matched to WoS-indexed source items if the following conditions were true: exact match on author family name and given name (or initials), publication year of cited reference earlier or equal to indexed source item, cited reference title and indexed source item title similar with Jaro-Winkler similarity > 0.9 and Levenshtein similarity > 0.88. These values were experimentally found to result in very high precision (practically no false positives) but the false negative rate is unknown.

For translated journals, original and translated titles were obtained from a 2005 list by Bruno Voisin and internet searches. Most of the found titles are Russian in the original language and published by Pleiades Publishing and distributed by Springer Nature, but the Chinese Journal

of Analytical Chemistry was also included. Search patterns for various versions of either language journal title were created, particularly abbreviations, based on exploratory searches in the WoS online cited references search in order to find as many valid cited references as possible. Matches for these patterns which were not linked by WoS to any indexed source item were stored as candidates. These candidates were accepted as citations to indexed items if they matched in journal identity, publication year ± 1 , exact match of volume and page number, and first author name trigram similarity > 0.5 . Candidates that did not match in volume and page but had a cited reference title highly similar to an otherwise matching source item were also included. We tested out a more lenient method which did not use the strict volume and page condition but only tried to identify the most similar indexed source item using volume, first author name, and page similarity (with still exact source title match and publication year ± 1 match). This was tried because the translated and original versions sometimes use different volume numbering and pagination (cf. Fig. 1). While this method found many more correct matches, there were too many false positive matches, mainly from the same authors publishing repeatedly in the same journals. Thus, the first, more strict matching is used here, with the caveat that the reported figures here are known to be incomplete but also fairly free of any false negatives.

3. Results

The search and matching for preprint citations found c. 806,000 citations to preprints of 303,000 WoS-indexed items which were not included and counted as citations in WoS and would be overlooked in conventional citation analysis. This figure is lower than that of Aman (2015) for a much smaller target set. The difference comes about because in contrast to that study, in the present study we have additionally matched cited references to published journal items. For comparison, our preliminary identified candidate pool of possible references to preprints is 3.4 million items, of which 2.1 million are to arXiv. This means that about 23 % of preprint references could be matched to a published paper indexed in WoS.

These 303,000 items with preprint citations had accrued ca. 13,033,000 regular WoS citations, thus on average 6.2 % of the items' citations were not counted. This problem is especially severe for papers which attracted more preprint citations than regular citations. These make up 15.5 % of the cases. Table 1 shows for illustration the 10 publications with the highest identified preprint citation counts, with their regular WoS citation counts and basic bibliographic data. It can be seen that these are all highly cited recent papers, mostly from computer science.

Table 1: 10 publications with highest preprint citation count

preprint citations	WoS citations	first author	year	title
9532	3983	Devlin, Jacob	2019	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
5005	6556	Ioffe, Sergey	2015	Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift
2488	14545	Vaswani, Ashish	2017	Attention Is All You Need
2077	47745	He, Kaiming	2016	Deep Residual Learning for Image Recognition
2035	6130	Jia, Yangqing	2014	Caffe: Convolutional Architecture for Fast Feature Embedding

966	42285	Bates, Douglas	2015	Fitting Linear Mixed-Effects Models Using lme4
877	4382	Isola, Phillip	2017	Image-to-Image Translation with Conditional Adversarial Networks
751	2131	Pesaran, M. Hashem	2021	General diagnostic tests for cross-sectional dependence in panels
747	21704	Goodfellow, Ian	2020	Generative Adversarial Networks
747	3899	Redmon, Joseph	2017	YOLO9000: Better, Faster, Stronger

For the translated journals, our –very strict and thus incomplete– matching procedure identified 18,646 unmatched citations to more than 8000 papers across 29 journals, see Table 2. Most of these, more than 10,000, are to publications from the 1980s, while the 1990s, 2000s, and 2010s each have around 2300–3200 extra citations. Most papers had just one identified extra citation (~5000) but three had more than 100 each. For about 30 % of these papers, the number of newly identified citations was more than their regular WoS citation count. The papers received 61,120 WoS citations, so about 23 % of their citations were not identified and counted.

Table 2: Identified unmatched citations to translated journals

journal	additional citations
DOKLADY AKADEMII NAUK SSSR	8564
DIFFERENTIAL EQUATIONS	2492
MATHEMATICS OF THE USSR-IZVESTIYA	1386
DOKLADY AKADEMII NAUK	911
CHINESE JOURNAL OF ANALYTICAL CHEMISTRY	816
IZVESTIYA MATHEMATICS	607
MEASUREMENT TECHNIQUES	346
CHEMISTRY OF NATURAL COMPOUNDS	341
BULLETIN OF EXPERIMENTAL BIOLOGY AND MEDICINE	299
SOVIET PHYSICS ACOUSTICS-USSR	267
ASTRONOMY LETTERS-A JOURNAL OF ASTRONOMY AND SPACE ASTROPHYSICS	255
IZVESTIYA AKADEMII NAUK FIZIKA ATMOSFERY I OKEANA	233
CHEMISTRY OF HETEROCYCLIC COMPOUNDS	218
IZVESTIYA-PHYSICS OF THE SOLID EARTH	199
DOKLADY EARTH SCIENCES	198
IZVESTIYA AKADEMII NAUK SSSR FIZIKA ATMOSFERY I OKEANA	169
DOKLADY MATHEMATICS	162
MECHANICS OF COMPOSITE MATERIALS	150
STRENGTH OF MATERIALS	145
FLUID DYNAMICS	140
IZVESTIYA VYSSHIKH UCHEBNYKH ZAVEDENII RADIOFIZIKA	133
MECHANICS OF SOLIDS	133
DOKLADY PHYSICS	111

ACOUSTICAL PHYSICS	106
ASTROPHYSICS	92
MAGNETOHYDRODYNAMICS	64
SOVIET ASTRONOMY LETTERS	46
RADIOPHYSICS AND QUANTUM ELECTRONICS	43
ASTRONOMY REPORTS	20

4. Discussion and conclusion

This is an exploratory study with several limitations. As mentioned, we have used relatively strict matching procedures in order to rule out false positives. This means we are missing many additional citations, so these figures are underestimates of the real numbers.

We have studied two underappreciated sources of uncounted citations, preprint and translated paper citations. The extent of uncounted preprint citations is vast in the number of affected papers and can be expected to further increase in the future. However, the average share of missed citations for these papers is modest. We can say that a lower limit for the preprint open access citation penalty is 6 %. As for translated journals, the number of affected papers is small and unlikely to increase, but the average shares of uncounted citations are large.

Especially for the phenomenon of citations to preprint versions of published papers, routinely overlooked in conventional citation analysis, the incentive structure of the present system is misaligned with the goals of the open science and open access movements. *Ceteris paribus*, if preprint citations are not counted in bibliometrics-supported evaluations, researchers will be more reluctant to post early and open versions of their papers on preprint servers. Developers of citation indexing databases should improve their systems by offering users the *option* to include citations to different versions of one work, e.g. preprints and papers in translations, into citation count figures.

Open science practices

This study uses proprietary primary data. Query code can be obtained from the author upon request.

Competing interests

The author has no competing interests.

Funding information

This research used infrastructure funded by the German Federal Ministry of Education and Research, project 16WIK2101A.

References

Aksenteva, M. (2015). Some Features of the Citation Counts from Journals Indexed in Web of Science to Publications from Russian Translation Journals. *ISSI 2015 Proceedings*, 1220-1221.

Aman, V. (2015). Citing e-prints on arXiv. A study of cited references in WoS-indexed journals from 1991-2013. *ISSI 2015 Proceedings*, 1107-1119.

Langham-Putrow, A., Bakker, C., & Riegelman, A. (2021). Is the open access citation advantage real? A systematic review of the citation of open access and subscription-based articles. *PloS ONE*, 16(6), e0253129.

Traag, V. A. (2021). Inferring the causal effect of journals on citations. *Quantitative Science Studies*, 2(2), 496-504.