

GROWTH AND OBSOLESCENCE OF LITERATURE IN THEORETICAL POPULATION GENETICS

B. M. GUPTA

*Scientometrics and Informetrics Group, National Institute of Science, Technology,
and Development Studies, Dr. K. S. Krishnan Marg, New Delhi 110012 (India)*

(Received April 7, 1998)

Studies the relation between growth rates and obsolescence rates and half-life of theoretical population genetics literature. Explores the application of lognormal distribution in age distribution of citations over a period of time.

Introduction

Growth and aging (or obsolescence) are two topics which have attracted world-wide attention in the informetrics literature. Extensive reviews on obsolescence has been published by *Gapen* and *Milner*¹ and *Line*² in the past. Obsolescence is the phenomenon of "the reduced use or decline in the use of information (on a certain topic) with time". Here we have to clear whether the concern is with documents as such or with the information they contain representing knowledge. If the obsolescence of documents (decline in its use) with time is concerned, it is of practical interest in the probability that an item will be required as a guide to such questions as: (a) when to discard older volumes, (b) how long to keep new ones, (c) what sort of retrospective storage and access in information retrieval system should provide, and so on. If the concern is with knowledge, obsolescence may be defined as a decline over time in the validity or utility of information. Such decline may occur for several reasons according to *Line*:²⁻³ (1) the information is valid, but incorporated in later work; (2) the information is valid, but superseded by later work; (3) the information is valid, but is in a field of declining interest; (4) the information is no longer valid.

Evidence of the obsolescence of publications is presumed, if the use of these publications declines with age. Decline in the use literature over time (aging or decay) may be ascertained through studies of library use or by studying age of citations in publications or articles cited.

The bases for obsolescence studies is the citation habits of authors (in the forms of reference lists), in so far as their citations are in order (there are of course many misuses or drawbacks of citations).

Obsolescence implies a relation between use and time, but the effects of time are revealed in different ways. First is the influence of the time past, as expressed in the current ages of the item of literature when they are studied. Secondly, the passage of time present increase in the age of each item. Thirdly, there is the effect of time future which we try to forecast by extrapolations from the effects of time past and time present. These effects can be investigated in two main ways- synchronous and diachronous.

Synchronous studies are made on records of uses or references at one point of time and compares the uses against the age distribution of materials used or cited. In other words, one fixes the citing literature, and age of cited references are counted backwards. Half-life, annual ageing factors, utility factors, etc. are studied in synchronous approach.

Diachronous studies are related to the use of a particular items through successive observations at different dates. In lending libraries such studies can be carried out by analysing the previous use of books as recorded by the dates in data labels or in use records. In diachronous study, citations counts are made in forward directions. Very few studies exists on this area.

Rate of obsolescence can be plotted as a curve. Alternately some numerical measures can be used. For synchronous studies this may be the "half-life", the period of time needed to account for one half of all the citations received by a group of publications. For diachronous studies, the comparable measure is the medium citation age. The obsolescence function (describing the numbers of citation over time) has been studied by *Egghe* and *Rao*⁴ and few other scholars.⁶⁻⁹

Growth is an increase in size, implying "a change of state". In science and technology, do we consider growth in the literature (measured by the number of publications) or growth in the number of authors) or both. In this paper, we restrict ourself to the growth of the study of literature.

Objectives

The main objectives of this study are:

- (a) to study the patterns of obsolescence of literature;
- (b) to study the relations between the growth of literature and obsolescence of literature.

Source

The basic source of information in the field of theoretical population genetics is *Bibliography of Theoretical Population Genetics* compiled by Felsenstein, 1981.⁵ It covers literature in this speciality since its inception in 1870 till the 1980. Based on the analysis of this bibliography, a set of core journals based on their article output, has been identified and selected for study. Two types of data have been collected. First is the growth of the total number of publications for 1931 to 1980, as reflected in the bibliography and is shown in Table 1.

Table 1
Growth in the number of publications in theoretical population genetics

Year	NP	Year	NP	Year	NP
1930	29	1945	7	1962	128
1931	33	1946	16	1963	146
1932	27	1947	21	1964	167
1933	27	1948	23	1965	209
1934	32	1949	45	1966	234
1935	26	1950	39	1967	301
1936	24	1951	32	1968	283
1937	28	1952	49	1969	284
1938	36	1953	61	1970	324
1939	30	1954	72	1971	285
1940	19	1955	75	1972	342
1941	19	1956	78	1973	418
1942	28	1957	71	1974	421
1943	11	1958	77	1975	452
1944	13	1959	91	1976	465
		1960	141	1977	508
		1961	122	1978	517
				1979	483

NP = Number of publications.

Table 2
Number of source articles and citations in different periods

Year	Number of source articles	Number of citations
1931	10	96
1937	13	92
1943	8	173
1949	26	328
1955	38	683
1961	37	478
1967	78	997
1973	74	1258
1979	146	3197

Secondly, we have selected the following years as the base years for study: 1931, 1937, 1943, 1949, 1955, and 1979 representing different phases of the development of theoretical population genetics. For each of these base years, a minimum number of source articles, depending upon their coverage in the bibliography, were selected through detailed analysis from core journals. The actual number of source articles selected and the number of citations made in these source articles are shown in Table 2. The distribution of citations appearing in source articles from various years are given in Tables 3 and 4.

Obsolescence of literature

As we have indicated earlier that there are two different approaches to collect data to measure obsolescence: Synchronous and diachronous. In the first approach one fixes the citing literature (articles from one journal or more journals in the same year), and study the age distribution of references therein. In this paper synchronous approach is used to collect the data and then compute obsolescence rates.

Data analysis

Many scholars have tried to fit the following statistical distributions in the data describing the age distribution of references.

1. Lognormal distribution.
2. Negative binomial distribution.
3. Exponential distribution.
4. Weibull distribution.

It is generally seen through many studies that the age distribution of cited references are normally described by lognormal distribution. As a result we also tried to fit lognormal distribution in the age distribution of citations of source articles for the source years: 1931, 1937, 1943, 1949, 1955,, 1979. The probability density function of the lognormal distribution is given by the following equation:

$$Y_t = \frac{1}{t\sqrt{2\pi\sigma}} e^{-1/2(\log t - \mu/\sigma)^2}$$

where μ and σ are mean and standard deviations with respect to the variable $\log t$.

Table 3
Age distribution of citations in source articles

Age	Number of citations (Years)						
	1931	1937	1943	1949	1955	1961	1967
0	8	3	6	17	75	23	45
1	13	14	18	38	99	51	91
2	16	19	17	26	67	42	110
3	10	10	20	20	63	29	95
4	6	3	14	13	29	40	90
5	3	7	15	12	36	31	45
6	9	6	11	13	45	37	49
7	2	8	7	9	28	40	60
8	2	2	7	18	47	19	44
9	3	5	9	10	14	18	47
10	4	2	5	11	12	7	29
11	1	3	7	11	11	9	31
12	1	0	7	4	14	16	32
13	4	2	3	13	6	6	30
14	2	0	4	10	8	9	19
15	2	1	1	2	7	7	17
16	0	3	3	8	9	4	7
17	2	0	1	10	7	7	9
18	2	1	1	10	10	13	16
19	3	0	1	14	5	5	7
20	2	1	0	2	12	4	8
21	0	0	3	3	4	3	2
22	0	0	4	3	8	5	0
23	1	0	0	1	7	6	2
24		0	0	3	11	3	4
25		0	2	2	11	6	7
26		0	1	6	1	5	5
27		1	0	6	1	0	4
28		0	0	1	4	2	2
29		1	0	1	2	1	6
29			1	2	2	6	8
30			1	5	1	6	0
31			0	5	0	3	2
32			0	1	5	2	2

Table 3 (cont.)

Age	Number of citations (Years)						
	1931	1937	1943	1949	1955	1961	1967
33			1	0	5	2	0
34			0	1	0	2	8
35			0	1	0	2	5
36			0	0	5	1	9
37			0	0	0	1	
38			0	0	0	0	2
39			0	4	1	0	5
40			0	1	0	0	4
41			0	0	1	0	1
42			0	0	0	3	4
43			0	1	1	0	0
44			0	1	2	0	3
45			0	0	0	0	2
46			0	1	1	0	2
47			0	1	1	0	0
48			0	1	1	0	2
49			1	1	0	0	1
50			0	0	1	1	1
52			0	0	0	0	1
53			0	0	0	0	0
54			0	0	0	0	1
55			0	0	1	0	1
58			0	0	0	0	1
59			0	1	0	0	0
60			0	0	0	1	1
62			0	0	0	0	0
63			0	0	0	1	1
64			0	1	0		1
66			1	0	0		1
67			1	0	0		1
68				0	0		1
69				0	0		1
72				1	0		2
79				0	1		0
80				0			0
82				1	0		0
83				1	0		0
84					1		0
85							0
86							0
102							1
105							0
108							1
110							1
138							
166							

Table 4
Age distribution of citations in source articles

Age	Number of citations (Years)	
	1973	1979
0	51	149
1	100	263
2	161	335
3	146	277
4	112	269
5	89	235
6	62	149
7	61	157
8	40	147
9	61	131
10	39	89
11	28	74
12	19	64
13	24	50
14	20	64
15	11	77
16	12	47
17	18	49
18	18	31
19	21	43
20	17	28
21	5	30
22	10	27
23	12	29
24	7	27
25	10	30
26	7	23
27	4	16
28	4	12
29	1	26
30	3	20
31	1	15
32	4	7
33	2	10
34	1	8
35	4	7
36	6	11
37	2	6
38	3	14
39	3	8
40	3	10
41	5	6
42	6	4

Table 4 (cont.)

Age	Number of citations (Years)	
	1973	1979
43	12	7
44	1	7
45	1	6
46	2	6
47	2	8
48	1	14
49	1	14
50	1	2
51	1	7
52	3	1
53	1	5
54	0	0
55	7	3
56	0	3
57	2	5
58	0	11
59	2	2
60	0	1
61	1	9
62	1	1
65	0	1
66	1	0
67	0	0
68	0	0
71	0	1
72	1	1
74	0	1
81	0	0
84	1	0
87	0	0
97	0	1
98	1	0
99	0	0
100	0	1
102	2	1
103	0	1
105	0	2
114	1	0
117	0	0
120	0	1
129	1	0
145	0	0

Application of lognormal distribution

The lognormal distribution is applied to 9 data sets on the age distribution of citations appearing in source papers for the years 1931, 1937, 1943,, 1979. The numerical value of the mean and medium of the lognormal distribution, as applied to these 9 observed distributions of citations are given in Table 6. A K-S statistical test was then applied to test the applicability of the lognormal distribution in each of the 9 observed distributions of the age of citations and the results obtained are given in Table 5.

Table 5
Value of Dmax obtained through application of lognormal distribution

Year	Value of Dmax obtained	K-S Stat.*	Results
1931	0.1254	0.1664	Accepted
1937	0.0735	0.0728	Rejected
1943	0.0607	0.1261	Accepted
1949	0.0716	0.0924	Accepted
1955	0.0495	0.0661	Accepted
1961	0.0408	0.0764	Accepted
1967	0.0321	0.0507	Accepted
1973	0.0362	0.0513	Accepted
1979	0.0193	0.0301	Accepted

*At 0.0001 level of significance.

The results indicate that lognormal distribution showed positive fits in 8 out of 9 data sets of the age distribution of citations appearing in source articles of various years. The statistics obtained from the application of the lognormal distribution is given in Table 6. On the basis of these results, we can say that a more natural estimate of the age of citation is not the number of years (*t*) elapsed since its publication, but the logarithm of this number (*t*).

Table 6
Statistics obtained from the application of the lognormal distribution

Year	Mean	SD	Year	Mean	SD
1931	1.3010	1.0548	1961	1.8147	1.0105
1937	1.3657	0.9021	1967	1.8197	1.0512
1943	1.6913	0.9558	1973	1.7736	1.0287
1949	2.0238	0.1326	1979	1.8761	1.0475
1955	1.6639	1.0838			

Computation of obsolescence rates

For this purpose, we have used the methodology followed by Rao.⁶ If we assume that the age distribution of cited references follows an exponential distribution, then it is presumed that computation of obsolescence rate is very trivial. In such a situation, the obsolescence factor is given by:

$$a = Y_{t+1}/Y_t = e^{-\theta}$$

where Y_t is the number of cited references t years old, θ is the exponential distribution parameter.

The obsolescent rate, is then given by $(a-1)\times 100$. Since the exponential distribution fits hardly to any of the data sets, the method (from now onwards will be referred as exponential method) does not give accurate results. In order to get accurate value of a , one may have to use graphical method. Since there are too many data sets, the graphical method is certainly cumbersome and hence, simple regression analysis is used to compute obsolescence rates.

For each of data sets efforts were made to fit the following equation;

$$\text{Log } Y_t = a + b t$$

where Y_t is the number of cited references which are t years old.

Since Y_0 is very small, when compared to the total number of observations in each of the data sets, only those values of Y_t for $t \geq 1$ are considered while fitting the data. The constant rate of obsolescence implicit in a fitted exponential curve, denoted by a is given by:

$$a = \{(\text{antilog } b) - 1\} \times 100$$

The obsolescence rates for all the sets of citations made in source articles are computed using both the methods (exponential and regression method), and the results are given in Table 7. In addition, half-life of the citation data sets for difference source articles from different years are computed and given in Table 7.

Growth rates of literature in theoretical population genetics covered in the bibliography are also computed and shown in Table 7.

Table 7
Growth rates, obsolescence rates, and half-life of theoretical population
genetics literature

Period (1)	Growth rate (2)	Obsolescence rates		Half-life (5)
		RM (3)	EM (4)	
1930-31	-13.79	-18.15	-19.05	3.17
1930-37	0.62	-19.27	-17.10	3.00
1930-43	-2.13	-15.95	-18.00	4.77
1930-49	11.02	-14.84	-17.98	7.89
1930-55	11.12	-14.70	-17.37	4.24
1930-61	11.00	-13.64	-16.67	5.62
1930-67	11.90	-13.15	-15.13	5.48
1930-73	11.13	-12.32	-13.80	4.66
1930-79	10.00	-12.07	-12.75	5.48

RM = Regression method.

EM = Exponential method.

Relation between growth and obsolescence

A few scholars have studied the relation between growth rate and the obsolescence rate and half-life.^{6,10,11} We have followed the methodology suggested by *Rao*. In order to study the relationship between growth and obsolescence, it is first hypothesised that:

“higher the growth of literature, higher the obsolescence”.

Half-life of articles is an indicator which through light on the extent of the usage of articles. It is then further hypothesised that:

“higher the growth of literature, higher the half-life of literature”.

In order, to test these two hypotheses, the following correlation coefficients between:

- (1) growth rates and obsolescence rates;
- (2) growth rates and half-life of articles

are computed and then tested ($H_0 : p = 0$, $H_1 : p \neq 0$) for their linearity. The correlation coefficients for different population after computation are shown in Table 8.

The different variables which have been studied are:

- A1: Growth rates of literature in theoretical population genetics for different periods, such as 1930 to 1931, 1930 to 1937, 1930 to 1943, 1930 to 1949 and so on (for data, see column (2) in Table 7).
- A2: Average obsolescence rates for theoretical population genetics literature, computed using regression methods, for the period corresponding to the growth rates (for 1930 to 1931, 1930 to 1937, etc.) (for data see column 3 in Table 7).
- A3: Average obsolescence rates for theoretical population genetics literature, computed using exponential method, for the period corresponding to the growth rates (for 1930 to 1931, 1930 to 1937, etc.) (for data, see column 4 in Table 7).
- A4: Average half-life of theoretical population genetics literature for the period corresponding to the growth rates (for data, see column 5 in Table 7).

The critical values and t-statistics are given in Table 8.

Table 8
Correlation coefficient and the t-value

SN	Variable	n	Correlation coefficient r	t-value t= $r\sqrt{n-2/1-r^2}$	Remarks
Growth vs Obsolescence					
1.	A1,A2	9	0.1965	0.5303	Accepted
2.	A1,A3	9	-0.1145		Rejected
Growth vs Half-life					
3.	A1,A4	9	0.2805	0.7713	Accepted

Note: Critical value for 1 to 3. $t_{\alpha/2}$ (n = 9, $\alpha = 0.05$) = 2.365 (df = 7).

Results

In the first case (i.e. when correlation coefficient between growth rates and average obsolescence rates using regression methods was considered), the null hypothesis ($H_0 : P = 0$) has been accepted. On the basis of this we can now say that there may not be any relation between growth and obsolescence. Therefore, one may not believe that "higher the growth of literature, higher the obsolescence of literature".

In the second case, when correlation between growth of literature and obsolescence rates (using exponential method) was considered, it is found to be negative. In this case, we can again say that there may not be any relation between growth rate and obsolescence rate.

In the third case, when the correlation between growth of literature and half-life was considered, the null hypothesis ($H_0 : P = 0$) has been accepted. We can, therefore, conclude that there may not be any relationship between growth of literature and half-life of literature.

References

1. D. K. GAPEN, S. P. MILNER, Obsolescence, *Library Trends*, 30 (1981) 107–124.
2. M. B. LINE, Changes in the use of literature with time, *Library Trends*, 41 (1993) 665–683.
3. M. B. LINE, A. SANDISON, “Obsolescence” and changes in the use of literature with time, *Journal of Documentation*, 30 (1974) 283–349.
4. L. EGGHE, I. K. R. RAO, Citation age data and the obsolescence function: Fits and explanations, *Information Processing and Management*, 28 (1992) 201–217.
5. J. FELSENSTEIN, *Bibliography of Theoretical Population Genetics*, Dowden, Hutchinson & Ross, Inc., 1981.
6. I. K. R. RAO, B. M. MEERA, Growth and obsolescence of literature: An empirical study. In: *Informetrics-91*, I. K. R. RAO (Ed.), Bangalore; Sarada Ranganathan Ranganathan Endowment for Library Science; 377–394.
7. E. MATRICCIANI, The probability distribution of the age of references in engineering papers, *IEEE Transactions on Professional Communications*, 34 (March 1991) 7–12.
8. C. D. HURT, J. M. BUDD, Modeling the literature of superstring theory, *Scientometrics*, 24 (1992) 471–480.
9. B. M. GUPTA, Analysis of the distribution of the age of citations in theoretical population genetics, *Scientometrics*, 40 (1997) 139–162.
10. L. EGGHE, On the influence of growth on obsolescence. *Scientometrics*, 27 (1993) 195–214.
11. E. EGGHE, A theory of continuous growth rates and applications to the theory of growth and obsolescence rates, *Information Processing and Management*, 30 (1994) 279–292.