Library and Information Services in Astronomy IV
July 2-5, 2002, Prague, Czech Republic
B. Corbin, E. Bryson, and M. Wolf (eds)

The NASA Astrophysics Data System: Obsolescence of Reads and Cites

Michael J. Kurtz, Guenther Eichhorn, Alberto Accomazzi, Carolyn S. Grant, Donna M. Thompson, Elizabeth H. Bohlen, and Stephen S. Murray

Harvard-Smithsonian Center for Astrophysics,
Cambridge, MA 02138 USA
kurtz@cfa.harvard.edu

Abstract.

The obsolescence of an article, how its use declines as it ages, has long been a central element of bibliometric studies. Normally this is determined using the citations to an article. We determine this function using the reads an article receives and then compare this with the function determined from a citation study. There are both similarities and differences. The similarities are strong enough that the normative theory of citations must be true in the mean.

1. Readership as a function of age

Because the use of the Astrophysics Data System (ADS) is now the dominant means by which astronomers access the technical literature the ADS usage logs can provide a uniquely powerful view of the way an entire discipline (astronomy) uses the technical literature. Here we will examine the obsolescence (e.g. White and McCain (1989), Line and Sandison (1975)) of the technical literature of astronomy as a function of article age based on the actual readership of an article. This is an extension and reexamination of the work done in Kurtz et al. (2000).

We use, as our basic data source, the log of all article "reads" using the ADS between January first and August 20th, 2001. We define a "read" as every time a user, who has access to a list of articles, their dates, journal names, titles and authors, chooses to view more information about an article. Currently 50% of these "reads" are of the abstract, 38% are of one of the forms of whole text, 8% are of the citation list, and the rest are distributed amongst the ten other options. There are more than 4.2 million "reads" in this log.

For this study we extracted those for any of the three major U.S. astronomy journals *The Astrophysical Journal, The Astronomical Journal,* and *The Publications of the Astronomical Society of the Pacific.* All three of these journals have been stable over the past century, are currently among the most important astronomy journals, and have had their full text versions beginning with their first issues available on-line through ADS since well before the beginning of the

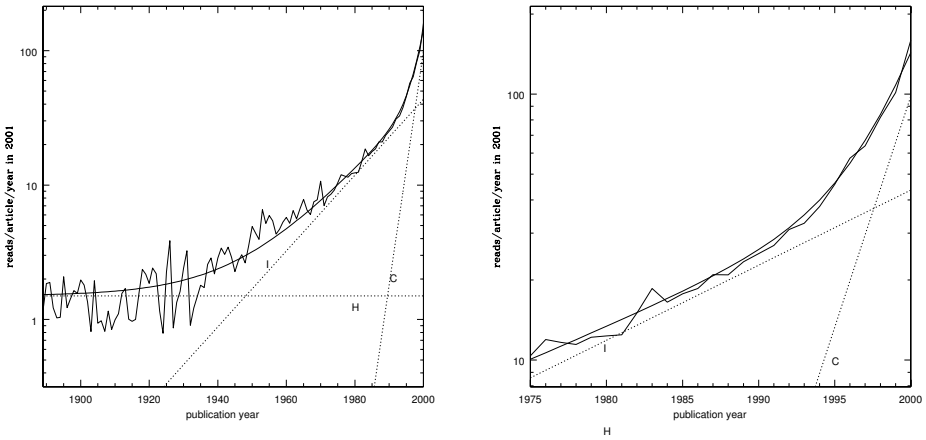

Figure 1. LEFT: The average number of reads per article per year for three U.S. astronomy journals. The thin line represents the actual data, the thick line is the model in the text, and the three dotted lines represent the three components of the model. RIGHT: An expanded view showing the most recent 25 years. Note that the model fits the actual data very well.

reporting period. These journals accounted for slightly more than 1.8 million reads in the first 7.66 months of 2001.

1.1. The obsolescence model for reads

Figure 1—LEFT shows the average number of reads per article per year for these three journals as a function of publication year. This shows more than a full century from the first issue of *The Publications of the Astronomical Society of the Pacific* in 1889 through 2000.

Figure 1—RIGHT shows an expanded view of the last 25 years of data from figure 1—LEFT. The dotted lines show the relevant three components of the four component readership model of Kurtz et al. (2000), as modified here. In this model research article readership (R) is parameterized by the sum of four exponential functions with very different time constants; we associate these four functions with four different modes of readership: Historical (R_H), Interesting (R_I), Current (R_C) and New (R_N). The New (R_N) mode, which corresponds to the newly arrived (either on-line or in the mail) issue, cannot be parameterized by the data in figures 1—LEFT and 1—RIGHT. The Historical (R_H) mode we actually parameterize as a constant, H_0. We leave the exponential form in equation C (with $k_H = 0$) because other combinations of multiplicative and time constants can also be found which fit the data well, including some combinations with $k_H \neq 0$.

$$R = R_H + R_I + R_C + R_N \qquad (C)$$

where

$$R_H = H_0 e^{-k_H T}$$

$$R_I = I_0 e^{-k_I T}$$
$$R_C = C_0 e^{-k_C T}$$
$$R_N = N_0 e^{-k_N T}$$

and

$$H_0 = 1.5; k_H = 0$$
$$I_0 = 45; k_I = 0.065$$
$$C_0 = 110; k_C = 0.4$$
$$N_0 = 1600; k_N = 16$$

The three longer term functions, R_H, R_I, and R_C are parameterized to fit the data shown in figures 1—LEFT and 1—RIGHT. The R_N function is included for completeness; k_N is taken from Kurtz et al. (2000). N_0 is obtained by assuming k_N is correct, and ascribing all readership of the *Astrophysical Journal* electronic edition which does not originate with ADS to the N mode. This is a very crude approximation, but the three component (R_H, R_I, and R_C) model for archival readership is not effected by the N mode usage, which fades very rapidly following publication.

The three mode model is not unique but does provide a very good fit to the existing data, as figures 1—LEFT and 1—RIGHT show. No model consisting of only two exponential functions can fit both the recent and historical data, as comparing the two figures makes clear.

Most studies of obsolescence find that the use of the literature declines exponentially with age, and parameterize this with a single number, often called the "half-life," which is related to the coefficient in the exponent by half-life = $log_e(2)/k$, the point where the use of an article drops to half the use of a newly published article. There are several other definitions of half-life in the literature, we use this one. Thus the Historical (R_H) component of equation C does not have a half-life; the half-life of the Interesting (R_I) component is 10.7 years; the half-life of the Current (R_C) component is 1.7 years. Kurtz et al. (2000) estimate the half-life of the New (R_N) component at 16 days.

Several studies (e.g. Egghe (1993) and references therein) decompose the exponential decay in use into the product of an intrinsic decay and the general growth of the literature. The results presented here are for the mean current use per article published as a function of time since the present, thus we measure directly the intrinsic decay. Kurtz et al. (2000) show the growth of the astronomy literature has been 3.7% per year, measured in terms of number of papers published over the past 22 years.

The total number of papers read over time in each mode is just the integral of the function from zero to infinity, which for a negative exponent is just the ratio of the two constants: $H_0/k_H = \infty$ reads (one and a half reads per year forever); $I_0/k_I = 818$ reads; $C_0/k_C = 275$ reads; $N_0/k_N = 100$ reads. This assumes no growth in the number of reads, If the number of reads per year increases long term at the 3.7% at which the number of publications is now increasing the constants in the exponents would all be increased by 0.037; this would have very little effect on the integrals of the R_N and R_C functions, but would more than triple the articles read in the R_I mode; and the R_H mode would grow apace with the growth in the number of reads.

1.2. Discussion

Beginning with Burton and Kebler (1960) there have been a number of studies (see White and McCain (1989) for a review) which suggest that the obsolescence function consists of the sum of two exponentials, which Burton and Kebler (1960) attribute to "classic" and "ephemeral" papers; parameterizations (e.g. Price (1965)) tend to be similar to our $R_H + R_I$ functions.

If we ignore the R_N component, which neither this study, nor any of the other studies of obsolescence could see, we still very clearly find three separate components to the obsolescence function. Why have these three components not been seen til now?

We suggest that the data available to previous studies has not been adequate to see these subtle effects. Most studies have used citation data to determine the obsolescence function. Because it takes time after a paper is published for it to be cited (e.g. section 2) the peak in the R_C mode is obscured in citation data. Also citation studies have substantial problems accounting for the growth of the literature, which has not been at all constant over the past century. Related to this is the determination of the size of the sample universe (the number of relevant papers to the study) at past times.

There are certainly other possibilities, perhaps the obsolescence function is different for reads and cites, and perhaps the very existence of the ADS has changed the way the literature is used. We will explore these questions further in section 2.

The reason why readership studies have not seen the three component nature of archival readership which we see, we suggest, is that the data available in such studies has been too sparse. The largest astronomy library, the Center for Astrophysics Library, has a reshelve rate of about 1000/month (Coletti (1999)) , which is less than 0.2% of the rate of reads in ADS. Additionally many astronomers keep (and use) their own paper copies of recent journals, which would suppress the R_C mode in library use.

2. The relationship between reads and cites

Central to bibliometrics is the study of citations (Garfield (1979)), and central to the study of citations is the so called normative assumption (Liu (1993)) that "the number of times a document is cited ... reflects how much it has been used..." (White and McCain (1989)) . There have been many articles suggesting problems with citations studies (e.g. MacRoberts and MacRoberts (1989)), and many articles defending them (e.g. Small (1987)). White (2001) and Phelan (1999) discuss these issues.

The readership data discussed in section 1 provide a totally independent, direct new measure of "how much (an article) is used." Comparing the readership statistics with citation measures will show exactly the similarities and differences between citations, which are an indirect measure of use, but, some would argue, a direct measure of usefulness and reads, which are a direct measure of use, but perhaps an indirect measure of usefulness. Here we expand considerably on the comparison presented in Kurtz et al. (2000).

2.1. The mean relationship between reads and cites

While there have been many dozens of studies on obsolescence using citations, and many dozen more using readership as determined by using library circulation statistics (see White and McCain (1989) for review), there are very few studies comparing the two methodologies over the same data. Tsay (1998) compared the readership obsolescence function (obtained by reshelving statistics) for a number of medical journals with the citation obsolescence function for the same journals. He found the half-life of the readership function was significantly shorter than the half-life of the citation function. Tsay (1998) reviewed the literature and found only one previous comparable study: Guitard, in 1985 (discussed in Line (1993)), using photocopy requests as the use proxy, found the citation half-life shorter than the readership half-life.

We have only found two other studies. Cooper and McGregor (1994), also using photocopy data, find citation half-life substantially longer than the use half life; also they find "no correlation between obsolescence measured by photocopy demand and obsolescence measured by citation frequency." Satariano (1978) used the questionnaire method to find "citation patterns reflect a cross-disciplinary focus that is not found in the journals most often read."

We believe Kurtz et al. (2000) contains the first study using a data-set large enough to show the similarities and differences between the two obsolescence functions. Here we use a substantially improved data-set; a complete discussion will be published in Kurtz et al. (2003).

Synchronous relation Kurtz et al. (2000) found that the instantaneous obsolescence function for articles from the recent technical astronomy literature as measured by citations is simply equal to a proportionality constant times the function measured by reads times an exponential ramp-up to account for the time delay from when an article is first published to when an article which cites that article is published:

$$C = cR(1 - e^{-k_D T}) \qquad (D)$$

where

$$c \approx 0.05; k_D = 0.7$$

The proportionality constant, c, represents the number of reads per citation. This changes with the (always incomplete) citation databases and with time, as the ADS use increases. Currently we estimate that the average paper is read about twenty times using ADS for every time it is cited. In comparing the citation and reads obsolescence functions we have adjusted c to provide the best fit to the samples.

Figure 2—LEFT compares the reads and cites obsolescence functions for recent articles. The readership data is for articles from *The Astrophysical Journal, The Astronomical Journal, The Publications of the Astronomical Society of the Pacific, and The Monthly Notices of the Royal Astronomical Society* which were read between 1 January 2001 and 20 August 2001. The citation data are taken from those four journals and both *Astronomy & Astrophysics and Nature*, where the publication date was also between 1 January 2001 and 20 August 2001. Only citations to one of the four journals in the readership sample were taken; the data contain 45,000 citations.

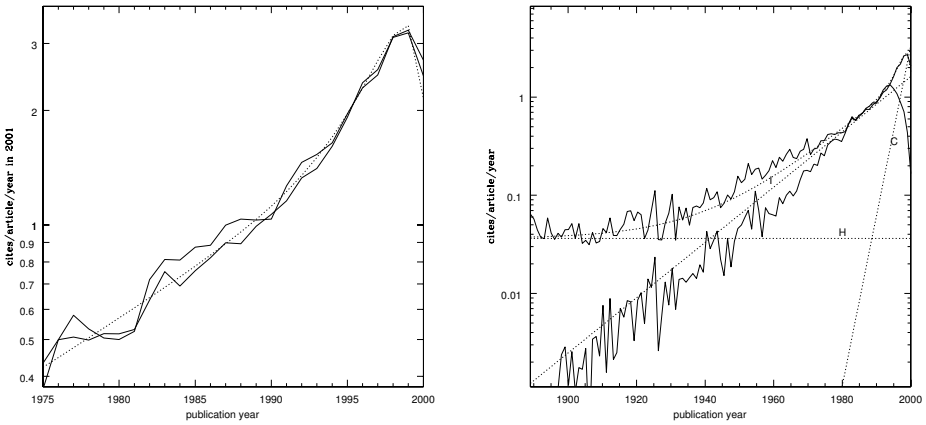

Figure 2. LEFT: A comparison of the $C = cR(1 - e^{-k_D T})$ model with the actual citations for papers from the 2001 sample for the most recent 25 years. The thick line shows the actual citations, the thin line is the model, using the actual reads, and the dotted line is the model using the reads model, equation C. RIGHT: A comparison of the $C = cR(1-e^{-k_D T})$ model with the actual citations for papers from the six year sample for the last 111 years. The thick line shows the actual citations, the thin line is the model, using the actual reads, and the dotted lines are the modified reads model from equation C and its components.

As can be seen from figure 2—LEFT the citation function follows the reads function very closely. In particular the R_C function clearly has an analog in the citation data; despite the suppression of the steep increase compared with the raw reads due to the exponential ramp-up. The change in slope in the citation function seen beginning about 1994 is exactly what is predicted from the reads function; the number of citations in 1998 and 1999 are more than 40% above that expected by an extrapolation of the exponential decay seen between 1975 and 1990, a decay which corresponds very closely to the R_I function. We suggest this shows that the citation derived obsolescence function has two components with exactly the same parameters as the two mid-range (in time) readership functions.

To examine the obsolescence function over a longer time period we use a different dataset of citations. We take all citations to the four journals in the readership sample from articles published between 1 January 1995 and 20 August 2001 in the ADS database. The data contain 625,000 citations.

We continue to use as our comparison the 2001 reads sample. Clearly papers published in 1995 could not have cited papers published in 2000, so comparison with recent obsolescence is impossible. This comparison is in figure 2—LEFT. We use these data exclusively to examine the long term behavior of the obsolescence function.

Figure 2—RIGHT shows the long term obsolescence function obtained from citation data compared with the readership function. They clearly are **not**

the same. The citation function follows the R_I function but not the $R_H + R_I$ function. This is not a statistical fluke based on having a small number of citations; the number of citations in the period from 1889 to 1940 which are "missing" from the citation function exceed 5,000. In the year 1900, for example, there were 18 citations in the six year sample, about 150 would be expected, were the R_H mode to produce citations at the same amplitude as the R_I mode.

We are therefore driven to the conclusion that:

$$C = c(R_C + R_I)(1 - e^{-k_D T}) \qquad (E)$$

where

$$c \approx 0.05; k_D = 0.7$$

for research articles in the astronomy literature.

There are a number of possible reasons for the citation obsolescence function to be different from the reads function. There are also a number of possible reasons why the citation obsolescence function measured here does not show the R_H component, whereas this component is seen in other citation based obsolescence functions, beginning with Price (1965). We see no clear candidate explanation which accounts for both differences, however.

Discussion We have shown that the citation rate as a function of time is equal to a constant times the sum of two modes of the readership function. There is no *a priori* reason why the constant c in equation E should not actually be a function of time; why should the number of citations per read (about 0.05) be constant, independent of the age of the article?

Examining figures 2—LEFT and 2—RIGHT we see that if c is a function of time it cannot change by more than about 1% per year. This is an extraordinary result, it says that within the (small) measurement error the C function and the $R_C + R_I$ function must be measuring exactly the same thing, the mean usefulness of journal articles as a function of time.

Because the private act of reading an article entails none of the various sociological influences as the public act of citing an article (Seglen (1997) lists several of these factors) this suggests that in the mean these factors do not influence the citation rate.

Unless the sum of all the various sociological influences as a function of time is exactly the same as the usefulness of articles as a function of time the existence of these influences would cause c not to be constant. That c is constant means that at every age the total effect of these various influences is zero.

We therefore assert that we have proven that the normative theory of citing (Liu (1993)) is true in the mean.

References

Burton, R.E. and Kebler, R.W. 1960, The Half-Life of some Scientific and Technical Literature, American Documentation, 11, 18.

Coletti, D.J. 1999, Report from the Librarian, Harvard-Smithsonian Center for Astrophysics, John G. Wolbach Library.

Cooper, M.D. and McGregor, G.F. 1994, Using Article Photocopy Data in Bibliographic Models for Journal Collection Management. Library Quarterly 64, 386.

Egghe, L. 1993, On the Influence of Growth on Obsolescence, Scientometrics 27, 195.

Garfield, E. 1979, Citation Indexing: Its Theory and Application in Science, Technology, and Humanities, New York: Wiley.

Kurtz, M. J., G. Eichhorn, A. Accomazzi, C. S. Grant, S. S. Murray, and J. M. Watson 2000. The NASA Astrophysics Data System: Overview. A&AS 143, 41-59.

Kurtz, M. J., G. Eichhorn, A. Accomazzi, C. S. Grant, M. Demleitner, and S. S. Murray 2003. The NASA Astrophysics Data System: Sociology, Bibliometrics, and Impact. Journal of the American Society for Information Science, to appear

Liu, M. 1993, Progress in Documentation The Complexities of Citation Practice: A Review of Citation Studies. Journal of Documentation 49, 340.

Line, M.B. 1993, Changes in the Use of Literature with Time—Obsolescence Revisited, Library Trends, 41, 665.

Line, M.B. and Sandison, A. 1975, Progress in Documentation - Obsolescence and Changes in Use of Literature with Time, Journal of Documentation 30, 283.

MacRoberts, M.H. and MacRoberts, B.R. 1989, Problems in Citation Analysis, Journal of the American Society for Information Science, 40, 342.

Phelan, T.J. 1999, A Compendium of Issues for Citation Analysis, Scientometrics, 45, 117.

Price, D.J. de Solla 1965, Networks of Scientific Papers, Science, 149, 510.

Satariano, W.A. 1978, Journal Usein Sociology: Citation Analysis versus Readership Patterns. Library Quarterly 48, 293.

Seglen, P. O. 1997, Citations and Journal Impact Factors: Questionable Indicators of Research Quality. Allergy 52, 1050.

Small, H.G. 1987, The Significance of Bibliographic References, Scientometrics, 12, 339.

Tsay M.-Y. 1998, Library Journal Use and Citation Half-Life in Medical Science. Journal of the American Society for Information Science 49, 1283.

White, H.D. 2001, Authors as Citers over Time. Journal of the American Society for Information Science, 52, 87.

White, H.D. and McCain, K.W. 1989, Bibliometrics, Annual Review of Information Science and Technology, 24, 119.