The Discovery of Discoveries
Exploring the Dissemination of Major Findings in the Life Sciences

Hampus Rabow

**Hampus.Rabow@lub.lu.se*
Lund University, Lund University Libraries, Head Office, Box 134, SE-221 00 Lund (Sweden)

Abstract
The citations of highly cited papers in the life sciences published in 1986 and 2000 was examined to determine the extent to which early impact predicts long term impact. Linear correlation between early and post-peak impact was found to be about 0.7

Introduction and purpose
This study is based on research that explores the characteristics of highly cited scientific papers and the ageing of scientific information. We have examined the most cited papers in the life sciences to see when, how and why they were cited. This is work in progress and only some preliminary results will be presented here.

The main purpose at this stage has been to investigate when different types of papers are discovered, reach the peak citation rate and become obsolete. Another questions was the existence of 'sleeping beauties' (van Raan, 2004) and their frequency. We also examined highly cited papers that quickly became obsolete.
More generally, we tried to see how soon it becomes possible to predict the total number of citations for a paper. The correlations between earlier and later citations were recorded at various stages. Simonton's (1997) model of creative productivity could preclude a strong correlation between very early and long term impact of a publication. (See Lee & al., 2003 and the discussion below.)

We have also used this study as a pilot to work out a tentative classification of citations that will be used in our future research.

Finally we wanted to check whether the highly cited papers in 2000 were "discovered" more quickly than the 1986 papers.

Methods and data
All papers are from the life sciences, taken in a broad sense, and have been published in journals indexed by ISI™.
Set 1 consists of the union of the 50 papers published in 1986 that were most cited after 32 months and the 50 articles most cited after 72 months (n=65). Set 2 similarly consists of the year 2000 papers most cited after 32 and 48 months (n=58). Set 3 consists of the 1986 papers that had received more than 500 citations by 1998[1] (n=309). (2 from set 1 not included). Set 4 consists of the year 2000 papers that had more than 320 citations by 2004 (n=244). (Includes set 2.) For an illustration of how these papers compare to "normal papers", see figure 1.

[1] In what follows "by XXXX", where XXXX is a year, should be read as "by the end of XXXX".

Proceedings of ISSI 2005, pages 322–326.
Edited by P. Ingwersen and B. Larsen.

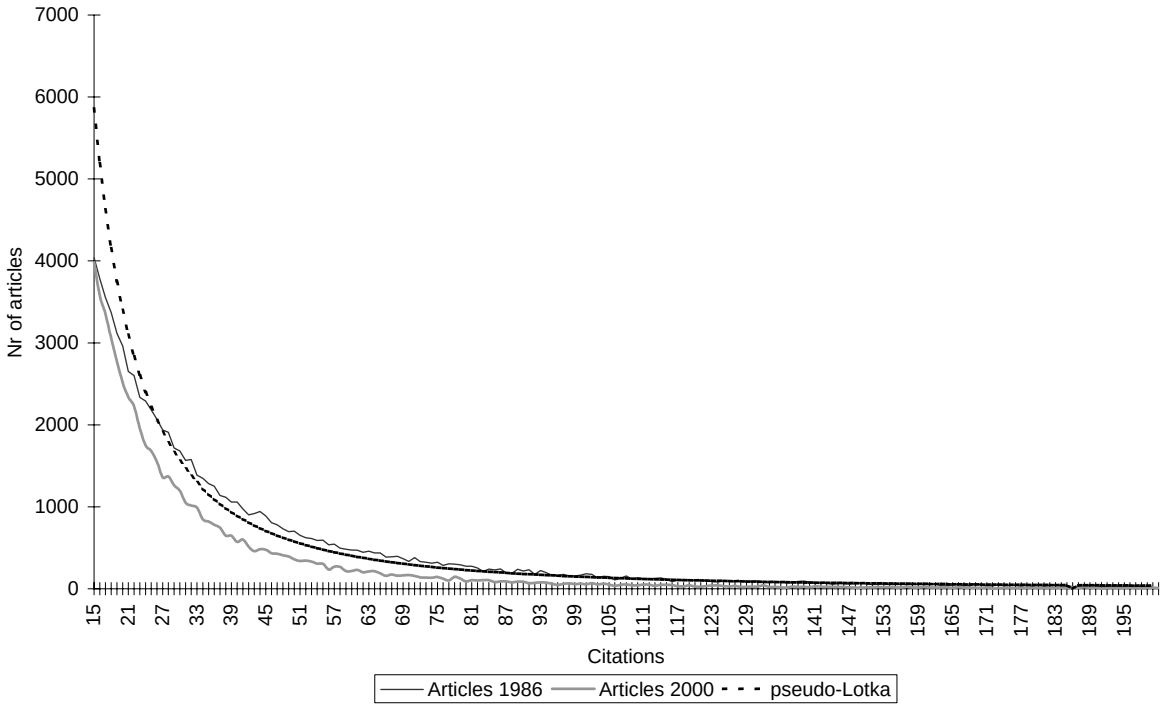

Figure 1. Samples of 300000 articles published 1986 and 2000. The thin black (1986) and thick grey (2000) lines show the number of articles cited n times.
The dotted line shows a modified Lotka's law for the citation of articles. The number of articles cited n times is approximately proportional to $1/(n+1)^2$. For smaller numbers of citations, the relationship is approximately proportional to $1/(n+1)$.

Results

The citation intensity for set 1 peaked 3 years after publication and remained fairly constant for about a year, after which a slow gradual decline began. In the year 1998, the papers still received on average 74 citations. Set 2 appears similar, but it is too early to say definitely.

No papers were found with a very high initial citation intensity and premature obsolescence. All papers continued to receive citations well above average for a number of years. This was expected, since an earlier study has shown that even retracted articles continue to be cited (Rabow & Rabow, 1992). The most significant drop was found in a paper entitled "Human Immunodeficiency Viruses" (Coffin & al.), published in *Science* in May 1986. It had received 173 citations by the end of 1988, and by 1998 it had received only 92 additional citations. See figure 2, where this paper is third from the bottom. The reason for the swift decline is fairly obvious. The paper, by an official committee, established the nomenclature for the HIV virus. Once the nomenclature had become established, the paper became obsolete.

Of the most cited papers in 1987, many of those whose citation rate declined quickly were about Aids or HIV, which was the single "hottest" topic in 1986 (Garfield, 1988). It seems that the topic is usually more significant in shaping the citation curve than the characteristics of the individual article. This is illustrated in figure 3, where the citation patterns for two highly cited Aids reviews are compared.

The presence or non-presence of "sleeping beauties" will partly depend on the definition. It is difficult to use a general definition that is independent of the topic and type of article. The ideal type would be a paper that received only a few citations the first two or more years and then received a high number of citations for an extended period. However, some papers, notably papers with a considerable mathematical content, would possibly require a longer "sleep time" to classify. The least equivocal "sleeping beauty" is probably a paper about preconditioning with ischemia (Murry & al., 1986). It

was published in *Circulation* in November 1986. By 1988 it had only been cited 7 times, of which 1 was a self-citation. 4 years from publication, when most articles citation rate had peaked, this paper had only received 51 citations. Yet by 1998 it had received more than 1000 citations, and it is currently the 11th most cited of the 1986 life science papers[2]. Another interesting example, although not receiving the same amount of citations, is the paper "Tumor Chemosensitivity Conferred by Inserted Herpes Thymidine Kinase Genes…" (Moolten, 1986). By 1992 it had received only 11 citations. 6 of those were self-citations. It appears to have been "discovered" in June 1992, and it received the most citations (81) in the year 2000. Generally speaking, however, these kinds of papers appear to be very rare. Of the papers that had received more than 500 citations by 1998 (set 3), slightly less than 3% had less than 20 citations by 1988. No clear cases were found in set 4.

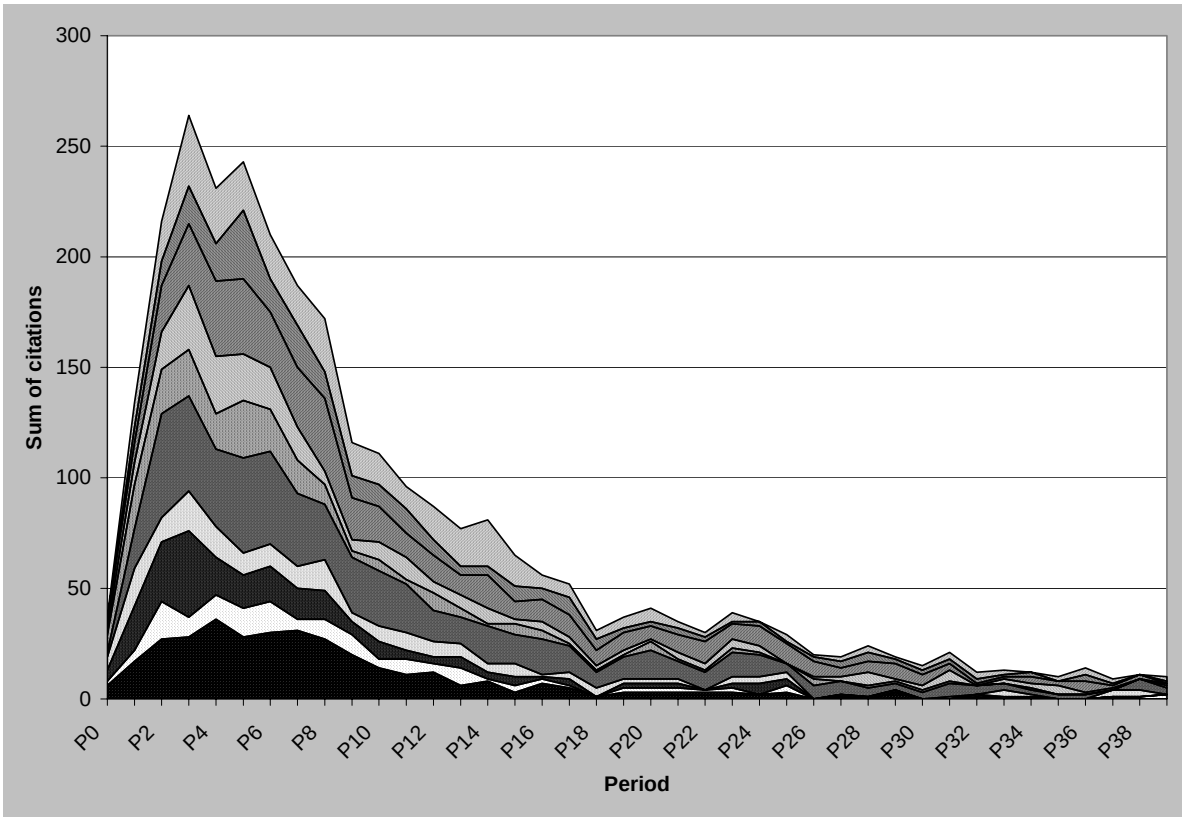

Figure 2. Drop in citation rate for 10 articles that were among the most cited the year after publication. The paper with the sharpest decline is third from the bottom. 1 period is 4 months

Table 1. Citations after 12 and 16 months have been correlated (Pearson) with citations after 48 months. For set 1 also 12, 16, 32 with 72 m. After 72 months only a few of the papers continue to receive citations at very high levels. These papers are often "slow starters".

Cited pubs	**12 → 48 m**	**16 → 48 m**	**12 → 72 m**	**16 → 72 m**	**32 → 72 m**
Set 1 r=	0.56	0.68	0.32	0.45	0.85
Set 2 r=	0.61	0.71			

[2] Possibly 12th if a paper in biostatistics is seen as a "life science" paper.

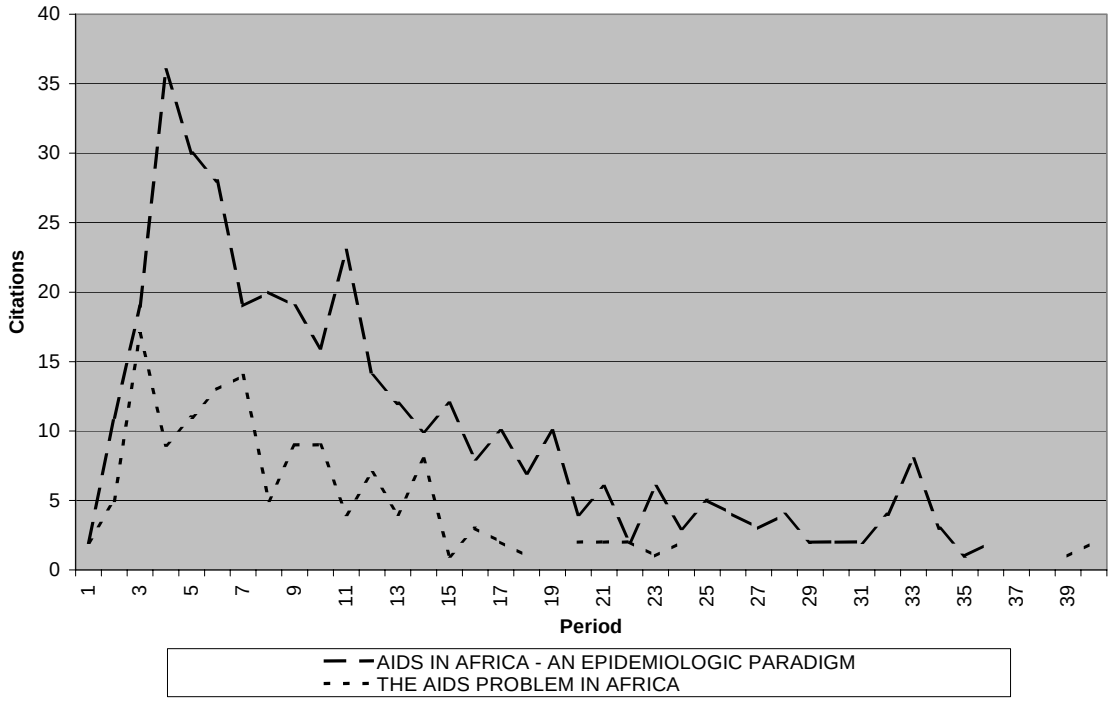

Figure 3: Comparison of 2 Aids reviews. *Aids in Africa* was published in *Science* in November 1986 . (Quinn & al.) *The Aids Problem in Africa* was published in *Lancet* in January 1986 (Biggar). A possible difference is the spike in period 4 (12-15 months after publication) for the *Science* article. Perhaps the later article "stole" some of the citations that the earlier article would otherwise have received.

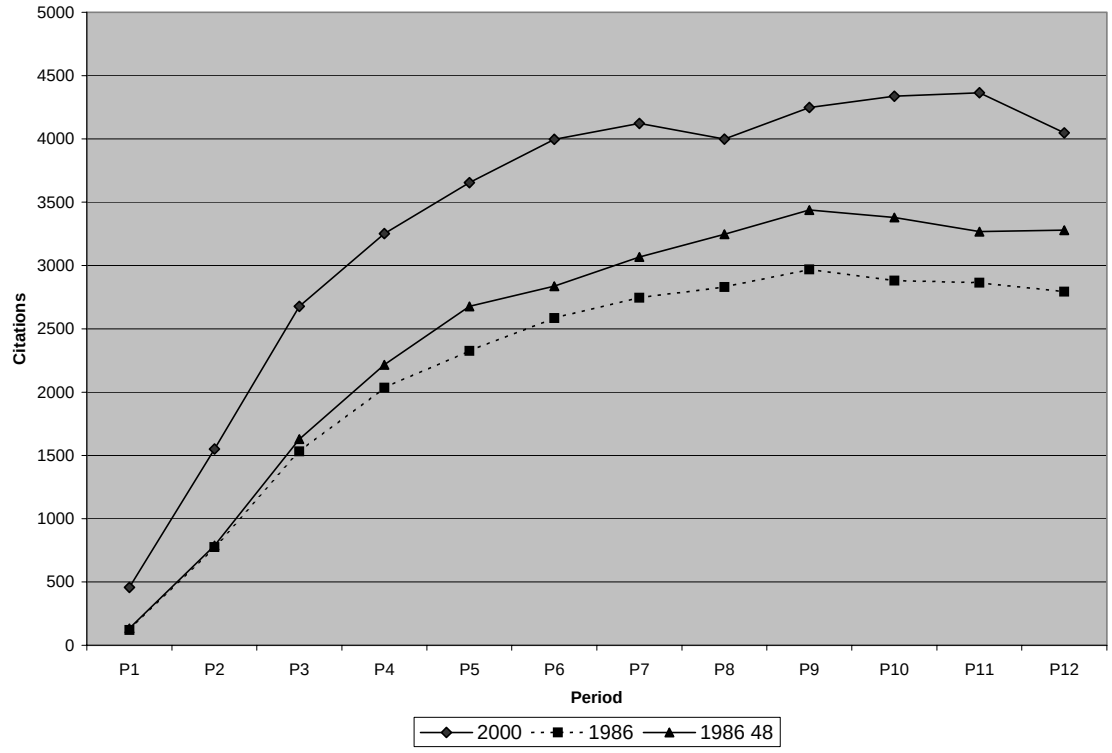

Figure 4. Articles from 2000 (set 2) are cited quicker than those from 1986 (set 1). To see that the different time frames don't explain the difference, the most cited 1986 articles after 48 months have also been plotted, but with citations "normalized" by multiplication by 2.

Discussion

Almost all of the papers were discovered within 20 months. The peak was reached after 3 years for the 1986 papers and probably for the 2000 papers as well. The citation impact after 16 months was strongly correlated (r≈0.7) with citation impact after 48 months (past peak for most articles). A single linear correlation is a rather crude measure of predictability because of the complexity of citation curves. It is unclear if this is enough to contradict Simonton (1997) To achieve better prediction rates it is necessary to separate different types of articles and categories of citations. A problem with this study is that the papers in the samples are atypical. Highly cited papers are known to maintain a more stable citation rate than 'normal' papers (Price, 1976). These issues will be discussed in a future paper.

References

Biggar, R. J. (1986). "The Aids Problem in Africa." *Lancet* 1(8472): 79-83.

Coffin, J., A. Haase, et al. (1986). "Human Immunodeficiency Viruses." *Science* 232(4751): 697-697.

Egghe, L. & Rousseau, R. (2000). Ageing, Obsolescence, Impact, Growth, and Utilization: Definitions and Relations. *Journal of the American Society for Information Science*, 51, 1004-1017.

Garfield, E. (1988). The Most-Cited 1986 Life-sciences Articles Highlight Cell-Surface Receptors, Tumor Necrosis Factor, and AIDS Research. *Current Contents*, 50, 3-16

Glänzel, W. & Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature. *Journal of Information Science*, 21, 37-53.

Griffith, B., Servi, P., Anker, A.L. & Drott, M. (1979). The aging of scientific literature: A citation analysis. *Journal of Documentation*, 26, 46-52.

Lee, J. D., Vicente, K. J., Cassano, A. & Shearer, A. (2003). Can scientific impact be judged prospectively?A bibliometric test of Simonton's model of creative productivity. *Scientometrics*, 56, 223-233.

Line, M., Sandison, A. (1974). 'Obsolescence' and changes in the use of literature withtime. *Journal of Documentation*, 30, 283-350.

Moolten, F. L. (1986). "Tumor Chemosensitivity Conferred by Inserted Herpes Thymidine Kinase Genes - Paradigm for a Prospective Cancer Control Strategy." *Cancer Research* 46(10): 5276-5281.

Murry, C. E., R. B. Jennings, et al. (1986). "Preconditioning with Ischemia - a Delay of Lethal Cell Injury in Ischemic Myocardium." *Circulation* 74(5): 1124-1136.

Oppenheim, C., Renn, S. (1978). Highly cited papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science*, 29, 225-231.

Price, D. J. de Solla (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27, 292-306.

Quinn, T. C., J. M. Mann, et al. (1986). "Aids in Africa - an Epidemiologic Paradigm." *Science* 234(4779): 955-963.

van Raan, A. F. (2004). Sleeping Beauties in science. *Scientometrics*, 59, 467-472.

Rabow H, Rabow I (1999) Retraction of articles from the perspective of peer review and quality control. *European Science Editing* 25 3-6

Simonton, D. K. (1997). Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review*, 104, 66-89.