Information Processing & Management Vol. 28, No. 2, pp. 201-217, 1992
Printed in Great Britain.

0306-4573/92 $5.00 + .00

CITATION AGE DATA AND THE OBSOLESCENCE FUNCTION: FITS AND EXPLANATIONS

L. EGGHE
LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium*
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium

I. K. RAVICHANDRA RAO
DRTC, 8th Mile, Mysore Road, RV College, P.O., Bangalore 560059, India*
LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium

(*Received* 6 *April* 1991; *accepted in final form* 1 *July* 1991)

Abstract—The paper deals with the shape of the obsolescence function, which one can construct, based on the age data of reference lists. This paper shows that the obsolescence factor (aging factor) a is not a constant but merely a function of time. This jeopardizes this factor as a useful measure. We show (by experiment and also mathematically) that the function a has a minimum, which is obtained at a time t later than the time at which the maximum of the number of citations is reached. We then fit sets of citation data by using the lognormal distribution (other distributions do not fit well). Analytical calculations with this function a are indeed valid for these data. These arguments also yield a description of the utility function u and the total utility U. The latter can be used in comparing the "total lives" of the literature in various subjects.

I. INTRODUCTION

Several attempts have been made in the past to compute the obsolescence (also called aging) factor (see below), half-life (the time in which 50% of the material has already been used), or the utility factor (see section IV) for periodical publications (Brookes, 1970a, 1970b, 1971; Rao, 1973; and many others). We note that only the obsolescence factor a must be determined, since both half-life and the utility factor are simple functions of a. Most of these studies were based on an assumption that the distribution of age of the cited journals follows an exponential distribution; that is, if t represents the discrete age of journals cited and $c(t)$ is the relative number of journals whose age is t years, then

$$c(t) = \theta e^{-\theta t}, \quad t \geqq 0, \tag{1}$$

where $\theta > 0$ is a parameter (see Fig. 1 for a graph of (1)). Based on this assumption, it is straightforward to define the obsolescence (aging) factor a as

$$a(t) = a = \frac{c(t+1)}{c(t)}. \tag{2}$$

Indeed, in case of (1), $a(t)$ is independent of t:

$$a(t) = \frac{\theta e^{-\theta(t+1)}}{\theta e^{-\theta t}} = e^{-\theta},$$

which we define as the aging factor

$$a = e^{-\theta}. \tag{3}$$

*Permanent address.

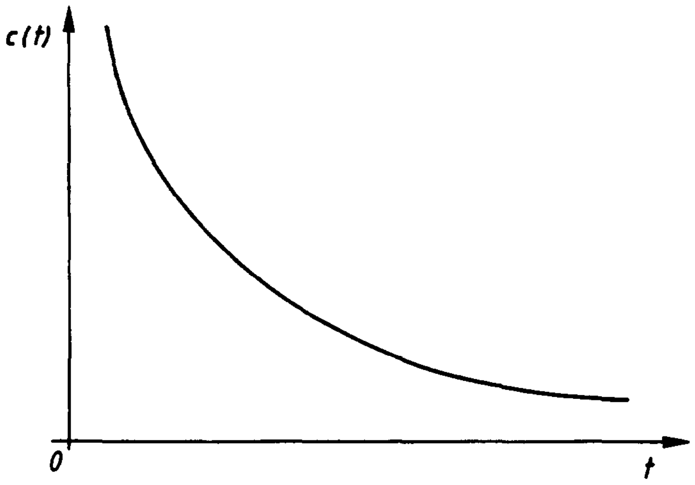

Fig. 1. Graph of the exponential distribution.

Formula (2) above is not used in practice, since

- in practical citation data, a lot of values $c(t) = 0$ occur, and hence (2) fluctuates a lot;
- even when $c(t) \neq 0$, formula (2) allows for an irregular set of numbers $a(t)$, since we do not use grouped citation data; or we use $c(t)$ for t small (in which case we have a lot of citations), but then we are troubled by the initial increase in numbers of citations; or we use $c(t)$ for t large (but then we deal with only few data).

A smart way to overcome this problem was offered by Brookes (1973). The method can be described as follows: Assuming (1) and (3), we can write

$$c(t) = \theta a^t. \tag{4}$$

Let m denote the total number of citations to publications that are t years old, or older. Hence

$$\begin{aligned} m &= \theta a^t + \theta a^{t+1} + \dots \\ &= a^t(\theta + \theta a + \theta a^2 + \dots) \\ &= a^t T, \end{aligned}$$

where T denotes the total number of citations. Hence

$$a = \left(\frac{m}{T}\right)^{1/t}. \tag{5}$$

This formula is very good in practice since we use only grouped data. Note that, as in (3), the result is always a constant, independent of time, if we assume the exponential distribution (1).

General citation data, however, do not conform with (1), nor with Fig. 1. The common graph encountered in practice is as in Fig. 2: There is an initial increase of citations (in this period the article or journal volume is distributed and the use increases), followed by a "sort of" exponential decay (Griffith *et al.*, 1979; Brookes, 1970a; Avramescu, 1973, 1979; Geller & de Cani, 1981; Stinson & Lancaster, 1987; Motylev, 1981, 1989; Rao, 1973).

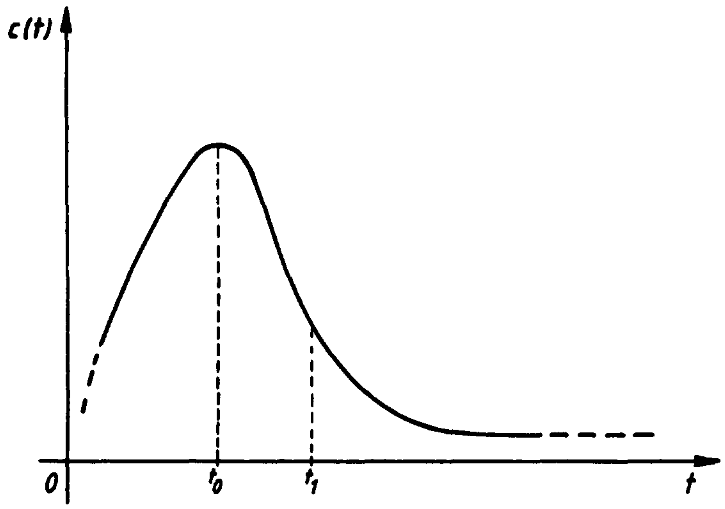

Fig. 2. Schematic form of the citation age distribution $c(t)$.

As a consequence, there is no way to find an aging factor a independent of t: a is only independent of t in case of the exponential distribution; hence, in any case conforming with Fig. 2, a is a function of t.

In the next section we will give several citation data and we will see that the function a, defined as in (2) (which can be used as a general definition of *aging function*), always has a minimum. This behavior of a will be explained in a mathematically qualitative way, only using the form of the graph in Fig. 2.

The third section studies (both analytically as well as statistically) possible fits of citation data. We conclude very clearly that only the lognormal distribution can be used as the underlying distribution (contrary to Weibull's, Avramescu's, or the Negative Binomial Distribution).

We therefore stress the impossibility of using the aging function as a constant and advocate the use of half-life instead. We also make a note on the total utility factor U which is directly related to the aging factor a via the formula:

$$U = \frac{1}{1-a}, \quad t \geqq 0. \tag{6}$$

The formula (6) is true only if $c(t)$ follows an exponential distribution. We study the extension of the utility as a function of t in case of other distributions.

II. THE AGING FUNCTION a: FORM AND QUALITATIVE EXPLANATION

II.1 *Data*

We start by giving details on a number of citation data that were obtained for the analysis of different books. The reason we used books is simply because we wanted to have a large amount of homogeneous data. We analysed the citation data appearing in the following books:

1. L. Egghe and R. Rousseau. Introduction to informetrics, Elsevier, Amsterdam, 1990. This is a new book on informetrics, dealing with a lot of informetric topics and having a large bibliography (508 references to the journals).
2. L. Egghe. Stopping time techniques for analysts and probabilists, London Mathematical Society Lecture Notes series, 100, Cambridge University Press, 1984. This is a book on a former specialty of the first named author comprising 193 references to the journals.

3. R.B. Cairns. Social development: The origins and plasticity of interchanges, W.H. Freeman & Co., San Francisco, 1979. This book contains 634 references to the journals.

We stress the fact that the observations we will make in this paper were also encountered in all other sets of citation data that we investigated.

The data are found in Tables 1, 2 and 3 (in the same order as above). They yield t (in years), $c(t)$ (the number of references to publications that are t years old, $t = 0,1,2,\ldots$), and $a(t)$, the obsolescence (aging) function. As mentioned in the introduction, for practical reasons, Brookes' method was used in order to have smoother values for $a(t)$. Here (5) becomes

$$a(t) = \left(\frac{m(t)}{T}\right)^{1/t}, \tag{7}$$

where $m(t)$ denotes the total number of citations to publications that are more than or equal to t years old, and T denotes the total number of citations (hence $T = m(0)$). (Disregard for the moment the columns with the headings $F(t)$, $G(t)$, and $D(t)$; we will deal with them in the next section). Based on these tables, we could also graph the frequency polygons for the age distribution c of citation data. For this, see Fig. 3.

It is clear that we have a citation graph of the form of Fig. 2. It is very striking from the above tables that the function $a(t)$ decreases (when t increases) until a minimum is reached, after which a (slow) increase starts (of course, for very high t, we encounter a few irregularities due to the small number of citations).

No matter what is the exact mathematical form of the function $c(t)$, we will now show—in a mathematically qualitative way—that a citation curve of the form of Fig. 2 implies that the aging function a must have the qualitative properties as described above.

Table 1. Age distribution of journals cited in informetrics

t	$Tc(t)$ (observed)	$a(t)$	$F(t)$	$G(t)$	$D(t)$
1	28		0.055118	0.019142	0.035976
2	46	0.972050	0.145669	0.086511	0.059159
3	38	0.948874	0.220472	0.171245	0.049227
4	48	0.939632	0.314961	0.256001	0.058960
5	37	0.927135	0.387795	0.334400	0.053395
6	21	0.921473	0.429134	0.404610	0.024524
7	16	0.923037	0.460630	0.466604	0.005974
8	18	0.925733	0.496063	0.521033	0.024970
9	22	0.926682	0.539370	0.568750	0.029380
10	17	0.925412	0.572835	0.610612	0.037778
11	21	0.925518	0.614173	0.647411	0.033238
12	16	0.923704	0.645669	0.679842	0.034172
13	22	0.923292	0.688976	0.708506	0.019530
14	19	0.919964	0.726378	0.733919	0.007541
15	10	0.917227	0.746063	0.756519	0.010456
16	18	0.917900	0.781496	0.776677	0.004819
17	13	0.914418	0.807087	0.794712	0.012375
18	12	0.912637	0.830709	0.810892	0.019817
19	10	0.910756	0.750394	0.825448	0.024945
20	8	0.909384	0.866142	0.838579	0.027563
21	8	0.908681	0.881890	0.850453	0.031437
22	2	0.907468	0.885827	0.861216	0.024611
23	4	0.909965	0.893701	0.870995	0.022706
24	4	0.910833	0.901575	0.879900	0.021675
25	2	0.911432	0.905512	0.888025	0.017487

(continued)

Table 1 continued.

t	$Tc(t)$ (observed)	$a(t)$	$F(t)$	$G(t)$	$D(t)$
26	4	0.913254	0.913386	0.895454	0.017932
27	6	0.913380	0.925197	0.902259	0.022938
28	6	0.911555	0.937008	0.908504	0.028504
29	1	0.909067	0.938976	0.914245	0.024741
30	3	0.910996	0.944882	0.919532	0.025350
31	1	0.910745	0.946850	0.924408	0.022442
33	1	0.914911	0.948819	0.933080	0.015739
34	1	0.916289	0.950787	0.936940	0.013848
35	1	0.917552	0.952756	0.840521	0.012235
36	2	0.918706	0.956693	0.943847	0.012846
39	2	0.922656	0.960630	0.952507	0.008123
41	3	0.924136	0.966535	0.957354	0.009181
42	1	0.922298	0.968504	0.959544	0.008960
43	2	0.922732	0.972441	0.961594	0.010847
45	1	0.923292	0.974409	0.965317	0.009092
49	1	0.927923	0.976378	0.971500	0.004878
50	1	0.927826	0.978346	0.972825	0.005521
51	1	0.927605	0.980315	0.974075	0.006240
56	1	0.932261	0.982283	0.979351	0.002933
57	2	0.931687	0.986220	0.980241	0.005980
68	1	0.928936	0.988189	0.987487	0.000702
70	2	0.928558	0.992126	0.988428	0.003698
81	1	0.941948	0.994094	0.992302	0.001793
83	1	0.940043	0.996063	0.992825	0.003238
89	1	0.929679	0.998031	0.994157	0.003875
145	1	0.957655	1.000000	0.998820	0.001180

Statistics of Table 1

Mean	11.946850
Var	180.227490
St. Dev.	13.424883
Total Number of Citations	508

Mean ($\log t$)	2.027863
Var ($\log t$)	0.958368
St. Dev. ($\log t$)	0.978962

Kolmogorov-Smirnov Statistic:	0.05 level	0.060340
	0.01 level	0.072320

Legend of Table 1

t: Age of the journal

$Tc(t)$: Number of Citations (observed) (T = total number of citations)

$a(t)$: Aging function (formula (7))*

$F(t)$: Cumulative (observed) (based on c(t))

$G(t)$: $$\int_0^t \frac{1}{\sqrt{2\pi}\sigma t'} \exp\left[-\frac{1}{2}\left(\frac{\log t'-\mu}{\sigma}\right)^2\right]dt' = \int_{-\infty}^{\log t}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right]du$$

where $u = \log t'$ and where μ and σ are the mean resp. the standard deviation w.r.t. $\log t$.

$D(t)$: $|F(t) - G(t)|$

*We used formula (7) instead of (2) in order to show better the behavior of the function a. Formula (7) is the "smoothened" version of formula (2) and, as is apparent from the tables, inherits all the properties of a as expressed in Fig. 5. The theoretical relation between (2) and (7) remains open, however.

Table 2. Age distribution of journals cited in analysis

t	$Tc(t)$ (observed)	$a(t)$	$F(t)$	$G(t)$	$D(t)$
1	5[a]		0.025907	0.005515	0.020391
2	13	.986962	0.093264	0.063303	0.029961
3	24	.967892	0.217617	0.173322	0.044294
4	33	.940492	0.388601	0.299963	0.088638
5	17	.906285	0.476684	0.420386	0.056298
6	20	.897692	0.580311	0.525371	0.054940
7	13	.883349	0.647668	0.613053	0.034616
8	17	.872746	0.735751	0.684714	0.051037
9	9	.862540	0.782383	0.742674	0.039710
10	12	.858557	0.844560	0.789357	0.055203
11	5	.844318	0.870466	0.826938	0.043528
12	3	.843397	0.886010	0.857242	0.028768
13	4	.846157	0.906736	0.881748	0.024988
14	0	.844128	0.906736	0.901638	0.005098
15	3	.853718	0.922280	0.917846	0.004434
16	1	.852429	0.927461	0.931109	0.003648
17	2	.856988	0.937824	0.942010	0.004186
18	0	.856997	0.937824	0.951006	0.013182
19	2	.863986	0.948187	0.958460	0.010274
21	2	.862526	0.958549	0.969843	0.011294
22	0	.865288	0.958549	0.974186	0.015637
23	2	.870748	0.968912	0.977840	0.008928
24	2	.865349	0.979275	0.980924	0.001649
28	1	.870713	0.984456	0.989259	0.004803
34	2	.884730	0.994819	0.995163	0.000345
41	1	.879538	1.000000	0.997441	0.002059

Mean:	7.3420.	Mean ($\log t$):	1.747902.
Variance:	36.3183.	Variance ($\log t$):	0.476030.
St. Dev.:	6.0265.	St. Dev. ($\log t$):	0.689950.

Total number of citations: 193.

Kolmogorov-Smirnov statistic: .05 level: 0.097895.
.01 level: 0.117330.

t, $c(t)$, $a(t)$, $F(t)$, $G(t)$, and $D(t)$ are as in Table 1.

[a]includes two citations to the zero-year-old journals.

II.2 *Analysis of the aging function*

Let us use a continuous time variable t and assume that our citation data conform with a density function of the form as in Fig. 2. Mathematically such a curve can be characterized as: c'' is continuous and there is a unique $t_0 > 0$, such that $c'(t_0) = 0$ and $c''(t_0) < 0$ and a unique $t_1 > t_0$, such that $c''(t_1) = 0$ (see Fig. 2). Furthermore, $\lim_{t\to\infty} c(t) = 0$ and $c(t) > 0$, for all t. Now using (2),

$$a(t) = \frac{c(t+1)}{c(t)} \tag{2}$$

for $t > 0$. The assumption of a continuous setting of time is quite natural; in fact, only due to publication restrictions, t is usually restricted to whole years, but in reality, time is continuous. So

$$a'(t) = \frac{c(t)c'(t+1) - c(t+1)c'(t)}{c^2(t)} \tag{8}$$

(a) *Suppose* $0 < t < t_0$. Then, based on our assumptions, $c(t) < c(t+1)$ and $c'(t) > c'(t+1)$. Hence (since all numbers $c(t)$, $c(t+1)$, $c'(t)$, $c'(t+1)$ are positive)

$$c(t)c'(t+1) < c(t+1)c'(t),$$

Table 3. Age distribution of journals in social development

t	*Tc*(*t*) (observed)	*a*(*t*)	*F*(*t*)	*G*(*t*)	*D*(*t*)
1	4[a]		0.006309	0.004715	0.001594
2	38	0.996840	0.066246	0.040864	0.025382
3	49	0.977412	0.143533	0.106123	0.037410
4	66	0.962006	0.247634	0.186730	0.062961
5	48	0.944682	0.323344	0.265606	0.057738
6	42	0.926975	0.389590	0.343083	0.046507
7	47	0.931911	0.463722	0.414432	0.049291
8	46	0.925068	0.536278	0.478732	0.057546
9	25	0.918158	0.575710	0.535974	0.039736
10	20	0.917839	0.607256	0.586584	0.020672
11	22	0.918546	0.641956	0.631171	0.010785
12	24	0.917969	0.679811	0.670392	0.009419
13	23	0.916124	0.716088	0.704887	0.011201
14	22	0.913990	0.750789	0.735247	0.015542
15	14	0.911530	0.772871	0.762000	0.010871
16	20	0.911522	0.804416	0.785613	0.018804
17	10	0.908476	0.820189	0.806492	0.013697
18	12	0.909077	0.839117	0.824992	0.014124
19	9	0.908317	0.853312	0.841417	0.011895
20	7	0.908489	0.864353	0.856031	0.008322
21	10	0.909256	0.880126	0.869060	0.011066
22	6	0.908079	0.889590	0.880701	0.008889
23	5	0.908640	0.897476	0.891123	0.006353
24	2	0.909462	0.900631	0.900472	0.000158
25	3	0.911780	0.905363	0.908876	0.003513
26	5	0.913304	0.913249	0.916443	0.003194
27	2	0.913433	0.916404	0.923271	0.006867
28	1	0.915180	0.917984	0.929442	0.011461
29	3	0.917379	0.922713	0.935029	0.012316
30	2	0.918199	0.925868	0.940096	0.014228
31	1	0.919494	0.927445	0.944698	0.017254
32	1	0.921289	0.929022	0.948886	0.019864
33	3	0.922966	0.933754	0.952701	0.018947
34	4	0.923269	0.940063	0.956183	0.016120
35	2	0.922735	0.943218	0.959364	0.016147
36	0	0.923411	0.943218	0.962276	0.019058
37	4	0.925401	0.949527	0.964943	0.015416
38	3	0.924421	0.954259	0.967390	0.013131
39	1	0.923951	0.955836	0.969637	0.013801
40	5	0.924968	0.963722	0.971704	0.007982
41	1	0.922284	0.865300	0.973607	0.008308
42	3	0.923094	0.970032	0.975361	0.005330
43	1	0.921666	0.971609	0.976980	0.005371
44	1	0.922420	0.973186	0.978475	0.005288
45	3	0.922730	0.977918	0.979857	0.001939
46	2	0.920451	0.981073	0.981136	0.000063
47	2	0.919057	0.984227	0.982321	0.001906
48	0	0.917184	0.984227	0.983420	0.000807
49	2	0.918804	0.987382	0.984440	0.002942
50	0	0.916263	0.987382	0.985387	0.001995
51	2	0.917835	0.990536	0.986267	0.004269
52	2	0.914278	0.993691	0.987087	0.006604
65	1	0.925025	0.995268	0.983894	0.001375
83	1	0.937537	0.996845	0.997553	0.000707
100	1	0.944039	0.998423	0.998865	0.000442
106	1	0.940947	1.000000	0.999119	0.000881

Mean: 11.8454. Mean (log *t*): 2.123179.
Variance: 138.6796. Variance (log *t*): 0.673986.
St. Dev.: 11.7762. St. Dev. (log *t*): 0.820967.

Total number of citations: 634.

Kolmogorov-Smirnov statistic: .05 level: 0.054013.
.01 level: 0.064736.

t, *c*(*t*), *a*(*t*), *F*(*t*), *G*(*t*), and *D*(*t*) are as in Table 1.

[a] includes three citations to the zero-year-old journals.

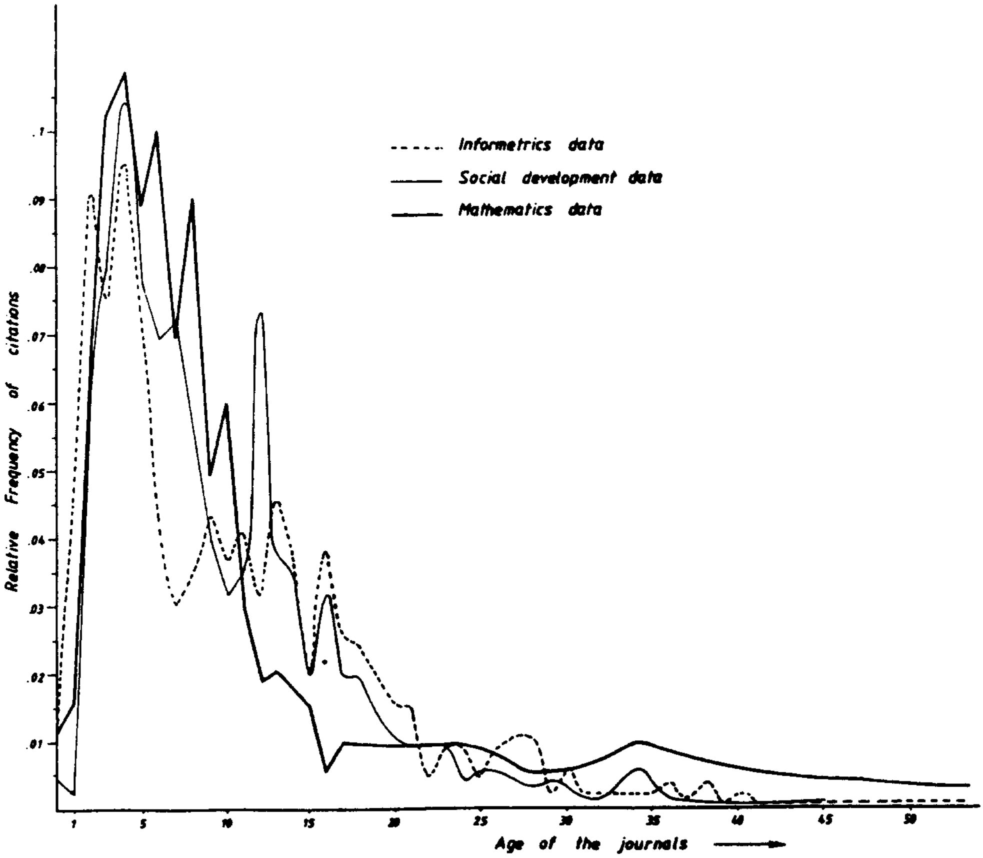

Fig. 3. Age distributions of journals in informetrics, social developments, and mathematics—frequency polygons.

implying, using (8), that

$$a'(t) < 0.$$

(b) *Suppose* $t_0 < t < t_1$. Now $c(t) > c(t+1) > 0$ and $0 > c'(t) > c'(t+1)$ so that $0 < -c'(t) < -c'(t+1)$, and hence

$$-c(t)c'(t+1) > -c(t+1)c'(t).$$

Consequently, again

$$a'(t) < 0.$$

(c) *Suppose that* $t > t_1$. Now $c(t) > c(t+1) > 0$, but $c'(t) < c'(t+1) < 0$, and hence there remain two cases:

(i) $|c(t)c'(t+1)| > |c'(t)c(t+1)|$. In this case

$$c(t)c'(t+1) < c'(t)c(t+1),$$

and hence

$$a'(t) < 0.$$

(ii) $|c(t)c'(t+1)| < |c'(t)c(t+1)|$. Now we have, evidently,

$$a'(t) > 0.$$

In case (c)(i) a keeps decreasing and hence (since $a > 0$) must have a horizontal asymptote at value

$$\lim_{t \to \infty} a(t) \geqq 0.$$

We are now in the case of Fig. 4.

In case (c)(ii), a' changes sign; hence there is a $t_2 > t_1$ such that

$$a'(t_2) = 0.$$

In this case we have described Fig. 5.

This gives a full qualitative analysis of the form of the function a, based on the qualitative assumptions on the function c.

Of course, in practice, Fig. 4 can be considered as a special case of Fig. 5, where t_2 is very high. We note, however, that almost all the citation data that we have investigated show a minimum for the aging function (as in Fig. 5). The reason for this will be explained in the next section. Note that our qualitative argument not only describes the form of Fig. 5 (as encountered in Tables 1,2,3), but also the fact that $t_2 > t_1 > t_0$; that is, the maximum of c is attained much earlier (in t_0) than the minimum of a (in t_2). For Table 1 we have, for exmaple, $t_0 \approx 4$ and $t_2 \approx 22$; for Table 2, $t_0 \approx 4$ and $t_2 \approx 12$; and for Table 3, $t_0 \approx 4$ and $t_2 \approx 19$.

Based on these results, we might wonder what function c best represents (fits) the citation data. This will be studied in the next section.

III. THE CITATION DISTRIBUTION FUNCTION c: STATISTICAL FIT AND MATHEMATICAL EXPLANATION

Based on the results of the previous section, we are now looking for a distribution function c that fits citation data, shows a graph as in Fig. 2, and agrees with the qualitative study of the aging function a of subsection II.2. Especially, a probability distribution c that allows an aging function a of the form in Fig. 5 interests us since, as is clear from Tables 1, 2, and 3 (and most other examples not discussed here), Fig. 5 is encountered in most cases.

Note that we combine two requirements here: (i) the statistical fit (including the behavior of the c curve as in Fig. 2) and (ii) the behavior of the a curve as in Fig. 5 (where a is derived from c via formula (2)). Indeed, previous studies have indicated good statistical fits of the exponential distribution (1), even in cases (as is most often so) where there

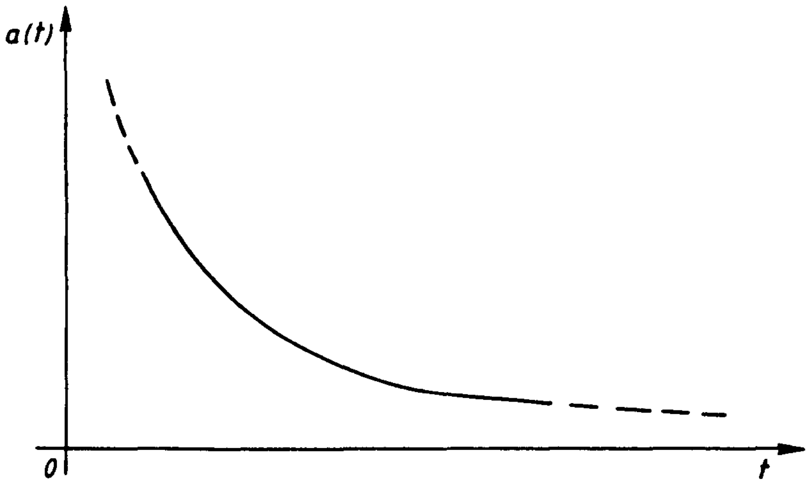

Fig. 4. a in case (c)(i).

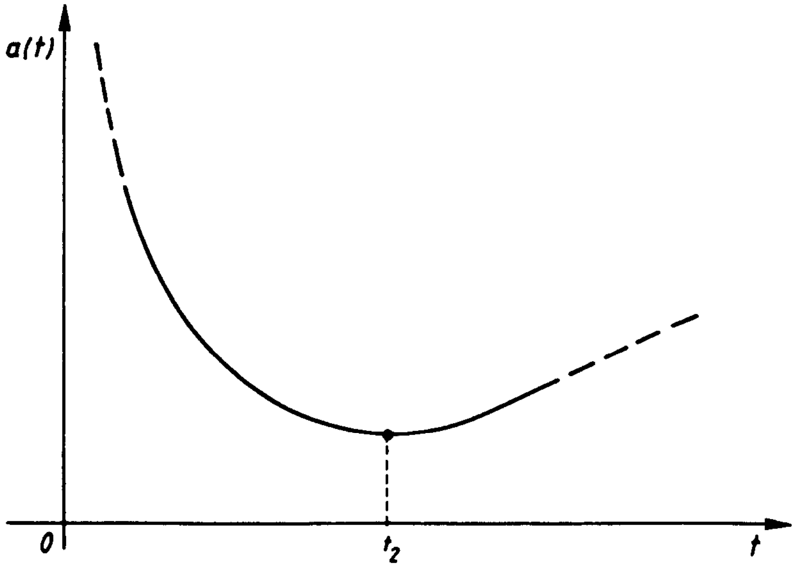

Fig. 5. a in case (c)(ii). The minimum occurs for $t_2 > t_1 > t_0$ (t_0, t_1 are the respective maximum and inflection point of the c curve).

is an initial increase in $c(t)$. So, requiring only (i) is not enough to guarantee that we deal with the right underlying function.

Next we study a few candidates.

III.1 *The model of Avramescu*

Avramescu (1973, 1979) proposes the following mathematical function:

$$c(t) = c(e^{-\theta t} - e^{-m\theta t}), \tag{9}$$

where $m \gg 1$ and $c > 0$ is a constant such that c is a probability distribution. Note the extension of formula (1), which can be derived from (9) for t large: t large $\Rightarrow mt \gg t$ and so

$$c(e^{-\theta t} - e^{-m\theta t}) \approx ce^{-\theta t} = ca^t.$$

Only if t is small, the term $e^{-m\theta t}$ is important and is responsible for the initial increase. It can indeed be verified that c has the form of Fig. 2. The aging function a for (9) is then

$$a(t) = \frac{c(t+1)}{c(t)} = \frac{e^{-\theta(t+1)} - e^{-m\theta(t+1)}}{e^{-\theta t} - e^{-m\theta t}}. \tag{10}$$

Using (8) (or directly (10)), it can readily be verified that

$$a'(t) = \frac{(-\theta)\,[me^{-\theta(mt+t)}(e^{-\theta} - e^{-m\theta}) + e^{-\theta(mt+t)}(e^{-m\theta} - e^{-\theta})]}{(e^{-\theta t} - e^{-m\theta t})^2}.$$

Since $\theta > 0$ and $m \gg 1$, we hence see that $a'(t) < 0$ for all $t > 0$, and hence a is of the form of Fig. 4 *only*. Hence, Avramescu's function is not able to cover the most frequently occurring cases of Fig. 5. We therefore conclude that Avramescu's model is not the right citation age model. Note that function (1) can be regarded as a special case of (9) (as indicated above); here a is a constant.

III.2 *The Negative Binomial Distribution (NBD)*

Another "natural" candidate for c is the NBD. The mathematical form is

$$c(t) = \frac{\Gamma(K+t)}{\Gamma(K)\Gamma(t+1)}\, p^K q^t, \tag{11}$$

$t = 0,1,2,\ldots,0 \leqq p,\ q \leqq 1$, and $K > 0$. Also here

$$a(t) = \frac{(K+t)q}{t+1} \tag{12}$$

and

$$a'(t) = \frac{q - Kq}{(t+1)^2}\begin{cases} < 0 \text{ for } K > 1 \\ > 0 \text{ for } 0 < K < 1 \\ = 0 \text{ for } K = 1 \end{cases}$$

for all $t > 0$, and hence NBD cannot be the citation age distribution (since Fig. 5 is not covered except for $K = 1$, but this represents the exponential model (1) (discrete case), whose deficiencies were already mentioned in section I).

III.3 *The Weibull distribution*

This distribution has the following mathematical form:

$$c(t) = \frac{ct^{c-1}}{b^c}\exp\left[-\left(\frac{t}{b}\right)^c\right], \tag{13}$$

where $c > 0$. Now this function has the form of Fig. 2 *only* for $c > 1$; indeed $c'(t) = 0$ for

$$t = b\left(\frac{c-1}{c}\right)^{1/c}. \tag{14}$$

That is the reason we include it in this paper, although we believe that this distribution is seldom (or never) used in informetrics. Now, if $c > 1$ we then see that

$$a'(t) = \left(\frac{t+1}{t}\right)^{c-2}\exp\left[\frac{t^c - (t+1)^c}{b^c}\right]\cdot\frac{1}{t^2}\left[-(c-1) - t(t+1)\frac{c}{b^c}\right] \tag{15}$$

is never zero.

So, based on our model of section II, we are inclined to say that we are not likely to have an a function of the form of Fig. 5. R. Rousseau (oral communication) kindly informed us of some weak fits he was able to establish of a few cases with the Weibull distribution but—as he admits—the results are not very convincing. It is furthermore well known that sometimes statistical fits are possible with models that are wrong (in a model-theoretic sense). Even the simple function (1) can fit citation data in some cases!

III.4 *The lognormal distribution*

III.4.1 *Definition.* If $z = \log t$ is normally distributed, then the distribution of t is said to be lognormal. The probability density function of the lognormal distribution (of t) is given by

$$c(t) = \frac{1}{t\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left[\frac{\log t - \mu}{\sigma}\right]^2\right\}. \tag{16}$$

The normal density function of $\log t$ is given by

$$f(\log t) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left[\frac{\log t - \mu}{\sigma}\right]^2\right\} \tag{17}$$

where μ and σ are the mean and the standard deviation with respect to the variable $\log t$. As is easily seen (and well known), function c (equation (16)) above has the required form as in Fig. 2. Indeed $c'(t) = 0$, for $t = e^{\mu-\sigma^2} > 0$, $\lim_{t\to\infty} c(t) = 0$ and $\lim_{\substack{t\to 0\\ >}} c(t) = 0$ as can be easily seen. Now

$$a(t) = \frac{c(t+1)}{c(t)} = \frac{t}{t+1} \exp\left(\frac{1}{2\sigma^2} \{[\log t - \mu]^2 - [\log(t+1) - \mu]^2\}\right).$$

Hence

$$\log a(t) = (\log t - \log(t+1))\left(\frac{\log t + \log(t+1)}{2\sigma^2} - \left(\frac{\mu}{\sigma^2} - 1\right)\right).$$

Hence

$$\lim_{\substack{t\to 0\\ >}} a(t) = +\infty. \tag{18}$$

Furthermore

$$\lim_{t\to\infty} a(t) = 1 \tag{19}$$

as is seen by applying de l'Hôspital's rule three times to the expression

$$\begin{aligned}(\log t - \mu)^2 &= (\log(t+1) - \mu)^2 \\ &= (\log t(t+1) - 2\mu)\log \frac{t}{t+1}.\end{aligned}$$

Finally, if we can show that $a(t) < 1$ for a certain $t > 0$, then we are sure of the fact that a has a minimum (based on our assumptions for c). Now the condition $a(t) < 1$ is equivalent to

$$\log t(t+1) > -2\sigma^2 + 2\mu,$$

which is always true from a certain $t > 0$ on. Hence, Fig. 5 in this concrete situation is as in Fig. 6.

That $a(t)$ can be larger than 1 is no surprise; $a(t)$ measures aging: high values imply no aging and low values imply high aging. Since, for low t, c increases we have the opposite of aging at that time, and hence $a(t) > 1$. $a(t) \approx 1$ means $c(t) \approx$ constant and $a(t) < 1$ means c decreasing. Also $a(t) \approx 1$ for high t is logical; the obsolescence (aging) stops since the (low) use of old material remains more or less the same for different t (note again that $c(t) \approx$ constant, for high t).

These mathematical considerations are confirmed by very good statistical fits, as is seen in the next subsection.

III.4.2 *Statistical fit of the lognormal distribution to the citation age data.* We refer to our three large homogeneous data of section II. $F(t)$ denotes the cumulative observed data and $G(t)$ the cumulative normal distribution (w.r.t. $\log t$):

$$G(t) = \int_{t'=0}^{t'=t} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\log t' - \mu}{\sigma}\right)^2\right] d\log t'. \tag{20}$$

$D(t)$ is then the absolute difference $|F(t) - G(t)|$, and we performed a Kolmogorov-Smirnov test of fit. In all cases we have a very good fit. Note also that we did not "cut away" any citation age data, and hence we worked with the full initial increase (t small)

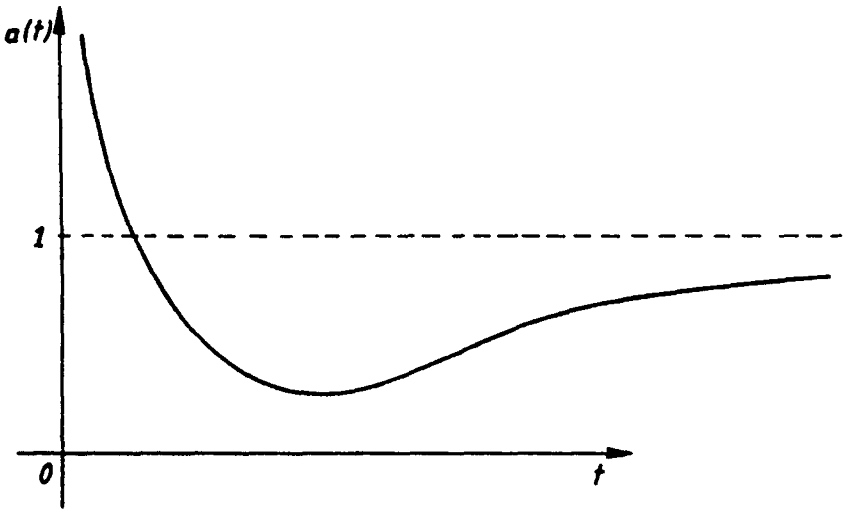

Fig. 6. The *a* curve, in case of the lognormal distribution.

and the long tail (t large) of c. Hence, $\log t$ follows a normal distribution (i.e., t follows a lognormal distribution).

III.4.3 *Explanation of the lognormal distribution.* The following argument, which can be read in Bartholomew (1982), Matricciani (1991) and Rao (1988), can be used as an explanation why the lognormal distribution is the underlying "logical" distribution for citation age data.

The lognormal distribution may be derived from the law of proportionate effect. That is, the citation to a journal at any stage is a random proportion of the citations it received in the immediately previous stage. If the journal has received x_j citations at the jth interval, x_{j+1} (the citations at the $(j+1)$th interval) is given by:

$$x_{j+1} - x_j = \epsilon_j x_j. \tag{21}$$

The ϵ_j's are mutually independent, and further they are all independent of all x_js. After a sequence of n proportionate "random shocks," citations to the journal will be

$$x_n = x_0(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3)\ldots(1+\epsilon_n) \tag{22}$$

where x_0 is the initial number of citations at some arbitrary origin of time. By taking the logarithm in (22), we have:

$$\log x_n = \log x_0 + \log(1+\epsilon_1) + \log(1+\epsilon_2) + \ldots + \log(1+\epsilon_n). \tag{23}$$

For sufficiently large n, from the Central Limit Theorem, it can be shown that $\log x_n$ is normally distributed, as each of the terms on the right-hand side of the equation is an independent random variable. This explains the lognormal distribution.

IV. A NOTE ON THE UTILITY FUNCTIONS $u(t)$ AND TOTAL UTILITY U

Practical informetrists are more interested in the "total utility" U of publications than in the "aging" of these documents. The two measures are, however, linked in a one-to-one way by a formula under the assumption that $c(t)$ is an exponential distribution (1):

$$U = \frac{1}{1-a} \tag{24}$$

(see also Brookes, 1970a, 1970b, 1971; Rao, 1973). An attempt has been made here to derive a similar formula under the assumption that $c(t)$ is a lognormal distribution.

Let $u(t)$ be the utility of a journal in the tth year ($t = 1,2,\dots$). If $a(t)$ is the aging factor at the tth year, $u(t)$ may be defined as

$$u(t) = u(t-1)a(t) = \dots = u(0)a(1)\dots a(t),$$

where $u(0)$ is the initial utility of the journal ($u(0) > 0$ is a parameter whose definition is unclear, but we do not need it in the applications—see further on). $u(t)$ is thus given by (for $t = 1,2,\dots$)

$$u(t) = u(0)a(1)\dots a(t) = u(0)\prod_{i=1}^{t} a(i). \tag{25}$$

This formula agrees with (24) in case c conforms with (1). Indeed, then, using (25)

$$u(t) = u(0)a^t,$$

where $a = a(t)$ is the constant aging factor ($0 < a < 1$). Hence the total utility U is given by

$$U = \sum_{t=0}^{\infty} u(t)$$

$$U = \frac{u(0)}{1-a}.$$

Putting $u(0) = 1$ (scale of measurement in case (1)), then we again find (24).

Hence $u(t)$ is an acceptable definition of "utility in the tth year." A study of U will be done in the sequel. We commence with the study of $u(t)$, for the lognormal function c.

For the lognormal distribution, $\prod_{i=1}^{t} a(i)$ is given by (cf. section III.4):

$$\frac{1}{t+1}\exp\left\{\frac{1}{2\sigma^2}\left[\mu^2 - (\log(t+1)-\mu)^2\right]\right\}. \tag{26}$$

Hence, according to (25):

$$u(t) = \frac{u(0)}{t+1}\exp\left\{\frac{1}{2\sigma^2}\left[\mu^2 - (\log(t+1)-\mu)^2\right]\right\} \tag{27}$$

for $t = 1,2,\dots$.

Now we study the total utility U. The total utility of the journal may be computed as:

$$U = \int_0^{\infty} u(i)\,di.$$

From (27), we thus have,

$$U = U(0)\int_0^{\infty}\frac{1}{i+1}\exp\left\{\frac{1}{2\sigma^2}\left[\mu^2 - (\log(i+1)-\mu)^2\right]\right\}di$$

$$= u(0)\int_0^{\infty}\frac{1}{i+1}\exp\left\{\frac{1}{2\sigma^2}\left[2\mu\log(i+1) - \log^2(i+1)\right]\right\}di. \tag{28}$$

On substituting $e^x = i + 1$ in (28), we have

$$U = u(0)\int_0^\infty \exp\left[\frac{1}{2\sigma^2}\,(2\mu x - x^2)\right]dx$$

$$= u(0)\int_0^\infty \exp\left(\frac{\mu}{\sigma^2}\,x - \frac{x^2}{2\sigma^2}\right)dx. \tag{29}$$

The equation (29) is of the form

$$U = u(0)\int_0^\infty \exp\left[\left(-\frac{x^2}{4\beta} - \gamma x\right)\right]dx \tag{30}$$

where

$$\beta = \frac{\sigma^2}{2} \quad \text{and} \quad \gamma = -\frac{\mu}{\sigma^2}. \tag{31}$$

The solution of the integral (30) is given by Gradshteyn and Ryzhik (1965, p. 307 (3.322.2)).

$$U = u(0)\{\sqrt{\pi\beta}e^{\beta\gamma^2}[1 - \Phi(\gamma\sqrt{\beta})]\}. \tag{32}$$

In (32), $\Phi(x)$ is the probability integral that is equal to

$$\Phi(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-z^2}\,dz.$$

Thus, if

$$F(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-y^2/2}\,dy + 0.5 = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^x e^{-y^2/2}\,dy,$$

then $\Phi(x) = 2F(x\sqrt{2}) - 1$.

The values of $F(x)$ can be obtained from the tables of the normal distribution. Thus substituting for $\Phi(x)$ in (32), we have

$$U = u(0)\sqrt{\pi\beta}e^{\beta\gamma^2}2(1 - F(\gamma\sqrt{2\beta})). \tag{33}$$

Since $\gamma < 0$ (from (31)), and since $F(-x) = 1 - F(x)$, (33) can be rewritten as

$$U = u(0)[\sqrt{\pi\beta}e^{\beta\gamma^2}2F(-\gamma\sqrt{2\beta})], \tag{34}$$

(34) and (31) together yield:

$$U = u(0)\sqrt{2\pi}\sigma\exp\left(\frac{\mu^2}{2\sigma^2}\right)F\left(\frac{\mu}{\sigma}\right). \tag{35}$$

There is an interesting alternative calculation of (35), as provided by Q. Burrell:

$$U = \int_0^\infty u(i)\,di \longrightarrow U = \frac{u(0)}{c(1)}\int_0^\infty c(i+1)\,di \longrightarrow$$

$$U = \frac{u(0)}{c(1)}\int_1^\infty c(i)\,di \longrightarrow U = \frac{u(0)}{c(1)}\,P(T > 1),$$

where T is the "lifetime" random variable. Now

$$P(T > 1) = P(\log T > 0).$$

Putting $Z = (\log T - \mu)/\sigma$ and remembering that Z is normally distributed, we find

$$P(T > 1) = P\left(Z > -\frac{\mu}{\sigma}\right)$$

$$= F\left(\frac{\mu}{\sigma}\right),$$

and so,

$$U = \frac{u(0)}{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{\mu^2}{\sigma^2}\right)} F\left(\frac{\mu}{\sigma}\right).$$

Hence we again find (35).

Now (35) can easily be calculated, using the table of the standard normal distribution. From our examples we find the estimates $U = 21\ u(0)$ for informetrics, $U = 43\ u(0)$ for mathematics, and $U = 58\ u(0)$ for social development (rounded off to the nearest whole number). These data are a real basis to compare the "lives" of the different disciplines (the determination of $u(0)$ is not needed here; U relates to the initial utility, which can be taken equal for all disciplines—for example, $u(0) = 1$; this is only a matter of standardization).

That papers in social science have a longer "life" than mathematics, for instance, is well known but here we provide an example of how much longer. Here we estimate about $\frac{58}{43} \approx 1.35$ times longer, based on our examples (which are indeed large samples of the total population of citations in the respective subjects). Of course, many more data should be analyzed to verify whether the number 1.35 is a kind of constant in this comparison.

We want to conclude with the philosophical argument that leads us back to the beginning: formula (1), the exponential decay. We think we have shown that the lognormal distribution is good for describing obsolescence data (initial increase *as well as* the decay later on). Yet, it is known that when we truncate the data so that they only show the decay, the exponential distribution (1) fits very well. Whether or not we should consider the increase and the decay as two separate phenomena (which then should be modelled separately) is unclear to us. If we do (as is for instance also the case for the Avramescu distribution), we keep the historically respected notion of "exponential decay." Although, in our paper, we try to fit everything with one distribution, we are not promoting the "lognormal decay" to replace the "exponential decay" notion. We think the latter can still be used when looking only at the decay part; on the other hand, when we try to understand the complete process (increase plus decay) we advocate the lognormal distribution.

These remarks could also be made for the "opposite" study—the study of growth processes. Here also one has the historically respected "law of exponential growth." Yet, as in life, nothing can grow forever, and more sophisticated models are in order (see, for example, Egghe & Rao, 1991). Do we consider these more intricate models (e.g., the logistic curve) as one phenomenon or as composed of at least two components, to be described separately? We think that both approaches have their value; it only depends on the topics one wants to study.

Note—After this paper was written Prof. Dr. R. Rousseau (to whom we give our sincerest thanks) pointed out to us that a new article (Matricciani, 1991) confirmed our findings on the lognormal distribution. The experiments were made in the collections of some *IEEE* publications and comprehensive parts in the human sciences (see also the *Cambridge Economic History of Europe*, vol. VI, 1965). As in our study, the author hence has examined homogeneous data and concludes the same way.

Acknowledgement—The second named author is grateful to the Belgian National Science Foundation (NFWO) for financial support in the period when he was a visiting professor in LUC (February-April, 1990).

The authors thank Prof. Q.L. Burrell for pointing out some mistakes in an earlier version of this paper. The authors are grateful to an anonymous referee for several interesting remarks which led to a substantial improvement of this paper.

REFERENCES

Avramescu, A. (1973). Science citation distribution and obsolescence. *St. Cerc. Doc. 15*, 345-356.

Avramescu, A. (1979). Actuality and obsolescence of scientific literature. *Journal of the American Society for Information Science*, *30*, 296-303.

Bartholomew, D.J. (1982). Stochastic models for social processes. New York: Wiley & Sons.

Brookes, B.C. (1970a). Obsolescence of special library periodicals: Sampling errors and utility contours. *Journal of the American Society for Information Science*, *21*, 320-329.

Brookes, B.C. (1970b). The growth, utility, and obsolescence of scientific periodical literature. *Journal of Documentation*, *26*, 283-294.

Brookes, B.C. (1971). Optimum *P%* library scientific periodicals. *Nature*, *232*, 458-461.

Brookes, B.C. (1973). Numerical methods of bibliographical analysis. *Library Tends*, 18-43.

Cairns, R.B. (1979). *Social development: The origins and plasticity of interchanges.* San Francisco: W.H. Freeman & Co.

Egghe, L. (1984). Stopping time techniques for analysts and probabilists. *London Mathematical Society Lecture Notes Series 100.* Cambridge: Cambridge University Press.

Egghe, L., & Ravichandra Rao, I.K. (in press). Classification of growth models based on growth rates and its applications. *Scientometrics.*

Egghe, L., & Rousseau, R. (Eds.) (1990). Informetrics 89/90. *Proc. 2nd International Conference on Bibliometrics, Scientometrics and Informetrics*, U. Western Ontario, Canada, 1989. Amsterdam: Elsevier.

Geller, N.L., & De Cani, J.S. (1981). Lifetime-citation rates: A mathematical model to compare scientists' work. *Journal of the American Society for Information Science*, *32*, 3-15.

Gradshteyn, I.S., Ryzhik, I.M. (1965). Table of Integrals, Series, and Products. New York: Academic Press. (Fourth edition prepared by Y.V. Geronimus and M.Y. Tseytlin. Translated from the Russian by Scripta Technica, Inc. Translation edited by Alan Jeffrey).

Griffith, B.C., Servi, P.N., Anker, A.L. & Drott, M.C. (1979). The aging of scientific literature: A citation analysis. *Journal of Documentation*, *35*(3), 179-196.

Matricciani, E. (1991). The probability distribution of the age of references in engineering papers. *IEEE Transactions of Professional Communication*, *34*, 7-12.

Motylev, V.M. (1981). Study into the stochastic process of change in the literature citation pattern and possible approaches to literature obsolescence estimation. *International Forum on Information and Documentation*, *6*, 3-12.

Motylev, V.M. (1989). The main problems of studying literature aging. *Scientometrics*, *15*, 97-109.

Ravichandra Rao, I.K. (1973). Obsolescence and utility factors of periodical publications: A case study. *Library Science*, *10*, 297-307.

Ravichandra Rao, I.K. (1988). Probability distributions and inequality measures for analysis of circulation data. In L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88*, *Proc. First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval*, LUC, Belgium, 1987 (pp. 231-248). Amsterdam: Elsevier.

Stinson, E.R., & Lancaster, F.W. (1987). Synchronous versus diachronous methods in the measurement of obsolescence by citation studies. *Journal of Information Science*, *13*, 65-74.