How do matchkeys affect citation counts? First steps towards an error calculus for bibliometric indicators

William Peter Dinkel

dinkel@forschungsinfo.de
Institute for Research Information and Quality Assurance (iFQ)

Abstract
While the methodological pitfalls of citation analysis are subject to intense debate in bibliometric research the more technical aspects of generating citation counts are sometimes lacking attention. However, as citation counts play a central role both in evaluative and descriptive bibliometrics they deserve more thorough consideration. In our contribution we want to present first results of an ongoing research project aiming at developing an error calculus for bibliometric methods. As a first step we compared the outcome of different algorithms for matching references with target documents using Web of Science data from 2007. This research in progress paper serves as a first exploration into the distribution of errors in citations rates. The preliminary results of our ongoing research suggest that the dispersion pattern of the resulting citation counts differ according to object of analyses. The extent of dispersion can be used as a simple measure for indicating the robustness of citation counts.

Background and purpose

The assumption that citations rates directly reflect the impact of publications has been challenged from a number of perspectives. From a theoretical point of view there is the general question of why researchers cite publications. From a methodological point of view there is a debate on how to count and attribute citations to objects of analyses have been raised. The use of full or fractional counting methods influences the results of citation analyses and also the possible scope of interpretation (e.g. Gauffriau et al. 2008). Additionally, there is the question on how to judge the context specific meaning of quantities of citations (eg. Bank & Delavalle 2008). Taken all this, most authors agree that the question of what citation counts actually measure is far from being resolved. However, a lively debate is going addressing these problems. Widely communicated but often unheard are concerns with regard to the reliability of citation data. A number of studies have analysed the quality of reference lists, which are the raw material for generating citation data, by quantifying errors in full reference entries. The reported error rates differ between range from 7% of errors as severe to render the identification of the target publication impossible (Sweetland 1989) up to 56% when counting every reported error (McLellan, Case & Barnett 1992). When narrowing the concept of error to errors in the minimal elements required for matching references with target documents the reported error rates range from an average 3% to maximum values up to 15% (Lok, Chan & Martinson 2002). Besides errors in reference lists the technical procedure applied for matching references with target documents has an impact on citation counts. Vriens and Moed (1989) thoroughly analyzed inaccuracies in the matching procedure applied in creating citation data to be used in the Web of Science (WoS) and distinguished a number of threats to the reliability of citation data, albeit only for a limited number of publications. Buchanan (2006) did a comparative case study on the interplay of errors in references lists and database errors matching the references of 204 citing articles to the cited documents based on WoS data. The result was an overall 10% of references being erroneous. These studies are highly informative with regard to the general accuracy of reference matching but of a limited value when it comes to judging the relevance of these errors for the reliability of citation analyses. Given that most bibliometricians rely on citation data provided in the WoS and Scopus, surprisingly little attention has been paid to how the database producers deal with errors in references list and to potential errors created by matching algorithms.

In this contribution we want to take a first step in this direction and examine to which extent citation counts change when using different methods for matching references with target document. Using the example of Web of Science (WoS) data for 2007 we focus on two research questions: Which impact do different methods for matching references with target documents have on citation counts? How do dispersion patterns of citation counts differ according to objects of bibliometric analysis? In order to answer these question we will in the first step define the concept of measurement error for citation counts, secondly present a method for calculating the measurement error of citation counts and thirdly, calculate the measurement error for various objects of analysis. We conclude by outlining limitations and next steps of our ongoing research.

Data and Method

Citation counts are calculated by algorithms which identify entries in reference lists that match with the metadata of indexed publications (target documents). In the design of matching algorithms two complementary approaches can be distinguished. Deterministic matching algorithms are based on the application of a set of rules for the identification of identical records. Rules are applied sequentially stepwise narrowing down potential matches. Probabilistic matching algorithms on the other hand rely on the calculation of similarities between records. Records are defined as matching if their similarity score is above a threshold (Elmagrid, Panagiotis & Vassilios 2007). In citation databases the most prominent approach for creating citation data is the use of matchkeys. Matchkeys are distinctive identifiers for references and publications which are created based on transforming the metadata of publications and the reference strings by removing and altering its elements. A number of matchkeys has been suggested each highlighting another element in the reference string (Lawrence, Giles & Bollacker 1999, Braun, Glänzel & Schubert 1985, overview in Synnestvedt 2007:16). Depeding on which matchkey one chooses to effects can be observed: Either the number of false positives (overmatch) increases or the number of false negatives decreases. The extent to which either one of the two is the case depends on the overall distribution of the elements of the reference in the set of target documents. For example: Assuming that there is a total of three publications written by authors named Kanciewicz in 2007, little additional information is required for identifying which one could possible be targeted by the reference. On the other hand, assuming that authors named Smith have published an overall of 300 publications in 2007 even small errors can reduce the chance of correctly attributing citations to publications from these authors. Consequently, it is justified to assume that the impact of the applied matchkey on citation counts varies between objects of analyses. In order to analyse if and to which extent this actually is the case we compared citation counts resulting from the use of a number of different matchkeys. Based on the comparison we calculated a relative measurement error for each citation count.

In order to do so, we created two datasets. The first dataset included the metadata of all publications from the WoS (SCIE, SSCI, A&HCI, excluding publications from proceedings) with publication year 2007 (target items, 1.5 Million records). This data was matched against a second dataset consisting of references referring to publications published in 2007 (references, 9 Million records) also gathered from the WoS (References of documents in SCIE, SSCI, A&HCI, excluding references from proceedings). The datasets included last name and initial of the first author (A), the abbreviated name of the source (S), begin page (P), volume number (V) and publication year (Y). After basic data cleaning consisting of removing special characters, useless white spaces and a number of other strings (“in press”, “forthcoming”, etc.) we created a number of matchkeys for the linkage between references and target items. In the next step we matched the reference dataset against the target

documents dataset based on these matchkeys, with each of the matchings providing a specific citation rate.
For each item the citation counts based on a matchkey consisting of all information in the dataset and the citation counts resulting from the matchkeys including less information were retrieved. In cases where after removing information a reference could be related to more than one target item we decided to give a fractional citation to all potentially relevant target documents. Consequently we divided the citation count of this item by the number of potential target documents. This approach is somewhat contra intuitive as it assumes an equal distribution of the elements of bibliographic records. Further analyses should systematically question this approach, e.g. by attributing references based on relative probabilities. In order to represent the resulting measurement error the relative observational error was calculated. We used the following formula for calculating the measurement error on an item base:

$$Measurement\ Error = \frac{average\ citation\ count\ (all\ matchings) - citation\ count\ (exact\ matching)}{citation\ count\ (exact\ matching)}$$

The measurement error can be described as the extent to which the average result of all matching procedures differs from the result of the exact matching. For example: a measurement error of 5% would suggest that that the combination of alternative matching procedures increases the citation count by 5% on average. A measurement error of 0 suggests that the citation count does not change when applying an alternative matching procedure.

Preliminary findings

The analysis of the distribution of measurement errors in citation counts for the 25 most productive countries in terms of absolute publications (Table 1) reveals that citation counts tend to be underestimated. However, the pictures on country level are diverse. The results of the analysis suggest that the highest measurement error can be found in publications with authors from Russia, China, South Korea, India, Taiwan, and Japan: There is evidence that citation counts of publications with authors from these countries tend to be underestimated.

Table 1. Distribution of measurement error in citation counts by country (full count), top 25 most publishing countries

Country	No error	+10%	+20%	+30%	>30%
Russia	78.1%	10.8%	3.3%	2.2%	5.6%
Peoples R China	80.7%	8.4%	3.1%	2.2%	5.7%
South Korea	80.8%	8.1%	3.5%	1.9%	5.6%
India	82.9%	7.7%	2.9%	1.8%	4.7%
Taiwan	83.0%	7.6%	3.0%	1.6%	4.7%
Japan	84.8%	6.7%	2.6%	1.4%	4.5%
Germany	85.2%	6.9%	2.5%	1.3%	4.1%
Poland	85.3%	7.0%	2.5%	1.3%	3.9%
USA	85.7%	6.4%	2.3%	1.3%	4.2%
France	85.9%	6.8%	2.3%	1.3%	3.8%
Israel	86.0%	6.8%	2.3%	1.3%	3.7%
Italy	86.1%	6.6%	2.3%	1.2%	3.8%
Spain	86.4%	7.0%	2.0%	1.4%	3.3%
Switzerland	86.4%	6.6%	2.4%	1.0%	3.7%
UK	86.4%	6.6%	2.0%	1.2%	3.7%
Canada	86.7%	6.3%	2.3%	1.3%	3.5%
Australia	87.0%	6.6%	2.0%	1.2%	3.3%

Belgium	87.3%	6.9%	1.8%	0.8%	3.2%
Greece	87.4%	6.3%	2.0%	1.3%	3.0%
Sweden	87.5%	6.2%	2.1%	1.0%	3.3%
Austria	87.5%	6.6%	2.1%	0.9%	3.0%
Turkey	87.9%	6.2%	1.9%	1.0%	3.0%
Netherlands	88.1%	6.2%	1.9%	1.1%	2.8%
Denmark	88.1%	6.4%	1.8%	0.8%	2.9%
Brazil	88.2%	6.2%	1.8%	0.9%	2.9%

This finding is supported by data on author level. Table 2 shows the most frequent author names sorted in descending order being mostly East Asian author names. We can see that for those authors the share of publications with extended error rates is higher than the overall share of publications. Further analysis will reveal to which extent these errors results from transcription errors or difficulties of non-native speakers in distinguishing between first name and last name when citing authors.

Table 2. Distribution of measurement error in citation counts by name of author

Name	Share in publications with error >20%	Total share
Wang	2.46%	0.81%
Zhang	1.85%	0.64%
Li	1.83%	0.65%
Chen	1.44%	0.59%
Liu	1.32%	0.54%
Lee	1.30%	0.54%
Kim	1.29%	0.53%
Yang	0.70%	0.34%
Wu	0.60%	0.29%
Huang	0.47%	0.25%
Xu	0.46%	0.22%
Park	0.39%	0.20%
Zhao	0.37%	0.18%
Zhou	0.36%	0.18%
Lin	0.35%	0.21%
Kumar	0.33%	0.11%
Lu	0.29%	0.14%
Singh	0.28%	0.14%
Yu	0.27%	0.16%
Zhu	0.27%	0.14%

Table 3 shows the distribution of measurement errors by document type. We can observe that the relative errors for the document types “Editorial Material” and “Letter” are above the ones for articles and reviews.

Table 3. Distribution of measurement error in citation counts by document type (document types contributing more than 1%)

Document Type	no error	+10,0%	+20,0%	+30,0%	>+30%	Share
Article	85.1%	6.8%	2.4%	1.4%	4.3%	89%

Review	89.0%	5.6%	1.7%	1.0%	2.7%	6%
Editorial Material	79.5%	7.4%	2.8%	1.7%	8.6%	3%
Letter	79.0%	7.8%	3.0%	1.9%	8.4%	2%

On the source level a very diverse picture emerges. Measurement errors range from a rough 4% of citation counts increasing when altering the method for creating citation rates to some 70%. One reason for this might be found in journals using article identifiers instead of page sequences in order to identify articles within a volume. However, in order to determine to which extent the results show up due to systematic errors further in detail analyses are required.

Table 4. Distribution of measurement error in citation counts by publication source, 10 sources with most publications

Source Name	no error	+10%	+20%	+30%	>+30%
ANAL CHEM	91.8%	3.9%	1.3%	0.9%	2.1%
ANGEW CHEM INT EDIT	88.0%	5.7%	2.0%	0.9%	3.4%
APPL PHYS LETT	38.2%	17.5%	10.8%	7.1%	26.3%
APPL SURF SCI	84.8%	7.5%	2.2%	1.6%	3.9%
ASTRON ASTROPHYS	84.4%	8.9%	2.0%	1.1%	3.5%
BIOCHEM BIOPH RES CO	93.2%	3.8%	1.1%	0.5%	1.3%
BIOCHEMISTRY-US	93.2%	3.0%	1.3%	0.5%	2.0%
BIOORG MED CHEM LETT	84.1%	8.8%	2.6%	1.5%	3.1%
BLOOD	94.0%	3.2%	0.6%	0.3%	1.8%
CANCER RES	94.8%	2.3%	0.7%	0.7%	1.4%

Discussion

This research in progress paper serves as a first exploration into the distribution of errors in citations rates. In our ongoing research we examine two research questions: How do matching procedures affect citation counts? How does the measurement error of citation counts differ among objects of analyses? Our first question is addressed by developing a method that allows us to comparatively assess the impact of matching algorithms and errors in reference lists on citation counts. We assumed that the measurement error of citation counts on an item level is determined by the algorithm used for matching references with target documents, the characteristics of the objects included in the references and the characteristics of the objects included in the target documents. Based on our results it is justified to argue that the method used for linking reference to target documents affects citations counts. Consequently, a careful reflection of the method used for generating citation data helps improving the quality of bibliometric analyses. The second question is tackled by analysing the distribution of measurement errors for different aggregations. As a summary of our result we can state that the distribution of measurement errors differs. Particularly citation counts for publications with authors from the East Asian countries seem to have a high measurement error. Citation counts calculated for publications involving authors from those countries seem to be less robust against errors in reference lists.

Limitations and trajectories for further development of the approach

The outlined preliminary findings serve as a starting point for further research on potential causes of errors in citation counts. The following methodological and conceptual limitations

of our approach have to be taken into account when dealing with the present preliminary results:

- No additional measures of dispersion beyond the relative observational error were applied.
- In order to correct for outliers we excluded uncited publications. In further analysis methods for representing changes in citation counts for initially uncited publications should be evaluated.
- Deterministic approaches as the one we used are very sensitive towards minor deviations. The inclusion of methods for correcting errors in references lists will contribute to further increasing the quality of record linkage.
- No comparisons were made between citation counts provided by Thomson Reuters and citation counts resulting from our research. Further analyses can reveal to which extent the described measurement errors apply for citation counts provided in the Web of Science.

Based on the findings a number of additional research questions arise: Which systematic biases in citation counts can be determined by our approach? How does the result of rankings or evaluative studies change when incorporating measurement errors? Which impact do specific types of errors have on citation counts? These questions but also the limitations of our current approach will direct the further progress of the present research.

References

Banks, MA & Dellavalle, R (2008). Emerging alternatives to the impact factor. OCLC Systems & Services, 24(3), 167-173.

Buchanan, RA (2005). Accuracy of cited references: the role of citation databases. College & Research Libraries, 67(4), 292-303.

Bornmann, L & Daniel, HD (2008). What do citation counts measure? A review on citation behaviours. Journal of Documentation 64(1), 45-80.

Braun, T, Glänzel,W & Schubert, A (1985). Scientometric Indicators, World Scientific.

Elmagrid, AK, Panagiotis, GI & Vassilios, SV (2007). Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19(1), 1-16.

Gauffriau, M, Larsen, PO, Maye, I, Roulin-Perriard, A & von Ins, M (2007). Comparisons of the results of publication counting using different methods. Scientometrics 77(1), 147-176.

Lawerence, S, Giles, LC & Bollacker, KD (1999). Autonomous citation matching, Proceedings of the third international conference on autonomous agents, Seattle, Washington, May 1-5, ACM Press, New York, NY, 1999.

Lok, C, Chan, M & Martinson, IM (2001). Risk factors for citation errors in peer-reviewed nursing journals. Journal of Advanced Nursing 34(2), 223-229.

McLellan, MF, Case, LD & Barnett, MC (1992). Trust, but verify. The accuracy of references in four anesthesia journals. Anesthesiology 77(1), 185-188.

Moed, HF & Vriens, M (1989). Possible inaccuracies occurring in citation analysis. Journal of Information Science 15, 95-107.

Moed, HF (2005). Citation analysis in research evaluation. Springer: Dordrecht.

Sweetland, JH (1989). Errors in bibliographic citations: A continuing problem. Library Quarterly 59, 291–304.

Synnestvedt, MB (2007). Data preparation for biomedical knowledge domain visualization: a probabilistic record linkage and information fusion approach to citation data. Drexel Thesis and Dissertations. http://hdl.handle.net/1860/2532