A BIBLIOMETRIC APPROACH TO THE ROLE OF AUTHOR SELF-CITATIONS IN SCIENTIFIC COMMUNICATION

Wolfgang Glänzel[a,b], Bart Thijs[a], Balázs Schlemmer[a]

[a] *K.U. Leuven, Steunpunt O&O Statistieken, Dept. MSI, Leuven, Belgium*
[b] *Hungarian Academy of Sciences, IRPS, Budapest, Hungary*

Abstract. The present paper analyses the role of author self-citations aiming at finding basic regularities of self-citations within the process of documented scientific communication and thus laying the methodological groundwork for a possible critical view at self-citation patterns in empirical studies at any level of aggregation. The study consists of three parts; the first part of the study is concerned with the comparative analysis of the ageing of self-citations and of non-self citations, in the second part the possible interdependence between self-citations and foreign citations is analysed and in the third part the interrelation of the share of self-citations in all citations with other citation-based indicators is studied.

The outcomes of this study are two-fold; first, the results characterise author self-citations – at least at the macro level – as an organic part of the citation process obeying rules that can be measured and described with the help of mathematical models. Second, these rules can be used in evaluative micro and meso analyses to identify significant deviations from the reference standards.

1 INTRODUCTION

In the bibliometric literature, there is an ongoing debate on the interpretation and role of author self-citations in the process of scientific communication. This debate has resulted in a certain polarisation. Particularly, users in science policy, but sometimes even the researchers themselves are condemning author self-citations as possible means of artificially inflating citation rates and thus of strengthening the authors' own position in the scientific community. Bibliometricians are, on the other hand, inclined to regard a reasonable share of author self-citations as a natural part of scientific communication. According to this view, the almost absolute lack of self-citations over a longer period is just as pathological as an always-overwhelming share. MacRoberts and MacRoberts (1989) have given a first overview of the unsolved problem of self-citations in their critical review on problems of citation analysis.

Beside the discussion on the principles of the role of author self-citations, there is no real consensus concerning how this type of self-citations should be defined operatively. In practice, two different approaches to direct self-citations are in use. At the micro level, that is, on the level of individual authors, a direct self-citation for an author *A* occurs whenever *A* is also (co-)author of a paper citing a publication by *A*. This definition cannot, however, be applied to higher levels of aggregation, that is, when publications and citations are aggregated over sets of different (co-)authors, and the notion of self-citations is uncoupled from an individual author A. At the meso and macro level, other criteria have to be used to determine what is considered a self-citation.

The final publication is available at Springer via http://dx.doi.org/10.1023/B:SCIE.0000013299.38210.74

The present study provides a large-scale analysis of the share and the ageing of self citations, as well as a breakdown by science fields on the basis of the total publication output indexed in selected annual updates of the Web of Science®. The objective of this study is not finding arguments pro or contra excluding self-citations from bibliometric analyses; the aim is to understand basic regularities of self-citations within the process of documented scientific communication in order to pave the methodological way for a possible critical view at self-citation patterns in empirical studies.

2 DATA SOURCES AND DATA PROCESSING

The results of this study are based on raw bibliographic data extracted from the 1992–2001 annual cumulations of the *Web of Science®* (WoS) of the Institute for Scientific Information (ISI - Thomson Scientific, Philadelphia, PA, USA). The extracted data have been undergone a detailed cleaning and then processed to bibliometric indicators. All papers of the document type Articles, Letters, Notes and Reviews indexed in the 1992 and 1999 annual updates of the WoS have been taken into consideration. Citations received by these papers have been determined for the period beginning with the publication year till 2001 on the basis of an item-by-item procedure using special identification-keys (so-called cluster-keys) made up of bibliographic data elements. Papers were assigned to countries based on the corporate address given in the by-line of the publication. All countries indicated in the address field were thus taken into account.

Subject classification of publications was based on the field assignment of journals (in which the publications in question appeared) according to the twelve major fields of science and three fields of social sciences and humanities developed in Leuven and Budapest (see, for instance, Glänzel and Schubert, 2003). In particular, the following fields have been used: Agriculture & Environment, Biology (Organismic & Supraorganismic Level), Biosciences (General, Cellular & Subcellular Biology Genetics), Biomedical Research, Clinical and Experimental Medicine I (General & Internal Medicine), Clinical and Experimental Medicine II (Non-Internal Medicine Specialties), Neuroscience & Behavior, Chemistry, Physics, Geosciences & Space Sciences, Engineering, Mathematics and Social Sciences I (General, Regional & Community Issues), Social Sciences II (Economical & Political Issues) and Arts & Humanities, respectively.

3 METHODS

For the present study, the same definition of self-citations has been applied as was used by Snyder and Bonzi (1998) and Aksnes (2003). In verbal terms, a self-citation occurs whenever the set of co-authors of the citing paper and that of the cited one are not disjoint, that is, if these sets share at least one author. Although the reliability of this methodology is affected by homonyms (resulting in Type II errors by erroneous self-citation counting) and spelling variances/misspellings of author names (resulting in Type I errors by not recognising self-citation), at high levels of aggregation, that is at the meso and macro level, there is no feasible alternative to the method used in this study.

Three main aspects of self-citations have been studied. The first part of the study is devoted to the analysis of the ageing of self-citations in comparison with that of non-self citations. It is based

on a 10-year diachronous (prospective) citation analysis of the 1992 volume of the WoS. The deviation of the aging (obsolescence) of self-citations from the standard set by foreign citations results was analysed. The analysis by fields in this part of the study was aiming at uncovering subject-specific peculiarities.

Second, the possible interdependence between self-citations and foreign citations was analysed. This was done on the basis of conditional expectations, where the condition is given by the total number of citations received.

Third, the analysis of the share of self-citations in all citations and its interrelation with other bibliometrics standard indicators was used to uncover simple regularities that can be applied as expected self-citation indicators in empirical studies for research evaluation. This part is based on the 1999 volume of the WoS using a 3-year (self-)citation window.

4 THEORETICAL CONSIDERATIONS

Before the empirical data will be analysed, a short formalisation is given to make it possible to understand the mathematical background of above-mentioned relationship between the conditional expectation of self-citations with the share of self-citations in all citations received by the publications of a given unit of analysis.

First let $\xi(t)$ denote the number of self-citation a paper published at time $s = 0$ has received in the period $[s, t] = [0, t]$. Putting $s = 0$ does not result in any restriction of generality. The rate of foreign citations, i.e., on non-self-citations is then be denoted by $\zeta(t)$. Consequently, we have $\eta(t) = \xi(t) + \zeta(t)$ for the total citation rate of a paper published at time 0. It is clear that $\xi(t)$ and $\eta(t)$ will not be independent nor uncorrelated random variables since $0 \leq \xi(t) \leq \eta(t)$ and $\eta(t) = 0$ thus implies $\xi(t) = 0$. The increments of the non-self citation processes will be denoted by $\Delta\zeta(t) = \zeta(t) - \zeta(t-1)$, those of the self-citation processes by $\Delta\xi(t) = \xi(t) - \xi(t-1)$ and the increments of the total citation process finally by $\Delta\eta(t) = \eta(t) - \eta(t-1)$. Let $P(\xi(t) = i)$ and $P(\zeta(t) = k)$ for $i, k = 0, 1, 2 \ldots$ denote the probability distributions of self-citations and foreign citations, respectively. In the following we will focus on the conditional expectations and probabilities.

1. The expected self-citation rate under the condition that the foreign citation rate is exactly k ($k = 0, 1, 2 \ldots$). Then we can express this expectation as follows,

$$E(\xi(t)|\zeta(t) = k) = \left(\sum_i i \cdot P(\xi(t) = i, \zeta(t) = k)\right) / P(\zeta(t) = k), \quad k \in \mathbb{N} \cup \{0\}.$$

2. The life-time distributions of ξ and ζ reflect ageing properties of the corresponding citation processes. They can be defined in the following manner (see Glänzel and Schoepflin, 1994, Burrell, 2002)

$$F_\xi(t) = E\,\xi(t)/E\,\xi(\infty) \text{ and } F_\zeta = E\,\zeta(t)/E\,\zeta(\infty)\,; \quad t \geq 0\,.$$

It must be mentioned that in this study $\xi(10)$ is used instead of $\xi(\infty)$ since the citation rates beyond 2001 are unknown.

3. The probability $P(S_t)$ that a citation is a self-citation in the interval $[0, t]$ can be expressed similarly as in the above definition, particularly,

$$P(S_t) = E\,\xi(t)/E\,\eta(t) = E\,\xi(t)/(E\,\xi(t) + E\,\zeta(t))\,.$$

In particular, if $\xi(t)$ and $\zeta(t)$ are independent variables then $E(\xi(t)|\zeta(t)) \equiv E\xi(t)$ for all $k \in \mathbb{N} \cup \{0\}$. Otherwise, there might exist an appropriate real function f, so that $E(\xi(t)|\zeta(t)) = f(\zeta(t))$. If there exists such function f, the probability that a citation is a self-citation can be expressed with the help of the above conditional expectations and the distribution of foreign citation as follows.

$$\begin{aligned} P(S_t) &= E\,\xi(t)/(E\,\xi(t) + E\,\zeta(t)) \\ &= \big(\textstyle\sum_k E(\xi(t)|\zeta(t) = k)\cdot P(\zeta(t) = k)\big)/\big(E\xi(t) + \sum_k kP(\zeta(t) = k)\big) \\ &= \big(\textstyle\sum_k f(k)\cdot P(\zeta(t) = k)\big)/\big(\sum_k f(k)\cdot P(\zeta(t) = k) + \sum_k k\cdot P(\zeta(t) = k)\big) \\ &= \big(1 + \textstyle\sum_k k\cdot P(\zeta(t) = k)/\sum_k f(k)\cdot P(\zeta(t) = k)\big)^{-1} \end{aligned}$$

That is, the $P(S_t)$ can be expressed with the help of specific moments of the distribution of non-self-citations over publications. This model will be applied to the results obtained from the empirical part of the following section.

5 RESULTS

The analysis of the ageing and share of self-citations, which forms the first part of this study, was based on papers published in 1992. The annual change of both self-citations and foreign citations (i.e., non-self citations) have been determined for all science fields combined, the twelve subject fields in the sciences and the field Social sciences I (General, regional & community issues) in the period 1992 till 2001. Figure 1 presents the plot of the life time distributions of self-citations and foreign citations based on the 10-year period for all fields combined and five selected fields, in particular, Biomedical research, Chemistry, Physics, Mathematics and Social sciences I. The life-time distributions is here calculated on the basis of empirical values of increments $\Delta\xi(t)$ and $\Delta\zeta(t)$, respectively. In particular, the distributions are defined as the following empirical densities on the basis of the increments of the processes, namely $\bar{f}_\xi = \Delta\xi(t)/\xi(10)$ for self-citations and $\bar{f}_\zeta = \Delta\zeta(t)/\zeta(10)$ for foreign citations.

The ageing of self-citations is much faster than that of foreign citations. This observation applies to all science fields (cf. Figure 1). However, the deviation of the ageing patterns of individual subject fields from each other and from that of all fields combined is considerable. The $\bar{f}_\xi$ curves are relatively similar for all fields except for physics. The curves of Physics and Chemistry have their modes in the second year (that is the year after publication) whereas those of all other fields have their modes in the third year (i.e. the in second year after publication). By contrast, the ageing curves of foreign citations have deviating and characteristic shapes in the different fields. This shows that ageing of self-citations is somewhat less field-specific than that of non-self-citations.

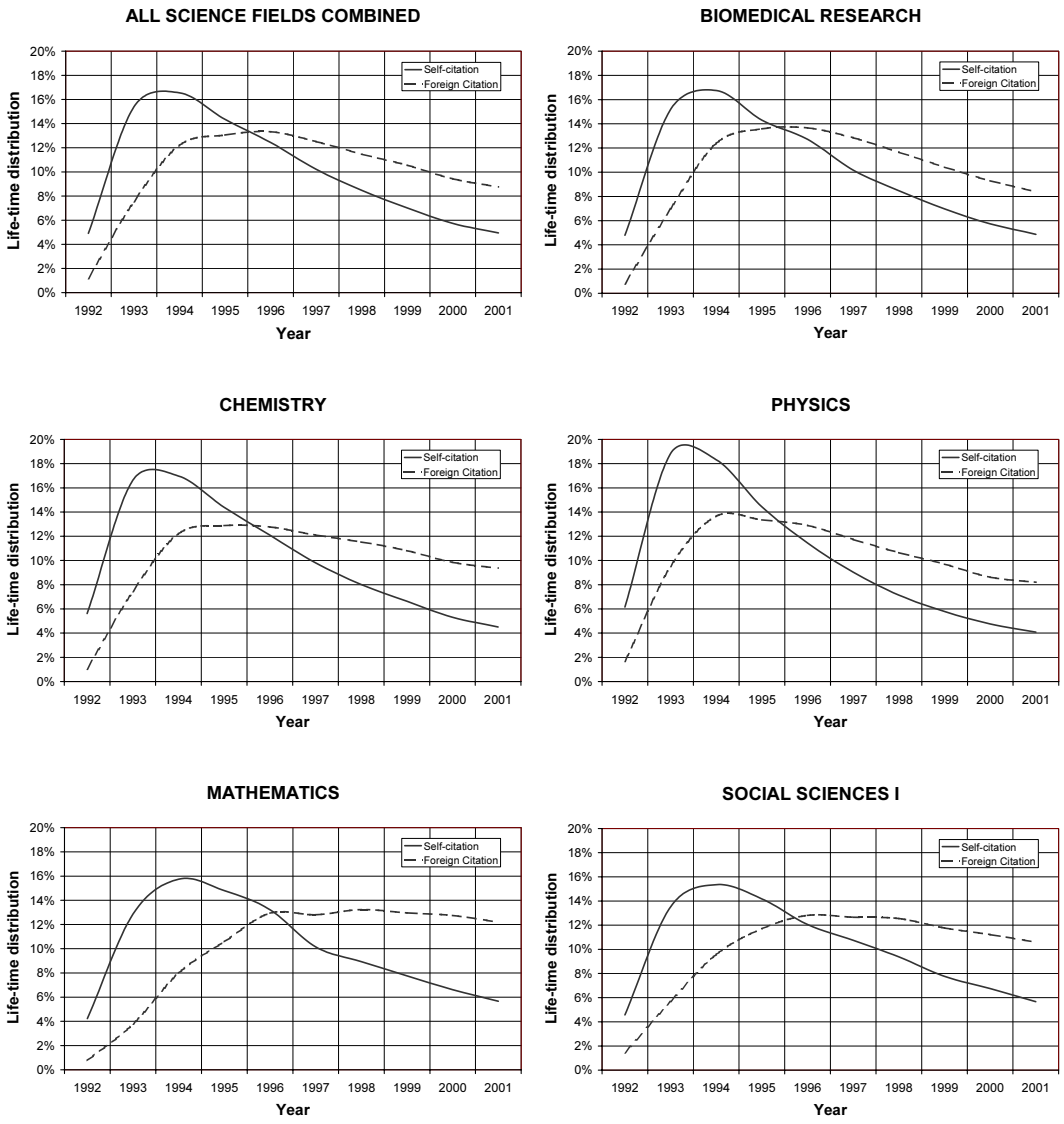

FIGURE 1. Empirical density of the life-time distribution $\bar{f}_\xi$ and $\bar{f}_\zeta$ for all science fields combined and five selected fields

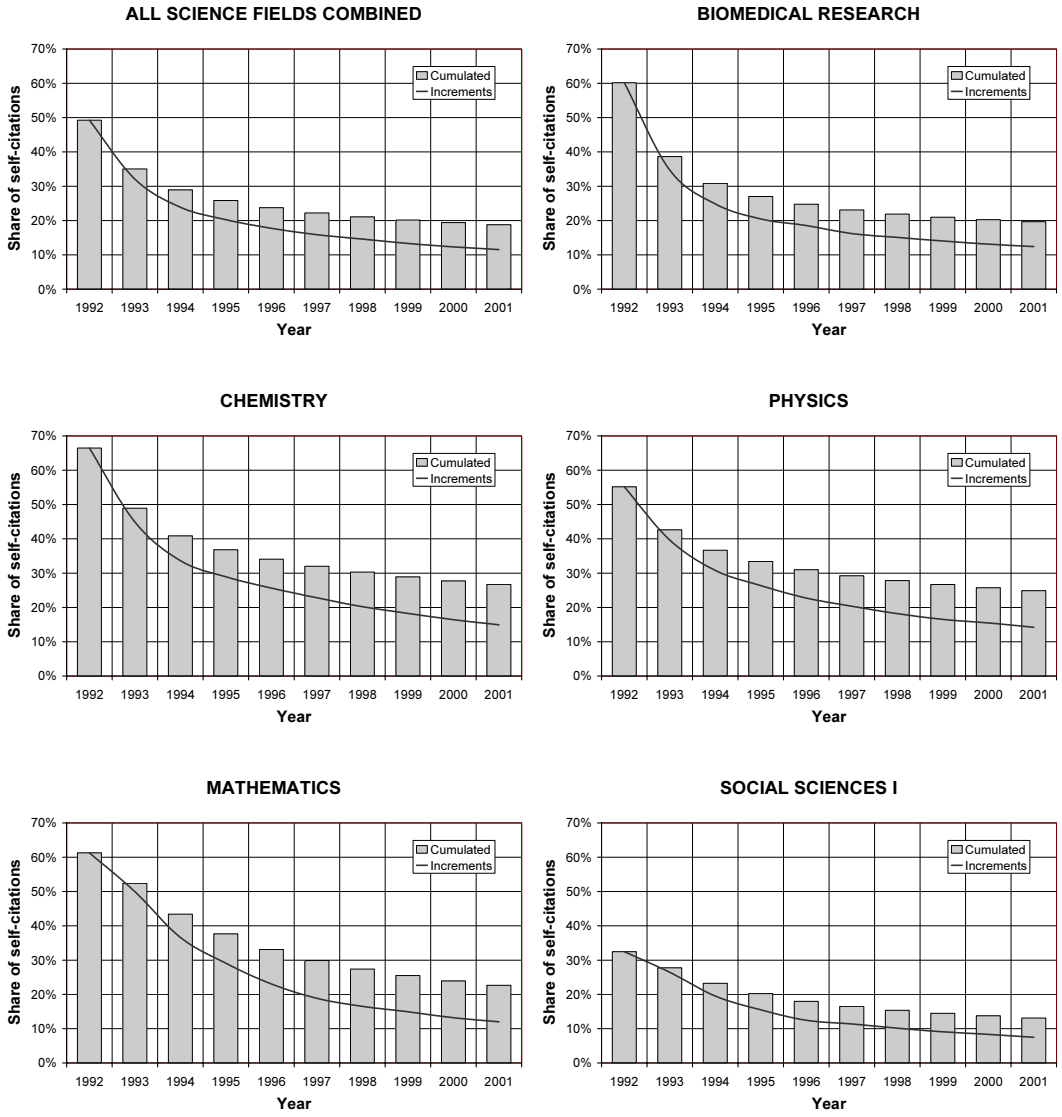

FIGURE 2. The annual change of the share of self-citations in all citations for all science fields combined and five selected fields (bars indicate shares of cumulated citation rates, lines that of increments)

The evolution of the share of self-citations in all citations is given in Figure 2. This share, which has been calculated for both, the process and its increments, is the estimate of the above-mentioned probability $P(S_t)$ and the corresponding density. We have $\overline{S}_t = \overline{\xi}(t)/\overline{\eta}(t)$ for the cumulative (self-)citation rates from 0 till t ($t = 0, 1, \ldots, 10$), where $t = 0$ corresponds to the year 1992 and $t = 10$ to 2001, and $\overline{S}_{\Delta t} = \overline{\xi}(\Delta t)/\overline{\eta}(\Delta t)$ for the share of self-citations received in the year t in all citations in the same year.

The decreasing share of self-citations for growing time windows is completely in keeping with the different ageing of the two processes. The share decreases from roughly 50% in the year of publication to less than 20% in the 10-year citation window. The values of this indicator considerably differ among the subject fields. Biomedical research, Chemistry and Mathematics have the highest share of self-citations ($\geq$ 60%) in the year of publication. The field 'Social sciences I' has by far the lowest one (32.5%). The decrease of this share over time is quite impressive in the first three fields, whereas the decrease in physics and above all in social sciences is less dramatic. The second research question we have addressed in the methodological section focuses on the possible interdependence between self-citations and foreign citations. Although this seems to be unlikely – the number of citations papers receive from their authors themselves might theoretically absolutely independent from the number of citations they receive from others. That is, self-citations and foreign citations might theoretically be quite different phenomena. If this were the case, self-citations should indeed be excluded from citation statistics. In order to be able to decide upon the hypotheses of possible independence of the two variables, a linear regression analysis will be used to check the hypothesis of independence, namely $H_0 : P(\xi(t) = i, \zeta(t) = k) = P(\xi(t) = i)P(\zeta(t) = k)$ for all pairs i and $k \geq 0$. We know that the random variable

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

has a Student distribution with parameter $n - 2$, where n is the sample size, i.e., the number of publications and r is the correlation coefficient. For the publication year 1992 and a 3-year citation windows (1992–1994) we have obtained $\xi(3) = 0.112 \cdot \zeta(3) + 0.680$ with $r^2 = 0.202$ and $n = 657,312$. Thus we have $t = 407.50$; this value is above the critical value at any reasonable confidence level. The same applies to the power-function model: $\xi(3) = 2.152\zeta(3)^{0.375}$ with $r^2 = 0.208$. In both cases we have to reject H_0, that is, self-citations and foreign citations cannot be considered independent for a 3-year citation window. Since citation processes are not homogenous, the rejection of independence can thus be generalised to larger windows. On the other hand, the correlation is quite weak ($r \approx 0.45$), so that neither the linear nor the power-function model can be accepted. In other words, individual self-citation rates cannot explicitly be expressed with the help of foreign citations alone.

In order to gain a deeper insight in the possible inter-dependence of self-citations and foreign citations, the conditional expectations $E(\xi(t)|\zeta(t) = k)$ have been calculated for all fields combined and for $t = 0, 1, \ldots, 10$. The condition $E(\xi(t)|\zeta(t)) \equiv E\,\xi(t)$ is necessary but not sufficient for independence. On the basis of the first regression analysis, we expect, of course, that this condition is not met for either citation window. Figure 3 presents the plot of foreign citations vs. mean self-citation rate for three selected citation windows, particularly 1992, 1992–1994 and 1992–2001 for all fields combined. The plots clearly reveal two basic properties.

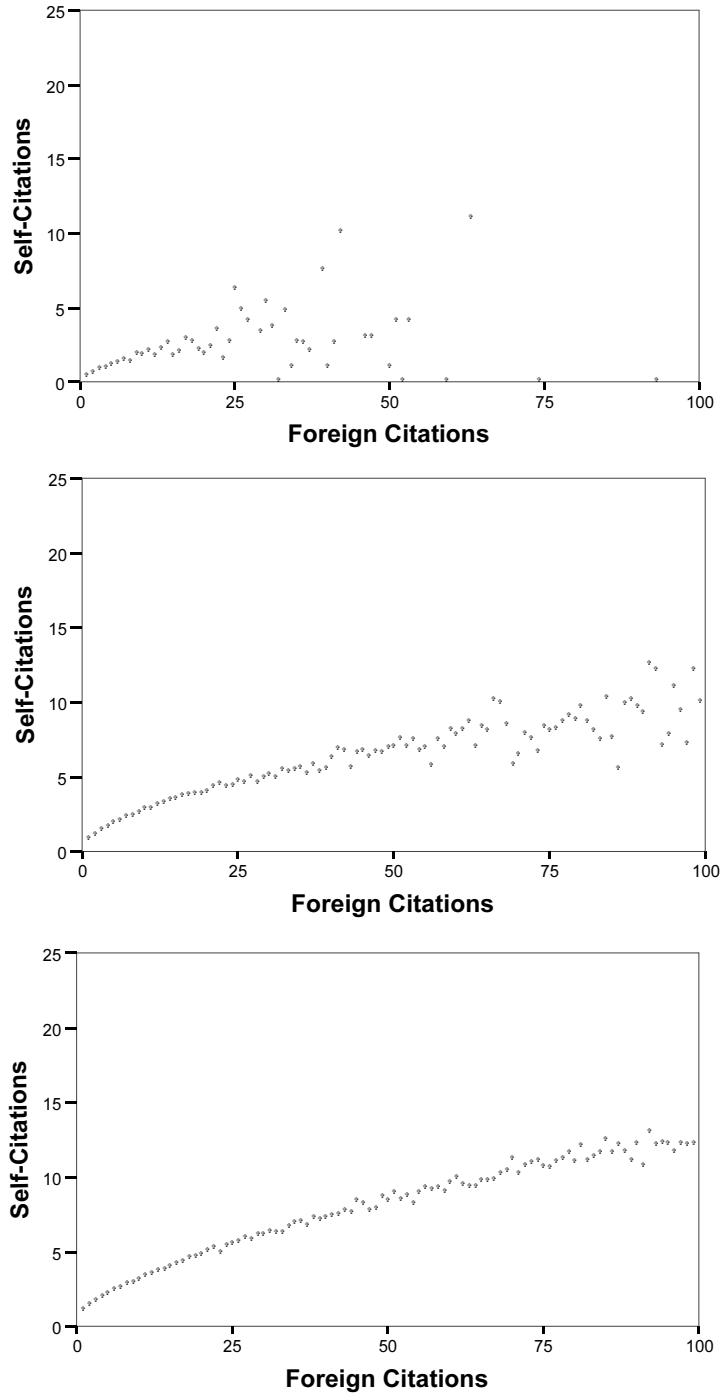

FIGURE 3. The plot of foreign citations vs. mean self-citation rate for three selected citation windows with at most 100 foreign citations for all fields combined (top = 1992, 1992–1994 and bottom = 1992–2001)

TABLE 1. Conditional self-citation means for papers published in 1992 with at most 50 foreign citations for 1-year to 10-year citation windows (k: number of foreign citations, $\bar{f}(k)$: estimated expectation)

k	1	2	3	4	5	6	7	8	9	10	$\bar{f}(k)$
0	0.10	0.29	0.39	0.44	0.47	0.48	0.49	0.49	0.49	0.48	0.47
1	0.35	0.62	0.79	0.89	0.95	0.98	1.00	1.01	1.02	1.02	1.13
2	0.57	0.90	1.10	1.20	1.27	1.31	1.35	1.35	1.37	1.38	1.56
3	0.80	1.15	1.37	1.48	1.55	1.60	1.64	1.65	1.64	1.66	1.91
4	0.93	1.36	1.60	1.74	1.82	1.85	1.88	1.90	1.90	1.91	2.22
5	1.10	1.58	1.83	1.97	2.05	2.09	2.10	2.14	2.16	2.14	2.49
6	1.23	1.70	2.04	2.22	2.26	2.30	2.33	2.35	2.37	2.36	2.74
7	1.39	1.92	2.25	2.39	2.48	2.53	2.57	2.54	2.54	2.56	2.97
8	1.34	2.05	2.34	2.56	2.69	2.74	2.73	2.78	2.77	2.77	3.19
9	1.84	2.17	2.56	2.76	2.92	2.89	2.95	2.95	2.94	2.92	3.40
10	1.76	2.41	2.77	2.99	3.06	3.17	3.15	3.15	3.16	3.10	3.60
11	2.01	2.51	2.79	3.12	3.22	3.34	3.31	3.29	3.28	3.33	3.79
12	1.64	2.60	3.02	3.25	3.37	3.44	3.48	3.48	3.48	3.49	3.97
13	2.19	2.52	3.24	3.40	3.57	3.60	3.67	3.61	3.67	3.63	4.14
14	2.52	2.93	3.38	3.68	3.62	3.72	3.87	3.86	3.76	3.73	4.31
15	1.63	2.96	3.46	3.73	3.97	3.93	4.00	4.06	3.90	3.92	4.48
16	1.91	3.13	3.71	3.88	4.19	4.20	4.16	4.07	4.14	4.13	4.63
17	2.78	2.98	3.74	4.07	4.20	4.31	4.27	4.36	4.33	4.25	4.79
18	2.60	3.10	3.87	4.01	4.29	4.42	4.44	4.51	4.54	4.56	4.94
19	2.10	3.35	3.83	4.42	4.33	4.54	4.68	4.67	4.61	4.62	5.09
20	1.86	3.38	3.90	4.54	4.66	4.64	4.72	4.77	4.79	4.74	5.23
21	2.25	3.67	4.28	4.42	4.85	4.88	4.86	4.89	4.88	4.93	5.37
22	3.38	3.38	4.44	4.68	4.65	4.90	5.11	5.00	5.01	5.22	5.51
23	1.50	3.68	4.31	4.85	5.05	5.14	5.17	5.16	5.26	4.90	5.64
24	2.60	4.15	4.38	4.91	5.25	5.21	5.25	5.32	5.34	5.34	5.78
25	6.17	3.91	4.68	5.03	5.16	5.46	5.53	5.45	5.43	5.46	5.91
26	4.83	3.65	4.52	5.24	5.59	5.59	5.61	5.66	5.65	5.57	6.03
27	4.00	3.90	4.98	5.48	5.44	5.56	5.72	5.63	5.77	5.85	6.16
28	—	3.40	4.57	5.27	5.38	5.99	5.74	6.02	5.81	5.78	6.28
29	3.33	4.23	4.88	5.49	5.90	5.75	5.88	6.09	6.11	6.11	6.40
30	5.33	4.69	5.03	5.42	5.85	6.04	5.99	6.08	6.10	6.08	6.52
31	3.67	4.59	4.85	5.76	6.07	6.01	6.31	5.99	5.87	6.24	6.64
32	0.00	4.68	5.39	5.99	5.97	6.30	6.25	6.23	6.47	6.14	6.76
33	4.67	4.16	5.36	6.15	5.93	6.46	6.46	6.70	6.55	6.22	6.87
34	1.00	4.92	5.36	5.68	6.58	6.55	6.63	6.53	6.45	6.61	6.98
35	2.67	5.56	5.60	5.93	6.88	6.54	6.89	7.00	6.68	6.89	7.09
36	2.50	4.21	5.18	6.15	6.82	6.55	6.89	6.76	6.84	6.96	7.20
37	2.00	5.02	5.73	6.11	6.72	6.78	6.82	6.89	7.06	6.71	7.31
38	—	7.00	5.32	6.62	6.67	6.67	6.65	7.04	6.67	7.19	7.42
39	7.50	4.75	5.45	6.76	6.61	6.85	7.22	6.82	7.24	7.07	7.53
40	1.00	5.45	6.18	6.18	6.98	7.12	7.27	7.34	7.35	7.19	7.63
41	2.50	4.93	6.83	6.59	6.90	7.50	7.03	7.44	7.71	7.32	7.74
42	10.00	4.91	6.67	7.00	6.65	7.58	7.30	7.47	7.62	7.37	7.84
43	—	4.71	5.59	6.77	7.35	7.35	7.44	7.82	7.53	7.62	7.94
44	—	6.27	6.52	6.51	6.85	7.60	7.55	7.50	7.92	7.59	8.04
45	—	4.72	6.69	7.35	7.39	7.62	8.21	7.75	7.31	8.31	8.14
46	3.00	5.55	6.29	7.00	7.90	8.42	8.06	8.23	8.24	8.14	8.24
47	3.00	4.64	6.65	7.50	8.17	7.69	7.78	7.78	8.06	7.66	8.34
48	—	4.93	6.53	7.22	7.33	7.63	7.88	8.19	8.13	7.87	8.43
49	—	4.68	6.89	7.60	7.41	7.74	8.70	8.46	8.52	8.61	8.53
50	1.00	5.07	6.98	7.19	7.83	8.19	8.43	8.43	8.18	8.32	8.62

TABLE 2. Statistics on the relationship between self-citation means and foreign citations for papers published in chemistry and mathematics in 1992 and 1999

Field	PY: 1992		PY: 1999	
	β	r	β	r
Chemistry	0.488	0.995	0.470	0.986
Mathematics	0.266	0.915	0.327	0.932

There is a strong interdependence between the two variables $E(\xi(t)|\zeta(t))$ and $\zeta(t)$. The second property reflects the increasing variance that is a consequence of the decreasing number of underlying publications for growing number of foreign citations. On the basis of the plots the following hypotheses has been tested. We have assumed that there exists an appropriate real function f, so that $E(\xi(t)|\zeta(t)) = f(\zeta(t))$, where we have chosen $f(k) = C\cdot(k+d)^{\beta}$ with C, d and β being appropriate positive real parameters. Since the distribution of foreign citations and the conditional expectations are changing over time we have to assume that one or more of the three parameters might be time-dependent.

Table 1 presents the empirical values of the conditional expectations of self-citations for papers with at most 50 foreign citations for all citation windows beginning with 1992–1992 up to 1992–2001. The process described by these expectations reaches a stationary limiting stage already several years after publication. Table 1 shows that the expectations do not essentially change for citations windows larger than 3 or 4 years. Therefore, a regression analysis has been applied to the stationary conditional means under the assumption that $C \approx 1$. In order to eliminate distortions caused by the extreme variation of papers with extremely high foreign citation rates, the upper 0.1% of papers representing the high end of the citation distribution has been omitted. The regression proved to be relatively insensitive to changes of the parameter d. In particular, for the parameter pair $C = 1$ and $d = 0.25$, the estimate $\overline{\beta} = 0.547$ with standard deviation 0.003 has been found. The correlation coefficient of $r = 0.996$ was quite high. The parameter $\overline{\beta}$ has been rounded to the value 0.55. The estimated $\overline{E}\,\xi(t)$ values on the basis of this parameter triple are given in the last column of Table 1. They do not guarantee a perfect fit, but $\overline{E}(\xi|\zeta)$ provides a quite good approximation to all citation windows basing on at least 3 years observation beginning with the year of publication. Even further simplifications such as $\overline{E}(\xi|\zeta) \approx \sqrt{\zeta + ¼}$ could be used as a rule of thumb for the stationary case for all fields combined.

The breakdown by subject fields reveals further interesting properties. As expected, the relationship between foreign and self-citations is dependent of the field, but the parameter estimates proved to be quite stable if the publication year and the appropriate citation window is shifted by a considerable period. The following statistics given in Table 2 that are based on 3-year citation windows may serve just as examples to illustrate this phenomenon.

The third and last part of the analysis is concerned with national (self-)citation patterns. This analysis is based on papers published in the sciences in 1999. Citations have been counted for the 3-year (self-)citation window 1999–2001. Besides the share of self-citations in all citation, two national standard indicators have been used, particularly, the Mean Expected Citation Rate (MECR) and the Relative Citation Rate (RCR) (see, for instance, Braun et al., 1985). MECR is

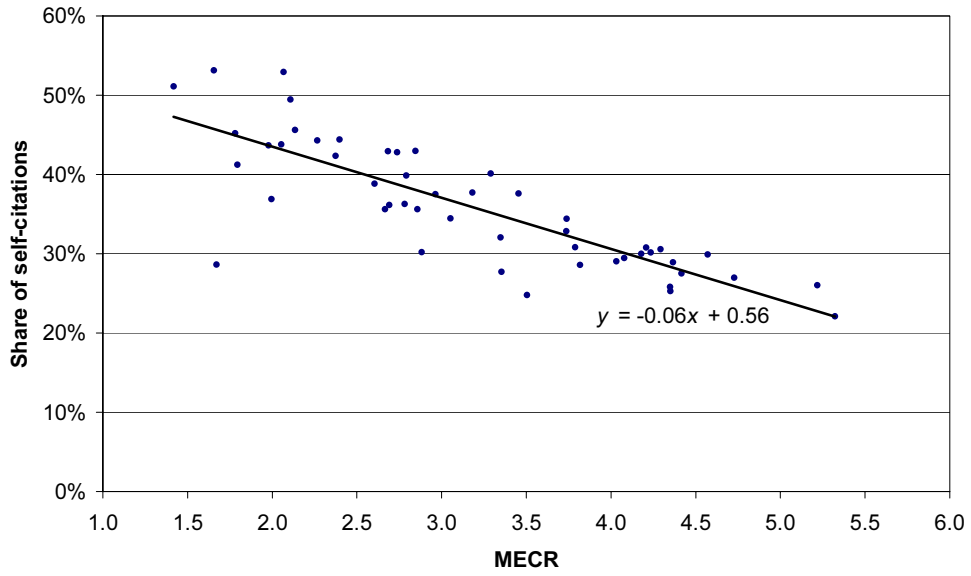

FIGURE 4. Plot of MECR vs. share of self-citations in all citations for the 50 most active countries in all fields combined (publications in 1999, citations between 1999–2001)

a journal-based indicator that expresses the expected citation rate of a given paper set. RCR compares the observed citation rate with the expected one in the same citation windows. RCR = 0 corresponds to uncitedness, RCR < 1 means lower-than-average, RCR > 1 higher-than-average citation rate, RCR = 1 if the set of papers in question attracts just the number of citations expected on the basis of the average citation rate of the publishing journals.

Table 3 presents the share of self-citations and the RCR indicator values for the 50 most active countries in all fields combined in 1999. The data are ranked in descending order by the share of self-citations. Although we have rather expected to detect biases by language-specific counting errors, that is, we have, for instance, expected to find countries with frequent author homonyms, at the top, and to find those the language of which might cause spelling variances or errors due to transliteration, at the bottom. The data, however, reveal rather negative correlation with the relative citation rate. Thus we find highly developed countries at the bottom of the list disregarding of the language spoken in these countries. The only exception here was Saudi Arabia the RCR value of which was quite low. International collaboration and mobility of scientists might, of course, blur language-specific peculiarities.

The correlation analysis of self-citation share vs. MECR makes the above-mentioned effect even clearer. The correlation coefficient r for the national MECR and $\overline{S}_3$ values amounts to 0.824. The interpretation might almost sound like a common-place: authors of those countries that are publishing on the average in journals with relatively low impact are more frequently cited by themselves than by others, or, in other words, low visibility increases the probability of self-citation.

TABLE 3. Publication count, share of self-citations and the RCR for the 50 most active countries in 1999 (ranked by self-citation share in descending order)

Country	Publication count	Share of self-citations	RCR
UKRAINE	4362	53.14%	0.77
IRAN	1108	52.92%	0.97
BELARUS	1241	51.10%	0.82
ROMANIA	1677	49.46%	0.89
CHINA	23843	45.62%	0.87
RUSSIA	27257	45.20%	0.96
BULGARIA	1624	44.43%	0.75
SLOVAKIA	1963	44.29%	0.96
INDIA	18080	43.79%	0.72
YUGOSLAVIA	1115	43.67%	0.65
CZECH REPUBLIC	4073	42.99%	0.95
POLAND	9480	42.91%	0.94
SLOVENIA	1274	42.81%	1.03
CROATIA	1137	42.35%	0.83
EGYPT	2328	41.26%	0.69
PORTUGAL	3034	40.12%	0.99
BRAZIL	10146	39.89%	0.80
HONG KONG	3520	38.84%	0.97
ARGENTINA	4216	37.72%	0.80
HUNGARY	4093	37.59%	0.90
MEXICO	4771	37.51%	0.78
TURKEY	5553	36.91%	0.72
SOUTH KOREA	12169	36.29%	0.90
SINGAPORE	3215	36.16%	0.93
TAIWAN	9421	35.63%	0.81
SOUTH AFRICA	3840	35.60%	0.94
GREECE	4652	34.46%	0.87
SPAIN	22801	34.41%	0.95
JAPAN	73641	32.86%	0.97
CHILE	1829	32.08%	0.97
NORWAY	4869	30.83%	1.09
ITALY	32104	30.77%	1.02
FINLAND	7247	30.57%	1.11
THAILAND	1117	30.19%	0.94
BELGIUM	10290	30.15%	1.12
GERMANY	67841	30.01%	1.11
DENMARK	7767	29.92%	1.17
AUSTRIA	7198	29.46%	1.09
FRANCE	50025	29.05%	1.03
SWEDEN	15377	28.94%	1.12
SAUDI ARABIA	1525	28.62%	0.70
AUSTRALIA	21730	28.60%	1.04
NEW ZEALAND	4289	27.74%	1.04
ISRAEL	9254	27.51%	0.98
NETHERLANDS	18975	26.99%	1.17
SWITZERLAND	14380	26.02%	1.21
CANADA	33714	25.84%	1.08
UNITED KINGDOM	72661	25.29%	1.09
IRELAND	2645	24.79%	1.22
USA	252150	22.10%	1.09

6 CONCLUSIONS

The large-scale analysis of author self-citations gives interesting insight into the mechanism of scientific communication. Although self-citations indicators are somewhat biased by errors of author identification, self-citation based indicators are valuable supplementary measures that can be used both in informetrics and research evaluation. Because of the already mentioned restriction concerning their reliability, self-citation indicators should be used in addition to traditional citation indicators, but not replace them.

The first important result characterises to the relationship between self-citations and foreign citations. The conditional expectation of self-citations for given number of foreign citation could be characterised as a square-root law. Even more important than finding a mathematical formulation was to show that there is from the statistical viewpoint nothing arbitrary in self-citations. Self-citations proved – at least at higher levels of aggregation – an essential part of scientific communication, indeed. The influence and weight of self-citations decreases rapidly. In the third after the year of publication, the expected number of self-citation for a given number of foreign citations becomes practically stationary. The results show once again that a citation window not smaller than three years but not larger than four years is sufficient for reliable bibliometric analyses since the share of self-citations is for such citation windows within acceptable limits ($\approx 25\%$ in the life sciences and $\approx 30\% - 40\%$ in the natural and engineering sciences). This choice also makes sure that literature still being recent is analysed.

The most striking observation was related to the relationship between self-citation shares and bibliometric standard indicators. The fact that low visibility goes with high self-citation shares seems, however, to be plausible.

The rules derived from the analysis have several implications for research evaluation; they can, for instance, be used to develop field-specific expected self-citation rates and shares within the framework of the evaluation of research performance in research groups and institutions.

REFERENCES

Aksnes D.W. (2003), A macro study of self-citation. *Scientometrics*, 56 (2), 235–246.

Braun, T., Glänzel, W., Schubert, A. (1985), *Scientometric indicators. A 32 country comparison of publication productivity and citation impact*. World Scientific, Singapore ∗ Philadelphia.

Burrell, Q.L. (2002), Modelling citation age data: Simple graphical methods from reliability theory. *Scientometrics*, 55 (2), 273–285.

Glänzel, W., Schoepflin, U. (1994), A stochastic model for the ageing analyses of scientific literature. *Scientometrics*, 30 (1), 49–64.

Glänzel, W., Schubert, A. (2003), A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, 56 (3), 357–367.

MacRoberts, M.H., MacRoberts. B.R. (1989), Problems of citation analysis: A critical review. *JASIS*, 40 (5), 342–349

Snyder, H., Bonzi, S. (1998), Patterns of self-citation across disciplines. *Journal of Information Science*, 24. 431–435.