The impact of core documents: a citation analysis of the 2003 Science Citation Index core-document population

Bo Jarneving[1]

[1] *bo.jarneving@ub.gu.se*
University of Gothenburg, Gothenburg University Library, Renströmsgatan 4, 40530 Gothenburg (Sweden)

Abstract

In the 1960s Kessler introduced bibliographic coupling as a method for grouping research papers, facilitating scientific information provision. Later research has verified the applicability of this method in various information science contexts, such as information retrieval and science mapping. In this study the impact of so called 'core-documents', previously highlighted in the context of research front mapping, was elaborated applying state of the art impact indicators and varying citation windows. Due to limited resources, a random sample from the 2003 core-document population from the *Science Citation Index* was applied for statistical inference. Results were analyzed at the 95 % confidence level, applying confidence intervals for the arithmetic mean, proportions and the regression line. Findings indicated that core-documents were well cited above baselines and that a large share belonged to the top-cited papers of the world. Findings, not contradicting previous results, but providing with considerably more detail, lay ground for a more nuanced interpretation of core-documents' citation impact, where previous claims of key-positions in the science communication system were moderated. Findings also indicated that core-documents may have a rate of obsolescence notably deviating from the world average.

Introduction

Bibliographic coupling (BC) was introduced by Kessler through a number of reports and research articles in the 60s' (Kessler 1960; 1962; 1963a; 1963b; 1965). A bibliographic coupling unit was defined as: "[a] single item of reference shared by two documents…" (1962). BC was basically presented as a method for grouping technical and scientific documents which would facilitate scientific information retrieval. The original experiments performed by Kessler were based on small data sets from the journal *Physical Review*, why only limited conclusions of the method's applicability could be drawn. It took about two decades before a large scale experiment in a multidisciplinary environment took place (Vladutz and Cook, 1984). Findings showed that strong bibliographic coupling links generally implied strong subject relatedness. About the same time, Sen and Gan (1983) elaborating on the relation between subject relatedness and BC from a theoretical point of view, suggested a measure of coupling strength, the Coupling Angle (CA). With the point of departure in a hypothetical Boolean matrix where elements indicated presence or absence of a relationship between citing documents (rows) and cited documents (columns), the CA was expressed as:

$$CA = \frac{(D_{oj} \bullet D_{ok})}{\sqrt{(D_{oj} \cdot D_{oj})(D_{ok} \cdot D_{ok})}} \ ,$$

where

D_{oj} and D_{ok} are the binary vectors of document *j* and *k*.

Specifically, the CA corresponds to the cosine of the angle for two vectors, *j* and *k*. The range is [0,1] where a cosine of 0 corresponds to an angle of 90° and a cosine of 1 to an angle of 0°. Using the CA as a measure of similarity between two documents, the minimum value (0) implies no common references whereas the maximum value (1) implies identical reference lists.

A more convenient way to express the same relation between document *j* and *k* is to calculate the ratio between the number of common references for *j* and *k* and the geometric mean of the number of references for *j* and *k*:

$$\frac{r_{jk}}{\left(n_j n_k\right)^{½}}$$

where

r_{jk} is the number of references common to both *j* and *k*

and

n is the number of references in document *j* or *k*.

Lacking empirical evidence of document-document similarity based on BC, Sen and Gan suggested a preliminary threshold of CA = 0.5 which corresponds to an angle $\theta = 60°$.

The relation between document-document similarity and BC was further elaborated by Peters, Braam and van Raan (1995) where the cognitive resemblance within groups of documents, bibliographically coupled by one and the same highly cited item, was explored using publications from the field of Chemical Engineering. Measuring word-profile similarities between the citing documents, it was found that word profile similarity within groups sharing a citation to a highly cited publication was significantly higher than between documents without such a relationship.

It may be concluded that empirical evidence of this method's ability to group similar papers, enough to warrant further investigations of plausible bibliometric areas of application, had been gathered at this point in time. In 1995 Glänzel and Czerwon presented a method applying BC for the identification of so called "hot research topics". Their method was based on the concept of "core-documents" which implied established thresholds for both the CA and the number of papers connected at the same set CA. Hence, a core-document would be defined as a paper connected with at least ten other papers with a minimum coupling strength of CA = 0.25. A limitation of document types was also done so that only articles, notes and reviews were included. In their empirical study, the whole annual accumulation of the 1992 volume of SCI was applied and about one percent of all publications of the preferred document types were identified as core-documents. In a sequel (1996) the same set of core-documents was analyzed with regard to the distribution of core-documents over journals, subfields and corporate addresses. A citation analysis was performed at the national level, applying a two-year citation window for all indicators.

Three main citation indicators were applied:

- *The relative citation rate* (RCR) which is the ratio of the mean observed citation rate (MOCR) to the mean expected citation rate (MECR). With regard to MECR, actual citations were substituted with journal impact factors.

- *Percentage of documents cited above average*. This indicator sums up the number of core-documents cited at least as many times as the corresponding journal impact factor and calculates the share.
- *Number of highly cited papers*. A core-document is considered "highly cited" if it has received at least 5 times as many citations as the corresponding journal impact factor.

Findings showed that core-documents, as defined, generally reflected "hot" research front topics, though the method seemed to have a bias towards the life sciences as most core documents were found in biomedical sub-fields. It was concluded that core-documents hold a key position in science communication on grounds of their high citation impact.

Research rationale

Later research on core-documents based on BC has involved cluster analytical approaches and network analysis (Jarneving 2007a, b; Glänzel and Thijs, 2011, 2012) as well as the combined application of textual information and citation data (Glänzel and Thijs, 2011, 2012). However, the citation impact of core-documents has not yet been exhaustively elaborated. Previous research on citations of core-documents has been limited to the use of journal impact factors for the expected citation rates, with a focus on geographical distributions. Hence, there is a need of investigating the citation of core-documents using current citation based performance indicators. In addition, a wider citation window would complement previous findings where a two year window was applied. In particular, the claim that core-document may be considered keys for the identification of outstanding research performance (Czerwon and Glänzel, 1996) should be elaborated on.

Data and methods

From the SCI volume 2003 on CDROM, 619,570 records of the document type article were downloaded. A delimitation of document types to genuine research articles was made on grounds that this document type best mirrors empirical research. This population is referred to as the 2003 SCI core-document population, though ten percent of the core-documents had a publication year other than 2003. A total of 17,674,944 references were processed and 6,060 core-documents identified, which is approximately one percent of the total population of articles. Limited resources implied that citation indicators could not be generated for the total population of core-documents, why a random sample substituted the population of core-documents and estimates were applied. In order to be representative of the population, the sample was based on proportionate stratified sampling where strata were constructed on basis of major fields of science as defined in *Essential Science Indicators* (Thomson Reuters). The appropriate sample size was computed with a point of departure in the standard error of a proportion:

$$0.05 = 2 \cdot 1.96 \cdot \sqrt{\frac{p(1-p)}{n}}$$

where

n = the sample size

p = the share of papers in the sample.

This means that a width of the confidence interval of 0.05 at the confidence level of 95 % was accepted. However, as we do not know the different shares, *p* must be guessed. Substituting *p* with 0.5 (which gives the largest value for *n*) gives the following equation after squaring and simplifying:

$$n = (2 \cdot 1.96)^2 \cdot \frac{0{,}5^2}{0{,}05^2}$$

Conclusively, a sample of 1,500 papers would probably work well. This means that approximately a quarter of all papers should be randomly drawn from the population of 6,060 core documents. This rather large share of a *finite* population as well as the fact that the sampling was performed without replacement requires a correction factor (Isserlis, 1918) for both proportions and means when computing the standard error,

$$\sqrt{\frac{N-n}{N-1}}$$

where *N* = the population size,

and *n* = the sample size.

A total of three citation based indicators was decided on:

The average field normalized citation score ($\bar{C_f}$), where the expected number of citations (*e*) was computed as the average number citations to publications of the same type, with the same publication year and from the same field. It is defined as:

$$\frac{\sum_{i=1}^{n} \frac{c_i}{e_i}}{n}$$

where

c_i = number of citations to publication *i*
e_i = the expected number of citations to publication *i*
n = number of publications

This indicator was presented by Lundberg (2007) as the "Item oriented field normalized citation score average".

The average journal normalized citation score $\bar{C_j}$ is calculated analogously but the expected citation frequency is calculated as the average citation frequency of the corresponding journal, considering document type and publication year.

Top *n %* is the percentage core-documents that belong to the *n %* most cited papers in the world, where papers are matched with regard to publication year, field and document type. In this study *n* assumes the values 5, 10 and 20.

The expected citation frequencies as well as the top *n %* indicator values were matched with each individual publication of the random sample by CWTS, Leiden University, using data from Thomson Scientific/ISI. All indicator values were computed with self citations excluded.

Findings

Before presenting the results from the citation analysis, some descriptive statistics should be commented on. Considering the distribution of core-documents over journals, 995 distinct journal titles out of 3,567 contained at least one core document, which means that 72 % of all journals in the 2003 SCI volume did not contain any core documents. The corresponding figure in Glänzel & Czerwon (1996) was 75 %. Another distribution of interest concerns co-authorships. The mean number of authors of a core-document was 6.0 and the maximum number of authors 255. The corresponding figures for the 1992 volume were 4.5 and 104 (Glänzel & Czerwon, 1996). For the whole 2003 volume the mean number of authors was 4.4.

Impact

For each core-document in the sample, citation data for 7 years was assembled. Counting whole publication years, the maximal error with regard to the publication date was thus less than but approximately one year. The first whole year after the publication year was considered to correspond to a (minimum) citation window of one year. In this way, three citation windows were applied:

- 2 years: three years after the publication year
- 4 years: five years after the publication year
- 6 years: seven years after the publication year

In Table 1, the arithmetic mean for $\bar{C}_f$ and $\bar{C}_j$ are displayed with confidence intervals at the 95 % confidence level. As can bee seen, both $\bar{C}_f$ and $\bar{C}_j$ decrease over time and $\bar{C}_f$ is notably higher.

Table 1. The $\bar{C}_f$ and $\bar{C}_j$ for three citation windows with confidence intervals at the 95 % confidence level.

Citation window	$\bar{C}_f$	CI	$\bar{C}_j$	CI
2 years	2.90	± 0.19	2.23	± 0.14
4 years	2.67	± 0.18	2.13	± 0.14
6 years	2.52	± 0.18	2.03	± 0.13

Considering the impact of core-documents on fields, the top *n* % indicators show the share of core-documents that belong to the world's top *n* %. Here, *n* assumes values of 5, 10 and 20, which are displayed over three citation windows along with confidence intervals at the 95 % confidence level (Table 2).

Table 2. The share of core-documents belonging to *n %* top-cited papers: 5, 10 and 20 percent levels are displayed for three citation windows with confidence intervals at the 95 % confidence level.

Citation window	top 5 %	CI	top 10 %	CI	top 20 %	CI
2 years	0.25	± 0.02	0.36	± 0.02	0.52	± 0.02
4 years	0.24	± 0.02	0.35	± 0.02	0.51	± 0.02

6 years	0.22	± 0.02	0.32	± 0.02	0.48	± 0.02

The impact profile for the sampled core-documents is displayed in Figure 1, providing with comprehensible class intervals (cf. Adams, Gurney & Marshall, 2007) for $\bar{C}_f$ over a 6-year citation window. Note that the upper bounds increase by a factor of two for each new class interval. The 6-year citation window was chosen in order to exhaustively assess the influence of the category "uncited".

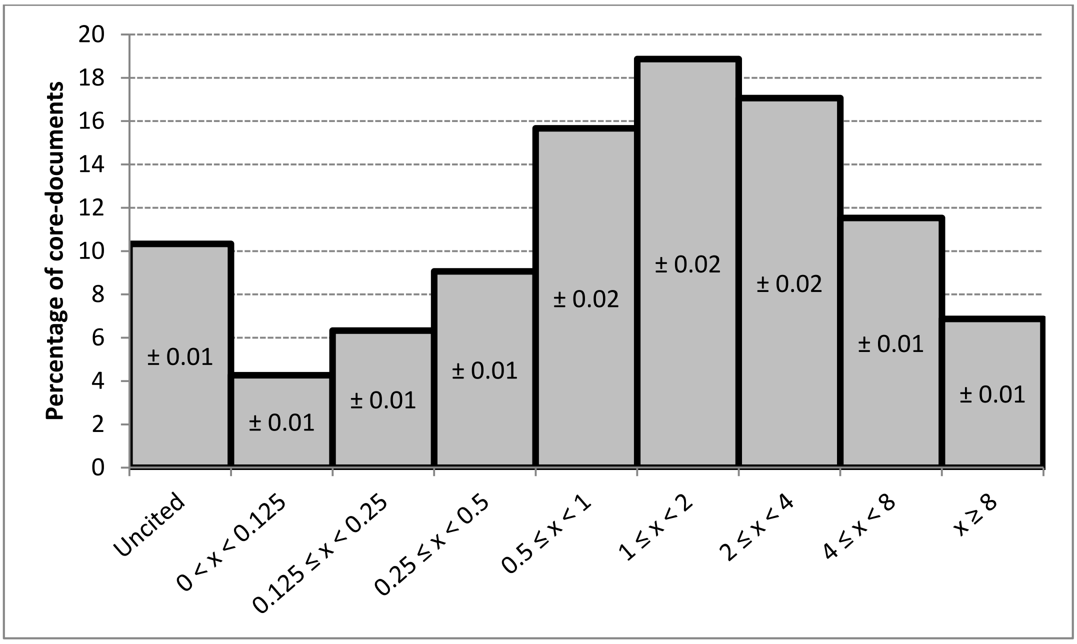

Figure 1. The 6-year impact-profile for 1500 core-documents: class intervals where $\bar{C}_f = x$ marked on the x-axis. Confidence intervals at the 95 % confidence level displayed within bars.

Growth of citations

Focusing on the relation between the length of the citation window and the number of observed citations, a regression analysis was performed. The graph in Figure 2 illustrates a near perfect linear relationship with confidence intervals at the 95 % confidence level for the number of observed citations. Given this growth model, the set of sampled core-documents receives an annual contribution of 8,618 citations, while the lower bound was 8,192 citations and the upper 9,045. The ratio of the annual number of expected citations to the lower bound of the observed citations was 1 to 2.5, reflecting a much faster accumulation of citations to the sampled core documents (Figure 2).

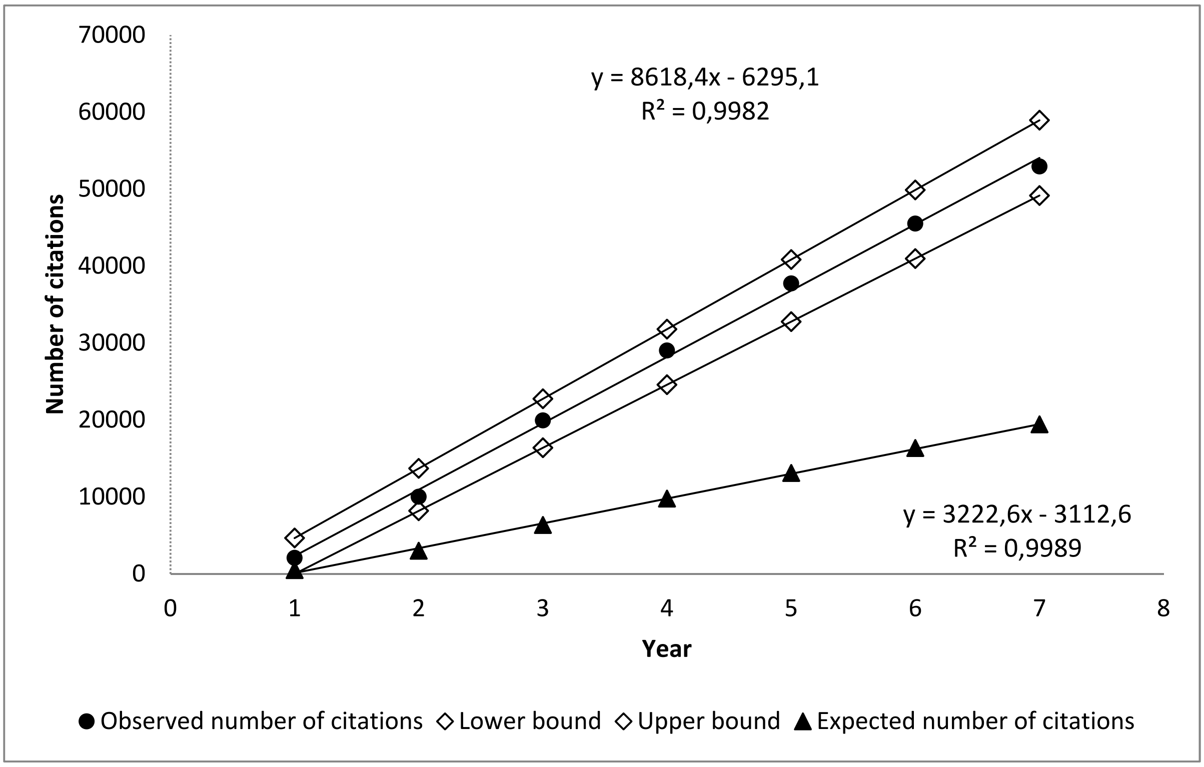

Figure 2. Number of accumulated citations for the sampled core-documents during a six-year citation window and number of accumulated expected citations. Lower and upper bounds at the 95 % confidence level are displayed for the observed citations.

However, for a 6-year citation window, the expected percentage growth was 3849 % and for the point estimate 2436 %, CI [2043 %, 2491 %].

Discussion

Results convincingly showed that the average field normalized citation score ($\bar{C_f}$) for core documents was well above the world standard (the expected). With regard to the two-year citation window, $\bar{C_f}$ was almost three times the expected according to the point estimate. Considering the *lower* bounds of the confidence intervals, over all three citation windows with 95 % confidence, the corresponding population parameter was within the interval 2.34 – 2.71. The corresponding interval for journal normalized citation counts ($\bar{C_j}$), was 1.89 – 2.09. These figures indicate a substantial difference between the expected and the observed. The different results arrived at when applying field normalization respectively journal normalization should reflect that core-documents are often published in high impact journals.

Mapping the impact of core-documents in terms of their percentage distribution over top *n* % categories is complementary to elaborations on averages. Given the narrowest citation window, a quarter of all sampled core-documents belonged to the top 5 % most cited publications with a margin of error of ± 2 %. With approximately the same margin of error, corresponding figures for top 10 % and top 20 % was 36 % and 52 % respectively. These findings are actually not in line with the claim that core-documents belong to the set of high impact papers of specialties (Glänzel & Czerwon, 1996), at least not in a general sense.

A more elaborated depiction of core-documents' citedness was provided by the histogram in Figure 1. We can appreciate that a majority of the sampled core-documents are cited above the world average (1.0) and a little less than half below. About ten percent is never to be cited during a six-year citation window, while approximately 18 percent have a citation frequency at least four times the world average. The modal group of sampled core-documents are cited above the world average but within the limit of a factor of two. The definition of core-documents as such suggests a close relationship with the research front, that is, with the portion of current papers within a field that is tied to a relatively small and select group of earlier papers by citation (cf. Price, 1965). However, this would not per see imply a high citation rate as other markers of high quality such as originality and immediacy play important roles. In fact, results arrived at here indicate that for every core-document cited more than twice the expected, we would find a core-document cited below the world average. Conclusively, core document attributes are not in themselves sufficient markers of "outstanding research performance" (cf. Glänzel and Czerwon, 1996). However, it would be complementary to explore to what extent high impact papers possess core document attributes.

Notably, the 1992 core-document population showed up with a considerably larger figure for the share of core-documents cited above average. In Glänzel and Czerwon (1996), 62.4 % were cited above average while the corresponding figure in this study was 54 %, ± 2.2 %. One may assume that there is a trend of an increasing number of core-documents of lower quality. Another, assumption is that the difference between the two populations is due to the fact that review papers generally have a higher citation impact than research articles (Glänzel and Moed, 2002).

Considering the accumulation of citations to core-documents, a much faster than expected growth during the 6 year citation-window was observed. This is in line with expectations and other findings. However, it was also observed that indicator values decline notably over time (cf. Table 1 and Table 2). This indicates that core-documents have a higher obsolescence than expected. Consequently, the percentage growth for core-documents was substantially lower than expected, also when considering the upper bound of the confidence interval.

Conclusions

In spite of the obvious limitations of basing inferences on a random sample, it has been feasible to map citation impact of core-documents at the 95 % confidence level. Findings indicate that core-documents are well cited above baselines and that a large share of the 2003 *Science Citation Index* core-document population belongs to the top-cited papers of the world. This is basically in line with previous findings, though considerably more detailed information with regard to relevant impact indicators lay ground for a more nuanced interpretation of core-documents' role in the scientific communication system. Hence, previous claims of core-documents key-position and impact should be moderated on grounds that the citation impact of core-documents is unevenly distributed.

References

Adams, J., Gurney, K. & Marshall, S: (2007). Profiling citation impact: A new methodology. *Scientometrics*, 72(2):325-344.

Glänzel, W. & Czerwon, H. J. (1995). A new methodological approach to bibliographic coupling and its application to research-front and other core documents, *Proceedings of 5th International Conference on scientometrics and Informetrics*, held in River Forest, Illinois, June 7-10: 167-176.

Glänzel, W. & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, 37(2):195-221.

Glänzel, W. & Moed, H. (2002). Journal impact measures in bibliometric research. *Scientometrics*, 53(2):171-193.

Glänzel W, Thijs B. (2011). Using 'core documents' for the representation of clusters and topics. *Scientometrics*, 88(1):297-309.

Glänzel W, Thijs B. (2012). Using 'core documents' for detecting and labeling new emerging topics. *Scientometrics*, 91(2):399-416.

Isserlis, L. (1918). On the value of a mean as calculated from a sample. *Journal of the Royal Statistical Society*: 75–81.

Jarneving B. (2007a). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics*, 1(4):287-307.

Jarneving B. (2007b). Complete graphs and bibliographic coupling: A test of the applicability of bibliographic coupling for the identification of cognitive cores on the field level. *Journal of Informetrics*, 1(4):338-56.

Kessler, M. M. (1960). An experimental communication center for scientific and technical information. Massachusetts Institute for Technology, Lincoln Laboratory.

Kessler, M. M. (1962). An experimental study of bibliographic coupling between technical papers. Massachusetts Institute for Technology, Lincoln Laboratory.

Kessler, M.M. (1963a). Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10-25.

Kessler, M.M. (1963b). Bibliographic coupling extended in time: Ten case histories. *Information Storage and Retrieval*, 1:169-187.

Kessler, M.M. (1965). Comparison of the results of bibliographic coupling and analytic subject indexing. *American Documentation*, 16(3):223-233.

Lundberg, J. (2007). Lifting the crown—citation z-score. *Journal of Informetrics*, 1(2):145–154.

Peters, H. P. F., Braam, R. R. and van Raan, A. F. J. (1995). Cognitive resemblance and citation relations in chemical engineering publications. *Journal of the American Society for Information Science*. 46(1):9-21.

Price, D. J. de Solla (1965), Networks of scientific papers. *Science*, 149(3683):510-515.

Sen, S. K. & Gan. S. K. (1983). A mathematical extension of the idea of bibliographic coupling and its applications. *Annals of Library Science and Documentation*, 30(2):78-82.

Vladutz, G. & Cook, J. (1984). Bibliographic coupling and subject relatedness. *Proceedings of the ASIS Annual Meeting*, 47: 204-207.